

***In-silico* Approach to Design a Novel Inhibitor of SHP1/2: Enhancement of HSCs Proliferation**

*A Major Project dissertation submitted in partial fulfilment*

*of the requirement for the degree of*

**Master of Technology**

**In**

**Bioinformatics**

*Submitted by*

**SAUMYA BHARTI**

**2K12/BIO/23**

**Delhi Technological University, Delhi, India**

*Under the supervision of*

**Dr. Gurudutta Gangenahalli,**

**FRSC (London) & Scientist 'F'**



**Institute of Nuclear Medicine and Allied Sciences (INMAS),  
Defence Research & Development Organisation (DRDO)**

*University Guide*

**Dr. Jai Gopal Sharma**

**Dr. Monica Sharma**



Department of Biotechnology  
Delhi Technological University  
(Formerly Delhi College of Engineering)  
Shahbad Daultapur, Main Bawana Road,  
Delhi-110042, INDIA



## **CERTIFICATE**

This is to certify that the M. Tech. dissertation entitled “***In-silico approach to design a novel inhibitor of SHP1/2: Enhancement of HSCs proliferation***”, submitted by **Saumya Bharti (2K12/BIO/23)** in partial fulfilment of the requirement for the award of the degree of Master of Engineering, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate’s own work carried out by him/her under my guidance.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

**Date:**

Dr. Jai Gopal Sharma

Dr. Monica Sharma

(Project Mentors)

Department of Bio-Technology

Delhi Technological University

(Formerly Delhi College of Engineering, University of Delhi)



# INSTITUTE OF NUCLEAR MEDICINE & ALLIED SCIENCES

Defence Research & Development Organization

MINISTRY OF DEFENCE

Brig. S K Mazumdar Marg, Delhi-110054

---

## CERTIFICATE

This is to certify that this project report entitled “*In-silico* approach to design a novel inhibitor of SHP1/2: Enhancement of HSCs proliferation” by Saumya Bharti (2K12/BIO/23), M.Tech (Bioinformatics), Delhi Technological University (Formerly Delhi College of Engineering), Shahbad Daulatpur, New Bawana Road, Delhi-110042 is a bonafide record of work carried out under my guidance and supervision in my laboratory.

**Dr. Gurudutta Gangenahalli**

*FRSC (London)*

**Scientist F (Indian Defence Research & Development Service Cadre)**

**Head, Stem Cell & Gene Therapy Research Group**

**Jt. Director**

**Tel: 011-23905144**

**Email: [gugdutta@rediffmail.com](mailto:gugdutta@rediffmail.com)**

**Fax: 011-2309509**

**Mobile: 9899136430**

**Date: 21-7-2014**

# **DECLARATION**

I **Ms. Saumya Bharti (2K12/BIO/23)**, student of M. Tech. (Bioinformatics), Delhi Technological University have completed the project titled “***In-silico* approach to design a novel inhibitor of SHP1/2: Enhancement of HSCs proliferation**” for the award of Degree of M. Tech. (Bioinformatics), for academic session 2012-14. The information given in this project is true to the best of my knowledge.

**Saumya Bharti**

**(2K12/BIO/23)**

# **ACKNOWLEDGEMENT**

**It gives me immense pleasure to thank all those people who have been instrumental in the completion of my project.**

In the first place I sincerely thank **Dr. R.P. Tripathi, Director of INMAS**, for allowing me to do this research work. I would like to extend my vote of thanks to **Dr. Rajiv Vij (Scientist 'F')** without whom this project would not have been possible in INMAS.

**I would like to express my heartfelt gratitude to my supervisor Dr. G.U. Gurudatta (Scientist 'F') for allowing me to work under his guidance and providing me with all the facilities during my project work.**

**I am deeply indebted to Mr. Pawan Kumar Raghav for his keen interest, constructive critical discussion, unceasing encouragement and guidance. I express my gratitude to him for sincerely helping me and boosting my morale to work hard.**

**I express my deep sense of thankfulness to Dr. Jai Gopal and Dr. Monica Sharma whose enthusiastic zeal boosted me for the successful completion of my work.**

**I gratefully acknowledge Dr. Yogesh Kumar Verma (Scientist 'C'), Mr. Neeraj Satija and Mr Vikas for their constant support.**

**Finally, yet importantly, I would like to express my heartfelt thanks to my beloved parents, for their blessings and my friends, for their help and wishes for the successful completion of this project.**

**SAUMYA BHARTI**

2K12/BIO/23

# CONTENTS

<u>TOPIC</u>	<u>PAGE NO.</u>
<i>LIST OF FIGURES</i>	<i>i</i>
<i>LIST OF TABLES</i>	<i>ii</i>
<i>LIST OF ABBREVIATIONS</i>	<i>iii</i>
1. ABSTRACT	1
2. INTRODUCTION	2
3. REVIEW OF LITERATURE	4
4. METHODOLOGY	11
5. RESULTS AND DISCUSSION	15
6. CONCLUSION AND FUTURE PERSPECTIVE	37
7. REFERENCES	39
8. APPENDIX	41

# LIST OF FIGURES

Fig. 1: Haematopoietic stem cell hierarchy.

Fig. 2: Structure of SCF receptor c-Kit (receptor tyrosine kinase).

Fig. 3: Proteins and signal transduction molecules interacting with activated c-Kit receptor.

Fig. 4: c-Mpl/TPO signaling pathway.

Fig. 5: Different developmental pathways involved in regulating HSC self renewal property.

Fig. 6: Secondary structure of c- kit.

Fig. 7: Tertiary (3D) structure of c-kit.

Fig. 8: Sequence alignment of SHP1 and SHP2.

Fig. 9: Superimposed structure of SHP-1 and HePTP.

Fig. 10: Common pharmacophore for ligands.

Fig. 11: QSAR visualization of various substituent's effect.

Fig. 12: Fitness graph between observed activity versus PHASE predicted activity.

# **LIST OF TABLES**

Table 1: Molecules/genes/drugs regulating HSC proliferation and their self-renewal characteristics along with their functions.

Table 2: RMSD values of the various members of the non- receptor PTPase

Table 3: CIDs and their corresponding IC<sub>50</sub> values under the assay ID (AID) 1059.

Table 4: CIDs and their corresponding IC<sub>50</sub> values under the assay ID (AID) 1077.

Table 5: Compounds from both the assay IDs (AID 1059 and AID1077) and their binding energy with SHP1 (PDB ID: 2B3O).

Table 6: Shows the structure, pIC<sub>50</sub> values and binding energy of compounds selected.

Table 7: Various PHASE hypotheses generated.

Table 8: Statistical result of 3D QSAR study.

Table 9: Fitness and PHASE predicted activity of the compounds.



## LIST OF ABBREVIATIONS

ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
AID	Assay ID
APS	Antigen Presenting Cells
BA	Binding Affinity
CID	Compound ID
COMFA	Comparative Molecular Field Analysis
COMSIA	Comparative Molecular Similarity Indices Analysis
DUSPs	Dual Specific Phosphatases
ESC	Embryonic Stem Cell
F	Fischer coefficient
HSC	Hematopoietic Stem Cells
HePTP	Hematopoietic Tyrosine Phosphatase
IC <sub>50</sub>	Inhibitory Concentration at 50%
LMW	Low-Molecular-Weight Phosphatases
LOO	Leave One Out Cross Validation
NRPTPs	Non Receptor Tyrosine Phosphatase
NSC 87877	SHP-1inhibitor
PTKs	Protein Tyrosine Kinases
PTPs	Protein Tyrosine Phosphatases
PTENs	Phosphatase and Tensin homologs
PTP1B	Protein Tyrosine Phosphatase 1B
Q <sup>2</sup>	Leave one out cross validation
QSAR	Quantitative Structure Activity Relationship
R <sup>2</sup>	Correlation Coefficient

RMSD	Root Mean Square Deviation
S	Standard Deviation
SCF	Stem Cell Factor
SHP-1	Src Homology 2 Domain Containing Phosphatase 1
SHP-2	Src homology 2 Domain Containing Phosphatase 2
STEP	Striatal-Enriched Protein Tyrosine Phosphatase
TCPTP	T-cell Protein Tyrosine Phosphatase

# ***In-silico* Approach to Design a Novel Inhibitor of SHP1/2: Enhancement of HSCs Proliferation**

Saumya Bharti

Delhi Technological University, Delhi, India

## **ABSTRACT**

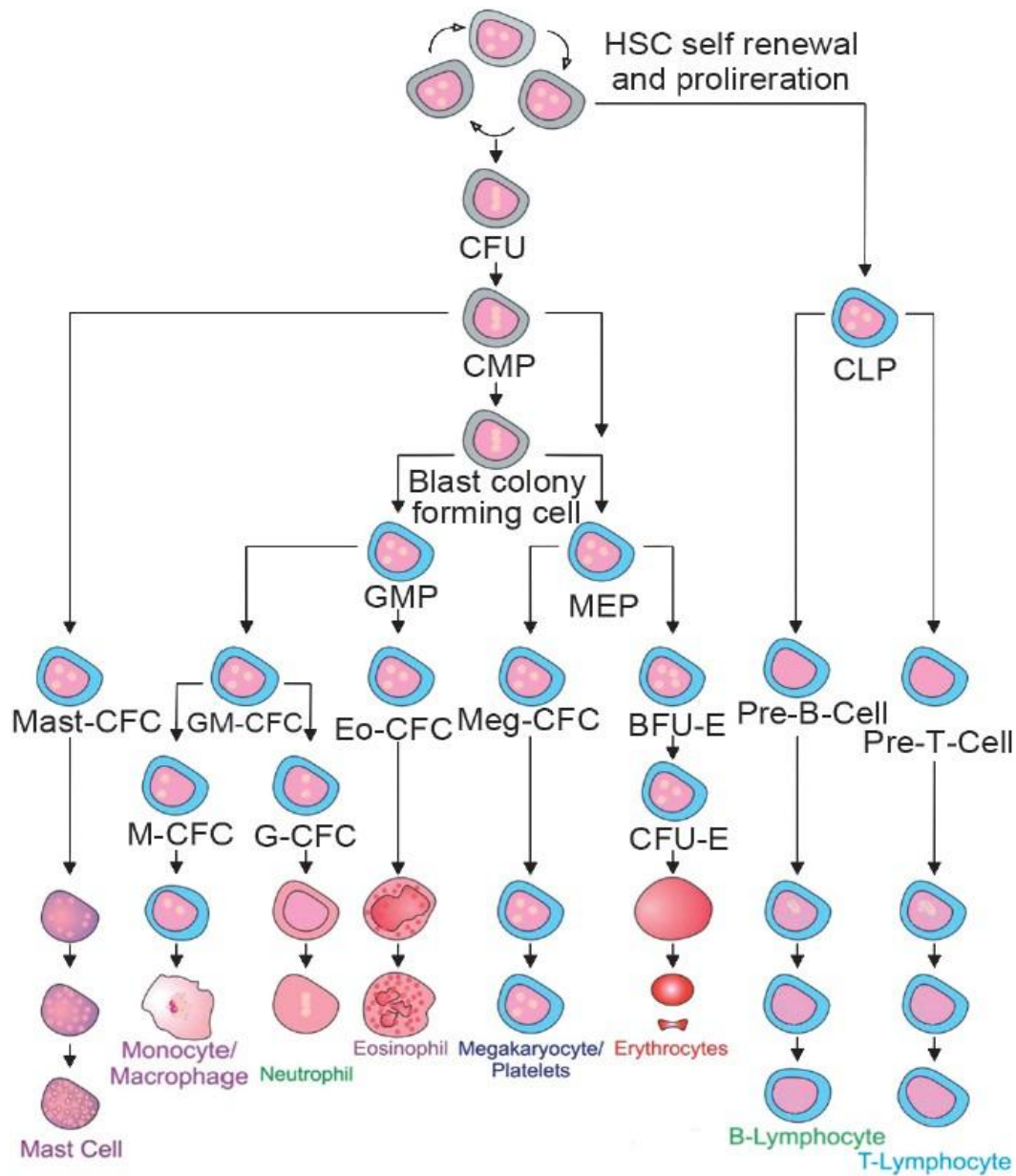
Hematopoiesis is a lifelong process of the production and maintenance of all the cells of the blood system from the hematopoietic stem cells (HSCs) in a hierarchical manner. In adult mammals the hematopoietic stem cells (HSCs) reside in the bone marrow cavity. HSCs give rise to all the types of blood cells of the lymphoid and myeloid lineages. Radiation therapy in cancer leads to loss of a large number of immune cells, so HSCs are transplanted into the bone marrow of irradiated patients but HSCs differentiate before reaching the bone marrow, so the main aim is to maintain the HSCs in their proliferative state until they reach bone marrow. HSCs have three main characteristics of proliferation, self-renewal and differentiation. The proliferative property of HSCs is regulated by a number of signaling pathways, ligands and molecules, but one of the main regulators is the c-kit/SCF signaling. The binding of the SCF (Stem Cell Factor) to the c-kit receptor results in receptor dimerization, thereby activating c-kit activity. Src homology 2 (SH2) domain containing phosphatase 1 (SHP1) negatively regulates the c-kit activity. Hence in this work we have identified the structural variations using structural alignment between the PTPases and observed that HePTP is having high identity with SHP1. Based on this result we have screened known inhibitors of HePTP against SHP1 and identified some ligands having higher binding affinity than NSC87877. In order to design SHP1 specific inhibitor a pharmacophore model was designed to search for a NCE (Novel Chemical Entity).

# **Introduction**

## 2.0 INTRODUCTION

Stem cells are a class of undifferentiated and unspecialized cells in the human body capable of proliferation, differentiation and self-renewing themselves. The concept of “stem cell” was first proposed by Till and McCulloch. This concept came into existence following the extensive studies on *in vivo* blood system regeneration (Seita *et al.*, 2010). Stem cells are mainly of two types i.e., embryonic and non- embryonic. Embryonic stem cells (ESCs) are pluripotent cells since they can differentiate into all cell types whereas non- embryonic stem cells (Non- ESCs) are multipotent cells due to their potential to differentiate into a number of cell types but not all cell types (Tuch, Bernard, 2006). Stem cells can be broadly classified into four types based on their origin, stem cells from embryos, stem cells from the fetus, stem cells from the umbilical cord, and stem cells from the adult (Forbes S *et al.*, 2002). Adult stem cells include the hematopoietic stem cells.

Hematopoietic stem cells have ability of haematopoiesis. Haematopoiesis is the production and maintenance of blood stem cells and their proliferation and differentiation into the cells of peripheral blood [Fig. 1]. Transplantation of stem cells from the original transplant recipient into secondary and tertiary irradiated recipients reconstitutes hematopoiesis with resultant normal life spans. Transplantation requires two essential properties proliferation to replenish the stem cell compartment (self-renewal) and lifelong production of blood (Pearce W *et al.*, 2008). During transplantation high number of HSC is needed as the cells reaching target eventually decreases. Various signaling pathways regulate the proliferation of HSCs, but the SCF and c-Kit signaling plays a very pivotal role in this process. SCF is a cytokine that exists both as a soluble protein and transmembrane protein which induces proliferation on binding with c-kit. The SCF receptor, c-kit is a receptor tyrosine kinase. c-Kit consists of an extracellular domain (where the ligand binds), a transmembrane segment, a juxtamembrane segment and a protein kinase domain (Roskoski, Robert. 2005, Ronnstrand, L. 2004). The binding of SCF to c-Kit results in receptor dimerization and activation of protein kinase activity. The activated receptor becomes autophosphorylated at tyrosine residues 568 and 570 in juxtamembrane region of c-Kit that serve as docking sites for signal transduction molecules containing SH2 domains SHP-1 and SHP-2. SHP-1 binds to the phosphotyrosine residue 570 of c-Kit and negatively regulates the proliferation of HSCs (Roskoski, Robert. 2005; Ronnstrand, L. 2004). NSC87877 (CID: 16654632) is the only single inhibitor known to inhibit SHP1/2 (from PubChem database). In this work we have identified the structural variations using structural alignment between the PTPases and observed that HePTP is having high identity with SHP1. Based on this result we have screened known inhibitors of HePTP against SHP1 and identified some ligands having higher binding affinity than NSC87877. In order to design SHP1 specific inhibitor a pharmacophore model was designed to search for a NCE (Novel Chemical Entity).



**Fig. 1: Haematopoietic stem cell hierarchy.** Self-renewing HSC give rise to several multipotent progenitors (colony forming units (CFU), common myeloid progenitor (CMP) and common lymphoid progenitors (CLP)), which, in turn, produce oligopotent progenitors, unipotent progenitors and eventually fully differentiated cells. The CMP is able to produce granulocyte-macrophage progenitors (GMP) and megakaryocyte/erythrocyte progenitors (MEP) giving rise to monocyte/macrophages/granulocytes and megakaryocytes/platelets/ erythrocytes, respectively. Erythroid burst forming unite (BFU-E) give rise to pro-erythroblast colony forming unit-erythroid (CFU-E) before erythrocytes are formed and the CLP gives rise to pre-B and pre-T cells which continue to mature into mature B and T lymphocytes.

# **Review of Literature**

### 3.0 REVIEW OF LITERATURE

Stem cells are a type of undeveloped and unspecialized cells in the human body capable of proliferation, differentiation and self-renewing themselves. The concept of “stem cell” was first proposed by Till and McCulloch. This concept came into existence following the extensive studies on *in vivo* blood system regeneration (Seita *et al.*, 2010). Stem cells are mainly of two types i.e., embryonic and non- embryonic. Embryonic stem cells (ESCs) are pluripotent cells since they can differentiate into all cell types whereas non- embryonic stem cells (Non- ESCs) are multipotent cells due to their potential to differentiate into a number of cell types but not all cell types (Tuch, Bernard, 2006). Stem cells can be broadly classified into four types based on their origin, stem cells from embryos, stem cells from the fetus, stem cells from the umbilical cord, and stem cells from the adult (Forbes S *et al.*, 2002). Adult stem cells are located in tissues throughout the body and functions as a reservoir to replace the damaged and aging cells. However, under physiological conditions they are traditionally thought to be restricted to differentiate into cell lineages of the organ system in which they are located such as the brain cells, blood cells, etc. Adult stem cells include the hematopoietic stem cells.

Hematopoiesis is a lifelong process of the production and maintenance of all the cells of the blood system from the hematopoietic stem cells (HSCs) in a hierarchical manner. In adult mammals the hematopoietic stem cells (HSCs) reside in the bone marrow cavity (Pietras *et al.*, 2011, Crusio *et al.*, 2012). All the mature blood cells in the body are derived from a small population of the hematopoietic stem cells (HSCs) and progenitors which becomes lineage restricted with each differentiation (Pietras *et al.*, 2011). HSCs give rise to all the types of blood cells of the lymphoid and myeloid lineages. Some of the important cell types of the lymphoid lineage are B- cells, T- cells and the natural killer (NK) cells whereas those of the myeloid lineage are monocytes, granulocytes, macrophages, microglial cells and dendritic cells. Each of these cell types of the lymphoid and myeloid lineages can be derived from a single HSC and each HSC has the potential of producing large numbers of mature blood cells over a long period (Smith, 2002). Transplantation of the HSCs into secondary and tertiary irradiated recipients restores normal hematopoiesis. The process of transplantation requires two essential properties, self- renewal capacity, and lifelong production of blood cells (Pearce W *et al.*, 2008). HSCs on the basis of their ability to self- renew can be divided into (1) Long term HSCs (LT- HSCs) and (2) Short term reconstituting HSCs (ST- HSCs). LT- HSCs have the ability of extensive self-renewal and sustaining lifelong hematopoiesis. ST- HSCs on the other hand have restricted self-renewal capacity (Blank *et al.*, 2007).

Stem Cell Factor (SCF, kit-ligand or steel factor) is a cytokine plays a pivotal role in the process of hematopoiesis. SCF is expressed at all the sites where hematopoiesis takes place such as the bone marrow and fetal liver (Gali *et al.*, 1994). SCF exists both as a soluble protein and transmembrane protein which induces proliferation on binding with c-kit.



The SCF receptor, c-kit is a receptor tyrosine kinase which is expressed in HSCs, mast cells, germ cells and melanocytes. The hematopoietic progenitor cells such as megakaryocytes, myeloblasts and erythroblasts also express the c-Kit receptor. c-Kit consists of an extracellular domain (where the ligand binds), a transmembrane segment, a juxtamembrane segment and a protein kinase domain [Fig.2]. The protein kinase domain contains an 80 amino acid residue insert (Roskoski, Robert. 2005, Ronnstrand, L. 2004). Stem cell factor and c-kit signaling plays a vital role in the process of hematopoiesis, gametopoiesis, melanogenesis and mast cell development and function. When stem cell factor binds to the c-kit, it leads to receptor dimerization and the protein kinase activity is activated via auto-phosphorylation at Tyr568 and Tyr570 of juxtamembrane domain. In the kinase insert domain three residues are phosphorylated. These three residues attract the adaptor protein Grb2 at Tyr703, phosphatidylinositol 3-kinase at Tyr721 and phospholipase C at Tyr730. In the distal kinase domain phosphotyrosine 900 binds phosphatidylinositol 3-kinase which in turn binds the adaptor protein Crk and the phosphotyrosine 936 binds the adaptor proteins APS, Grb2 and Grb7 (Roskoski, Robert. 2005) [Fig.3].

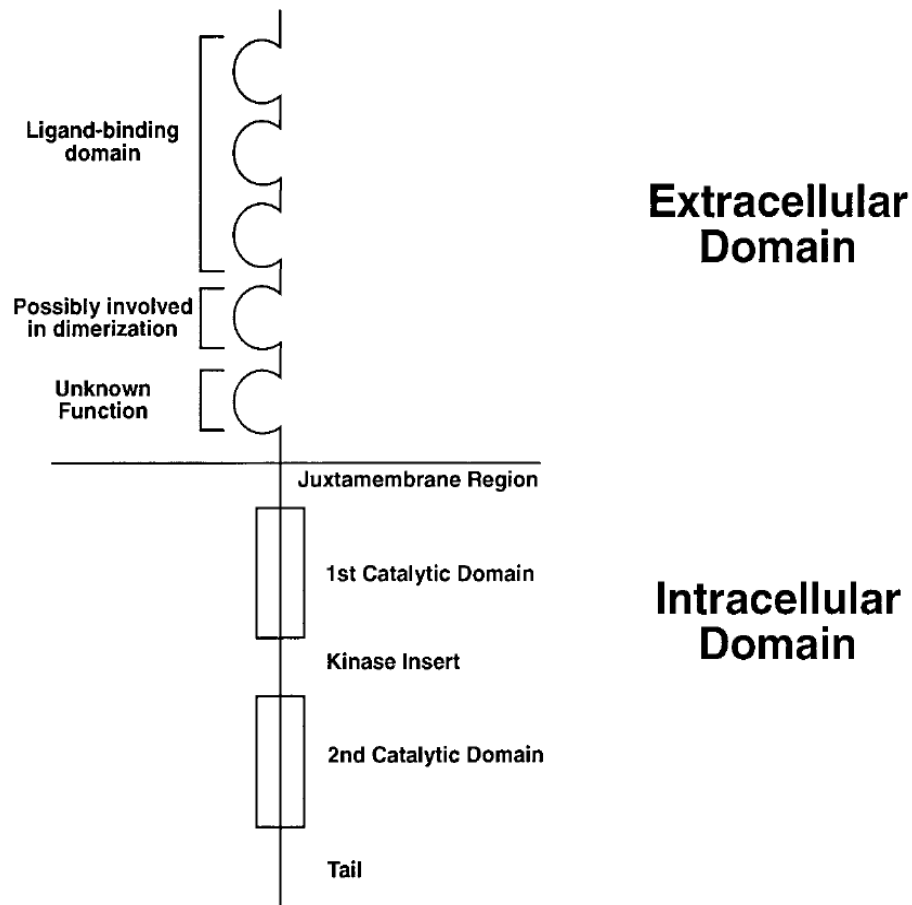
SHP1, a cytosolic phosphotyrosyl phosphatase (non-membrane protein tyrosine phosphatase) primarily occurs in the hematopoietic and epithelial cells. SHP1 contains two tandem SH2 domains, a phosphatase domain and a C-terminal tail. SHP1 negatively regulates the growth factor signaling. SHP1 reduces the growth promoting properties of the colony stimulating factor 1, erythropoietin and interleukin 3 receptors in addition to inhibiting Kit signaling. This effect is mediated either by direct receptor dephosphorylation or indirect dephosphorylation of the receptor-associated protein tyrosine kinases. The catalytic domain of SHP1 is blocked by its N-terminal SH2 domain thereby maintaining the enzyme in an inactive conformation (Kozlowski *et al.*, 1998).

SHP2 is a cytosolic phosphotyrosyl phosphatase (non-membrane protein tyrosine phosphatase) just like SHP1, but unlike SHP1 it occurs in many types of cells. SHP2 contains two tandem SH2 domains, a phosphatase domain and a C-terminal tail (Lawrence *et al.*, 2008). The SH2 domain is responsible for targeting SHP2 to phosphotyrosine residues of a variety of signaling molecules. Thus, SHP1 and SHP2 negatively regulates the Kit signaling due to interactions with specific phosphotyrosine residues present on the c-kit receptor (Kozlowski *et al.*, 1998). The cell signaling by the cytokines and growth factors is mediated by SHP2 by acting via the RAS/MAP Kinase pathway (Geronikaki *et al.*, 2008). Usually the enzyme SHP2 is found in its inactivated state due to the autoinhibition of its catalytic domain (Lawrence *et al.*, 2008, Yu *et al.*, 2008).

The property of proliferation and self-renewal of HSCs is regulated by a number of signaling pathways, ligands, macromolecules, genes and drugs. Table 1 shows a list of molecules, genes and drugs which regulate the proliferation and self-renewal properties of HSCs along with their functions. There are a number of animal models which have been used for the study of the different signaling pathways regulating the proliferative and self-renewal capacity

[Fig.5]. Some of the most studied signaling pathways are c-Mpl /TPO signaling pathway [Fig.4], Tie2/Ang signaling pathway, BMP signaling pathway, Hedgehog signaling pathway, Notch signaling pathway, Wnt signaling pathway, etc. (Chotinantakul *et al.*, 2012, Zon, 2008).

The binding of developmental regulators and chemical modulators to the appropriate cell-surface receptors, initiates the signalling pathways, which leads to the translocation of transcription factors from the cytoplasm to the nucleus. (The receptor for retinoic acid is in the nucleus.) The transcription factors in turn interact with other cell-specific transcription factors for regulating self-renewal property of HSCs. The competition between the various transcription factors is a simple mechanism for regulating self-renewal capacity, and another level of competition and control is provided by the chromatin-associated factors, such as MLL and BMI1 (Zon, 2008).



**Fig. 2: Structure of SCF receptor c-Kit (receptor tyrosine kinase).**

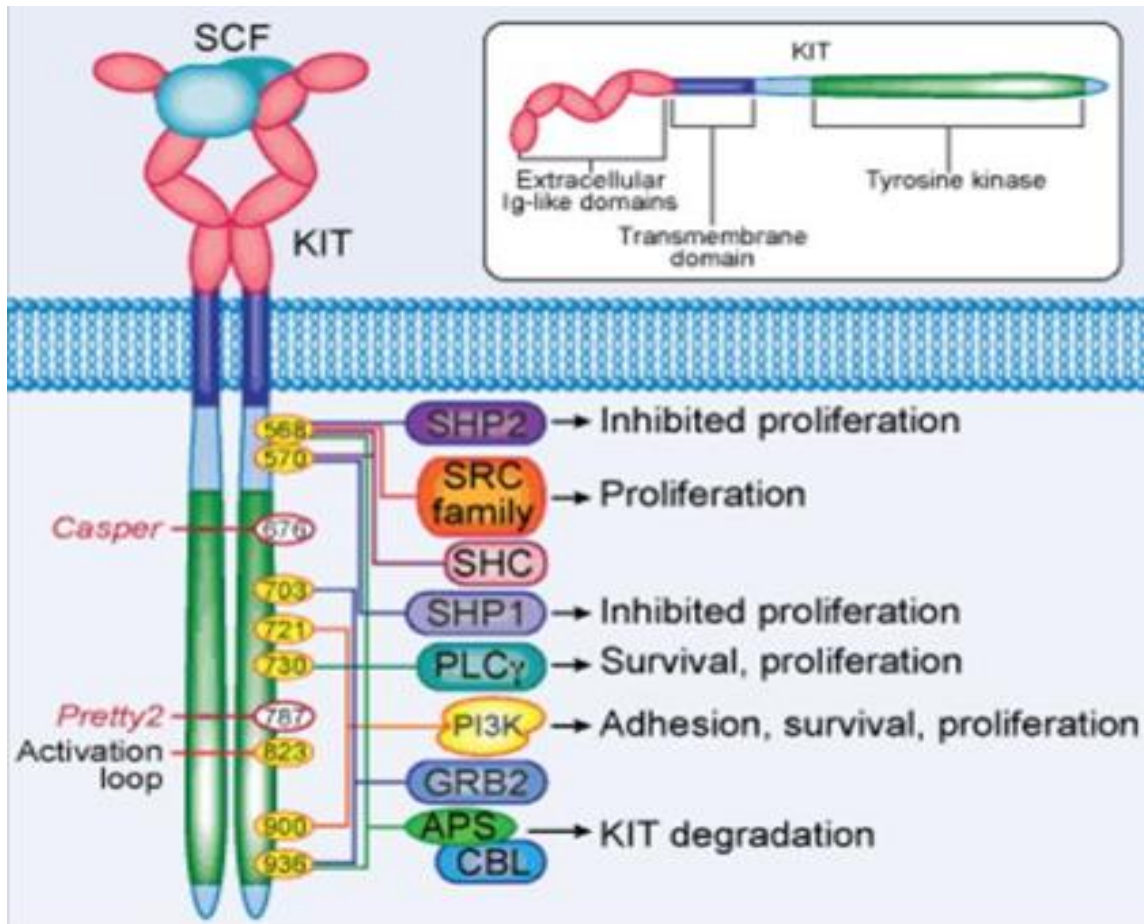


Fig. 3: Proteins and signal transduction molecules interacting with activated c-Kit receptor.

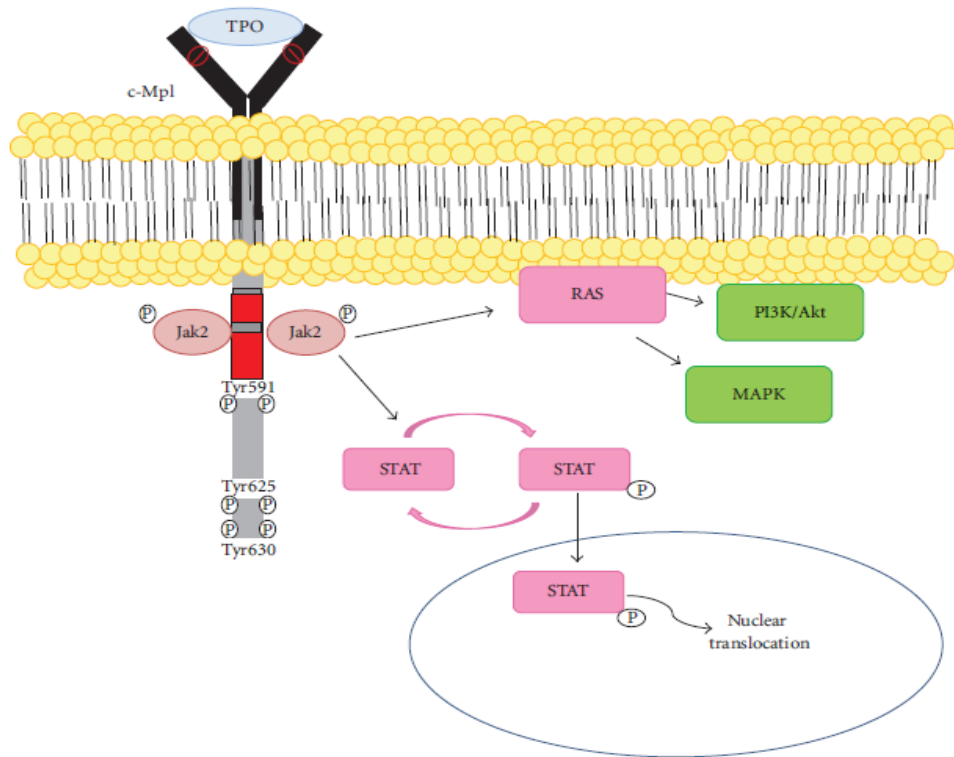


Fig. 4: c-Mpl/TPO signaling pathway (Source: Chotinantakul *et al.*, 2012).

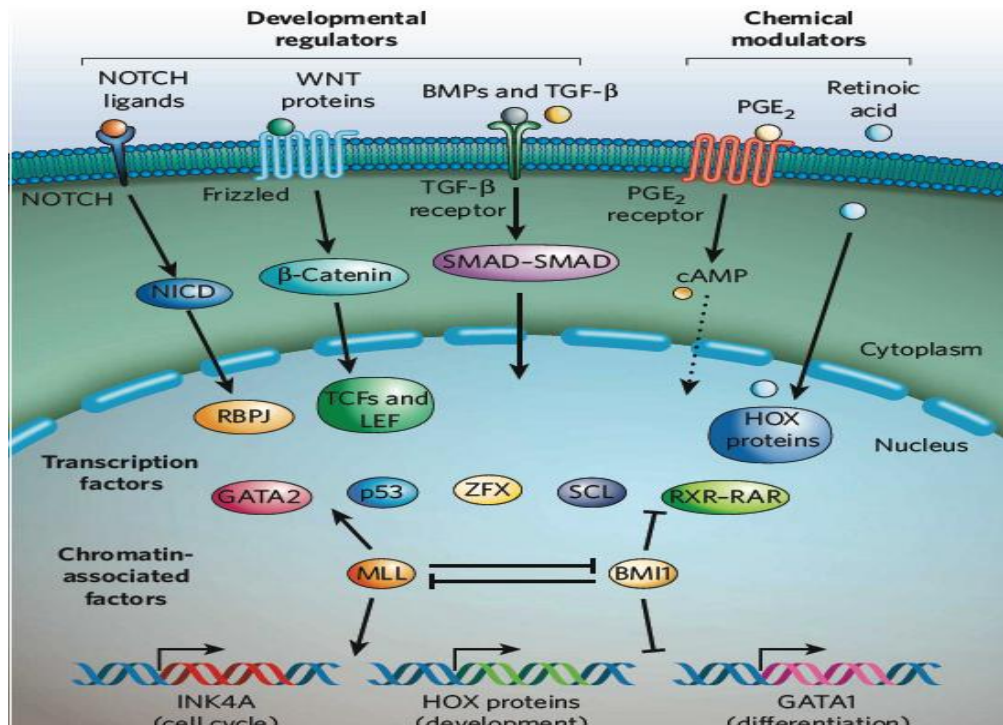


Fig. 5: Different developmental pathways involved in regulating HSC self renewal property. (Source: Zon, 2008).

**Table 1: Molecules/genes/drugs regulating HSC proliferation and their self-renewal characteristics along with their functions.**

<b>S. No.</b>	<b>Molecule/Gene/Drug</b>	<b>Function</b>
1.	Wnt3A	Increases self renewal 3- fold in culture
2.	Wnt3A	Induce proliferation of B- cell precursors in a LEF-1 dependent manner
3.	Wnt5	Suppresses tissue recovery
4.	Wnt8	Increased proliferation
5.	Hoxb4 and Hoxa9	Overexpression leads to increased self renewal in mouse bone marrow cells.
6.	Hoxa9	Increased myeloid lineage differentiation and leukaemogenesis
7.	Hoxb4	Normal bone marrow development and abnormal/aberrant expression increases the no. of transplantable HSCs both <i>in vivo</i> and <i>in vitro</i>
8.	Hoxa10	15- fold increase in self- renewal
9.	Hox gene	Important in regulation of self- renewal
10.	Bmi-1	Defeciency results in decreased self- renewal
11.	Bmi-1	Overexpression results in increased self- renewal
12.	Bmi-1	Represses the gene encoding cell cycle regulator INK4A
13.	MLL fusion proteins	Increased self-renewal whereas inactivation leads to decreased self- renewal
14.	Retenoic acid	Alterations in Hox gene expression and also modifies Wnt mediated signaling pathway
15.	Retenoic acid	Maintains HSCs in culture and can increase self- renewal in serial transplantation experiments

---

16.	DNMT3A and DNMT3B	Required for DNA methylatin in HSCs
17.	p21	Lack results in higher rate of HSC proliferation and differentiation, lower self- renewal capacity. Hence required for maintaining HSC quiescence
18.	GATA-2	Inactivation inhibits HSC self- renewal
19.	GFI-1	Inactivation inhibits HSC self- renewal
20.	Myc	Inactivation inhibits HSC self- renewal
21.	SMAD- 4	Inactivation inhibits HSC self- renewal
22.	Axin	Abnormal/abberant expression inhibits HSC proliferation, increased cell death of HSCs <i>in vitro</i> and reduced reconstitution <i>in vivo</i>
23.	Valproic acid	Increases both proliferation and self- renewal, accelerates cell cycle progression, down regulates p21cip1/waf1, inhibits GSK3 $\beta$ thereby activating Wnt signaling pathway, up regulates Hoxb4, induces differentiation or apoptosis in leukemic blasts, increases replating capacity of murine HSC
24.	Laq824	Properties similar to valproic acid on HSC
25.	CG152	Properties similar to valproic acid on HSC
26.	Stem Regenin (SR-1)	Increases <i>ex vivo</i> expansion of peripheral blood derived CD34+ cells by 50 fold

---

NSC87877 (CID: 16654632) is the only single inhibitor known to inhibit SHP1/2 (from PubChem database). This inhibitor non-specifically binds with other PTPases like Acid phosphatase 1 (ACP1) (Song *et al.*, 2009), Dual specificity phosphatase 14 (DUSP14) (Song *et al.*, 2009), Dual specificity phosphatase 23 (DUSP23) (Song *et al.*, 2009), Dual specificity phosphatase 26 (DUSP26) (Song *et al.*, 2009) and Vaccinia H1- related (VHR) phosphatase activity (Park *et al.*, 2009). NSC87877 has been found to have ten times more inhibitory effect on SHP2 as compared to DUSP14 (Song *et al.*, 2009). NSC87877 has also been seen to be more specific for DUSP26 in comparison to SHP1 as it has more inhibitory effect on DUSP26. The binding mechanism of NSC87877 to the catalytic cleft of DUSP26 is the same as that of SHP2 was suggested by kinetic studies with NSC87877 and DUSP26 (Song *et al.*, 2009).

**Materials**  
**&**  
**Methods**

## **4.0 MATERIALS AND METHODS**

### **4.0.1 Sequence Analysis of c-Kit and SHP1/2**

The sequence, secondary and tertiary structures of the c-Kit receptor was retrieved from the Protein Data Bank (PDB). The various domains of c-Kit in the tertiary structure were analyzed using Pymol. Sequence alignment of the SHP1 and SHP2 sequences was performed to determine the percentage identity and percentage similarity using EMBOSS.

### **4.0.2 Structural Analysis of Protein Tyrosine Phosphatase and Search of its Antagonist**

PubChem Compound database was used to search for the antagonists or inhibitors of SHP1. Only one compound, NSC87877 is reported as SHP1/2 PTPase inhibitor. In order to design a novel inhibitor, other member inhibitors of the non-receptor protein tyrosine phosphatase family which are closely related to SHP1 were searched. The closely related member of SHP1 and SHP2 were superimposed with different members of the non-receptor PTPase family and their RMSD values were calculated using Pymol.

### **4.0.3 Searching PubChem Database for SHP1/2 Inhibitors**

Superimposition of SHP1 with other members of the non-receptor PTPase family showed it to be closely related to HePTP with an RMSD value of 0.656. Hence, from PubChem two assay IDs (AID 1059 and AID 1077) were selected wherein IC<sub>50</sub> values of compounds were reported against HePTP. Since NSC87877 was tested against HePTP using two different assays hence the inhibitors along with their IC<sub>50</sub> values from both the assay IDs were collected as two different datasets.

### **4.0.4 Virtual Screening of Ligands and Dataset Preparation**

The inhibitors were collected in SDF format from both the assay IDs (AID 1059 and AID 1077) which were then screened against SHP-1 and subsequently binding energy of all the compounds were obtained using AUTODOCK Vina. Inhibitors which showed higher binding affinity than NSC87877 were further selected on the basis of similar structure and good biological activity. Hence the inhibitors with better binding energy, good biological activity and belonging to congeneric series were included in the dataset to be used for pharmacophore and QSAR studies. The clustering feature of PubChem was used to obtain the congeneric series. A dataset of 24 compounds (including NSC87877) with well-defined inhibitory activity given as IC<sub>50</sub> values in  $\mu\text{M}$  concentration was prepared for building 3D-QSAR model. For the correlation purpose IC<sub>50</sub> values were then converted to their molar values and subsequently calculated to free energy-related terms, i.e.,  $-\log(1/\text{IC}_{50})$ . This dataset was then chosen for generating common pharmacophore hypotheses and then performing QSAR analysis. PHASE-3.1 module



of Maestro-9 (Phase 3.1, Schrödinger, LLC, 2009) molecular modeling software was used to generate 3D pharmacophore models for selected series of inhibitors.

#### **4.0.5 Ligand Preparation and Conformation Generation**

The structures were sketched using maestro builder toolbar and were imported to develop pharmacophore model panel of the PHASE with their respective activity values. The ligands were assigned as actives and inactives by giving an appropriate activity threshold value 5.6.

The activity threshold value was selected in the basis of dataset activity distribution and active ligands are chosen to derive a set of suitable pharmacophores. Sketched structures were energy minimized/cleaned up by Ligprep module using OPLS\_2005 force field(LigPrep, Schrödinger, LLC, 2009) and proper protonation states were assigned with the ionizer subprogram at  $\text{pH } 7.2 \pm 0.2$ .

Conformation generation is an important step in PHASE. The conformations were generated with the help of ConfGen method taking GB/SA solvent model using OPLS\_2005 (MacroModel, Schrödinger, LLC, 2009) force field.

About 1000 conformers were generated per structure ensuring 50 step minimization. The minimized conformers were filtered using a relative energy parameter limitation of 10kcal/mol and a minimum atom deviation of 1Å. Thus lowest energy non-redundant conformers of a ligand were used for pharmacophore model development. A couple of conformer was defined as identical if the relative distance between them is below 1Å.

#### **4.0.6 Creating Pharmacophore Sites and Common Pharmacophore Hypothesis Generation**

According to the bioactivity, the molecules were divided into actives and inactives, setting the maximum and minimum values in the activity threshold window of PHASE. Pharmacophore sites of a ligand are represented in the 3D space by a set of points. These points coincide with various chemical characteristics with type, location, and directionality, which facilitate non-covalent bonding with the receptor sites. The pharmacophore features like hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobic/Non-polar group (H), negatively ionizable (N), positively ionizable (P), and aromatic ring (R) were used to create the pharmacophore sites for the energy-calculated ligands. Tree-based partition algorithm is used by PHASE for detection of common pharmacophore from a set of variants taking maximum tree depth 3. To find common pharmacophore, PHASE algorithm use an exhaustive analysis of k-point pharmacophore match picked from the conformations of a set of active ligands on the basis of inter site distances, and then find all spatial arrangements of pharmacophore features those are common to at least 8 of 10 active ligands. The generated pharmacophores match different set of actives eliminating the chance of its exclusiveness toward a small subset of ligands. The different pharmacophore hypotheses were further examined by using a scoring function, so that it produced the best alignment of the ligands.

#### 4.0.7 Scoring Pharmacophore Hypothesis

The scoring of the pharmacophore hypotheses was done in relation to the information from the active ligands considering various geometric and heuristic factors. The alignment to a reference pharmacophore is considered according to RMSD of the site points and the average cosine of the vectors keeping their tolerance 1.2 Å and 0.5, respectively. To get the reference ligand from the most active set, upper 10% was considered for score calculation. For further refinement, volume scoring was performed in order to measure quantitatively of how each non-reference ligand is superimposing with the reference ligand. Here, the cutoff for volume scoring was kept at 1.00 for the non-reference pharmacophores.

The resulting pharmacophore was then scored and ranked. The scoring was done to identify the best candidate hypothesis, and which provided an overall ranking of all the hypotheses. The scoring algorithm included the contributions from the alignment of site points and vectors, volume overlap, selectivity, number of ligands matched, relative conformational energy, and activity. Among which best hypothesis AAADRRR.190 was selected on the basis of score and discrimination of active and nonactive molecules i.e if active molecules score well, the hypothesis could be invalid as it does not discriminate between active and inactive.

#### 4.0.8 Building of QSAR model

To produce a statistically significant 3D-QSAR model, the first and the foremost requirement is the alignment of ligands; therefore, to execute the QSAR study, a pharmacophore-based alignment was considered. The PHASE algorithm uses a very flexible approach for the development of 3D QSAR model. It considers a rectangular grid of 1 Å grid distance in a 3D space. Thus, it creates cubes of said dimension in the 3D space. The atoms of the molecules which are considered as overlapping Vander Waal spheres fall inside these cubes depending on the volume of the atomic spheres. These occupied cube spaces are termed as volume bits. A volume bit is allocated for each different class of atom that occupies a cube. There are six atom classes, viz. two hydrogen bond acceptor (A), one positively ionizable (P), and two aromatic ring (R) used for classifying the atom characteristics. The total number of volume bits consigned to a specified cube is based on how many training set molecules occupy that cube. A single cube may represent the occupation by one or various atoms or sites, and even those from the same molecule or may be from unlike molecules of the training set. Thus, a molecule may be represented by a binary string concurrent to the occupied cubes, and also the various types of atomic sites that exist in those cubes. To create an atom-based QSAR model, these volume bits which encode the geometrics and chemical characteristics of the molecule are regarded as independent variables in PLS (Partial Least square) regression analysis. The maximum PLS factor that can be taken is  $N/5$ , where  $N$  is the number of ligands present in the training set. In this study, a significant 3D-QSAR model was generated using AAADRRR-190 hypothesis. For QSAR model generation, training and test partition was done by random selection method. Atom-based model selection criterion was chosen for model building. PLS factor was set as 03,

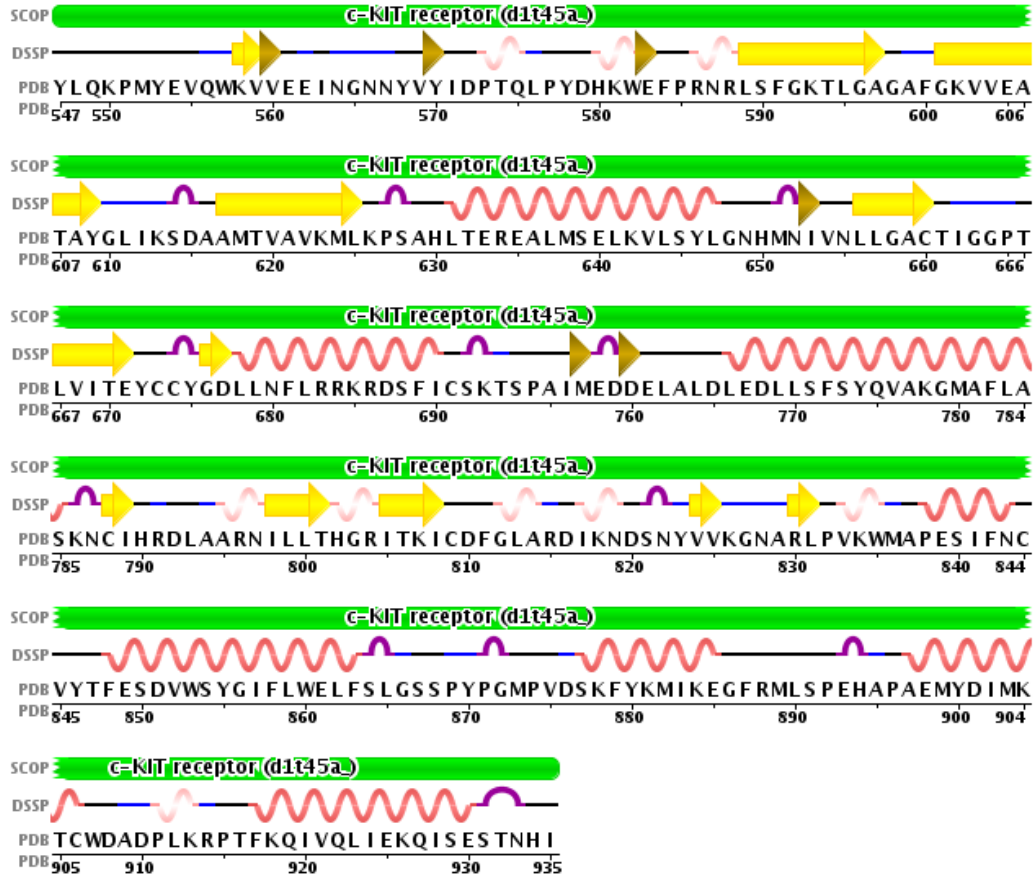
the maximum number of PLS factors in each model can be 1/5 the total number of training set molecules. More the PLS factor value, more will be the reliability of models. Various models have been generated and the best model was selected on the basis of the statistical significance.

# **Results & Discussion**

## 5.0 RESULTS AND DISCUSSION

### 5.0.1 Sequence Analysis of c-Kit and SHP1/2

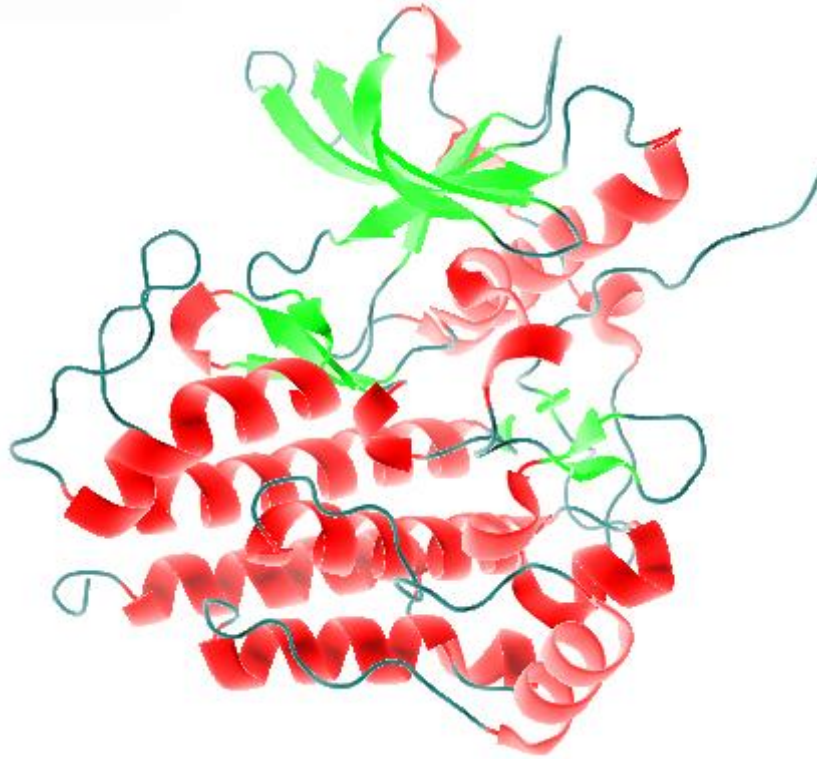
The secondary structure of c-kit shows structural rigidity of active site for SHP1.



**Fig. 6: Secondary structure of c- kit wherein arc depicts turn, yellow arrows depicts beta strand, black line depicts no assigned secondary structure, blue line depicts bend, arrow represents beta bridge and coils depict helix.**

The secondary structure of c-Kit shows that the receptor is 39% helical that is there are nearly 17 helices and 18% beta sheet that is nearly 18 beta sheet strands.

The various domains of c-kit receptor were analyzed in the tertiary structure.



**Fig. 7: Tertiary (3D) structure of c-kit.**

```

Query 3   RWFHRDLSGLDAETLLKGRGVHGSFLARPSRKNQGFSLSVRVGDQVTHIRIQNSGDFYD 62
Sbjct 1   RWFH +++G++AE LL  RGV GSFLARPS+ N GD +LSVR  VTHI+IQN+GD+YD
RWFHPNITGVEAENLLLTRGVDGGSFLARPSKSNPGDLTLSVRRNGAVTHIKIQNTGDYYD 60

Query 63  LYGGEKFATLTELVEYYTQQQGVLQDRDGTIIHLKYPLNCSDPTSERWYHGHS GGQ AET 122
Sbjct 61  LYGGEKFATLAE LVQYYMEHHGQLKEKNGDVIELKYPLNCADPTSERW F HG H L S G K E A E K 120

Query 123 LLQAKGEPWTFVLVRESLSQPGDFVLSVLS--DQPKAGPGSPLRVTHIKVMCEGGRYTVGG 180
Sbjct 121 LLTEKKGKHSFLVRESQSHPGDFVLSVRTGDDKGESNDGKS-KVTHVMIRCQELKYDVGG 179

Query 181 LETFDSLTDLVEHFKKTGIEEASGAFVYLRQPYATRVNAADIENRVLELNKKQES EDTA 240
Sbjct 180 GERFDSLTDLVEHYKKNPMVETLGTVLQLKQPLNTRINAAEIESRVRELSKLAETTDKV 239

Query 241 KAGFWEEFESLQKQEVKNLHQRLLEGQRPENKGNRYKNILPFDHSRVILQGRDSNIPGSD 300
Sbjct 240 KQGFWEEFETLQQQECKLLYSRKEGQRQENKKNRYKNILPFDHTRVVLHDGDPNEPVSD 299

Query 301 YINANYIKNQLLGPDENAK---TYIASQGCLEATVNDFWQMAWQENS RVIVMTTREV EKG 357
Sbjct 300 YINAN I + N+K +YIA+QGCL+ TVNDFW+M +QENS RVIVMTT+EVE+G 359

Query 358 RNKCVPYWPEVGMQRAYGPYSVINCGEHDTTEYKLR TLQVSP LDNGDLIREIWHYQYLSW 417
Sbjct 360 KSKCVKYWPDEYALKEYGVMRVRNVKESAAHDYTLRELKLSKVGGQNTERTVWQYHFRTW 419

Query 418 PDHGVPSPPGGVLSFLDQINQRQESLPHAGPIIVHCSAGIGRTGTIIVIDMLMENISTKG 477
Sbjct 420 PDHGVPSDPGGVLD FLEEVH HKQESIMDAGPVVVHCSAGIGRTGTFIVIDILIDIIREKG 479

Query 478 LDCDIDIQKTIQMVRQRSGMVQTEAQYKFIYVAIAQFIETTKK 521
Sbjct 480 +DCDID+ KTIQMVR+QRSGMVQTEAQY+ IY+A+ +IET+++
VDCDIDVPKTIQMVRQRSGMVQTEAQYRSIYMAVQHYIETSRR 523

```

**Fig. 8: Sequence alignment of SHP1 and SHP2 using BLAST to determine the sequence conservation.**

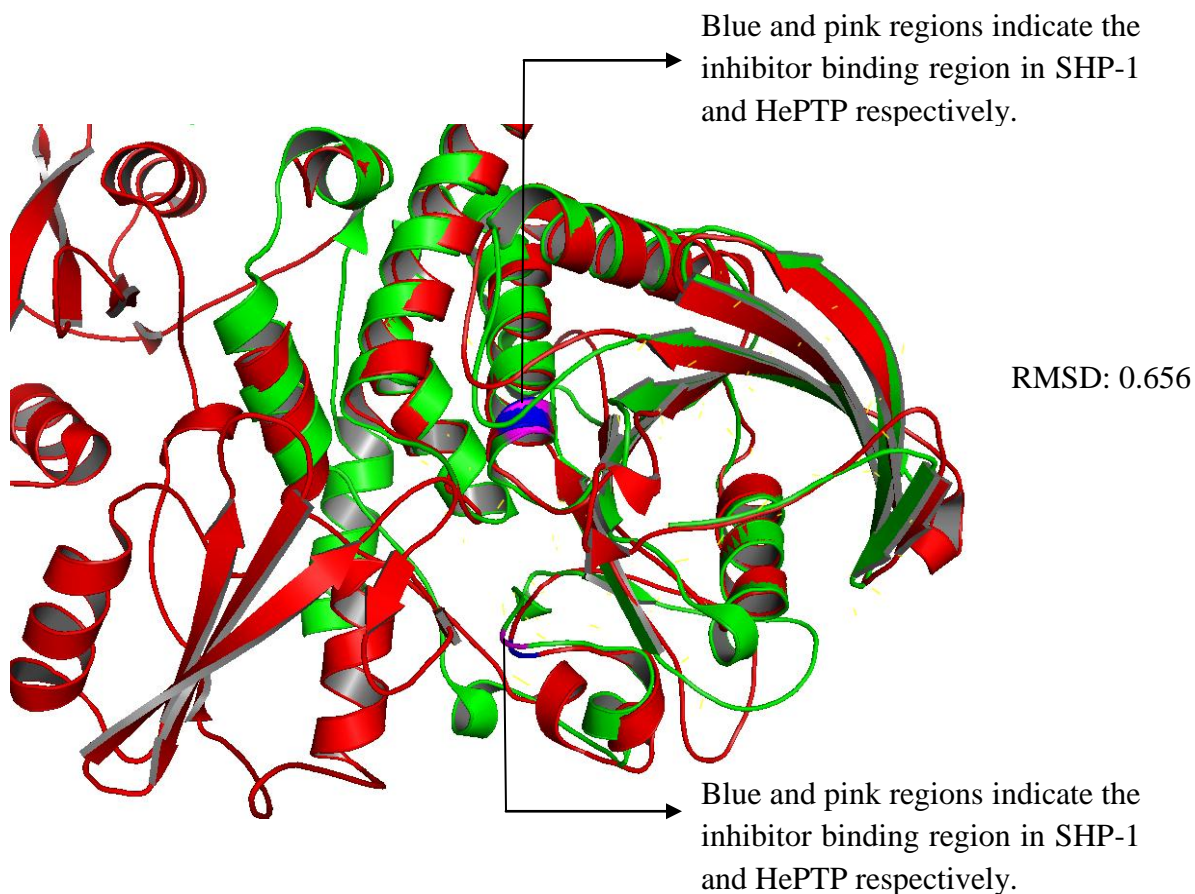
The % identity as determined by the local sequence alignment using BLAST was found to be 59% with query coverage of 97%.

## 5.0.2 Structural Analysis of Protein Tyrosine Phosphatase and Search of its Antagonist

SHP1 and SHP2 were superimposed with other members of the non-receptor PTPase family such as PTP1B, PTPN9 and HePTP and their RMSD values were calculated. The RMSD values of the various members of the non-receptor PTPase family are tabulated in Table 3.

**Table 2: RMSD values of the various members of the non- receptor PTPase**

Members	SHP1/PTPN6	SHP2/PTPN11
PTP1B/PTPN1	0.729	0.665
PTPN9	0.962	0.799
HePTP/PTPN7	0.656	0.673



**Fig. 9: Superimposed structure of SHP-1 (shown in red color) PDB\_ID:2B30 and HePTP (shown in green color) PDB\_ID 1ZCO. The RMSD value obtained from the superimposed structure of SHP-1 and HePTP was found to be 0.656.**



### 5.0.3 Dataset Collection

Superimposition results showed that SHP1 is closely related to HePTP with an RMSD value of 0.656. Hence from the PubChem Compound database, two assay IDs AID 1059 and AID 1077 in which NSC 87877 was tested against HePTP were selected. The inhibitors from both the assay IDs along with their  $IC_{50}$  values were collected (Table 3 and 4).

**Table 3: CIDs and their corresponding IC<sub>50</sub> values under the assay ID (AID) 1059**

S.No.	CID	IC <sub>50</sub> Value	S.No.	CID	IC <sub>50</sub> Value
1.	2928673	2.38	37.	2940938	13.3
2.	16654891	2.41	38.	654761	28
3.	16654890	2.63	39.	16654690	28.3
4.	4039540	2.51	40.	16654689	15.2
5.	16654893	3.87	41.	16654688	0.792
6.	652912	11.3	42.	1357397	1.26
7.	3136927	1.63	43.	5341934	43.9
8.	617227	5.06	44.	5341943	2.4
9.	654089	0.503	45.	3112185	7.58
10.	2859888	2.18	46.	2868734	47.3
11.	3124342	6.05	47.	6456506	0.543
12.	2269367	5.11	48.	3439114	71.4
13.	16654892	11	49.	2276396	3.88
14.	4715351	5.4	50.	3122746	13.2
15.	2869196	3	51.	2901613	5.62
16.	16654632	0.483	52.	2173774	3.63
17.	646406	5.43	53.	3122747	70.2
18.	2914536	2.04	54.	2872935	92.9
19.	2243732	3.26	55.	2920908	44.8
20.	2925672	3.3	56.	2921964	10.4
21.	646096	15.6	57.	3124366	13.7
22.	2865731	12.9			
23.	6492412	8.11			
24.	3115075	21			
25.	717599	3.93			
26.	2207086	6.28			
27.	2930528	2.8			
28.	2975102	9.63			
29.	976017	88			
30.	3453217	3.84			
31.	16654597	3.35			
32.	3053058	41.5			
33.	6400942	10.4			
34.	1780	15.4			
35.	16654691	0.801			

36.	1589738	12.8			
-----	---------	------	--	--	--

**Table 4: CIDs and their corresponding IC<sub>50</sub> values under the assay ID (AID) 1077**

S.No.	CID	IC <sub>50</sub> Value	S.No.	CID	IC <sub>50</sub> Value
1.	24761488	1.172	36.	2266660	5.46
2.	24178237	1.42	37.	24178233	2.56
3.	24178227	0.966	38.	3053058	2.84
4.	892446	0.555	39.	44182133	36.3
5.	889983	1.129	40.	5766720	1.051
6.	1331726	1.71	41.	3157646	2.17
7.	1331726	1.148	42.	1228861	0.44
8.	16330874	0.05	43.	1796598	1.195
9.	24178226	0.814	44.	2258411	35
10.	1324805	1.007	45.	5076888	3.1
11.	24178231	0.968	46.	901652	15.65
12.	2301472	4.625	47.	652912	9.51
13.	2925154	0.923	48.	2585712	2.59
14.	1299158	9.44	49.	6000533	4.637
15.	889170	2.225	50.	1072900	1.605
16.	1209230	9.985	51.	2243732	2.7
17.	654089	2.17	52.	2545524	0.066
18.	24178215	3.458	53.	891589	14.3
19.	16654688	1.76	54.	1213466	8.46
20.	3883207	9.21	55.	20110352	4.253
21.	24178235	2.52	56.	44182131	81.2
22.	2301472	3.085	57.	3000187	4.1
23.	24178230	2.04	58.	2240797	25.099
24.	762708	5.055	59.	4715351	20.5
25.	4039540	4.03	60.	2229326	2.286
26.	2925555	3.01	61.	3239711	6.47
27.	818221	22.2	62.	3157647	0.278
28.	2826665	0.593	63.	1072898	2.705
29.	5336454	1.037	64.	2214811	3.455
30.	2214811	3.655	65.	2924978	2.935
31.	6492412	16.071	66.	3792955	4.9
32.	1587127	1.267	67.	5989418	7.51
33.	1516220	7.081	68.	617227	20.8
34.	3453217	7.98	69.	2928673	4.56

35.	9595032	0.924	70.	3136927	4.17
71.	2859888	7.45	110.	2230267	33.093
72.	2878586	1.346	111.	1282000	27.329
73.	1780	20	112.	1356098	31.4
74.	16654632	2.67	113.	1209211	2.79
75.	3266419	1.051	114.	2901613	12.9
76.	3112185	16.6	115.	2545473	1.745
77.	1357397	0.24	116.	1299058	2.465
78.	646406	4.81	117.	2865731	42
79.	2914536	4.68	118.	1092683	2.3
80.	2869196	13.8	119.	3124366	32.5
81.	2230291	35.111	120.	5765582	4.065
82.	2226406	39.56	121.	3122746	29.5
83.	16654890	31.7	122.	24178232	24.1
84.	16654891	28.5	123.	3115075	58.9
85.	1737079	3.1	124.	4969416	1.38
86.	16654893	24.7	125.	3182456	84.779
87.	2975102	26.6	126.	16654689	73.3
88.	2930528	8.98	127.	16654690	72.1
89.	5341934	4.94	128.	9512029	0.276
90.	2921964	3.49	129.	1435211	6.378
91.	1329592	10.85	130.	5765581	0.32
92.	2300608	5.7	131.	2269367	22.2
93.	4302116	1.838	132.	1328767	0.274
94.	5341943	1.8	133.	44182134	4.167
95.	2940938	31.675	134.	6400942	30.64
96.	1589738	26.1	135.	5504142	75.4
97.	3124342	23.4	136.	2260301	3.955
98.	646096	6.26	137.	16654597	11.6
99.	6456506	0.949	138.	2062730	66.934
100.	2173774	3.568	139.	16654892	40.4
101.	16217011	48.1	140.	2975144	27.611
102.	24178225	2.14	141.	2545467	0.321
103.	5346285	3.677	142.	2925672	10.8
104.	44229065	3.515	143.	2207086	35
105.	44229061	14.999	144.	717599	25
106.	16654691	0.22	145.	2213073	10.775
107.	2012947	4.513	146.	1756795	4.845
108.	8853383	16.665	147.	2276396	2.738

109.	1228895	7.47	148.	3236343	7.38
149.	2924768	15.7			

#### 5.0.4 Virtual Screening

The binding affinity of all the compounds was calculated and tabulated in Table 6.

The inhibitors with their high negative binding affinity greater than NSC87877 can be considered as more specific towards SHP1 and can be considered as more potential inhibitor or antagonist of SHP1.

**Table 5: Compounds from both the assay IDs (AID 1059 and AID1077) and their binding energy with SHP1 (PDB ID: 2B3O)**

S. No.	Compound ID	Binding Affinity	S. No.	Compound ID	Binding Affinity
1.	cid 24178230	-9.5	28.	cid 762708	-8.3
2.	cid 1209211	-9.5	29.	cid 1072900	-8.3
3.	cid 1299058	-9.4	30.	cid 5076888	-8.2
4.	cid 24178225	-9.3	31.	cid 3453217	-8.2
5.	cid 2214811	-9.3	32.	cid 2258411	-8.2
6.	cid 654089	-9.3	33.	cid 901652	-8.2
7.	cid 1331726	-9.3	34.	cid 20110352	-8.2
8.	cid 4715351	-9.3	35.	cid 2240797	-8.2
9.	cid 1789	-9.3	36.	cid 403950	-7.8
10.	cid 4715351	-9.3	37.	cid 6492412	-7.7
11.	cid 128895	-9.2	38.	cid 5766720	-7.7
12.	cid 372955	-9.0	39.	cid 6000533	-7.7
13.	cid 24178231	-8.9	40.	cid 24178237	-7.6
14.	cid 2925154	-8.9	41.	cid 24178227	-7.6
15.	cid 892446	-8.9	42.	cid 3239711	-7.6
16.	cid 2924978	-8.9	43.	cid 3157647	-7.6
17.	cid 3239711	-8.9	44.	cid 1228861	-7.4
18.	cid 24178215	-8.9	45.	cid 5346285	-7.4
19.	cid 24178232	-8.8	46.	cid 2266660	-7.4
20.	cid 901652	-8.8	47.	cid 44182133	-7.3
21.	cid 1329592	-8.7	48.	cid 889983	-7.3
22.	cid 1331766	-8.7	49.	cid 1516220	-7.2
23.	cid 24178226	-8.5	50.	cid 3157646	-7.2
24.	cid 176598	-8.4	51.	cid 16654632	-7.1
25.	cid 24178233	-8.4	52.	cid 2229326	-6.9

26.	cid 2301472	-8.3	53.	cid 3000187	-6.9
27.	cid 1092683	-8.3	54.	cid 3266419	-6.9
55.	cid 2878586	-6.9	94.	cid 3883207	-5.8
56.	cid 1213466	-6.9	95.	cid 5336454	-5.8
57.	cid 891589	-6.9	96.	cid 2826665	-5.8
58.	cid 3157647	-6.9	97.	cid 2266660	-5.7
59.	cid 1072898	-6.9	98.	cid 9595032	-5.6
60.	cid 1357397	-6.9	99.	cid 5765582	-5.6
61.	cid 646406	-6.8	100.	cid 1756795	-5.6
62.	cid 2928673	-6.8	101.	cid 2260301	-5.6
63.	cid 3136927	-6.5	102.	cid 1435211	-5.6
64.	cid 44229061	-6.5	103.	cid 3182456	-5.6
65.	cid 2012947	-6.5	104.	cid 44182134	-5.5
66.	cid 8853383	-6.5	105.	cid 2975144	-5.5
67.	cid 2230267	-6.3	106.	cid 254573	-5.5
68.	cid 1282000	-6.3	107.	cid 5504142	-5.5
69.	cid 2230291	-6.3	108.	cid 2062730	-5.5
70.	cid 2226406	-6.3	109.	cid 4969416	-5.4
71.	cid 4302116	-6.3	110.	cid 2545467	-5.4
72.	cid 1329592	-6.3	111.	cid 5765581	-5.3
73.	cid 16654893	-6.2	112.	cid 9512029	-5.2
74.	cid 16654891	-6.1	113.	cid 1328767	-5.1
75.	cid 16654890	-6.1	114.	cid 24178232	-5.0
76.	cid 16654691	-6.1	115.	cid 71599	-5.0
77.	cid 646096	-6.1	116.	cid 2207086	-5.0
78.	cid 2940938	-6.1	117.	cid 16654969	-5.0
79.	cid 1589738	-6.1	118.	cid 2269367	-5.0
80.	cid 16217011	-6.0			
81.	cid 2869196	-6.0			
82.	cid 2975102	-6.0			
83.	cid 2930528	-6.0			
84.	cid 2300608	-6.0			
85.	cid 2921964	-6.0			
86.	cid 2914536	-5.9			
87.	cid 2173774	-5.9			
88.	cid 5341943	-5.9			
89.	cid 65456566	-5.9			
90.	cid 24761488	-5.9			
91.	cid 818221	-5.9			

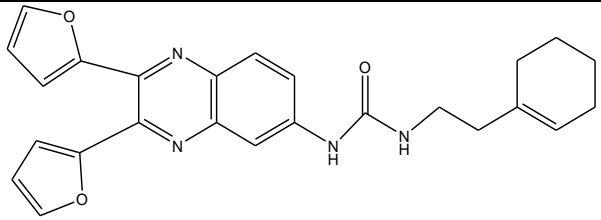
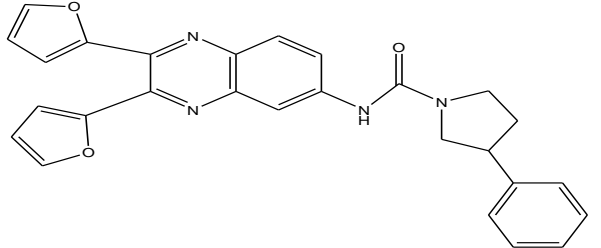
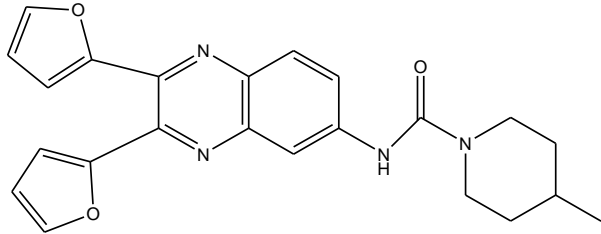
92.	cid 889170	-5.9			
93.	cid 1209230	-5.9			

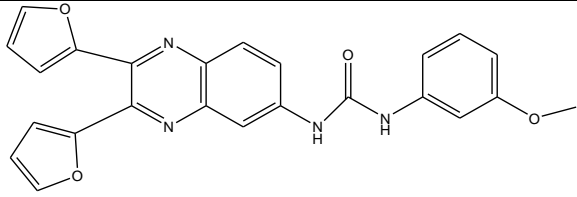
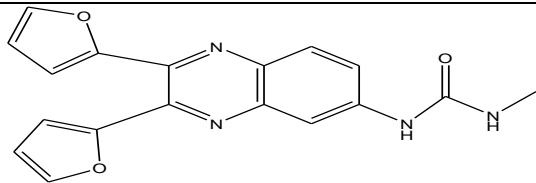
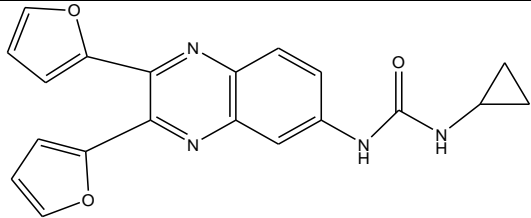
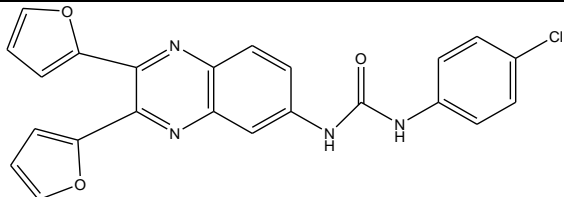
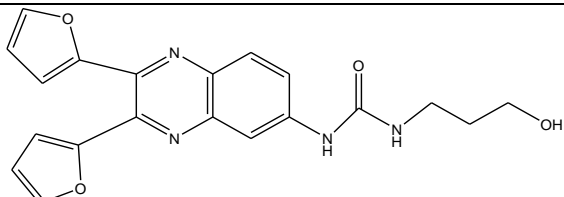
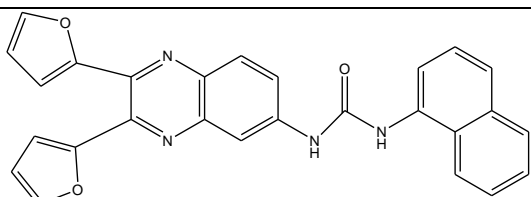
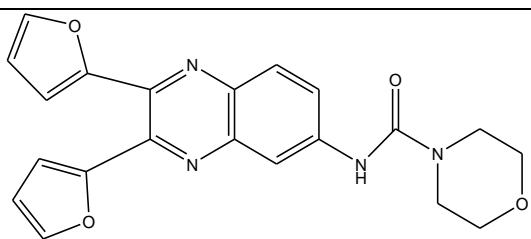
Table 6 shows that compounds with binding affinity less than NSC87877 (CID: 16654632) are more specific for SHP1. Hence it can be concluded that the compounds with better binding efficiency than NSC87877 can prove to be more potent inhibitors of SHP1.

### 5.0.5 Dataset Preparation

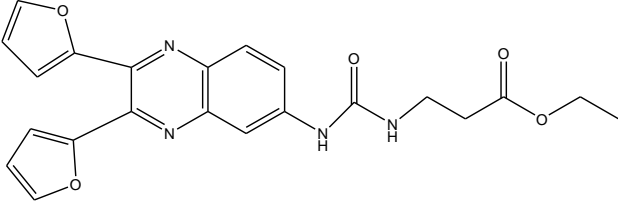
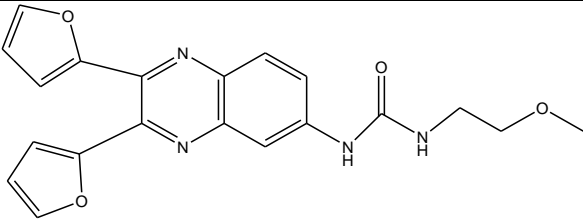
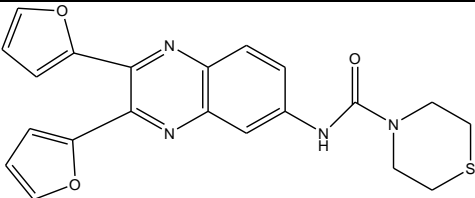
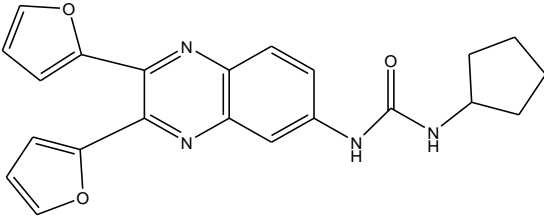
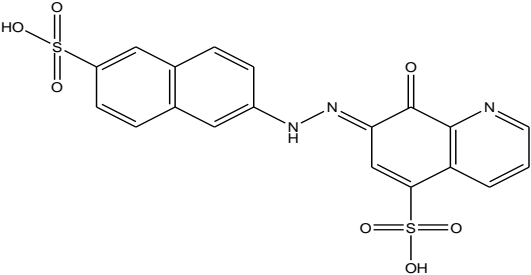
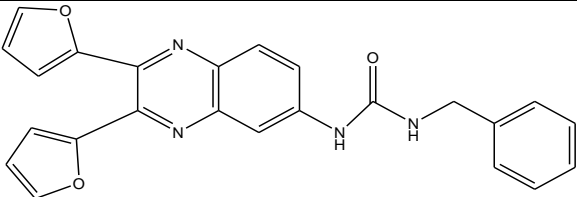
The inhibitors having higher binding affinity than NSC87877 further underwent pharmacophore modeling to identify activity of inhibitors on the basis of structural similarity. Table 7 shows the compounds that were selected for pharmacophore and QSAR studies along with their structure,  $-\log IC_{50}$  values and binding energy.

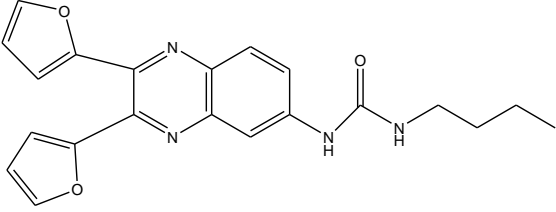
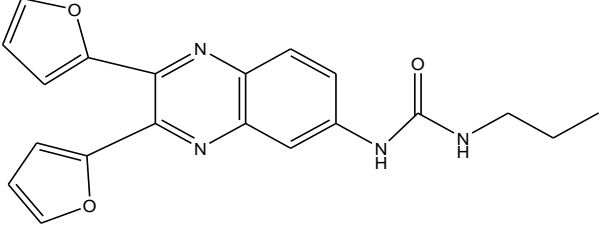
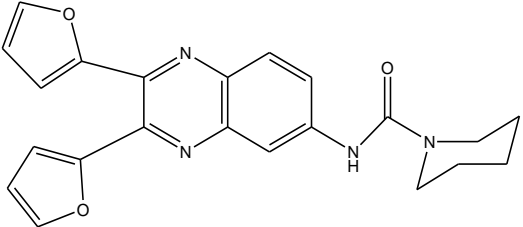
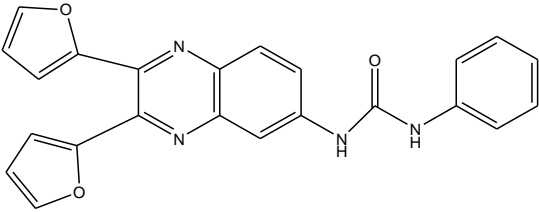
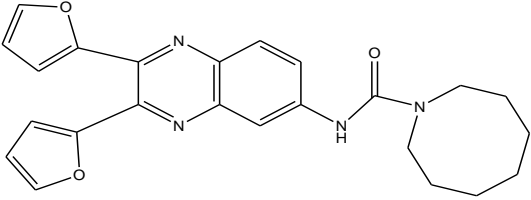
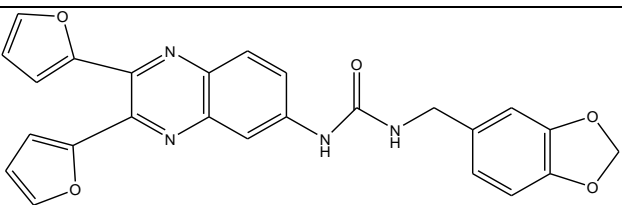
**Table 6: Compounds which were selected for pharmacophore and QSAR studies along with their structure,  $pIC_{50}$  values and binding energy.**

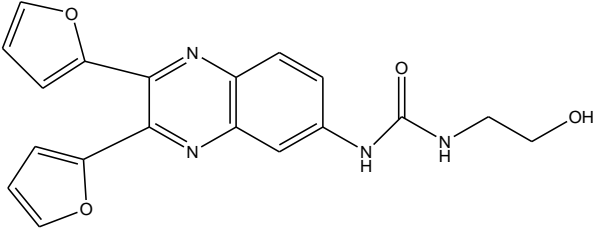
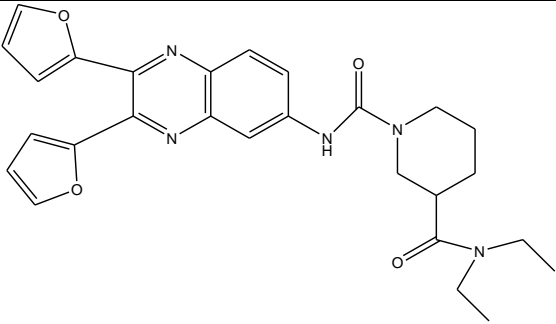
S. No.	Compound ID	Structure	$-\log IC_{50}$	B.E
1.	CID 1209211		5.560	-9.5
2.	CID 2925154		6.250	-8.9
3.	CID 2924978		6.070	-8.9

4.	CID 762708		5.550	-8.3
5.	CID 1092683		6.080	-8.3
6.	CID 1329592		5.400	-8.7
7.	CID 24178225		6.020	-9.3
8.	CID 2301472		6.070	-8.3
9.	CID 24178230		5.890	-9.5
10.	CID 1072900		6.150	-8.3



11.	CID 24178237		6.160	-7.6
12.	CID 3157647		6.700	-7.6
13.	CID 24178215		5.470	-8.9
14.	CID 24178233		5.950	-8.4
15.	CID 16654632		5.530	-7.1
16.	CID 24178232		5.810	-8.8

17.	CID 24178227		6.400	-7.6
18.	CID 3239711		5.400	-7.6
19.	CID 372955		5.530	-9.0
20.	CID 24178231		6.240	-8.9
21.	CID 2214811		5.730	-9.3
22.	CID 1299058		5.670	-9.4

23.	CID 1331766		6.000	-9.3
24.	CID 24178226		6.320	-8.5

The B.A and the  $\log IC_{50}$  value of most of the compounds in table 7 are in correlation to each other which means that the B.A of a compound with lower  $IC_{50}$  value is high. This indicates that the particular compound/inhibitor is interacting with the residues in the active site for the inhibitor with higher affinity. Some of the other compounds have B.A higher or lower in comparison to their high or low  $IC_{50}$  value. Higher B.A and high  $IC_{50}$  value of a compound may be due to the reason that the compound does not bind at the active site for the inhibitor and instead binds to a region other than the inhibitor binding site. While a lower B.A and low  $IC_{50}$  value may be attributed to the fact that the compound binds to the inhibitor binding site but with lesser affinity and thus has loose interaction with the residues in the binding site.

### 5.0.6 Pharmacophore Modeling

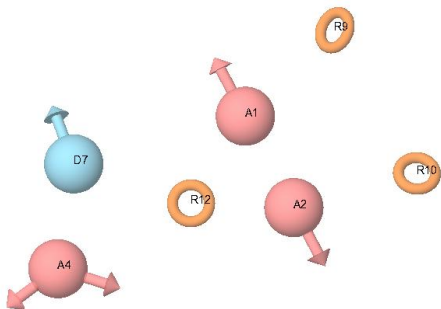
Several seven- point common pharmacophore hypotheses were generated using active molecules with various combinations of site. Minimum and maximum sites considered to obtain optimum combinations of sites or features common to active compounds were 4 and 7 respectively. Identification of the pharmacophore features taking into consideration the highest active molecule was done by classifying the compounds into active and inactive categories based on the activity threshold. The survival scoring(s) function was used which identifies the best candidate from the generated models and assigns an overall ranking of all the hypotheses. The scoring algorithm includes contributions from the alignment of site points and vectors, volume overlap, selectivity, relative conformational energy, activity and number of ligands matched. However, the model should be able to discriminate between active and inactive molecules. Hypothesis generated and their scores are listed in Table 8. The hypotheses AAADRRR.190 was selected from among the various hypotheses generated based on the score and discrimination of active and inactive molecules. The best pharmacophore hypotheses AAADRRR.190 was selected for further QSAR study. The 3D pharmacophore hypotheses shows the following features: 3

hydrogen bond acceptor (A) in pink color, 1 hydrogen bond donor (D) in blue color and 3 aromatic rings (R) in yellow color. The best hypotheses was selected based on the survival score. A higher value of survival score indicates better fitness of the active ligands on the common pharmacophore and validates the model. Table 8 shows that AAADRRR.190 has the best survival score (3.581) amongst the hypotheses generated. The inactive survival score, vector score, site score, volume score and selectivity of the selected hypotheses AAADRRR.190 was calculated to be 1.787, 0.948, 0.880, 0.757 and 2.579 respectively.

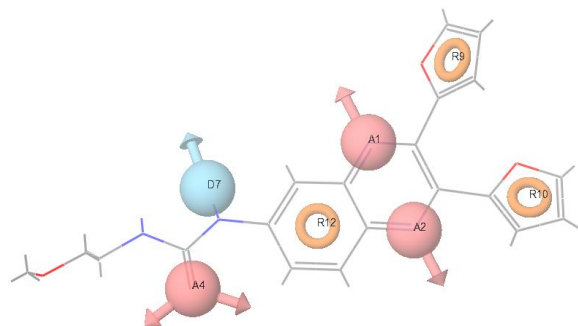
**Table 7: Various PHASE hypotheses generated**

Phase hypothesis	Survival	Survival inactive	Vector	Site	Volume	Selectivity
AAADRRR.190	3.581	1.787	.948	.88	.757	2.579
AAADRRR.263	3.543	1.769	.947	.87	.730	2.476
AAADRRR.235	3.540	1.755	.947	.87	.720	2.475

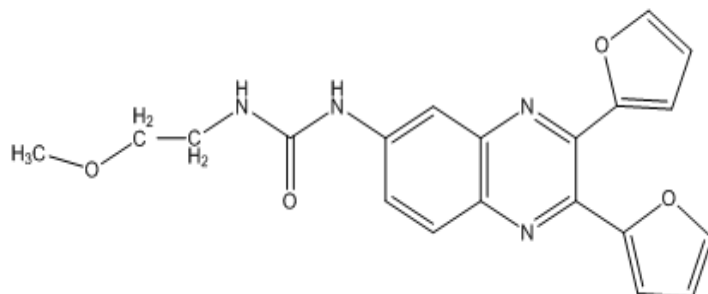
10a)



10b)



10c)



**Fig. 10:** 10a and 10b shows the common pharmacophore for active ligands which has the following features: 3 hydrogen bond acceptors (A) in pink color, 1 hydrogen bond donor (D) in blue color and 3 aromatic rings (R) in yellow color. 10c shows the 2D representation of common pharmacophore.

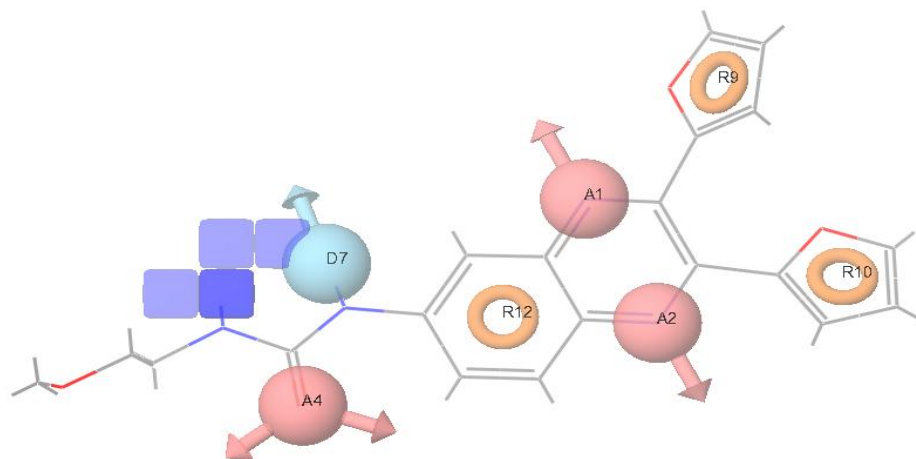
### 5.0.7 Building of 3D QSAR model

3D QSAR study was successfully performed on a series of inhibitors selected in order to understand the effect of the spatial arrangement of the structural features on the biological activity of the molecules selected. Figure 10 shows the results of the 3D QSAR study performed.

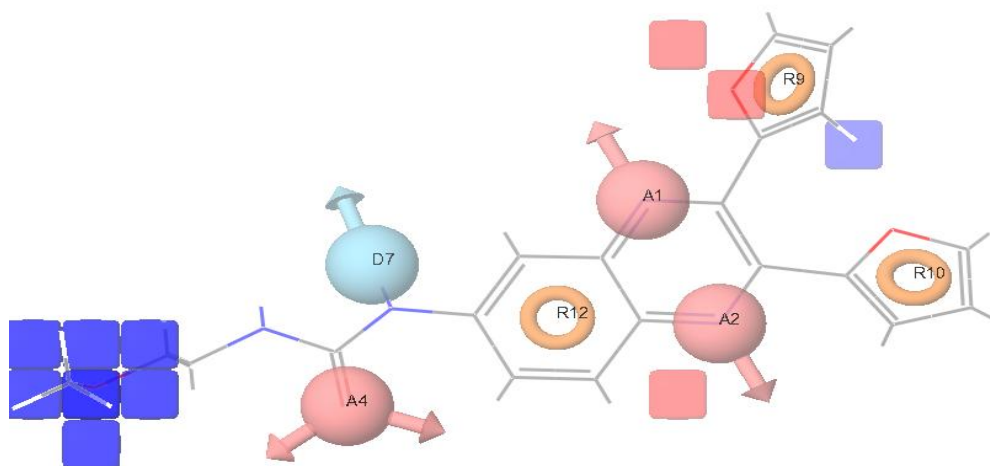
The blue cubes in the 3D plot of the 3D representation of the 3D pharmacophore refers to the sites or regions of the ligands in which a particular feature is favorable for the biological activity of the molecule, whereas the red cubes indicates those sites or regions which are unfavorable for the biological activity of the molecules with respect to a particular feature. Figure 11a shows that addition of a donor group is favorable at the carbon just adjacent to the hydrogen bond acceptor A4 and hence can lead to increase in the biological activity. Figure 11b shows that the addition of an electron withdrawing group to the carbon chain and near the aromatic ring R9 is favorable. Figure 11c shows that the addition of a hydrophobic group near the aromatic rings R9 and R10 as well as to the carbon chain is favorable, whereas the addition of a hydrophobic group in the region near the hydrogen bond donor D7 is unfavorable.

The reliability of the 3D QSAR analysis on the selected hypotheses is justified from the fact that all the statistical parameters calculated are significant. Table 9 shows the statistical results of the 3D QSAR study.

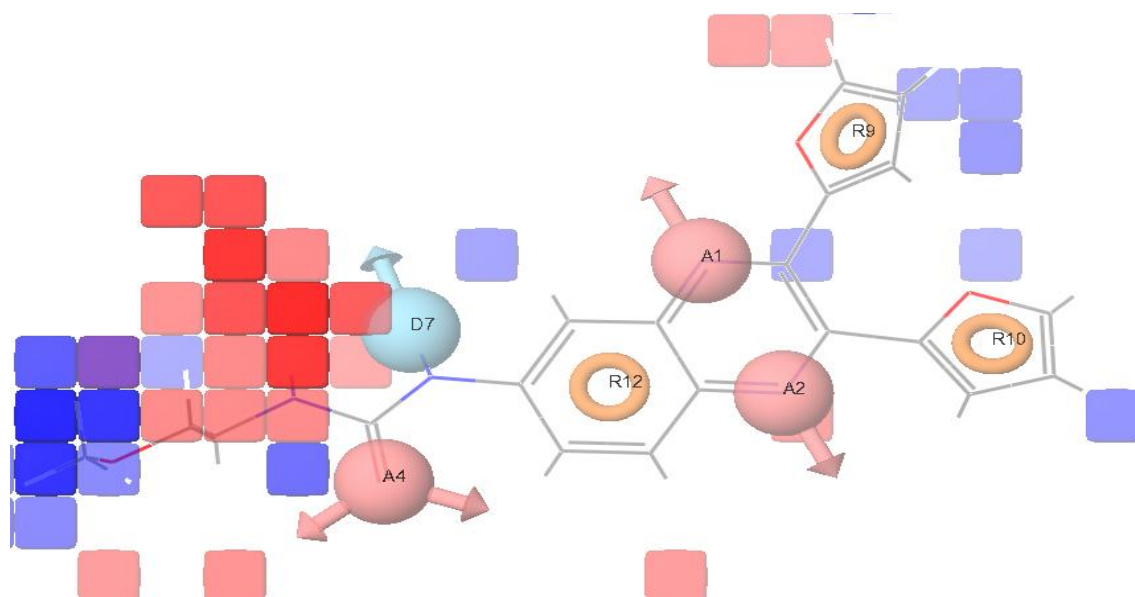
11a)



11b)



11c)



**Fig. 11: QSAR visualization of various substituent's effect: 11a) hydrogen-bond donor effect, 11b) electron withdrawing feature, 11c) hydrophobic effect.**

**Table 8: Statistical result of 3D QSAR study**

PLS factors	SD	$r^2$	F	P	RMSE	$q^2$	Pearson R
1.	0.2271	0.6496	27.8	9.378e-005	0.1719	0.6242	0.9546
2.	0.1275	0.897	60.9	1.233e-007	0.1198	0.8173	0.9467
3.	0.0817	0.9607	105.9	2.184e-009	0.1749	0.6111	0.711

SD is standard deviation of the regression,  $r^2$  is regression, and F is variance ratio. Larger value of F indicates a statistically more significant regression; P is the significance level of variance ratio. Smaller value indicates a greater degree of confidence. RMSE is the root-mean-square error,  $q^2$  indicates the predicted activity, and Pearson R value indicates the correlation between the predicted and observed activity for the test set.

Validity of the model generated can be expressed by internal predictivity ( $q^2$ ) which is 0.81 in this case. The internal predictivity can be obtained by leave-one-out (LOO) method.  $q^2$  is a more reliable statistical parameter than  $r^2$  as it is obtained by external validation method by dividing

the dataset into training and test set. Larger value of F (60.9) indicates a statistically significant model. It is further supported by a smaller value of P (1.233e-007) which is an indication of a higher degree of confidence. Also the smaller values of SD (0.1275) and RMSE (0.1198) indicate that the model or data used for QSAR analysis is best. The PLS factor was taken as 3 in this study. PLS factor is also another parameter that confirms the reliability of the model generated. Table 10 shows the fitness and PHASE predicted activity of the training and the test set compounds.

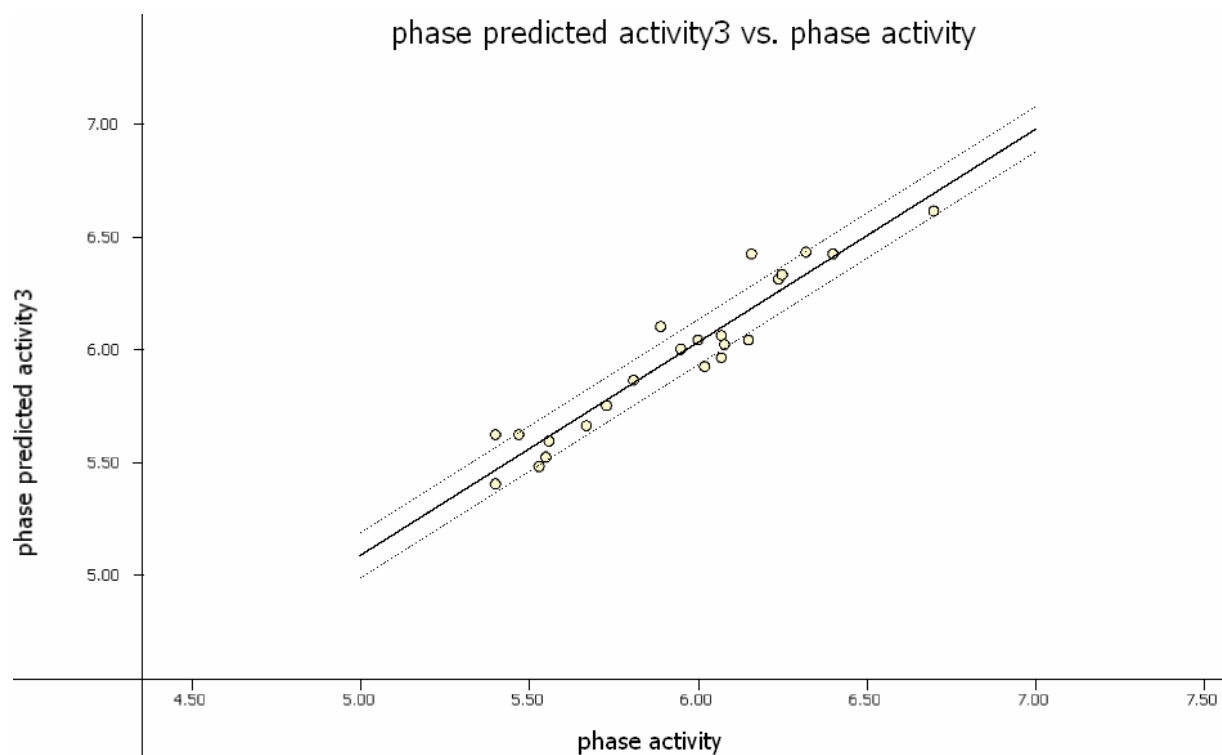
The fitness graph (Fig.12) between the observed activity and the PHASE predicted activity of the training and test set compounds indicates the active compounds are closely fitted to the regression line while the inactive compounds are scattered. The solid line in the graph indicates the hypothetical best fit line between the predicted and experimental activity.



**Table 9: Fitness and PHASE predicted activity of the compounds**

<b>CID</b>	<b>QSAR set</b>	<b>Activity</b>	<b>PLS factors</b>	<b>Phase predicted activity</b>	<b>Pharm set</b>	<b>Fitness</b>
1209211	Training	5.560	1 2 3	5.89,5.56,5.59	Inactive	2.75
2925154	Test	6.250	1 2 3	6.16,6.36,6.33	Active	2.57
2924978	Test	6.070	1 2 3	5.87,5.99,5.96	Active	2.16
762708	Training	5.550	1 2 3	5.62,5.52,5.60	Inactive	1.95
1092683	Training	6.080	1 2 3	6.06,6.12,6.02	Active	2.34
1329592	Training	5.400	1 2 3	5.47,5.34,5.40	Inactive	2.34
24178225	Training	6.020	1 2 3	6.22,6.09,5.92	Active	2.33
2301472	Training	6.070	1 2 3	6.21,6.14,6.06	Active	2.51
24178230	Test	5.890	1 2 3	6.12,6.04,6.10	Active	2.77
1072900	Training	6.150	1 2 3	5.72,5.99,6.04	Active	2.12
24178237	Test	6.160	1 2 3	6.24,6.27,6.42	Active	2.89
3157647	Training	6.700	1 2 3	6.26,6.39,6.61	Active	3.00
24178215	Training	5.470	1 2 3	5.60,5.69,5.62	Inactive	2.13
24178233	Training	5.950	1 2 3	5.95,5.95,6.00	Active	2.71
16654632	Training	5.530	1 2 3	5.44,5.44,5.48	Inactive	2.34
24178232	Test	5.810	1 2 3	6.02,5.81,5.86	Active	2.79
24178227	Training	6.400	1 2 3	6.28,6.43,6.42	Active	2.68
3239711	Test	5.400	1 2 3	5.56,5.59,5.62	Inactive	2.35
372955	Test	5.530	1 2 3	5.54,5.46,5.52	Inactive	2.66

24178231	Training	6.240	1 2 3	6.30,6.37,6.31	Active	2.71
2214811	Training	5.730	1 2 3	5.84,5.69,5.75	Active	2.77
1299058	Training	5.670	1 2 3	5.96,5.62,5.66	Inactive	2.73
1331726	Training	6.000	1 2 3	5.72,6.09,6.04	Active	2.22
24178226	Training	6.320	1 2 3	6.30,6.32,6.43	Active	2.84



**Figure 12: Fitness graph between observed activity versus PHASE predicted activity for training and test set compounds.**

**Conclusion**  
**&**  
**Future Perspective**

## 6.0 CONCLUSION AND FUTURE PERSPECTIVE

Radiation therapy in cancer leads to loss of a large number of immune cells, so HSCs are transplanted into the bone marrow of irradiated patients but HSCs differentiate before reaching the bone marrow. Hence, the focus of our study was to find and develop strategies by which HSCs can be maintained in their undifferentiated state and enhancing proliferation until they reach their target site i.e bone marrow. A number of signaling pathways, ligands and genes regulate the proliferative and self-renewal properties of the HSCs. One of the signaling pathways that plays a crucial role in the process of hematopoiesis is the SCF (Stem Cell Factor) and c-kit pathway. SCF binding induces c-kit receptor dimerization and a cascade of signal transduction is activated. As a result of the signaling cascade activation Tyr568 and Tyr570 residues present on the juxtamembrane domain of c-kit are phosphorylated. The phosphorylated Tyr568 and Tyr570 residues are the binding site of its negative regulators. SHP1 and SHP2 are members of the non-receptor tyrosine phosphatase which negatively regulates c-kit signaling. So to increase the proliferation through c-kit, we have targeted SHP-1. In this work both structure based (docking) and ligand based (pharmacophore modeling and 3D QSAR) approaches have been used to design SHP-1 specific inhibitor. A dataset of 24 compounds was built to carry out pharmacophore modeling and QSAR studies using PHASE module of Schrodinger. The dataset was classified into active and inactive molecules based on their  $IC_{50}$  values. Among the various hypotheses generated based on the score and discrimination of active and inactive molecules, hypotheses AAADRRR190 (survival score: 3.581) was selected as the best pharmacophore hypotheses on which further QSAR study was carried out. The 3D QSAR result of hypotheses AAADRRR190 was found to be statistically good and significant. All the molecules showed good alignment with good fitness ranging from 3.00 (for most active) to 1.95 (for least active). QSAR model generation was carried out by randomly dividing the dataset into training set of 17 compounds and test set of 7 compounds. For good model generation 70% of the compounds in the dataset should be in the training set and 30% in the test set. PLS factor was taken as 3 as the maximum number of PLS factors in each model can be 1/5 of the total number of training set compounds. More is the number of PLS factors, more is the reliability of the model. The reliability of the 3D QSAR model generated is proved by the fact all the statistical measures are significant. The model generated showed statistically good results with  $r^2$  (Correlation coefficient) as 0.897,  $q^2$  (Leave one out cross validation) as 0.8173. The correlation coefficient  $r$  signifies how closely the observed data tracks the fitted regression line. Leave one out cross validation involves using a single observation from original sample as the validation determinant and remaining observation as training data. This process is repeated such that each observation in the sample is used once as validation data. The reliability and the statistical significance of the model can also be confirmed by a higher value of Fischer ratio which is 60.9 in our case.

The future perspective of this study is that various combinations of hydrogen bond donor, electron withdrawing and hydrophobic groups can be added at various favorable sites of the generated model thereby building new substituents. The physiochemical properties of the new

substituents can be validated by Lipinski rule of 5. The substituents which pass the Lipinski rule filter can then be docked with SHP1. The substituents which satisfy both the above conditions (i.e. pass the Lipinski filter and exhibit B.A. better than the reference) can then be tested using 3D-QSAR model generation.

A good drug should exhibit a good ADMET profile. Once a compound clears the ADMET profile filter, wet lab synthesis along with the characterization of that compound can be carried out.

# References

## 7.0 REFERENCES

- 1) Alonso A., Sasin J., Bottini N., Friedberg I., Osterman A., Godzik A., Hunter T., Dixon J., Mustelin T (2004). Protein tyrosine phosphatases in the human genome. *Cell*.117, 699–711.
- 2) Anderson J.N., Mortensen O. H., Peters G. H., Drake P. G., Iversen L. F., Olsen O. H., Jansen P. G., Andersen H. S., Tonks N. K., Moller N. P. (2001). Structural and evolutionary relationships among protein tyrosine phosphatase domains. *Mol. Cell. Biol.* 21, 7117–7136.
- 3) Blank U., Karlsson G., Karlsson S. (January 2008). Signaling pathways governing stem cell fate. *Blood* Vol. 111 No. 2.
- 4) Bongso A., Eng Hin Lee (2005). Stem cells: their definition, classification and sources. *Stem Cells: From Benchtop to Bedside* .12, 8–12.
- 5) Bug G., Gül H., Schwarz K., Pfeifer H., Kampfmann M., Zheng X., Beissert T., Boehrer S., Hoelzer D., Ottmann O. G., Ruthardt M. (2005). Valproic Acid Stimulates Proliferation and Self-renewal of Hematopoietic Stem Cells. *Cancer Res* 2005; 65:2537-2541.
- 6) Chotinantakul K., Leraanaksiri W. (June 2012). Hematopoietic stem cell development, niches and signaling pathways. Hindawi Publishing Corporation Bone Marrow Research Vol. 2012.
- 7) Dixon J. E., Denu J. M.(1998). Protein tyrosine phosphatases: mechanisms of catalysis and regulation. *Curr Opin Chem Biol.* 5, 633-41.
- 8) Dror. O. (2006). Predicting molecular interactions in silico. An updated guide to pharmacophore identification and its applications to drug design. *Front Med.Chem.*3, 551-584.
- 9) Gall S, Zseb K.M., Geissler E.N. (1994). The kit ligand, stem cell factor. *Adv. Immun.* 55, 1-96.
- 10) Geronikaki A., Eleftheriou P., Vicini P., Dixit A., Saxena A. K. (2008). 2-Thiazolylimino/ Heteroarylimino-5-arylidene-4-thiazolidinones as new agents with SHP-2 inhibitory action. *J. Med. Chem.* 2008, 51, 5221-5228.
- 11) John G. Topliss. (1983). *Quantitative Structure-Activity Relationships of Drugs*, Academic Press, New York .337, 1–13.
- 12) Kaur K., Mirlashari M. R., Kvalheim G., Kjeldsen-Kragh J. (2013). 3',4'-Dimethoxyflavone and valproic acid promotes the proliferation of human hematopoietic stem cells. *Stem Cell Research & Therapy* 2013, 4:60.
- 13) Kozlowski M.L., Larose F., Lee D.M., Rottapel R., Siminovitch K.A. (1998). SHP-1 binds and negatively modulates the c-kit receptor by interaction with tyrosine 569 in the c-kit juxtamembrane domain. *Mol.Cell. Biol.* 18, 2089- 2099.
- 14) Lawrence H. R., Pireddu R., Chen L., Luo Y., Sung S- S., Szymanski A. M., Yip M. L. R., Guida W. C., Sebt S. M., Wu J., Lawrence N. J. (2008). Inhibitors of Src homology-2 domain containing protein tyrosine phosphatase- 2 (SHP-2) based on oxindole scaffolds. *J. Med. Chem.* 2008, 51, 4948- 4956.

- 15) Park S. J., Song M., Cho S. (2009). Regulation of Vaccinia H1- related (VHR) phosphatase activity by NSC- 87877. *Bull. Korean Chem. Soc.* 2009, Vol. 30, No. 12.
- 16) Pietras E. M., Warr M.R., Passegué E. (November 2011). Cell cycle regulation in hematopoietic stem cells. *J. Cell Biol.* Vol. 195 No. 5 709- 720.
- 17) Rönstrand L. (2004). Signal transduction via the Stem cell factor receptor/ c-kit. *Cell. Mol. Life Sci.* 61, 2535-48.
- 18) Roskoski R. (2005). Signaling by Kit protein tyrosine kinase- The stem cell factor receptor. *Biochemical and Biophysical Research Communications* 337 1-13.
- 19) Seita J., Weissman I. (2010). Hematopoietic Stem Cell: Self- renewal versus Differentiation. *Wiley Interdiscip Rev Syst Biol Med.* 2010; 2(6): 640- 653, doi: 10.1002/wsbm.86.
- 20) Simone. B. (2009). Pharmacophore modeling: A continuously evolving tool for computational drug design. *New perspectives in medicinal chemistry. 1*, 13-23.
- 21) Song M., Park J. E., Park S. G., Lee D. H., Choi H-K., Park B. C., Ryu S. E., Kim J. H., Cho S. (2009). NSC- 87877 inhibitor of SHP- 1/2 PTPs, inhibits dual- specificity phosphatase 26 (DUSP26). *Biochemical and Biophysical Research Communications* 381, 491- 495.
- 22) Song M., Park S. J., Cho S. (2009). Inhibition of Acid phosphatase 1 (ACP1) activity by NSC- 87877. *Bull. Korean Chem. Soc.* 2009, Vol. 30, No. 1.
- 23) Song M., Cho S. (2009). Regulation of Dual- specificity phosphatase 14 (DUSP14) by NSC- 87877. *Bull. Korean Chem. Soc.* 2009, Vol. 30, No. 5.
- 24) Song M., Cho S. (2009). NSC- 87877 inhibits Dual- specificity phosphatase 23 (DUSP23) that regulates ERK. *Bull. Korean Chem. Soc.* 2009, Vol. 30, No. 8.
- 25) Song M., Cho S. (2010). NSC- 87877 inhibits cell growth by suppressing ERK Signaling. *Bull. Korean Chem. Soc.* 2010, Vol. 31, No. 9.
- 26) Tuch B. E (2006). Stem cells- A Clinical Update. *Australian Family Physician* 35 719-21.
- 27) Yu WM., Guvench O., MacKerell A. D., Qu CK. (2008). Identification of small molecular weight inhibitors of Src Homology 2 domain containing tyrosine phosphatase 2 (SHP-2) via in silico database screening combined with experimental assay. *J. Med. Chem.* 2008, 51, 7396- 7404.
- 28) Zhang X., He Y., Liu S., Yu Z., Jiang ZX., Yang Z., Dong Y., Nabinger S. C., Wu L., Gunawan A. M., Wang L., Chan R. J., Zhang ZY. (2010). Salicylic acid based small molecule inhibitor for the oncogenic Src Homology- 2 domain containing protein tyrosine phosphatase- 2 (SHP- 2). *J. Med. Chem.* 2010, 53, 2482- 2493.
- 29) Zon L.I. (2008). Intrinsic and extrinsic control of haematopoietic stem-cell self-renewal. *Nature* Vol. 453.
- 30) Yvonne C., Martin A. (1981). Practitioner's Perspective of the Role of Quantitative Structure Activity Analysis in Medicinal Chemistry, *J. Med. Chem.*, 24, 229.



# Appendix

## 8.0 APPENDIX

### Protein Tyrosine Phosphatases

A cornerstone of many cell-signalling events rests on reversible phosphorylation of tyrosine residues on proteins. The reversibility relies on the co-ordinated actions of protein tyrosine kinases and protein tyrosine phosphatases (PTPs), both of which exist as large protein families (Stoker *et al.*, 2005). PTPs regulate a wide range of signalling pathways. PTPs work antagonistically with Protein Tyrosine Kinases (PTKs) and inhibit cell proliferation. PTPs are a group of enzymes that remove phosphate groups from phosphorylated tyrosine residues on proteins. Phosphorylation of proteins is one of the posttranslational modifications, which is reversible and plays a critical role in the regulation of many cellular functions. As a consequence, maintaining an appropriate level of protein tyrosine phosphorylation is essential for many cellular functions (Anderson *et al.*, 2001). Tyrosine-specific protein phosphatases catalyse the removal of a phosphate group attached to a tyrosine residue. These enzymes are key regulatory components in signal transduction pathways cell cycle control, and are important in the control of cell growth, proliferation, differentiation and transformation. PTPs have been implicated in regulation of many cellular processes, such as cell growth, cellular differentiation, Mitotic cycles, Oncogenic transformation (Anderson *et al.*, 2001).

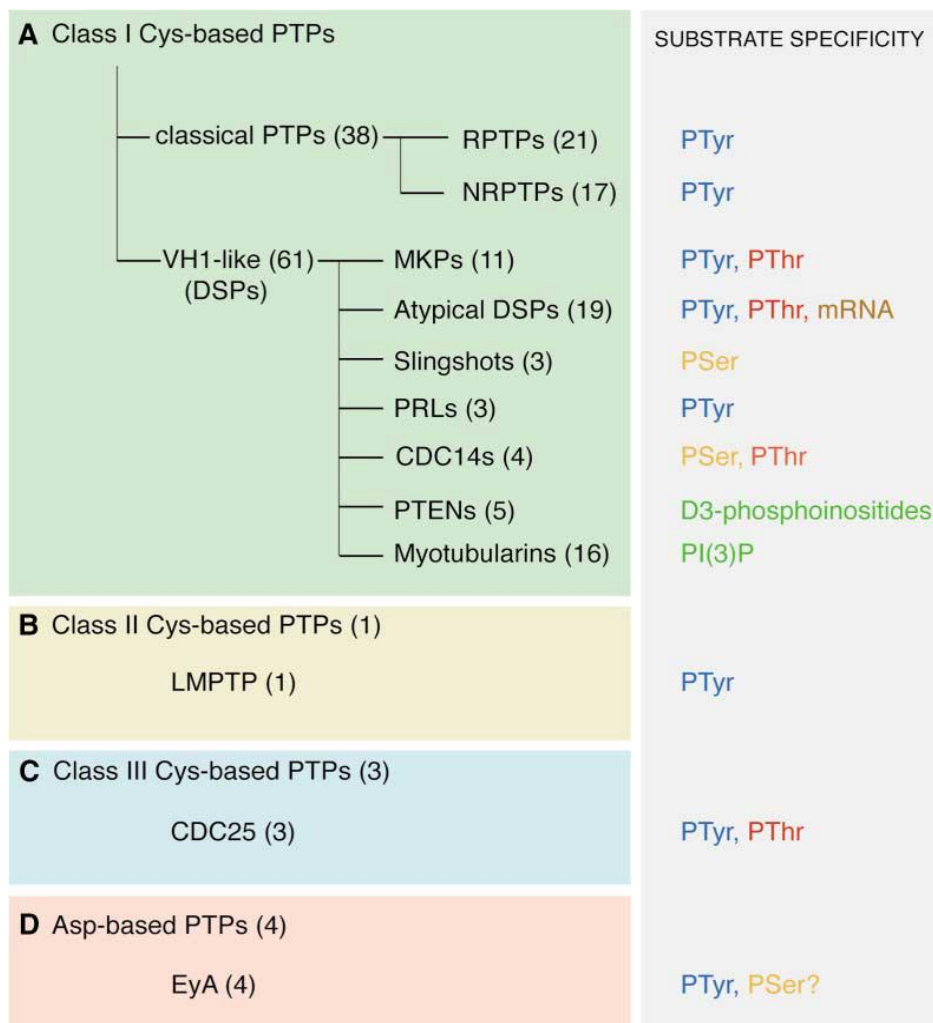
All PTPases carry the highly conserved active site motif C(X)5R (PTP signature motif), employ a common catalytic mechanism, and possess a similar core structure made of a central parallel beta-sheet with flanking alpha-helices containing a beta-loop-alpha-loop that encompasses the PTP signature motif (Dixon *et al.*, 1998)

Individual PTPs may be expressed by all cell types, or their expression may be strictly tissue-specific. Most cells express 30% to 60% of all the PTPs, however hematopoietic and neuronal cells express a higher number of PTPs in comparison to other cell types. T cells and B cells of hematopoietic origin express around 60 to 70 different PTPs. The expression of several PTPs is restricted to hematopoietic cells, for example, LYP, SHP1, CD45, and HePTP (Alonso *et al.*, 2004). Of the 107 PTP genes, 11 are catalytically inactive 2 dephosphorylate mRNA and 13 dephosphorylate inositol phospholipids. Thus, 81 PTPs are active protein phosphatases with the ability to dephosphorylate phosphotyrosine (Alonso *et al.*, 2004).

### Classification of PTPs

The **class I** PTPs, are the largest group of PTPs with 99 members, which can be further subdivided into 38 classical PTPs and 61 VH-1-like or dual-specific phosphatases (DSPs) Classical PTPs can be further divided into 21 receptor tyrosine phosphatase and 17 nonreceptor-type PTPs. Dual-specific phosphatases (DSPs) can be further divided into 11 MAPK phosphatases (MPKs), 3 Slingshots, 3 PRLs, 4 CDC14s, 19 atypical DSPs, 5 Phosphatase and tensin homologs (PTENs), 16 Myotubularins

Dual-specificity phosphatases (dTyr and dSer/dThr) dual-specificity protein-tyrosine phosphatases. Ser/Thr and Tyr dual-specificity phosphatases are a group of enzymes with both Ser/Thr and tyrosine-specific protein phosphatase activity able to remove the serine/threonine or the tyrosine-bound phosphate group from a wide range of phosphoproteins, including a number of enzymes that have been phosphorylated under the action of a kinase. Dual-specificity protein phosphatases (DSPs) regulate mitogenic signal transduction and control the cell cycle, **Class II** LMW (low-molecular-weight) phosphatases, or acidphosphatases, act on tyrosine phosphorylated proteins, low-MW aryl phosphates and natural and synthetic acyl phosphates. The class II PTPs contain only one member, low-molecular-weight phosphotyrosine phosphatase (LMPTP). The **Class III** Cdc25 phosphatases (dTyr and/or dThr) PTPs contains three members, CDC25 A, B, and C.



**Fig. 12 Classification and Substrate Specificity of PTPs .The PTP families are color coded: class I Cysbased PTPs (green), class II Cys-based PTPs(pale yellow), class III Cys-based PTPs (pale blue), and Asp-based PTPs (pink). The substrate specificity of each group or class of PTPs is listed (Alonso *et al.*,2004).**

## Docking

Docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules. Molecular docking is thought of as an optimization problem, which describes the “best-fit” Orientation of a ligand that binds to a particular protein of interest. It is similar to “*lock-and-key*” model, where one is interested in finding the correct relative orientation of the “*key*” which will open up the “*lock*”. Thus the protein can be thought of as the “*lock*” and the ligand can be thought of as a “*key*”. Docking is important as a binding interaction between a small molecule ligand and an enzyme protein may result in activation or inhibition of the enzyme. If the protein is a receptor, ligand binding may result in agonism or antagonism. Docking is the most commonly used in the field of drug design-most drugs are small organic molecules, and docking may be applied to:

**Hit identification**-docking combined with a scoring function can be used to quickly screen large databases of potential drugs in silico to identify molecules that are likely to bind to protein or target of interest.

**Lead optimization**-docking can be used to predict in where and in which relative orientation a ligand binds to a protein (also referred to as the binding mode or pose). This information may be used to design more potent and selective analogs. (Donald *et al.*, vol1)

PyRx is a Virtual Screening software for Computational Drug Discovery that can be used to screen libraries of compounds against potential drug targets. PyRx enables Medicinal Chemists to run Virtual Screening from any platform and helps users in every step of this process - from data preparation to job submission and analysis of the results

AutoDock is a suite of automated docking tools. It is designed to predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure.

## Pharmacophore modelling

The concept of pharmacophore was first introduced in 1990 by Paul Ehrlich, who defined the pharmacophore as “a molecular framework that carries (phoros) the essential features responsible for a drugs (pharmacon) biological activity”. **Ligand-based pharmacophore modeling** has become a key computational strategy for facilitating drug discovery in the absence of a macromolecular target structure. It is usually carried out by extracting common chemical features from 3D structures of a set of known ligands representative of essential interactions between ligands and a specific macromolecular target. In general, pharmacophore generation from multiple ligands (usually called training set compounds ) involves two main steps:

Creating the conformational space for each ligand in the training set to represent conformational flexibility of ligands

Aligning the multiple ligands in the training set and determining the essential common features to construct pharmacophore models (Simone, B.2009).

**Molecular alignment** is the major challenging issue in ligand based pharmacophore modeling. The alignment methods can be classified into two categories in terms of their fundamental nature: point –based and property-based approaches (Wolber, G.2008).

The point (**in the point-based method**) can be further differentiated as atoms, fragments or chemical features. In point-based algorithm, pairs of atoms, fragments or chemical feature points are usually superimposed using a least square fitting. The biggest limitation of these approaches is the need for predefined anchor points because the generation of these points can become problematic in the case of dissimilar ligands.

The property-based algorithms make use of molecular field descriptors, usually represented by set of Gaussian functions, to generate alignments. The alignment optimization is carried out with some variant of similarity measure of the intermolecular overlap of the Gaussians as the objective function.

Another challenging problem lies in the practical task of proper selection of training set compounds. This problem, apparently being simple and non technical, often confuses users, even experienced ones. It has been demonstrated that the type of ligand molecules, the size of the dataset and its chemical diversity affect the final generated pharmacophore model considerably (Dror, O. 2006).

### **Quantitative Structure Activity Relationship (QSAR)**

**QSAR** stands for “quantitative structure-activity relationships”, is a method that relates chemical structure to biological or chemical activity using mathematical models. If the activity of a set of ligands can be determined, a model can be constructed to describe this relationship. Unlike a pharmacophore model, which encodes only the essential features of an active ligand, the QSAR model allows one to determine the effect of a certain property on the activity of a molecule. For example, the QSAR model may reveal a property to have a highly negative, or alternatively a weak positive effect on ligand activity. Such information is not available using a pharmacophore model (Perkinson *et al.*, 2003).

Quantifying the structure and activity of a ligand is important in the modeling process. Structure quantification is not a trivial problem, since a structure cannot be represented by a mere value. Instead, a set of properties, usually known as the “descriptors”, is computed from the structure and used to quantify it. By using structural descriptors as independent variables and activity as a dependent variable, a model can be built to describe the relationship between the two.

### **Building a QSAR Model:**

The process of constructing a QSAR model can be summarized as follows: First, ligands and their activities are collected. Descriptors are calculated and selected before a mathematical modeling method is chosen and the ligand data are then used to construct the QSAR models. After the models are completed, they are tested by internal and external validation procedures. Only then can a QSAR model be used in any practical **applications, such as predicting the activity of a novel compound**. As is the case when building a pharmacophore model, the active ligand set must be gathered from molecular databases or from literature searches before QSAR modeling begins. The process requires not only the collection of ligand structures but also of their activities. Generally, IC<sub>50</sub>s (half maximal inhibitory concentration), EC<sub>50</sub>s (half maximal effective concentration) and *K<sub>i</sub>* values (inhibition constant) are commonly used to quantify drug activity. However, the quantification of ligand activity as used in QSAR is not limited to pharmacokinetic parameters. Other activity indexes can also be incorporated into model depending on the phenomena one wishes to predict. In addition to structure verification as described in the section on pharmacophore model construction, ligand activity data should also be checked. All activity data should come from the same experimental procedure or assay, and it is preferable if the data comes from the same laboratory, and even the same researcher (Yvonne *et al.*, 1981)

Before a QSAR model can be built, ligand structure descriptors should be ascertained or calculated. Some descriptors obtained directly from data sources or calculated using simple arithmetic operations take into account the specific number of atoms, molecular chain length, molecular mass, *etc.* However, other descriptors may require complex computation, for example pharmacophore-based descriptors molecular field descriptors, which are derived from the interaction of probes and molecules and used in **CoMFA** and **CoMSIA**. It is important that the descriptors are related to the biological or chemical activity which the model will be used to predict. In other words, if a descriptor is not related to activity, one should avoid incorporating the descriptor into the modeling process.

After the activity index (the dependent variable) and descriptors (the independent variables) are prepared for each ligand, a variable selection method and a modeling method can be selected, and a model is built. The selection process If two descriptors represent a similar biological or chemical parameter, one of them should be disregarded. In order to select descriptors, genetic algorithms principle component analysis, artificial neural networks and *k*-nearest neighbor approaches can all be used. If a linear model is assumed, some conventional statistical methods, such as the partial least squares method and multiple linear regression can be used. If a nonlinear model is preferred on the other hand, machine learning methods like artificial neural networks or support vector machines can be applied. The main differences among the frequently used QSAR algorithms reside in their means of descriptor generation. For example, most QSAR algorithms, like CoMFA, CoMSIA use similar linear statistical models to explore the relationship

between activity and descriptors, which are calculated by different processes. In CoMFA and CoMSIA, pre-aligned molecules are put onto a grid, or lattice. The descriptors are calculated by the interaction of the molecule and a probe is placed at each intersection of the lattice. The differences between CoMFA and CoMSIA are in the use of different probes and interaction-calculating functions.

In CoMFA, only probes representing steric and electrostatic interactions can be used. In CoMSIA, probes representing hydrophobic and hydrogen bond interactions, in addition to CoMFA probes may be selected. In addition, CoMSIA uses a Gaussian-type function for calculating prober-molecule interaction. By using such a smooth function, the result value is more reasonable than the function used in CoMFA, and defining a cut-off limit to remove invalid values is no longer required.. The fit value describes the goodness of alignment between a ligand and a pharmacophore model and is obtained from a pharmacophore model generated and optimized using known structure and activity data. The model must then be validated before it can be used to predict activity. There are some popular methods used to validate a QSAR model including internal validation approaches (such as the “leave-one-out” or “leave-n-out” cross validation methods , and external validation approaches. In cross validation, one (leave-one-out) or more (leave-n-out) ligand of the training set is excluded. The excluded data is predicted by the model constructed with reduced training set data. These steps are repeated until all data has been excluded and predicted, and the power of a model is determined by the accuracy of prediction. External validation is a widely used method, and is considered important in the QSAR building pipeline. In external validation, the capability of the model is tested using data which is not included in the training set, in contrast to internal validation, which utilizes data taken from the training set to validate the model. In most of the studies, both internal and external validations are performed to ensure the reliability of the model. After the model has passed these strict validation tests, it can be used to predict the activity of novel molecules (John *et al.*, 1983).

### **Statistical concepts**

A QSAR generally takes the form of a linear equation

$$\text{Biological Activity} = \text{Const} + (C1 P1) + (C2 P2) + (C3 P3) + \dots$$

where the parameters P1 through Pn are computed for each molecule in the series and the coefficients C1 through Cn are calculated by fitting variations in the parameters and the biological activity

#### **a) Standard deviation s:**

The standard deviation of the data, *s*, shows how far the activity values are spread about their average. This value provides an indication of the quality of the guess by showing the amount of variability inherent in the data The standard deviation is calculated as shown below

$$s = \sqrt{(\text{compound activity} - \text{average activity})^2 + (\text{compound activity} - \text{average activity})^2 + \dots \div (n-1)}$$

In the above equation n represents no of compounds. This formula gives how to calculate standard deviation

**b) Correlation coefficient r:**

Variation in the data can be quantified by correlation coefficient r which measures how closely the observed data tracks the fitted regression line

$$r^2 = \frac{\text{Sum-of-Squares of the deviations from the regression line}}{\text{Sum-of-Squares of the deviations from the mean}}$$

$$r^2 = \frac{\text{Regression Variance}}{\text{Original Variance}}$$

Where original variance = (compound activity-average activity)<sup>2</sup>+(compound activity-average activity).....

And regression variance = original variance-variance around the line

r<sup>2</sup>= 0 and 1. r<sup>2</sup>of 0 means that there is no relationship between activity and parameter  
r<sup>2</sup>=1mean there is perfect correlation.

**c) F statistics:**

While the fit of the data to the regression line is excellent, how can one decide if this correlation is based purely on chance

F statistics is calculated as

$$F_{1,n} = (n-2) \frac{r^2}{1-r^2}$$

This value can be checked in statistical table to determine the significance of regression equation.

**d) Leave one out cross validation q<sup>2</sup>:**

It involves using a single observation from original sample as the validation determinant and remaining observation as training data. This is repeated such that each observation in the sample is used once as validation data.

$$q^2 = (1 - \frac{\sum(Y_{\text{predicted}} - Y_{\text{actual}})^2}{\sum(Y_{\text{actual}} - Y_{\text{mea}})^2})^2$$

where Y is the activity.q<sup>2</sup> should be close to r<sup>2</sup>.