Robust Modeling and Scaffold Hopping: Case Study Based on HIV Reverse Transcriptase Inhibitors Type-1 Data

Girinath G. Pillai*,a,b, Laznier Mederosb, Chandramukhi S. Pandab, Amber Gronskib, Peeter Burka, Charles D. Hallb, Alan R. Katritzky#,b, Kaido Tämma and Mati Karelson*,a

*Department of Chemistry, University of Tartu, Ravila 14a, Tartu 50411, Estonia; bDepartment of Chemistry, University of Florida, PO 117200, Gainesville, FL. 32611, USA

Abstract: Background: Human immunodeficiency virus type 1 (HIV-1) is the causative agent of AIDS occurs across mucosal surfaces or by direct inoculation. Objective: The objective of this study was to consider chemically diverse scaffold sets of HIV-1 Reverse Transcriptase Inhibitors (HIV-1 RTI) subjected to ideal oriented QSAR with large descriptor space.

Method: We generated a four-parameter QSAR model based on 111 data points, which provided an optimum prediction of HIV-1 RTI for overall 367 experimentally measured compounds.

Results: The robustness of the model is demonstrated by its statistical validation (Ntraining = 111, R2 = 0.85, Q2lmo = 0.84) and by the prediction of HIV-1 inhibition activity for experimentally measured compounds.

Conclusion: Finally, 5 novel hit compounds were designed in silico by using a virtual screening approach. The new hits met all the pharmacophore constraints and predicted pIC50 values within the binding ability of HIV-1 RT protein targets.

Keywords: Global QSAR, applicability domain, diverse scaffolds, field points, scaffold hopping.

1. INTRODUCTION

The aim of this study was to collect a large amount of experimental data and obtain a minimum optimum set for QSAR, which can be applied adequately with statistical relevance to a large number of compounds. To implement the comprehensive use of large descriptor space for modelling QSAR, we used HIV-type 1 Reverse Transcriptase inhibitor bioactivity data. AIDS (Acquired Immunodeficiency Syndrome) categorized by infections is caused by the retrovirus human immunodeficiency virus type 1 (HIV-1), and is one of the leading causes of death around the world. Based on the WHO 2014 report, as of 2013, there are an estimated 35.0 million people living globally with HIV [1]. HIV type 1 behaves as a selective measure in response to stimulus for human helper T lymphocytes, which express the Cluster of Differentiation 4 (CD4) which is a surface glycoprotein [2]. In viral replication, three enzymes are involved, including reverse transcriptase (RT), protease (PR), and integrase (IN).

The approved drugs for treating HIV-1 infected patients usually target RT or PR [3]. However, the efficiency of the approved drugs is declining mainly because of drug resistance and immune response [4]. Targeting the HIV-RT [5-7] and the design of HIV protease inhibitors [8-10] are the most commonly applied strategies to find new drugs against AIDS. Non-nucleoside reverse transcriptase inhibitors (NNRTIs) reduce the ability of enzymes to perform RNA dependent DNA polymerase and are important anti-retroviral drugs used for the treatment of HIV [11]. The efficiency of HIV drugs within the same class was studied in clinical trials [12]. Research efforts focus on the identification of NNRTIs with suitable dosage regimes, better genetic barriers to avoid resistance, and enhanced safety profiles [13]. In parallel with experimental efforts, computational studies such as molecular modeling analysis [14-18] and other in-silico designs are powerful methods to obtain potential NNRTIs. In particular, QSARs are significant chemometric tools, for the development of predictive models, which can be used in several areas of chemical science including agriculture, therapeutic, ecological or material science including medicinal chemistry [19]. QSAR’s main task is to find novel potential drug candidates for experimental tests and therefore provide some relief to the overall time and budget required for synthesizing new chemical entity and bioassays. QSAR correlates molecular descriptors (physicochemical, topological, thermodynamic, constitutional and quantum chemical) with chemical or biological properties (e.g. EC50, IC50, ED50, Ki, etc.).

Therefore QSAR models have the ability to predict the biological activity of a novel chemical compound to estimate activity and design novel drugs [20-22]. Since the discovery
of anti-retroviral drugs, many experimental and computational efforts have been dedicated to the discovery of superior inhibitors. Classical QSAR/QSPR approaches have limited prediction capability, partly due to their applicability domain. In the current approach, we implemented a new protocol in combining different descriptor spaces to develop robust and generalized QSAR models.

The current approach focuses on modeling of global QSAR with a diverse set of scaffolds to predict activity for novel scaffolds in HIV-1 RT. Combinations of global and local QSAR approaches are implemented for dataset trimming and for validation [23]. The respective mechanism of action was studied earlier using molecular docking and molecular dynamics [24]. NNRTIs were targeted against a hydrophobic active site of HIV RT and the binding of NNRTIs alters the conformation of active site residues thus hampering normal enzymatic activity [25]. The technique to obtain a generalized structure-activity relationships for diverse sets of compounds used rationalized and statistical methods [26, 27]. The iterative sampling of inhibitors was used to optimize certain properties in QSAR including interactions with biological systems which provides noisy data. This limitation was been overcome by classifying the data and by techniques such as superimposition, clustering, similarity analysis, and conformational analysis of the inhibitors within the binding site of the target protein [28, 29]. This approach can be used for prediction of the activity of novel compounds by the scaffold hopping process [30].

2. DATA AND METHODOLOGY

2.1. Data Set Selection for QSAR Model Building

Commonly used inhibitors are Efavirenz, Delavirdine, Etravirine, Rilpivirine, and Nevirapine (Fig. 1) [31] which were approved by the FDA, and have been used for more than 10 years. Most of these studies focused on specific classes of compounds (local QSAR) and may not provide a generalized model for a diverse scaffold set of molecular structures.

In this study, we developed QSAR models for HIV-1 inhibitors using diverse classes of structural motifs, and combined different sets of descriptors and chemometric tools such as best multiple linear regression (BMLR), and genetic algorithm (GA). Experimental IC₅₀ data for 419 HIV-RT inhibitors with diverse scaffolds were obtained from the literature and are given in Figure 2 [31] References are added to the Supporting Information - SI]. The experimental bioassay data from literature was reported without any evidence on the bio-safety and bioavailability. Therefore, multi-target approaches have not been examined [32].

Bioavailability and lead-likeness was predicted computationally using Instant JChem [33-36]. Structural data, calculated properties and inhibition activity is given in Table T1 of SI. The selection of datasets was based on molecular similarity analysis with known active drugs like Nevirapine, Efavirenz, Rilpivirine, etc, including a similarity ensemble approach. In this study we used tanimoto method and simi-
larity ranging from 60% to 99%. The diverse set in Fig. 2 was used for QSAR modeling and was categorized into the training set, test set and external prediction set to show the practical validation of the QSAR model. The data in this study was analyzed by a non-linear regression logistic dose-response model, for IC50 values [37]. The logarithmic conversion of biological activity data is one of the requirements for QSAR analysis when free energy-related entities are involved, so we transformed from the equilibrium constant scale to the energy scale [38, 39].

2.2. Data Set Trimming

The distribution normalization of the training set was confirmed by the p value using Shapiro-Wilk [40] and Kolmogorov-Smirnov analysis [41]. The main reprimand for the QSARs is the over-parameterization and narrow applicability domain, therefore our aim was to come up with a robust model with better applicability domain and reproducibility. The statistical details of the trimmed dataset are given in Table S2 (SI). To estimate the homogeneity or similarity of the diverse set, superimposition matching and grid based techniques were used to trim the datasets for QSAR modeling [29, 42]. Cluster analysis and principal component analysis were performed with respect to molecular structure fingerprints and molecular descriptors to understand the classification groups. Based on the progressive 3D structure alignment analysis implemented in the Instant JChem 6.0 software [36] for datasets, structures were categorized and filtered with respect to five or more different classes of core fragments represented in Fig. 2. To confirm, we performed clustering and classification of structures with respect to the biological activity using the Scaffold Hunter tool [43]. The filtered molecules were then considered for 3D QSAR studies. From the literature data, there are 52 compounds out of 419 with indefinite numerical values (e.g., <1 or >1000 etc) and therefore they were removed from the set. Similar activity is probably caused by the complex nature of the non-nucleoside reverse transcriptase inhibitors, which were chosen for QSAR models based on chemical similarity and mechanism of action. Based on the trimming techniques (clustering, PCA, similarity analysis), the final model was developed using 111 diverse compounds including 5 scaffolds as training set. Twenty five compounds were considered as test set and 231 compounds as external prediction set (including 5 scaffolds). The training set was trimmed to minimum size in order to cover the predictions for maximum size of the test set with acceptable statistics. The schematic representation of the dataset selection is presented in Scheme 1.

2.3. Structure Optimization and Descriptor Calculation

A pool of 2D, 3D and 4D molecular descriptors including constitutional, quantum chemical (charge, dipole moment), physicochemical, geometrical, topological, electrostatic, thermodynamic, iso-density, and other quantum chemical molecular descriptors were calculated using Codessa-Pro [44], CDK [45], CODESSA III [46], Padel [47], and EDRAGON [48]. Chemical structures were drawn in Marvin Sketch [49], then the structures were cleaned to net neutral charge, standardized and optimized to 3D using Chem3D
Table 1. Data points for QSAR modeling based on clustering, 2D, 3D similarity analysis. SMILES and 2D Structures for each set is given in SI.

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>No. of Compounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Total Compounds</td>
<td>419</td>
</tr>
<tr>
<td>2.</td>
<td>Valid for QSAR</td>
<td>367</td>
</tr>
<tr>
<td>3.</td>
<td>Training Set</td>
<td>111</td>
</tr>
<tr>
<td>4.</td>
<td>Test Set</td>
<td>25</td>
</tr>
<tr>
<td>5.</td>
<td>External Set</td>
<td>231</td>
</tr>
</tbody>
</table>

[50]. In order to generate 3D & 4D descriptors, conformers for each and every molecule were calculated implementing a systematic method in MacroModel of Schrödinger 2013 Suite [51]. The lack of information on the conformer of highest bioactivity in the original literature forced us to perform conformational screening analysis with the target protein in order to determine the most stable conformer within the binding pocket and also in vacuum. The information of a receptor’s binding pocket for its ligand is crucial for drug design, particularly for conducting mutagenesis studies [52]. The receptor’s binding pocket for a ligand is usually defined by residues that have at least one heavy atom within a distance of 5Å from a heavy atom of the ligand. Such a criterion was originally used to define the binding pocket of ATP in the Cdk5-Nck5a complex [53] that has later proved to be quite useful in identifying functional domains and stimulating the relevant truncation experiments [54]. Similar approaches have also been used to define the binding pockets for many other receptor-ligand interactions [55-57]. Electrostatic interaction constraints were considered from the binding site of the target proteins a) wild strain of HIV-1 RT enzyme co-crystallized with TNK-651 (PDB ID: 1RT2), b) K103N and Y181C mutant strain of HIV-1 RT co-crystallized with TMC278 (Rilpivirine) (PDB ID: 3BGR), c) catalytic complex of HIV-1 RT protein (PDB ID: 1RTD) and d) wild-type and K103N HIV-1 RT with TMC125 (Etiravirine) and TMC278 (Rilpivirine) (PDB ID: 3MEE/3MEC) to replicate the mechanism of action for in-silico virtual screening studies [25, 58-60]. The binding site including co-crystal ligand representation with receptor spatial, hydrogen donor and acceptor constraints given in Fig. F1.A, B, C & D of SI. To maintain reasonable agreement between the quality of the QSAR models, computational parameterization semi-empirical methods were used to derive all possible molecular features i.e., molecular descriptors. High level of accuracy can be attained through ab-initio calculations (Hartree-Fock, Density Function Theory), but for large datasets the calculations are computationally demanding and most of the descriptors were derived manually. The best reorganized structures of lowest energy were chosen for further quantum chemical and geometry optimization using AM1 [61] parameterization in AMPAC 10 software [62] (for Codessa III), MOPAC 2006 [63, 64] (for Codessa-Pro), CDK (geometry is from AM1), PaDEL (geometry is from AM1) and EDRAGON3 to generate data for descriptor calculation. For the generation of geometry based descriptors, the major steps were (1) generation of conformers and their energy optimization; (2) prioritizing the lowest energy conformer; (3) selection of a reference molecule (from lowest energy conformation); (4) analyzing molecular superimposition using the overlap or cluster analysis method; (5) calculating molecular area similarity; (6) determination of other factors by generating spatial, electronic, and conformational parameters; (7) consideration of interaction energy of inhibitor binding site complex with the contribution of van der Waals and electrostatic interactions which had influence on the comparative molecular field analysis [65-67].

2.4. Multi-linear Modeling

Multiple linear regression (MLR) is one of the common statistical methods used for determining relationships between variables. It is used to find a linear model that predicts the best dependent variable from independent variables. We used this method to trim the large descriptor pool with more consistently correlating descriptors to the anti-HIV activity. However, these model diagnostics alone are insufficient to determine the best model. The quality of the regression is also reflected in the numerical values of several statistical parameters including the coefficient of determination ($R^2$) which is a measure of the quality of fit between model-calculated and experimental values, and does not reflect the predictive power [38, 39]. Small values of fitness functions $q^{\text{Log}}$ (Leave One Out) or $q^{\text{LMO}}$ (Leave Many Out) test indicate low prediction ability, but the converse is not necessarily true. It indicates robustness, but not the prediction ability of the model. Therefore, the model should be used to predict the same activity for an external dataset. The MLR model includes finding all orthogonal pairs of descriptors in the dataset and building two-descriptor models. With orthogonal pairs of descriptors found, additional descriptors are added to the previous two-descriptor equations until the additional descriptors cease to bring any improvement to F. Finally, the stepwise procedure ends according to the $R^2$, t and F values until the best equations are obtained. The Molecular Descriptor Analysis (MDA) tool developed by Ruslan O. Petrukhin implemented in Codessa-Pro was used to generate the (Best) MLR QSAR models [44]. MDA tool was used to establish the correlation between all descriptors and activity Log(IC50). Statistical parameters are reported in Table 2, and respective correlation plots are given in SI from Fig. F2 to Fig. F7. BMLR provided a list of descriptors correlated with the activity data, but in order to overcome the limitation of unreliability and poor predictions (reverse QSAR) of raw properties we used Genetic Algorithm (GA) for further generation of QSAR models [68, 69]. The schematic representation of the dataset selection is presented in Scheme 2.

2.5. Genetic Algorithm

The Genetic Algorithm (GA) method was used as a variable selection tool and MLR were employed to model the relationships between molecular descriptors and the biological activity of molecules. With the known MLR limitations, QSAR model renders unreliable and poor predictions for the raw property (IC50) which will affect the performance and reliability of the derived models. Therefore we used MLR as a tool for variable selection based on the orthogonal pairing algorithm to analyze all possible combinations of descriptor
sets. The descriptor variable search ranges from 1 to n (n is assigned by the user); in our case n = 12. Thus, the algorithm will search the combination of each descriptor and correlates with the property value. Genetic algorithm is a stochastic optimization machine learning technique that simulates natural selection and its advantages have been proven in several QSAR studies [70]. The genetic algorithm used in this study was presented by Leardi et al. for the first time [71]. The fitness function in the QSARINS program is the leave-one-out (LOO) cross-validation correlation coefficient (Q²) and was used for the selection of descriptors and the best model based on the applicability domain - William’s Plot, internal validation, external validation, and relevance of descriptor’s physical meaning to the inhibitors [72]. QSAR models ranging from 2 to 6 descriptors for 111 training (Table T1-TR in SI) and 25 test sets (Table T1-TS in SI) using the merged descriptor set are given in Table 3.

2.6. Selection Criteria for Descriptors

Feature selection was used to remove redundant or irrelevant molecular descriptors for improving model quality and reducing computational cost [73]. To generate a pool of scaled descriptors from different software we performed the following steps: (i) the descriptors with zero variance or near zero variance were removed from the initial set of descriptor screening for model consistency. This excludes semi-constant descriptors, i.e. those chemical structures with a constant value of more than a certain percentage (suggested value: 80%), (ii) The descriptors that are too inter-correlated

---

**Table 2. Statistical parameters from BMLR models for 367 compounds generated from software specific descriptors.**

<table>
<thead>
<tr>
<th>Tool[a]</th>
<th>D[b]</th>
<th>R²</th>
<th>R²&lt;sub&gt;cv&lt;/sub&gt;</th>
<th>F</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codessa-Pro</td>
<td>394/347/6</td>
<td>0.42</td>
<td>0.40</td>
<td>66.14</td>
<td>0.958</td>
</tr>
<tr>
<td>CDK</td>
<td>163/105/6</td>
<td>0.52</td>
<td>0.50</td>
<td>96.50</td>
<td>0.802</td>
</tr>
<tr>
<td>Codessa III</td>
<td>923/381/6</td>
<td>0.61</td>
<td>0.60</td>
<td>144.78</td>
<td>0.638</td>
</tr>
<tr>
<td>PADEL</td>
<td>411/261/6</td>
<td>0.64</td>
<td>0.62</td>
<td>158.69</td>
<td>0.602</td>
</tr>
<tr>
<td>E-DRAGON</td>
<td>539/375/6</td>
<td>0.58</td>
<td>0.57</td>
<td>125.96</td>
<td>0.693</td>
</tr>
<tr>
<td>Merged</td>
<td>2430/1469/6</td>
<td>0.65</td>
<td>0.64</td>
<td>166.09</td>
<td>0.585</td>
</tr>
</tbody>
</table>

[a] AM1 optimized structures are used to generate descriptors by different software package/tool. [b] No. of Total Descriptors Calculated/ Trimmed Descriptors/ Descriptors used in the Model

---

**Scheme 2. Approach for combining molecular descriptor and the detailed workflow is given in SI Fig. F4.**
Thereafter 3D-QSAR BMLR models were developed for the valid dataset of 367 compounds using individual descriptors from Codessa-Pro-S1, CDK-S2, Codessa-III-S3, PADEL-S4, and EDRAGON-S5. These models were internally validated using a cross-validation technique. In addition, a merged and trimmed pool of 1469 descriptors from all software was used to generate a comprehensive model. To conclude the selection of the descriptors, final model was developed using 111 structures with a total number of 1469 descriptors and 25 compounds (~20% of training set) which were used as the test set for external validation and 231 compounds for external test set.

2.7. Statistical Validation

A robust QSAR modeling workflow was followed to generate models, validate and predict activities for new data sets. The fitting ability of the model was internally validated based on the leave-one-out (LOO) cross-validation and leave-many-out (LMO) cross-validation techniques. In the LMO cross-validation technique, ~20% of training set compounds were deleted in different cycles based on outliers and heterogeneity of the compounds in the dataset. For all iterations, the biological activities of the excluded compounds were predicted using the model developed with the corresponding dataset of compounds. Training sets were further divided into multiple sets of descriptive training and test sets of different size, i.e., based on descriptor similarity and structure similarity using overlap analysis in DS Visualizer 4.0 [74]. The external predictive ability of the model was assessed based on the predictions of the test set compounds followed by the calculation of the R^2LOO & LMO parameter.

2.8. Theoretical Background of Molecular Descriptors

Three descriptors in four parameter (D_1, D_2, D_3) models are directly related to electrostatic properties of the molecules. Together with D_1 which can be considered as a molecular stability parameter, all the descriptors influence HIV-1 inhibition. D_1 signifies the surface area of inhibitors and the corresponding charge distribution which is important to the binding affinity. Because of the presence of dipole–dipole interactions within the active site, hydrophobic substituents are not favored [75]. The requirement of polar surface area is evident with electrostatic field. D_2, van der Waals molecular surfaces of each atom in the inhibitor uses empirically derived hydrophobic atom constants calculated from the solvent partition constants of small molecules [76]. Therefore, the increase in the compound’s hydrophobic surface area would be favorable for better activity and increased negatively charged surface area is better for inhibitory activity [77]. In the following, we also present the theoretical background of the descriptors.

2.8.1. D_1 - HACA-2/Sqrt(TMSA) – Zefirov PC

HACA-2 belongs to the class of Charged Polar Surface Area (CPSA) descriptors, which define the atomic electronegativities of hydrogen atoms [78, 79]. HACA-2 is the area-weighted surface charge of hydrogen bonding acceptor atoms and Sqrt(TMSA) is the mathematical square root of total molecular surface area.

$$HACA_2 = \sum_A \frac{q_A \sqrt{S_A}}{S_{tot}} \quad A \in X_{H-acceptor} \quad \text{Eq. 1}$$

$$D_1 = \frac{HACA_2}{\sqrt{TMSA}} \quad \text{Eq. 2}$$

where $S_A$ = solvent-accessible surface area of H-bonding acceptor atoms, selected by threshold charge, $q_A$ = partial charge on H-bonding acceptor atoms, selected by threshold charge, $S_{tot}$ = total solvent-accessible molecular surface area [80].

2.8.2. D_2 - Fractional Atomic Charge Weighted Partial Negative van der Waal's Surface Area

This descriptor assumes the partial atomic charges from molecular mechanics force fields to compute the electrostatic interaction energy for understanding of the structure and reactivity of molecules [81]. The descriptor is derived from FNSA3 of Codessa-Pro with van der Waals surface area instead of total molecular surface area [66, 88].

$$FN_{-SA} = \frac{PNSA_3}{vSA} \quad \text{Eq. 3}$$
where \( PNSA3 \) = total charge weighted partial negatively charged molecular surface area, \( vSA \) = van der Waals surface area [82].

2.8.3. \( D_3 \) – Average Bond Order C-C

Bond order and bond length indicate the type and strength of covalent bonds between atoms. The bond order is defined as follows:

\[
B.O. = \frac{\text{bonding electrons} - \text{antibonding electrons}}{2} \quad \text{Eq. 4}
\]

A higher bond order means that the atoms are held together tighter and therefore the whole molecule has a greater stability. In other words, a large positive value of bond order signifies strong bonding between the atoms of the molecular entity, whereas negative value imply that electrons are displaced away from the inter-atomic region and point to an anti-bonding interaction [83, 84].

2.8.4. \( D_4 \) - IsoDensity Surface – varElectrostatic Potential

The electron probability density is referred as electron density and it’s a function of \( r \) displayed as contour maps or isodensity surface. Isodensity surface is defined as the surface mapped by points where the density has the same value [85]. Total variance of electrostatic static potential is related to conservative force and derived as:

\[
\frac{1}{m} \sum_{i=1}^{m} \left[ V^+(r_i) - \bar{V}^+ \right]^2 + \frac{1}{n} \sum_{i=1}^{n} \left[ V^-(r_i) - \bar{V}^- \right]^2 \quad \text{Eq. 5}
\]

where \( V^+_i \) = average value of the positive electrostatic potential in the molecule, \( V^-_i \) = average value of the negative electrostatic potential in the molecule, \( V^+(r) \) = positive electrostatic potential in the molecule, \( V^-(r) \) = negative electrostatic potential in the molecule, \( m, n \) = number of integration points [86, 87].

2.9. Bioisosteric Group Replacement – Scaffold Hopping

Bioisosteres are chemical groups with identical physico-chemical properties which produce similar biochemical properties to other compounds. The process is known as scaffold hopping and this is performed to reduce toxicity, improve bioavailability and mode of action. To perform scaffold hopping pharmacophore constraints and attachment or replacement of chemical groups are assigned to pre-fragment databases [88-91]. For computational efficiency, we used Spark v10 from Cresset UK. The Cresset’s Field technology condenses the molecular Fields of a molecule into a set of points around the molecule, termed ‘Field points’ [92]. Field points are defined as the local extrema of the electrostatic, van der Waals and hydrophobic potentials of the molecule. The field points are calculated from the physical properties, and interaction point size/strength associated with molecules (not all H-bond donors are treated the same: some make stronger bonds than others) [93]. The four Field types are used to describe all the potential interaction points that a ligand can make to the target protein in a desired conformation. A chemical group or vector in a molecule at its bioactive conformation (3D bound structure to target protein) is selected and chosen for replacement. Spark searches a database of >600,000 fragments for bioisosteres that exhibit similar shape and electronic properties when placed in the context of the final molecule. Spark’s native fragment databases consist of fragments generated from commercially available screening compounds, ChEMBL, and VEHICLE databases [94, 95].

Each one of these first-pass fragments was fitted into the original molecule; fields and field points were calculated for the resulting molecules. The resulting molecules are minimized using the XED forcefield, and then the full field and shape of the molecule is scored against the full field and shape of starting molecule including the reference molecules [92]. By default, 50% shape and 50% fields are used in score calculation. Using Rivipirine, Etravine, Abacavir and Stavudine as a starting point, Spark was parameterized to search for bioisosteric replacements that would introduce novel compounds into the assigned vector area. Spark uses molecular interaction fields to represent the key binding interactions of a molecule by giving a close approximation to the HIV-1 RT protein (PDB ID: 3MEE) view of the potential ligand [96, 97]. The workflow scheme of scaffold replacement is given in Scheme S3A of SI and interpretation of field point technology is given in Scheme S3B of SI.

4. RESULTS AND DISCUSSIONS

4.1. QSAR Model

In this study, conformers based on protein binding site interactions were taken into account on calculation of molecular descriptors. Global structural activity relationship models were developed to cover wide range of chemo-types with a large applicability domain. Implementation of 1D, 2D, 3D and 4D descriptors delivered a meaningful correlation between the mechanism of inhibitors to HIV-1 target protein and mode of action. The functional groups like hydroxyl –OH, methoxy –OCH3 are important for the conformational change in the RT which distorts the position of protein residues. However, different RT inhibitors bind to the target protein in different modes depending on the conformation, which influences shape, size and chemical composition of the compound. All statistical parameters obtained for the models demonstrate the robustness and reliability of the models. The schematic workflow of the whole approach in comprehensive modeling is given in Scheme S2 of SI. The descriptors involved in regressions have definite physico-chemical background and can be interpreted in terms of HIV-1 inhibition. The potential drawback of this methodology is the limitation of specific sets of constitutional or atom-based descriptors in order to consider diverse sets of structures for the training set. Literature data was statistically and theoretically analyzed to generate a reproducible robust QSAR model and equation (Eq. 6) with all statistical parameters (Table 4).

For statistical validation the Y-scrambling method was used to estimate the performance of a model. Furthermore, the activity prediction for a test set of 25 compounds was confirmed by using the QSAR model given in Table 4. For external set validation, the diverse set of 231 compounds (Table T1-EX in SI) was used to predict the HIV-1 inhibition activity for experimentally measured compounds. Respective values are reported in supporting information (Table
Statistical parameters for external predictions are given in Table 5. These data suggest that the model we propose has robust predictive power.

In addition to the internal and external validation data, three sets of graphical data are given as follows: (i) Correlation plot for external dataset with experimental and predicted values (using model from Table 4) (Fig. 5), (ii) Williams plot – plot of standardized residuals versus the Hat leverage to visualize the reliability of the QSAR model and identify outliers (Fig. 6); (iii) Applicability domain which defines the training set space and prediction capability of the model, this systematic approach to represent the reliability to make predictions for new compounds from QSAR model generated from the training set.

Table 4. QSAR model for the anti-HIV-1 activity.

<table>
<thead>
<tr>
<th>Desc No.</th>
<th>Descriptor</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-test</th>
<th>Desc vs Log(IC50)</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>I0</td>
<td>Intercept</td>
<td>-11.812</td>
<td>1.93</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>HACA-2/Sqrt(TMSA) – Zefirov PC</td>
<td>-18.924</td>
<td>1.93</td>
<td>-14.57</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>Fractional atomic charge weighted partial Negative van der Waal’s Surface Area</td>
<td>4.987</td>
<td>0.99</td>
<td>4.79</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>Average Bond Order C-C</td>
<td>12.621</td>
<td>1.44</td>
<td>19.67</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>D4</td>
<td>IsoDensity Surface – Var Electrostatic Potential</td>
<td>-2445.014</td>
<td>616.22</td>
<td>-5.584</td>
<td>0.22</td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{Log(IC50)} = -11.812 + [(D1 \times -18.924) + (D2 \times 4.987) + (D3 \times 12.621) + (D4 \times -2445.014)] \quad \text{Eq. 6} \]

Table 5. Details on the model robustness.

<table>
<thead>
<tr>
<th>ID</th>
<th>Set</th>
<th>R2</th>
<th>R2*</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR</td>
<td>Training Set</td>
<td>0.78</td>
<td>0.85</td>
</tr>
<tr>
<td>TS</td>
<td>Test Set</td>
<td>0.62</td>
<td>0.76</td>
</tr>
<tr>
<td>EX</td>
<td>External Set</td>
<td>0.41</td>
<td>0.61</td>
</tr>
</tbody>
</table>

External Validation: RMSE\text{corr} = 65.25, R2\text{corr} = 0.76, Q2\text{F1} = 0.73, Q2\text{F2} = 0.72, Q2\text{F3} = 0.76, CCC\text{corr} = 86.10
Prediction by Model Eq. 6 for External Dataset: N\text{external} = 231, R2\text{external} = 0.61
*R2* = if outliers with error greater than 2.5 deviations excluded from Williams plot.

4.2. Scaffold Hopping

The use of field points represents a major advance in the computational methods available for finding novel bioisosteres. It analyzes and scores the replaced bioisosteres in the context of the final molecule rather than as an isolated fragment. The results are significantly more diverse with both...
known bioisosteres and novel replacements suggested in most cases. With reference to the requirement of polar surface area electrostatic field, van der Waals molecular surfaces, hydrophobic surface area and increased negatively charged surface area is reflected in the field points of the generated compounds generated from scaffold hopping process. Field point representation for Rilpivirine is given in Fig. 8A and the generated compounds after bioisostere group replacement is shown in Fig. 8B highlighted with a black sphere.

Antiviral activity prediction was performed for the new compounds generated from scaffold hopping using the QSAR model in Table 4. A selected list of compounds and their predicted activity is given in Table 7. The complete list
Fig. (7). Applicability domain of the QSAR model.

Table 6. Descriptor’s inter-correlation matrix.

<table>
<thead>
<tr>
<th>Desc</th>
<th>D₁</th>
<th>D₂</th>
<th>D₃</th>
<th>D₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>D₁</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D₂</td>
<td>-0.0046</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D₃</td>
<td>-0.528</td>
<td>0.255</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>D₄</td>
<td>0.254</td>
<td>-0.030</td>
<td>-0.417</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Mutual relation of two descriptors and the degree of correlation between two descriptors in a QSAR model progressively. Details of all list of descriptor are given in Table T2 of SI. Descriptor values of D₁, D₂, D₃, D₄ for training and test set is given in Table T3-TR and Table T3-TS respectively in SI.

Fig. (8A). Field-based template containing a single docked conformation of Rilpivirine based on their three-dimensional field point patterns. Negatively charged field points are shown in blue; positively charged field points are red; van der Waals/shape field points are displayed in yellow; centers of hydrophobicity are shown in orange. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper).
Fig. (8B). In silico bioisostere replacement (scaffold-hopping). 3 best new substituent conformations with its associated field points in the binding site of protein PDB: 3MEE was identified and highlighted with a black sphere.

Table 7. In-silico hits generated by scaffold hopping. Best 5 hits selected based on activity prediction using QSAR model given in Eq. 6.

<table>
<thead>
<tr>
<th>ID</th>
<th>Structure</th>
<th>Database</th>
<th>Log (IC50) Predicted</th>
<th>IC50 Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABAC095</td>
<td><img src="ABAC095" alt="Structure Image" /></td>
<td>ZINC7710510</td>
<td>0.465</td>
<td>2.923</td>
</tr>
<tr>
<td>RILP016</td>
<td><img src="RILP016" alt="Structure Image" /></td>
<td>ZINC7065337</td>
<td>1.707</td>
<td>50.98</td>
</tr>
<tr>
<td>RILP098</td>
<td><img src="RILP098" alt="Structure Image" /></td>
<td>ZINC8612194</td>
<td>1.822</td>
<td>66.40</td>
</tr>
<tr>
<td>ETRA027</td>
<td><img src="ETRA027" alt="Structure Image" /></td>
<td>ZINC9554652</td>
<td>0.947</td>
<td>8.863</td>
</tr>
<tr>
<td>ETRA016</td>
<td><img src="ETRA016" alt="Structure Image" /></td>
<td>ZINC3918098</td>
<td>1.323</td>
<td>21.03</td>
</tr>
</tbody>
</table>

ABAC=Compounds from Abacavir as starting molecule, RILP= Compounds from Rilpivirine as starting molecule, ERTA= Compounds from Etravirine as starting molecule.

Database= Compound ID from where the chemical group/fragment replacement was performed.

Structural data for 385 in-silico compounds are given in excel format (SI2) and activity prediction using QSAR Eq 6 & including molecular descriptor values are given in Table T5-NS in SI.

CONCLUSION

Analysis by QSAR revealed molecular determinants of antiviral action against HIV-1 RT, and this knowledge was translated into search queries for a scaffold hopping step.

We have demonstrated that a meaningfully reduced experimental data set can be used for prediction of antiviral properties of a diverse set of chemicals. The results obtained from the model can be considered for selection of novel chemicals for further experimental testing. The important role of the hydrophobic regions and vdW’s forces in the Rilpivirine and Etravirine scaffold have been verified by both QSAR and molecular binding site analysis of target protein PDB ID: 3MEE/3MEC. The past decade has seen an increased understanding of the crucial role of QSAR and scaffold hopping in lead optimization to develop novel inhibitors. QSAR studies on diverse classes of inhibitors have been thoroughly explored. The QSAR method in combination with a range of descriptors and other in-silico tools has increased the applicability of former improved their application to drug design and discovery. Based on theoretically obtained compounds from scaffold hopping we propose a number of novel and potentially active scaffolds for further assessment.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

AUTHOR CONTRIBUTION

Conceived and designed the experiments: GGP, KT, MK. Performed the experiments: GGP, LM, CSP, AG. Analyzed the data: PB, CDH, ARK, KT, MK. Contributed analysis tools: GGP, ARK, MK. Wrote the manuscript: GGP, LM, KT.

ACKNOWLEDGEMENTS

G.G. Pillai is grateful to the graduate school “Functional materials and technologies,” University of Tartu for funding from the European Social Fund under project 1.2.0401.09-0079, the EU European Regional Development Fund through the Center of Excellence in Chemical Biology, Estonia, by
targeted financing from the Estonian Ministry of Education and Research (SF0140031As09) Estonian Research Council grant (PUT582) and Florida Center for Heterocyclic Compounds, Kenan Foundation at University of Florida to carry out my research. Authors are grateful to Prof. Paola Gramatica for providing free licenses and scientific advices for Q SARINS software, Mr. Khalid and Dr. Alexander Oli-ferenko for scientific discussions. The authors are grateful to Dr. Rachel A. Jones for English checking.

SUPPLEMENTARY MATERIAL

Supplementary material contains the structure, SMILES and activity data of the whole dataset, including predicted activity and others statistical data.

REFERENCES

Robust Modeling and Scaffold Hopping

Medicinal Chemistry, 2016, Vol. 12, No. 6 525


