The development of a predictive model to identify potential HIV-1 attachment inhibitors

Amer Hosny[a], Mark Ashton[b], Yu Gong[c], Ken McGarry[d]

[a] *Faculty of Health Sciences and Well-being, University of Sunderland, City Campus, Sunderland, SR1 3SD, UK*
[b] *The School of Pharmacy, Faculty of Medical Sciences Newcastle University, UK*
[c] *Chengdu Yontino Tech Co. Ltd, Sichuan, China*
[d] *The School of Computer Science, University of Sunderland, St Peters Campus, Sunderland, SR6 ODD, UK*

Abstract

Despite the significant progress in managing patients infected with HIV through the development of Highly Active Anti-Retroviral Therapy (HAART), major challenges and opportunities remain to be explored. Of particular interest, is the binding of glycoprotein 120 (gp120) to the primary cellular receptor Cluster of Differentiation 4 (CD4). In this work we describe our two phased computational process to identify useful compounds capable of binding to the gp120 protein for therapeutic purposes. We identified 187 compounds from the literature that conform to active binding sites on these proteins and use these as training/test sets. The data in the form of quantitative structure-activity relationships (QSAR) is downloaded from the ZINC database and transformed using principal components analysis. In the first phase we developed a Radial Basis Function neural network model that identifies potential inhibitors from a virtual screen of a subset of the ZINC database. In the second phase we modelled the top performing compounds using the Discovery Studio docking and screening software. By employing this approach, we identified that those compounds with a LogP value of approx 2-4 performed well in the binding simulations while the lower scoring compounds do not bind well.

Keywords: HIV, gp120, CD4, QSAR, RBF, PCA

1. Introduction

The development of Highly Active Anti-Retroviral Therapy (HAART) (a combinatorial regime in which one or more of the three viral enzymes are targeted by the use of three therapeutic agents) has transformed the life expectancy of HIV positive individuals but major challenges remain to be solved [46]. Firstly, and perhaps most importantly, due to the poor fidelity of replication observed with retroviruses, they quickly develop resistance to antiviral agents and hence treatment failure results. [9] Secondly, due to the current requirement of daily dosing with HAART, it is imperative that agents have a good safety profile, no drug interactions and are easily administered. [25]. Thirdly, the issue of viral latency still needs to be addressed [50]. In order to address these challenges, there is a need to develop new agents and the targeting of viral entry has emerged as a viable approach. Viral entry into host cells is a complex phenomenon [16] involving many steps operating in sequence. Initial contact with a host cell by the virus is mediated via heparin sulfate [35] and leads to the specific interaction between the viral protein, glycoprotein 120 (gp120) and the primary cellular receptor Cluster of Differentiation 4 (CD4).

In this work we describe our approach to develop statistical models to identify compounds that can bind to the gp120 glycoprotein. Gp120 is derived from the heavily glycosylated protein Env gp160 and contains five conserved domains (C1-C5) and five variable loops (V1-V5). The binding of gp120 to CD4 induces a significant conformational change in gp120 that facilitates receptor engagement [19]. The binding to the chemokine co-receptor (CCR5 or CXCR4) by gp120 exposes the fusion peptide gp41 which tethers the viral and host membrane together. Following the fusion of the two membranes, the viral capsid is released into the cytosol of the cell, where the virus begins its replicative cycle.

Targeting the early stages of the replicative cycle is attractive because it precedes cellular infection and two agents are currently licensed; enfuvirtide [14] and maraviroc [30]. Over the last several years a number of papers have been published that have indicated that gp120 has the potential to be a good target to develop new anti-HIV inhibitors and examples of both peptide and small

Email address: mark.ashton@newcastle.ac.uk (Mark Ashton)

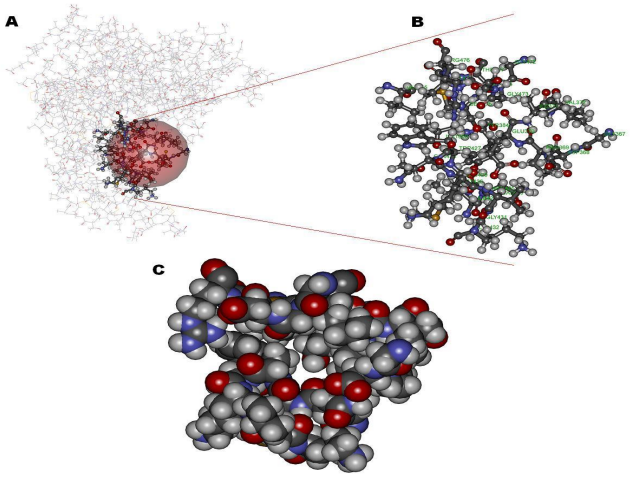


Figure 1: gp120 (PDB code: 1G9M) A: View of full gp120 protein (Chain G) in line view, with the defined binding site in ball and stick view. B: View of the binding site only with labels of amino acid residues defined within the binding sphere. C: View of the binding site in CPK from a different angle, showing the Phe 43 pocket.

molecule inhibitors have been reported.

Several families of small molecule HIV-1 entry inhibitors have been reported over the years. Bristol-Myers Squibb (BMS) first introduced BMS-378806 but this compound was later dropped due to unfavourable pharmacokinetic properties and focus moved to BMS-626529 (and its prodrug, BMS-663068) [33]. The BMS-378806 family of inhibitory compounds are thought to operate in a concentration dependent manner exhibiting multiple mechanisms to inhibit HIV-1 entry [34].

NBD-556, the archetype of the NBD family of compounds work by targeting the 'Phe43 cavity'; binding of NBD-556 in this highly conserved hydrophobic cavity in gp120 generates a substantial conformational change in the gp120 protein that in turn disrupts binding [55].

Owing to the well-defined nature of the Phe43 cavity, we sought to develop a neural network model that could identify potential inhibitors that display favorable physiochemical properties to interact with the Phe43 cavity from a virtual screen of a subset of the ZINC database (three datasets each containing 5000 compounds were selected for the current work).

A quantitative structure activity relationship (QSAR) is a mathematical relationship between the biological activity and the chemical and geometric characteristics of a molecular system to be used in predicting the activity of unknown compounds [12, 10]. QSAR modeling is widely accepted across several diverse organizations such as industry, academia and government departments worldwide for modeling the biological and physical properties of chemicals and compounds [17]. Computational modeling of QSAR data requires different molecular properties and structural fingerprints to be calculated and then one or more of statistical analysis models are applied to determine the relation between the activity and the structural features of these compounds in terms of equations that can be used as activity prediction tools.

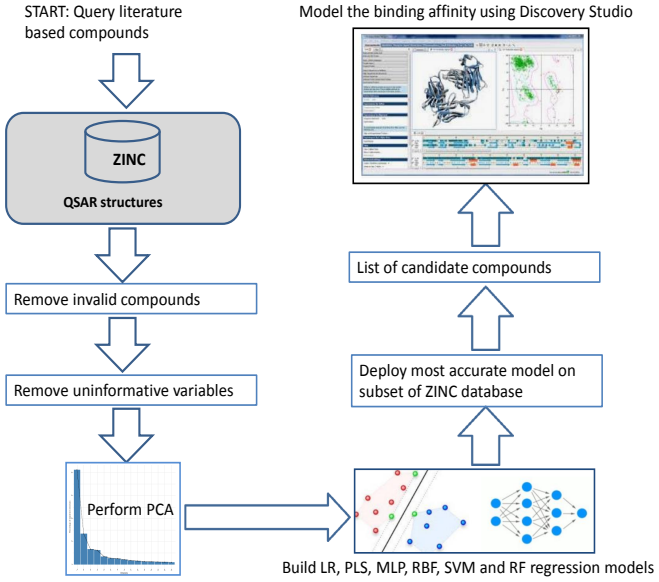


Figure 2: Overview of workflow operation

The overall data flow is presented in figure 2, here we show data access, manipulation and model building that is consistent with the QSAR data format [54]. Computationally speaking, the analysis is a regression solution whereby we model a continuous value (pIC50) as the dependent variable, influenced by the QSAR variables. The models are used to screen the ZINC database, models such as linear regression, partial least squares, SVM, MLP, Random Forest and RBFs are compared for accuracy but also for the usefulness of the detected compounds they identify from the ZINC database.

2. Related work

Previous work, which has seen drug development concentrate on targeting the interaction between gp120 and the cellular CD4 receptor has mostly affected by the sequence variation degree of the V1/V2 variable regions [29]. Thus, the diversity between different viral subtypes in these regions in addition to the intra-patient variability can negatively influence the efficacy of gp120-CD4 inhibitors [1]. Targeting the binding between the gp120 and the co-receptors which is the next step and is mediated by the conformational changes of gp120 especially V1 /V2 and V3, is theoretically expected to be less affected by the sequence variability of the viral glycoproteins because it is concerned with the host protein [22].

The use of virtual screening protocols in drug discovery is well established and the Kaggle competition in 2012 is recognised as the first demonstration of the capabilities

of Deep-Learning (DL) algorithms in the drug-discovery sphere [20]. Following that initial study, DL algorithms have been extended to explore a range of diverse QSAR datasets with a view to predicating the physiochemical properties of molecules and selecting molecules for particular protein targets. For example, Lusci *et-al* developed a neural network model that could predict a range of molecular properties, including the aqueous solubility of drug-like molecules [28]. Other algorithms that have been employed include the use of partial Least square (PLS), multiple linear regression (MLR), artificial neural networks (ANN) and principal component or factor analysis (PCA/FA); are all examples of the statistical procedures that are frequently used for building a QSAR model [43].

3. Methods

We selected all of the known compounds (187 at the time of this investigation) implicated in binding to the gp120 protein from the literature using PubMed [4, 31, 32, 39, 48, 49, 53, 51, 52]. The literature references for the 187 compounds were manually checked to ensure that the biological assays used to determine activity were comparable and the activity range (measured as IC50 values) was between 0.0054nM and 65.8μM. These compounds were downloaded from the ZINC database and divided randomly into a train and test set split (75%/25%). The ZINC database (http://zinc15.docking.org/) is a useful repository of compound data of commercially available materials for virtual screening and contains over 350 million compounds in ready-to-dock formats [45, 21]. The variables for each structure used in this study, both the initial dataset of 187 compounds and each of the screening batches (3x5000 batches of compounds labelled A,B and C obtained from the Zinc database) were calculated using the Molecular Operating Environment (MOE, version 2008.10; https://www.chemcomp.com/Products.htm) software. The MOE software calculated a mixture of both 2D and 3D descriptors for each compound (http://www.cadaster.eu/sites/cadaster.eu/files/challenge/descr.htm); the calculated variables were further processed.

We used the R statistics software to transform and analyze this data [37], other add-on packages included the ChemmineR package for handling SDF files and analyzing drug-like small molecule data[3]. We processed a matrix of 187 compounds with 316 variables, this was further reduced by removing any variable containing more than 90% zeros or ones (unlikely to be informative) leaving 269 variables. Principal components analysis (PCA) further transformed the data into a smaller number of variables. We kept 15 principal components containing 82% of the variance. The PCA data was scaled to unit variance before the analysis. This avoids any variable from becoming too dominant because of large magnitudes of their values. Detailed graphical analysis of the PCA results was provided by the FactomineR package[23]. Our R source code and data are available from: https://github.com/kenmcgarry/QSAR-hiv

3.1. Computational models for predicting activity

We built several regression models using Radial Basis Function (RBF) networks, Random Forests(RF), Multilayer Perceptrons (MLP) and Support Vector Machines (SVM), comparing their performance on the test data. The SVM is well suited for regression problems and SVM's have been successfully applied in bioinformatics [11, 44]. RBF networks provide a local solution to linear and nonlinear problems [26]. The MLP is the oldest and best known of the techniques and has seen many applications in a wide variety of fields [6]. Many improvements have been made to the original MLP algorithm in terms of speed, accuracy and memory storage [8]. We also trained Random Forests, these are a supervised ensemble regression technique where a group of weak models combine to form a powerful model [7]. Several decision trees are created (hence a forest) with random sampling used as the attribute to split on, there is a direct correlation between the number of trees in the forest and the accuracy. Another important advantage of the Random Forest method is that the predictor variables are ranked according to importance using the Gini impurity measure. The RBF and MLP proved to be the superior models in detecting viable drug-like compounds. We deployed the RBF to detect the potential compounds in the three random datasets, using the RBF to detect the potential compounds from each dataset (5000 each) selected at random from the ZINC database.

The Radial Basis Function model outperformed the other regression models when it came to identifying 'drug-like' leads, and in the case of both test datasets, the top ten compounds identified complied with Lipinski's rules [24]. The improved performance of the RBF (see figure 3a) can be explained by considering the loadings of the terms in both PC1 and PC2; 31.3% of the variance that is captured by PC1 generally relates to a molecule's polarity and based on the physiochemical properties of the target site on gp120, molecules with a Log P in the range 2-4 exhibit favourable binding characteristics.

3.2. Measures for determining model goodness of fit

The quality measures used to evaluate the models include R^2, MAE, RMSE and the Tropsha and Roy measures.

R-squared (R^2) is a statistical measure accounting for the proportion of the variance for a dependent variable that's explained by an independent variable or variables. The R^2 value is calculated by sum of squared residuals (SSR) divided by the total sum of squares (SSTO). The SSR are deviations predicted from actual empirical values of data. It is a measure of the discrepancy between the data and an estimation model. A small RSS values imply a good fit of the model to the data. The SSTO is defined as the sum of the squared differences, over all observations, between the observations and their overall mean.

$$r^2 = \frac{SSR}{SSTO}. \quad (1)$$

Where: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$; and $SSTO = \sum_{i=1}^{n}(y_i - \hat{y})^2$

The mean absolute error (MAE) measures the difference between two variables. Where X and Y are variables of paired observations we use for comparisons between predicted and observed values. Roy, further explains the MAE measures the actual prediction errors with respect to the total number of observations and also it's weaknesses [42].

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}. \quad (2)$$

The root mean squared error (RMSE) provides an indication of how much error there is between two sets of values. Like the other measures it compares a predicted value and an observed value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_j - O)^2}{n}} \quad (3)$$

Where: P= Predicted and O= observed values.

The Roy measure can be calculated by:

$$r_m^2 = r^2(1 - \sqrt{r^2 - r_0^2}). \quad (4)$$

Where: r^2 is the squared correlation coefficient between observed and predicted values, r_0^2, is the same parameter but with the intercept set to zero. The Roy value should be > 0.5 to confirm a good model [40, 41].

The Golbraik-Tropsha measure is defined by:

$$Tropsha = \frac{r^2 - r_0^2}{r^2}. \quad (5)$$

Where: similar to Roy measure, we have r^2 is the squared correlation coefficient between observed and predicted values, r_0^2, is the same parameter but with the intercept set to zero. The Tropsha measure should return a value of < 0.1 to confirm a robust model [15, 2].

3.3. Molecular docking and molecular dynamics

The top ten performing compounds from from datasets A,B and C were then screened using a standard molecular docking platform via Discovery Studio 4.5. Docking calculations were used to measure the binding affinity of the test compounds towards the target protein [18]. Owing to the fact that the viral glycoprotein gp120 is highly flexible [47], both the open (unbound) and closed (bound) forms were employed in the docking study. The open and closed forms were represented by the PDB-ID 4RZ8 and 1G9M respectively, and were downloaded from the protein data bank (www.rcsb.org) [5], having a resolution of 1.9 $\mathring{A}$ for 4RZ84 and 2.2 $\mathring{A}$ for 1G9M [13]. The docking calculations were validated by comparing the resulting docking energies of the selected compounds with that of NBD-556 which is a potent HIV-1 entry inhibitor [36].

For the docking simulation, water of crystallization was removed and protonation was carried out for both forms of the target protein. The active site was determined using the volume occupied by an existing co-crystalized ligand in case of 4RZ8N protein, while the Phe43 amino acid was used to generate a sphere with a diameter of 10 $\mathring{A}$ to represent the active site in the 1G9M protein structure [38]. Each compound was energy minimized, and the top 10 conformations generated during the docking process. The docked ligands were ranked and evaluated using the bonding interactions with the residues of the active site and their consensus scores which were calculated for each compound using the Consensus Score protocol [27]. Following docking of all compounds, CDOCKER energy and the CDOCKER INTERACTION energy scores of all compounds have been compared in order to evaluate the docking potential with the target viral protein.

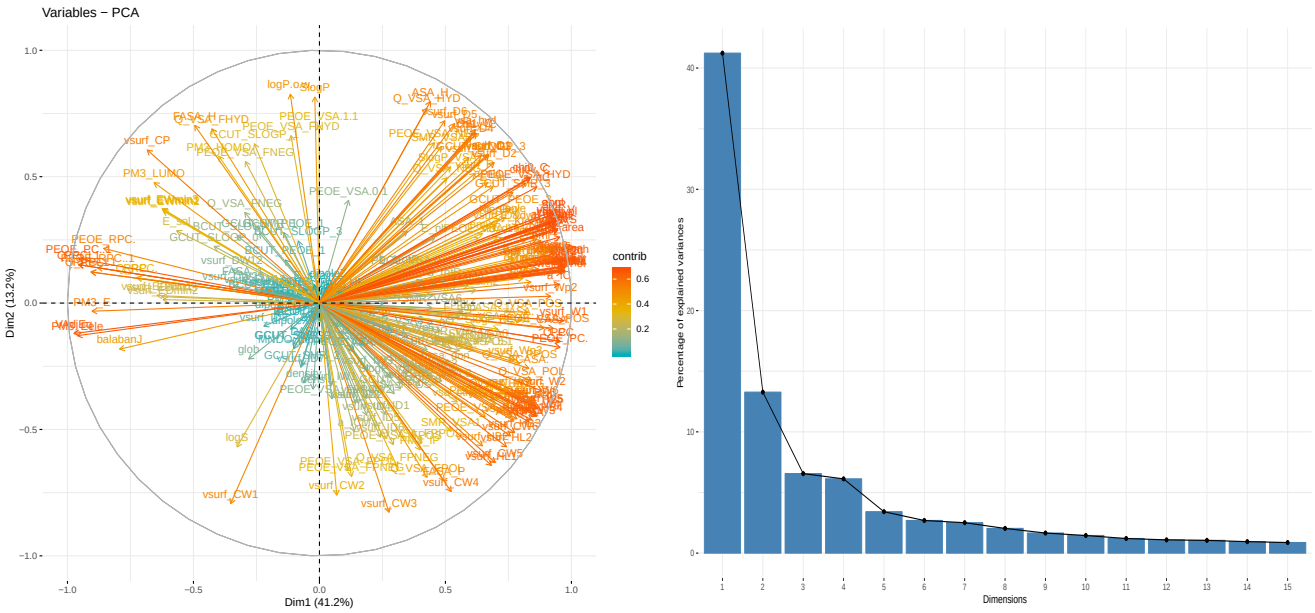
Molecular dynamics simulations were undertaken for one of the top performing leads, CMP2463 in order to assess the validity of the binding mode using Discovery Studio 3.1. The ligand was selected for the molecular dynamics on a pre-equilibrated molecule. The protein-ligand system was not solvated. Prior to running the simulations, (CHARMm) energy minimisations were run using the Standard Dynamics Cascade. The open form of the protein was used for the study.

4. Results

In figure 3b we present the results of the PCA analysis, 41.2% of the variance is captured by PC1 and 13.2% by PC2, thereafter each PC provides a smaller percentage of the variance as indicated by the screeplot. The biplot shows the loading's of each variable across PC1 and PC2, the variables colored cyan (found mainly in center of the plot) have low contributions while those colored red/orange indicate high contributions. The arrows point in the direction of the variables when they projected into the x-y axis of the biplot. The larger the arrow, the larger the loading of that particular variable on the first two principal components and hence its importance. Like all biplots, we are assuming both PC1 and PC2 contain a sufficient amount of the variance together (54%) to provide a meaningful visual representation of the structure of cases (compounds) and variables.

We trained an RBF neural network using the PCA reduced data. The RBF networks structure was determined empirically using 10-fold CV and with repeats until the optimum model was produced with a σ of 10, this is highlighted in figure 5.

Examining Figure 3b and table 2 it is evident that the RBF network has clustered a lot of terms that describe the 'polarity' of a molecule. Relating this to the character of the gp120 binding pocket, the more hydrophobic, the better in terms of an interaction. The neural network also



(a) Biplot of PC1 and PC2

(b) Screeplot

Figure 3: Details of PCA analysis, 15 principal components were retained, explaining 82% of variance

appears to be significantly better at selecting 'drug-like' molecules compared to the other models.

Table 1: Variable contributions to PC1

Descriptor	Definition
Vsurf-CW3	Capacity factor at -1.0
Logp.o.w.	Log octanol/water partition coefficient
Slogp	Log octanol/water partition coefficient
ASA-H	Total hydrophobic surface area
Vsurf-CW1	Capacity factor at -0.2
Q-VSA-HYD	Total hydrophobic vdw surface area
Vsurf-CW2	Capacity factor at -0.5
Vsurf-CW4	Capacity factor at -2.0
Vsurf-D6	Hydrophobic volume at -1.2
Vsurf-D5	Hydrophobic volume at -1.0
POEO-VSA+1	Total negative 1 vdw surface area
FASA-P	Fractional polar surface area
FASA-H	Fractional hydrophobic surface area
Q-VSA-FHYD	Fractional hydrophobic vdw surface area
Q-VSA-FPOL	Fractional polar vdw surface area
PEOE-VSA-FPNEG	Fractional polar negative vdw surface area
Vsa-hyd	VDW hydrophobe surface area
Chi1-C	Carbon connectivity index (order 1)
Chi1v-C	Carbon valence connectivity index (order 1)
PEOE-VSA-FPOL	Fractional polar vdw surface area

Table 2: Variable contributions to PC2

Descriptor	Definition
chi1	Aromatic connectivity index (order 1)
a-heavy	Number of heavy atoms
VDisMa	Vertex distance magnitude index
VAdjEq	Vertex adjacency information (equal)
chi0	Aromatic connectivity index (order 0)
Kier1	First kappa shape index
VAdjMa	Vertex adjacency information (mag)
b-heavy	Number of heavy-heavy bonds
Kier2	Second kappa shape index
vsurf-W1	Hydropholic volume at -0.2
vDistEq	Vertex distance equality index
Zagreb	Zagreb index
PM3-Eele	Electronic energy (Kcal/mol)
weinerPath	Weiner path number
weinerPol	Weiner polarity number
PC-	Total negative partial charge
PC+	Total positive partial charge
Vsurf-Wp1	Polar volume at -0.2

In figure 6 the residual error plot is shown, residual plots are useful for checking regression models performance and to indicate outlier values. In the top left we have the cumulative distribution plot, this is the proportion of values less than or equal to residual errors (on x-axis). It shows an increasing step function that has a vertical jump of 1/N at each value of x-axis equal to an observed value.

In the top right diagram we have the residual errors plotted against the fitted data points. The residuals on the y axis and fitted values (estimated responses) on the x axis. This plot is useful for highlighting outliers, non-linearity and unequal error variances. It indicates that compounds 32, 56 and 106 are marginally different compared with the other compounds in the training data. However, the data points are trended around the 0 line which indicates that the assumption that the relationship is linear is reasonable. The residuals also follow a horizontal shaped cloud around the zero line which in indicates the variances of the error terms are equal.

In the bottom left we have the residual Q-Q plot, if the data are normally distributed the residuals follow a straight line on this plot and for the RBF model it is a good indication they are normally distributed. We can see the same outliers (compounds 32, 56 and 106 identified). In the bottom right we have the absolute residual errors plotted against the fitted values, here the larger the residual absolute value becomes, the further that the point lies from the regression line.

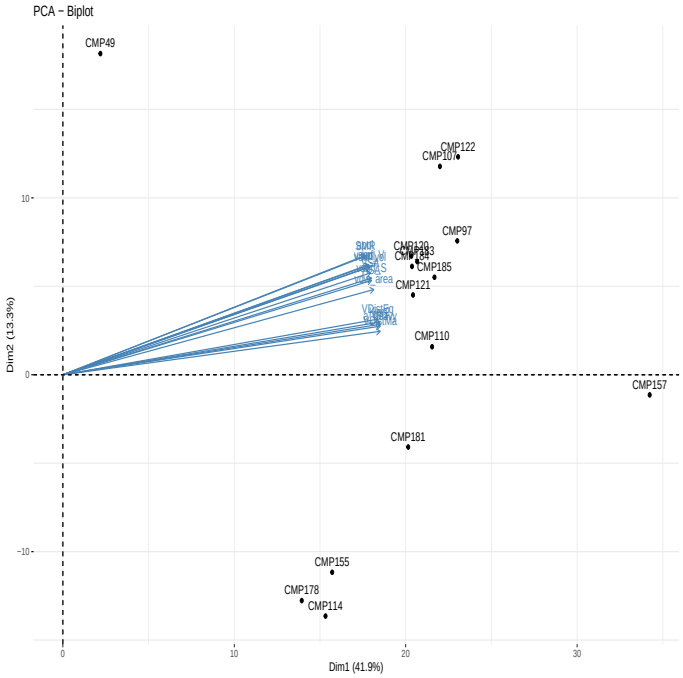


Figure 4: Top 20 contributing compounds and variables.

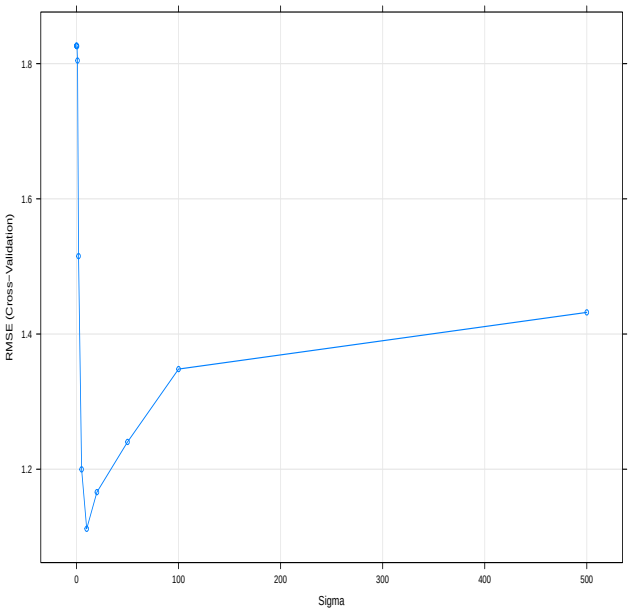


Figure 5: Cross validation tuning of σ parameter for RBF network

4.1. Database screening using trained computational models

All models were developed using 10-fold cross validation with repeats enabling a range of parameters (depending on the model) that were automatically tuned for optimum performance.

In table 3 we compare the RMSE accuracy and MAE accuracy on the training data, this approach is often used in QSAR analysis to ensure that a good model is theoretically possible. These values are derived by passing the training data through the model, hence accuracies are higher than those using only the test data. The Roy and Tropsha measures require the test data to be employed and hence cannot appear in table 3.

Table 3: Model accuracy on training data

Model	R2	RMSE	MAE
RBF	0.90	0.81	0.56
MLP	1.0	0.0003	0.001
SVM	0.93	2.33	1.85
Random Forest	0.97	0.51	0.39
Partial Least Squares	0.70	1.29	1.05
Linear Regression	0.70	1.28	1.05

In table 4 we compare the RMSE accuracy and MAE accuracy on the test data. The Partial Least Squares and the linear regression models had the least accuracy and they identified compounds that were clearly unsuitable as drug-targets. Slightly better results were returned by the Random Forest and SVM models. Only the RBF and MLP neural network had succeeded in identifying viable compounds. The final architecture of the RBF was a layout of 15 input units with 150 basis units, the output layer was a single neuron. The final architecture of the MLP consisted of 25 units in the 1st hidden layer and the 2nd hidden layer with 21 units. One output unit was used.

The RBF network used the Kernel-Based Regularized Least Squares (KRLS) which enables the solution of regression and classification problems. KRLS finds the best fitting function by minimizing a regularization problem with a squared loss. This is achieved by using Gaussian Kernels as the radial basis functions. KRLS reduces bias by adapting the functional form from the data and also allows for interpretability and inference in ways similar to ordinary regression models.

Table 4: Model accuracy on test data

Model	R2	RMSE	MAE	Roy	Tropsha
RBF	0.63	1.18	0.94	0.56	0.02
MLP	0.47	2.06	1.62	0.22	0.57
SVM	0.54	1.12	0.89	0.22	0.65
Random Forest	0.52	1.35	0.98	0.22	0.65
Partial Least Squares	0.47	1.40	1.10	0.39	0.0
Linear Regression	0.47	1.41	1.11	0.38	0.0

Having constructed and tested the regression models using the 187 literature curated compounds we downloaded three random datasets, labeled A, B, and C (each containing 5000 compounds) from the ZINC database website (http://zinc15.docking.org) and used the models to predict their activity against the gp120-CD4 interaction. For each of the datasets A,B and C, the compounds were ranked according to their predicted activity, with the top 10 compounds from A,B and C indicated in table 5.

Examining table 5 the CMP column is our own inter-

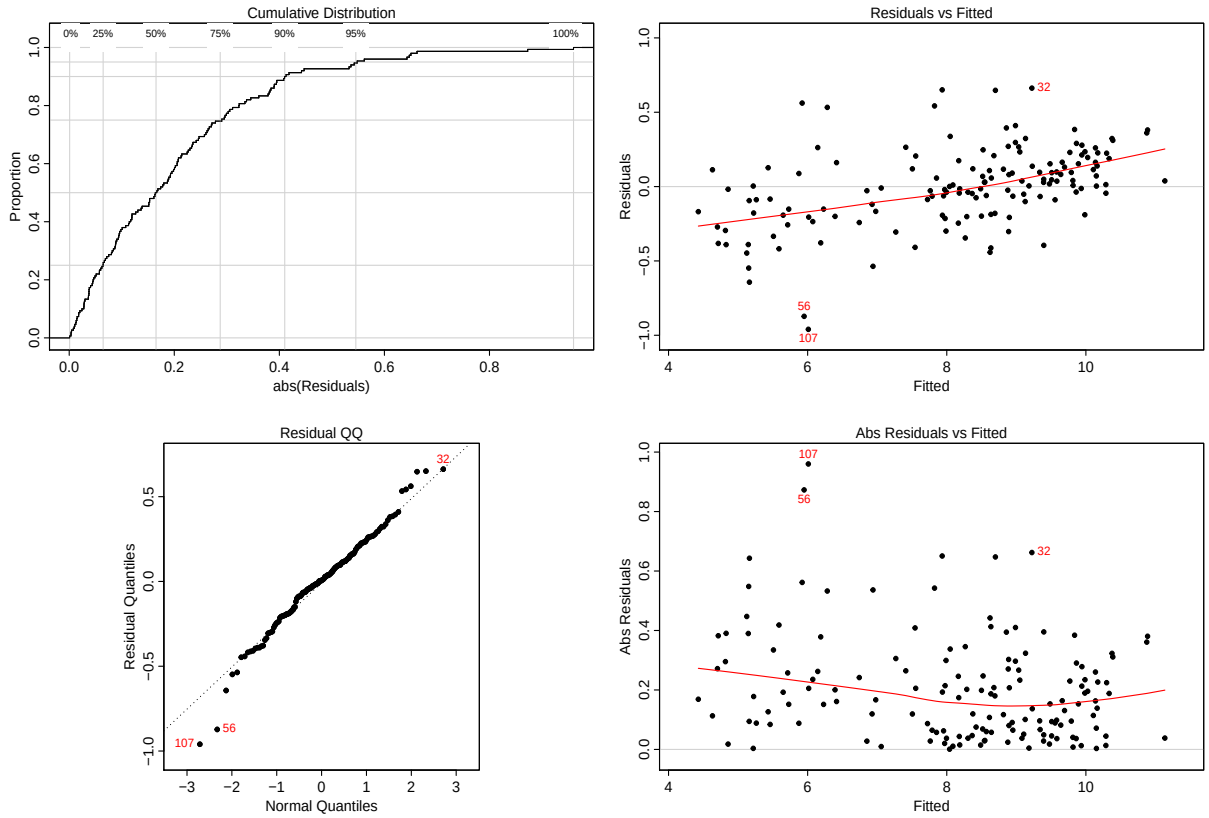


Figure 6: Residual plot of RBF network performance

nal identification for the compounds. The expected pIC50 column refers to the negative log of the IC50 value in molar, generally speaking the larger the pIC50 value the more useful/powerful the compound will be. The next column gives the generally accepted identification from the ZINC database. The LogP column provides predicted value for each compound and is one of Lipinski's rules [24]. More specifically, it is used to give an indication of the lipophilicity of the compounds, and can be defined as the ratio of the concentration of the unionized compound at equilibrium. A negative value for LogP suggests the compound is more hydrophilic. Values for logP approx zero, indicates the compound is equally partitioned between the lipid and aqueous phases. Whereas, a positive LogP value suggests a higher concentration in the lipid phase. The CDOCKER energy column provides information generated by the compound to protein docking process. It is an estimation based upon the internal ligand strain energy and receptor-ligand interaction energy. Similarly, the next column describes the CDOCKER interaction energy, large values implies a greater potential for binding between the protein and the ligand. The CDocker E (1G9M) and Interaction (1G9M) columns are diagnostic outputs used to measure the binding affinity of the test compounds towards the target protein.

In general, the docking energies of the top 30 compounds (comprised of the top 10 from batches A-C) were much higher (high CDOCKER energy) than that of NBD-556 against both the open and closed forms of the viral glycoprotein gp120. For example, compound 2463 has an energy score (CDOCKER energy)= -49.41 kcal mol^{-1} while the NBD-556 compound has a CDOCKER energy= -38.96 kcal mol^{-1} (Note: the more negative binding energy, the tighter the binding between the ligand and the biological target). Figure 8a shows the binding interactions observed with CMP2463.

To assist validation in the docking model we also selected compounds with low scores from the RBF network model. These compounds demonstrated very low binding affinity compared with the top scoring compounds and thus provides greater confidence in our model.

To further confirm the potential inhibitory potential of CMP2463, a MD simulation employing the open form of gp120 was run and the data from the last 1 ns of the equilibrium stage is shown in Table 6, where Time is nanoseconds; Energy values are kcal mol^{-1}; Temperature values listed in kelvin.

5. Discussion

The RBF model outperformed the other regression models when it came to identifying 'drug-like' leads, and in the

Table 5: Top ten compounds identified by the RBF model from data sets A, B and C. All CDOCKER energy values are in $kcal \cdot mol - 1$. CDOCKER energy relates to the energy from the protein-ligand interaction and the CDOCKER interaction energy relates to the nonbonding interaction energy from the protein-ligand interaction. Both parameters were calculated via the Discovery Studio standard protocols.

DATA SET A						
Compound	Predicted pIC50	Zinc Database ID	CDOCKER Energy	CDOCKER.Interaction.Energy	ALogP	Binding Energy
CMP1927	11.238	ZINC39971756	-18.270	-32.980	2.993	-52.755
CMP910	11.155	ZINC15784661	-24.768	-36.179	3.350	-48.050
CMP2554	11.072	ZINC79851413	-19.672	-31.580	3.264	-44.596
CMP3133	11.069	ZINC15784696	-28.888	-38.339	3.813	-55.536
CMP1537	10.958	ZINC73744134	-26.313	-35.303	3.489	-46.425
CMP1348	10.948	ZINC79851405	-11.084	-38.433	1.849	-32.172
CMP466	10.943	ZINC73744123	-26.013	-34.974	3.489	-72.320
CMP2067	10.917	ZINC79018773	-11.100	-34.444	0.680	-118.778
CMP1812	10.876	ZINC51150530	-25.562	-40.271	1.019	-115.241
CMP2385	10.858	ZINC51150533	-26.757	-40.869	1.019	-120.310
DATA SET B						
Compound	Predicted pIC50	Zinc Database ID	CDOCKER Energy	CDOCKER.Interaction.Energy	ALogP	Binding Energy
CMP2783	10.919	ZINC86366942	-29.716	-34.218	1.080	-113.818
CMP3191	10.897	ZINC93999713	-14.306	-40.833	1.064	-129.116
CMP582	10.825	ZINC72246203	-28.996	-42.549	1.196	-107.617
CMP4158	10.825	ZINC72773347	-16.252	-44.047	1.977	-137.570
CMP3257	10.751	ZINC94700306	-35.089	-35.816	-0.238	-139.741
CMP785	10.719	ZINC46503106	-26.045	-35.361	-0.160	-116.005
CMP4964	10.704	ZINC93974761	-25.594	-26.734	1.572	-53.793
CMP3158	10.671	ZINC93604656	-11.940	-29.602	1.416	-107.094
CMP2413	10.669	ZINC74334516	-17.550	-37.770	1.200	-130.524
CMP4271	10.668	ZINC77420512	-26.012	-33.503	-0.306	-140.540
DATA SET C						
Compound	Predicted pIC50	Zinc Database ID	CDOCKER Energy	CDOCKER.Interaction.Energy	ALogP	Binding Energy
CMP2463	10.979	ZINC11911190	-27.746	-49.410	2.146	-138.228
CMP660	10.948	ZINC11911189	-27.750	-51.580	2.146	-135.031
CMP3106	10.902	ZINC67012458	-18.542	-46.609	2.193	-119.507
CMP4524	10.856	ZINC02493577	-25.096	-41.377	2.893	-107.055
CMP1193	10.843	ZINC58295890	-18.598	-40.037	1.214	-122.927
CMP1321	10.830	ZINC67012457	-17.906	-45.928	2.193	-121.347
CMP945	10.773	ZINC37210364	-14.861	-32.838	1.476	-107.686
CMP3058	10.673	ZINC05656549	-14.498	-32.645	3.389	-91.071
CMP5	10.605	ZINC81998334	-8.132	-41.977	0.434	-133.282
CMP2744	10.598	ZINC37210365	-13.937	-31.316	1.476	-129.768

case of all three test datasets, the top ten compounds identified complied with Lipinski's rule. The improved performance of the neural network can be explained by considering the loading's of the terms in both PC1 and PC2; 31.3% of the variance that is captured by PC1 generally related to a molecule's polarity and based on the physicochemical properties of the of the target site on gp120, molecules with a logP in the range 2-4 exhibit favorable binding characteristics.

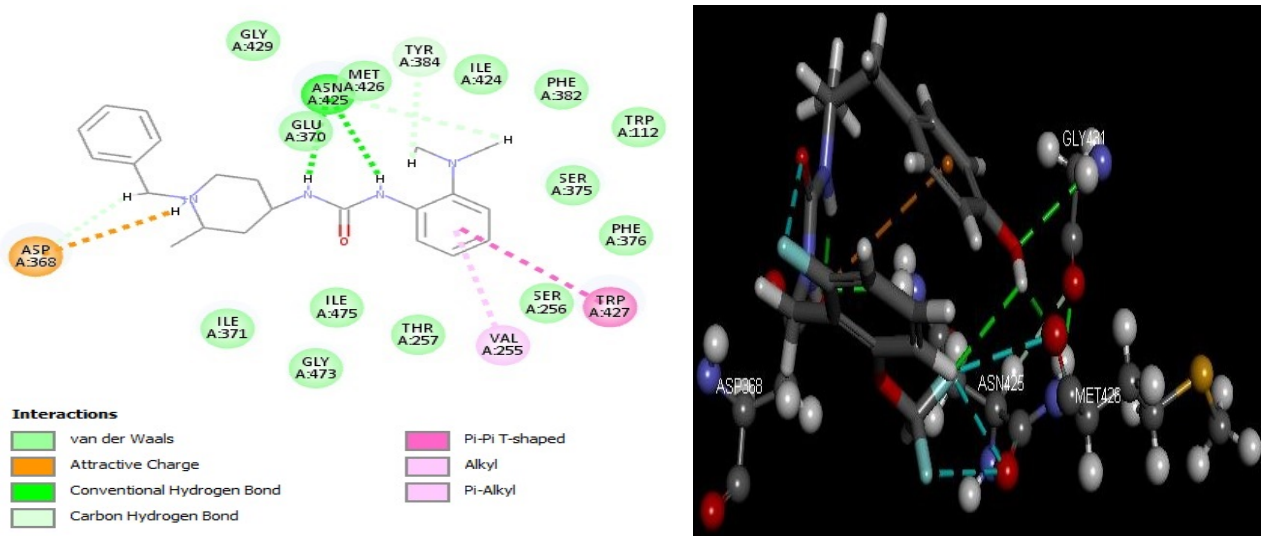
The top 10 compounds from the analysis of each of the batches A-C all targeted the highly conserved Phe43 cavity and displayed very favourable binding interactions with the key amino acid residues; key amino acid residues include Trp112, Val255, Ser 256, Thr 257, Asn425, Met426, Trp427, Val430, Gly473, Met475.

For example, CMP4364 (Figure 7), like NBD-556 can be seen to interact with several key amino acid residues including Val 255, Tyr 384, Asp368, Asn425 and Trp427; also, like NBD-556, the basic amino moiety forms a hydrogen bond to Asp368. In addition, the hydrophobic aromatic ring of CMP4364 pushes deep into the hydrophobic core of the Phe 43 pocket. Also of note with CMP4364 are the presence of two amine moieties in the molecule which would allow the preparation of pharmaceutical salts and hence improve the water solubility of CMP4364. This point is significant since although NBD-556 possess good antiviral activity, it has poor water solubility [55].

Compound 1321 (Figure 8) exhibits extensive binding interactions within the Phe43 cavity; the chlorosubstituted aryl ring makes a number of hydrophobic interactions with the following hydrophobic residues, Val255, Phe382, Ile424 and Trp427. In addition, several hydrogen bonds are formed with residues Asp368, Glu370 and Asn425. Compound 2463 is one of the best performing compounds identified by the RBF model, making many of the same binding interactions in the Phe43 cavity as is seen with NBD-556; Asp368, Asn425, Trp427 and Gly473.

Figure 9 shows the binding interactions of CMP2463 with the Phe43 cavity; the substituted benzyl ring has an electrostatic interaction with Asp368; the substituted piperidine ring has a Van der Waals type interaction with Ile371; the triazole ring makes π stacking and $\pi - \pi$ stacking interactions with Trp427 and Met426 respectively; the two methyl groups of the amide make a number of hydrophobic based interactions with Ile424,Phe382, Trp112, Ser256, Trp427 and Val255.

A MD simulation, which allows the calculation of the empirical free energy of binding was run for one of the top performing compounds identified, CMP2463 (Table 6); total binding energy from the docking simulation was -138.228 kcal mol^{-1} versus an average value of -150.2354 kcal mol^{-1} for the stable conformation generated during



(a) Schematic presentation of interactions between compound 4364 and residues of gp120 protein (PDB ID: 4RZ8)

(b) Molecular modelling image of compound 4364 showing bonding interactions and lowest energy conformation.

Figure 7: Molecular modelling images of compound 4364 generated in Discovery Studio.

Table 6: Binding Free Energy of CMP2463 with gp120 calculated by MD simulation. Key: Time is nanoseconds; Energy values are kcal/mol; Temperature values listed in kelvin.

Step	Time	Total Energy	Kinetic Energy	Potential Energy	Temp	Bond Energy	Angle Energy	Torsion Energy	Improper Torsion Energy	Van der Waals Energy
4100	4.1	150.1769	52.2287	97.9481	307.4	22.5481	31.2062	24.8694	1.681	-4.6673
4200	4.2	150.3716	46.0512	104.3204	271.041	25.8667	33.7893	23.8902	1.4811	-3.072
4300	4.3	150.4226	40.1788	110.2438	236.478	26.5536	37.3468	25.4329	1.3542	-3.7465
4400	4.4	149.8507	58.9242	90.9265	346.807	20.7757	26.04	24.245	0.8982	-3.9755
4500	4.5	150.1733	45.0114	105.1618	264.922	22.5248	36.7968	24.8954	0.8107	-3.5466
4600	4.6	150.1489	53.7562	96.3927	316.39	20.4215	32.7107	23.8365	1.4913	-5.3857
4700	4.7	150.2818	51.8958	98.386	305.44	21.1674	35.8577	22.5301	1.5722	-6.0508
4800	4.8	150.3356	49.5212	100.8144	291.464	22.7718	39.3319	22.4577	1.0156	-6.7101
4900	4.9	150.6016	42.4288	108.1728	249.721	27.5421	36.0916	23.7647	0.6644	-3.6401
5000	5.0	149.9905	54.5455	95.445	321.036	17.3497	34.3373	22.7916	2.6686	-3.4528

the last 1 ns of the MD simulation.

Further modelling work is planned for the best performing compounds identified by the RBF model in order to develop a SAR model for each potential lead. However, based on the observed similarity of binding between our leads and NBD-556 with the Phe43 cavity, we speculate that our compounds could also act as competitive inhibitors for CD4 and hence prevent cell-virus fusion.

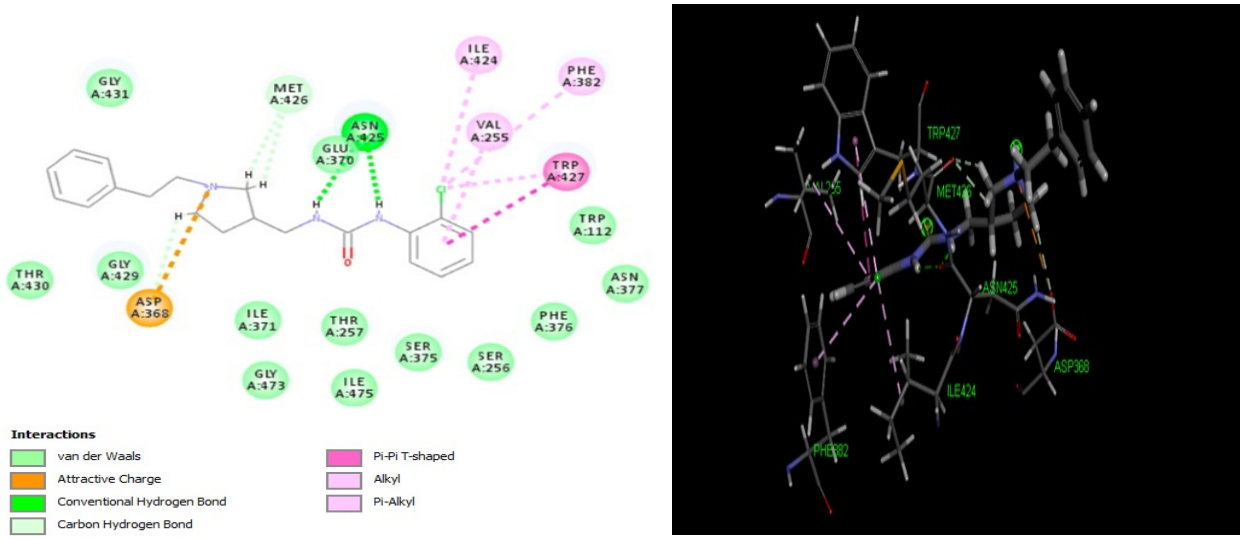
6. Conclusions

It is increasingly the case that modern drug discovery relies on the ability to rapidly run virtual high-throughput screening programmes that can screen as much of chemical space as the computing power will allow in order to identify new potential leads. It is therefore desirable to be able to pre-filter the large datasets involved and hence we have developed an RBF model that can be used to filter large structural databases such as the ZINC database in order to identify leads that target gp120. One of the main advantages of the RBF model we have developed is that once trained, it is using an additional and different set of algorithms compared to many of the commercially available molecular modelling packages to search chemical space (this should help to improve the diversity of the searches). We have also demonstrated that the RBF model is particularly adept at selecting 'drug- like' molecules that comply with Lipinski's rules and also target the Phe43 cavity and indeed, many of the top leads showed excellent binding to the target and made a number of binding interactions that mirrored those seen with NBD556. The ability to target HIV-1 gp120 and hence block viral entry is an attractive point of attack since such agents also offer the possibility of opening the virus via epitope exposure, to an immune response involving broadly neutralizing antibodies.

We are currently synthesising a number of the leads identified for biological screening and will report our results soon.

Conflict of interests

We declare that we have no conflicts of interest.



(a) Schematic representation of compound 1321 bound in the Phe43 cavity showing the key bonding interactions (PDB ID: 4RZ8)

(b) Molecular modelling image for compound 1321 binding to gp120 indicating both the key bonding interactions and the lowest energy conformation (PDB ID: 4RZ8)

Figure 8: Molecular modelling images of compound 1321 generated in Discovery Studio.

Acknowledgments

We would like to thank Professor Roy for providing his software which was used to calculate several quality measures implemented in this paper. We would also like to thank the anonymous reviewers for their helpful comments and suggestions towards improving our manuscript.

References

[1] A. Abdolmaleki, S. Pirhadi, F. Shiri, and J. Ghasemi. Application of multivariate linear and nonlinear calibration and classification methods in drug design. *Comb Chem High Throughput Screen*, 18(8):795–808, 2015.

[2] D Alexander, A Tropsha, and D. Winkler. Beware of R(2): simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of chemical information and modeling*, 55(7):1316–1322, 2015. doi: 10.1021/acs.jcim.5b00206.

[3] T. Backman, Y. Cao, and T. Girke. ChemmineR: A compound mining framework for R. *Bioinformatics*, 24(15):1733–4, 2008.

[4] A. Bender, Z. Yang, B. Eggers, Y. Gong, P. Lin, D. Parker, S. Rahematpura, M. Zheng, N. Meanwell, and J. Kadow. Inhibitors of hiv-1 attachment. part 11: The discovery and structure–activity relationships associated with 4,6-diazaindole cores. *Bioorganic & Medicinal Chemistry Letters*, 23(1): 218–222, 2013.

[5] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[6] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[7] L. Breiman. Random forests. *Machine Learning*, 45(1): 5–32, Oct 2001. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

[8] R. Brent. Fast training algorithms for multilayer neural nets. *IEEE Transactions on Neural Networks*, 2(3):346–354, 1991.

[9] R. Buckheit. Understanding hiv resistance, fitness, replication capacity and compensation: targeting viral fitness as a therapeutic strategy. *Expert Opin Investig Drugs*, 13(8):993–58, 2004.

[10] A. Cherkasov, E. Muratov, and D. Fourches. QSAR modeling: where have you been? where are you going to? *J Med Chem*, 57(12):4977–5010, 2014.

[11] F. Chu and L. Wang. Applications of support vectors machines to cancer classification with microarray data. *International Journal of Neural Systems*, 15(6):475–484, 2005.

[12] R. Cramer. The inevitable QSAR renaissance. *J Comput Aided Mol Des*, 26(1):35–38, 2011.

[13] F. Curreli, Y. Kwon, H. Zhang, D. Scacalossil, and D. Belov. Structure-based design of a small molecule cd4-antagonist with broad spectrum anti-hiv-1 activity. *Journal of Medicinal Chemistry*, 58(17):6909–6927, 2015.

[14] P. Dorr, M. Westby, S. Dobbs, P. Griffin, B. Irvine, M. MaCartney, Julie J. Mori, G. Rickett, C. Smith-Burchnell, and C. Napier. Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor ccr5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrobial Agents and Chemotherapy*, 49(11):4721–4732, 2005.

[15] A Golbraikh and A Tropsha. Beware of q2! *Journal of Molecular Graphics and Modelling*, 20(4):269 – 276, 2002. doi: https://doi.org/10.1016/S1093-3263(01)00123-1.

[16] A. Gross, K. Mobius, C. Haussner, N. Donhauser, B. Schmidt, and J. Eichler. Mimicking protein-protein interactions through peptide-peptide interactions: Hiv-1 gp120 and cxcr4. *Frontiers in Immunology*, 4(257):1–11, 2013.

[17] R. Guha and J. Van Drie. Structure-activity landscape index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model*, 48:646–658, 2008.

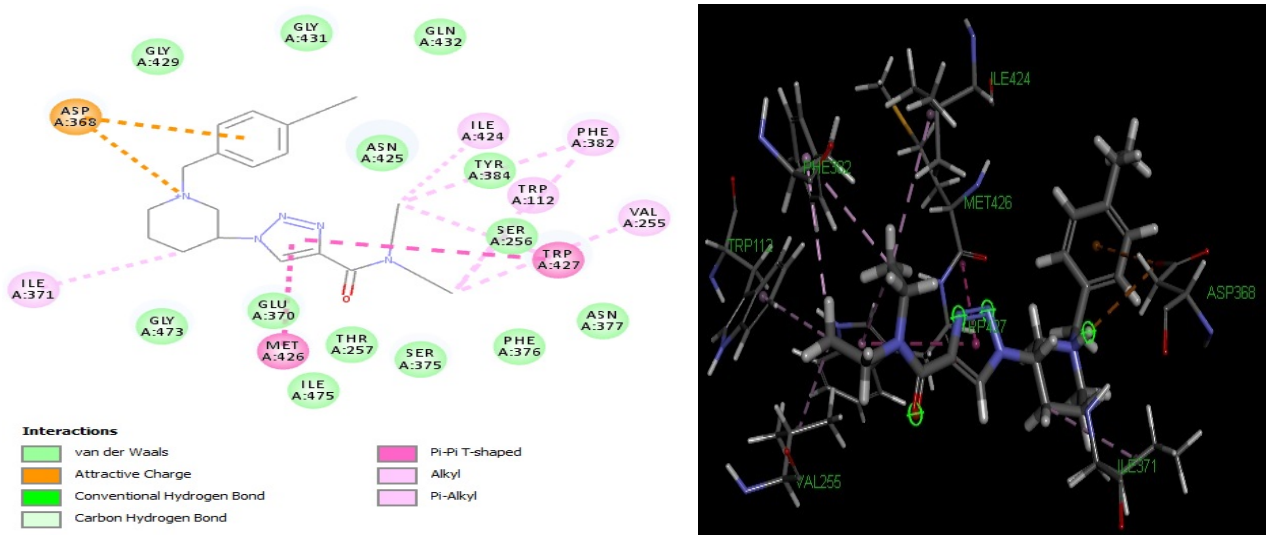
[18] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics*, 47(4):409–443, 2002. ISSN 1097-0134.

[19] O. Hartley, P. Klasse, and Q. Sattentau. V3: HIV's switch-hitter. *AIDS Res Hum Retroviruses*, 21(2):171–189, 2005.

[20] T. Hughes, G. Miller, and S. Swamidass. Modeling epoxidation of drug-like molecules with a deep machine learning network. *ACS Cent Sci*, 1(4):168–180, 2015.

[21] J. Irwin, T. Sterling, M. Mysinger, E. Bolstad, and R. Coleman. Zinc: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, 2012. doi: 10.1021/ci3001277.

[22] P. Kwong, R. Wyatt, J. Robinson, R. Sweet, J. Sodroski, and



(a) Schematic presentation of interactions between compound 2463 and residues of 4RZ8 protein (PDB ID: 4RZ8)

(b) Molecular modelling image of compound 2463 showing bonding interactions and lowest energy conformation.

Figure 9: Molecular modelling images of compound 2463 generated in Discovery Studio.

W. Hendrickson. Structure of an hiv gp120 envelope glycoprotein in complex with the cd4 receptor and a neutralizing human antibody. *Nature*, 393(6686):648–659, 1998.

[23] S. Le, J. Josse, and F. Husson. Factominer: An r package for multivariate analysis. *Journal of Statistical Software, Articles*, 25(1):1–18, 2008. ISSN 1548-7660. doi: 10.18637/jss.v025.i01. URL `https://www.jstatsoft.org/v025/i01`.

[24] C. Lipinski, C. Lombardo, F. Dominy, and B. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development. *Advanced Drug Delivery Reviews*, 23:3–25, 1997.

[25] M. Louie and M. Markowitz. Goals and milestones during treatment of hiv-1 infection with antiretroviral therapy: A pathogenesis-based perspective. *Antiviral Res*, 55(1):15–25, 2002.

[26] D. Lowe. On the use of nonlocal and non positive definite basis functions in radial basis function networks. In *Proceedings of the 3rd International Conference on Artificial Neural Networks*, pages 206–211, Cambridge, UK, 1995.

[27] S. Lu, J. Wu, H. Liu, J. Zhao, K. Liu, C. Chuang, H. Lini, W. Tsai, and Y. Ho. The discovery of potential acetylcholinesterase inhibitors: A combination of pharmacophore modeling, virtual screening, and molecular docking studies. *Journal of Biomedical Science*, 18:8, 2011.

[28] A. Lusci, G. Pollastri, and P. Baldi. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J Chem Inf Model*, 53(7): 1563–1575, 2013.

[29] N. Madani, A. Perdigoto, K. Srinivasan, J. Cox, J. Chruma, and J. LaLonde. Localized changes in the gp120 envelope glycoprotein confer resistance to human immunodeficiency virus entry inhibitors. *Journal of Virology*, 78:3742–3752, 2004.

[30] T. Matthews, M. Salgo, M. Greenberg, J. Chung, R. DeMasi, and D. Bolognesi. Enfuvirtide: the first therapy to inhibit the entry of hiv-1 into host cd4 lymphocytes. *Nature Reviews Drug Discovery*, 3(3):215–225, 2004.

[31] N. Meanwell, O. Wallace, H. Fang, H. Wang, M. Deshpande, T. Wang, Z. Yin, Z. Zhang, B. Pearce, and J. James. Inhibitors of hiv-1 attachment. part 2: An initial survey of indole substitution patterns. *Bioorganic & Medicinal Chemistry Letters*, 19 (7):1977–1981, 2009.

[32] N. Meanwell, O. Wallace, H. Fang, H. Wang, M. Deshpande, T. Wang, Z. Yin, Z. Zhang, B. Pearce, and J. James. Inhibitors of hiv-1 attachment. part 3: A preliminary survey of the effect of structural variation of the benzamide moiety on antiviral activity. *Bioorganic & Medicinal Chemistry Letters*, 19(17): 5136–5139, 2009.

[33] N. Meanwell, M. Krystal, B. Nowicka-Sans, D. Langley, D. Conlon, and M. Eastgate. Inhibitors of hiv-1 attachment: The discovery and development of temsavir and its prodrug fostemsavir. *J Med Chem*, 61(1):62–80, 2018.

[34] M. Pancera, Y. Lai, T. Bylund, A. Druz, S. Narpala, and S. O'Dell. Crystal structure of trimeric hiv envelope with entry inhibitors bms-378806 and bms-626529. *Nature Chemical Biology*, 13(10):1115–1122, 2017.

[35] M. Patel, M. Yanagishitama, and G. Roderiquez. Cell-surface heparan sulfate proteoglycan mediates hiv-1 infection of t-cell lines. *AIDS Res Hum Retroviruses*, 9(2):167–74, 1993.

[36] K. Qian, S. Morris-Natschke, and K. Lee. Hiv entry inhibitors and their potential in hiv therapy. *Medicinal research reviews*, 29(2):369–393, 2009.

[37] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `https://www.R-project.org/`.

[38] D. Rayalu, C. Selvaraj, S. Singh, R. Ganeshan, N. Kumar, and P. Seshapani. Homology modeling, active site prediction, and targeting the anti hypertension activity through molecular docking on endothelin – b receptor domain. *Bioinformation*, 8 (2):81–86, 2012.

[39] A. Regueiro-Ren, Q. Xue, J. Swidorski, Y. Gong M. Mathew, D. Parker, Z. Yang, B. Eggers, C. D'Arienzo, and Y. Sun. Inhibitors of human immunodeficiency virus type 1 (hiv-1) attachment. 12. structure-activity relationships associated with 4-fluoro-6-azaindole derivatives leading to the identification of (bms-585248). *Journal of Medicinal Chemistry*, 56(4): 1656–1669, 2013.

[40] K. Roy and A. Mandal. Development of linear and nonlinear predictive qsar models and their external validation using molecular similarity principle for anti-hiv indolyl aryl sulfones. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 23(6): 980–995, 2008. doi: 10.1080/14756360701811379.

[41] K Roy, P Chakraborty, I Mitra, P Ojha, S Kar, and R Das. Some case studies on application of "rm2" metrics for judging quality of quantitative structure–activity relationship predictions: Emphasis on scaling of response data. *Journal of Computational Chemistry*, 34(12):1071–1082, 2013. doi: 10.1002/jcc.23231.

[42] K. Roy, R. Das, P. Ambure, and R. Aher. Be aware of error measures. further studies on validation of predic-

tive qsar models. *Chemometrics and Intelligent Laboratory Systems*, 152:18 – 33, 2016. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2016.01.008.

[43] R. Sabet and A. Fassih. QSAR study of antimicrobial derivatives using different chemometric tools. *Int J Mol Sci*, 9(12): 2407–2423, 2008.

[44] H. Shah. Protein secondary structure prediction using support vector machines SVMs. In *Proceedings of the 2013 International Conference on Machine Intelligence and Research Advancement*, pages 594 – 598, Jammu, India, 2013.

[45] T. Sterling and J. Irwin. ZINC 15 – Ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11): 2324–2337, 2015. doi: 10.1021/acs.jcim.5b00559.

[46] J. Sterne, M. Hernan, and B. Ledergerber. Long-term effectiveness of potent antiretroviral therapy in preventing aidsand death: a prospective cohort study. *Lancet*, 366(9483):378–84, 2005.

[47] C. Tang, M. Zhang, S. Majeed, E. Montabana, R. Stanfield, D. Dimitrov, B. Korber, J. Sodroski, and I. Wilson. Structure of a v3-containing hiv-1 gp120 core. *Science (New York, N.Y.)*, 310(5750):1025–1028, 2005. ISSN 0036-8075.

[48] T. Wang, J. Kadow, Z. Zhang, Z. Yin, Q. Gao, D. Wu, D. Parker D. DiGiugno, Z. Yang, L. Zadjura, and B. Robinson. Inhibitors of hiv-1 attachment. part 4: A study of the effect of piperazine substitution patterns on antiviral potency in the context of indole-based derivatives. *Bioorganic & Medicinal Chemistry Letters*, 19(17):5140–5145, 2009.

[49] T. Wang, Z. Yang, Z. Zhang, Y. Gong, K. Riccardi, P. Lin, D. Parker, S. Rahematpura, M. Mathew, and Ming M. Zheng. Inhibitors of hiv-1 attachment. part 10. the discovery and structure–activity relationships of 4-azaindole cores. *Bioorganic & Medicinal Chemistry Letters*, 23(1):213–217, 2013.

[50] S. Xing and R. Siliciano. Targeting HIV latency: pharmacologic strategies toward eradication. *Drug Discovery Today*, 18(11): 541–545, 2012.

[51] K. Yeung, Z. Qiu, Q. Xue, H. Fang, Z. Yang, L. Zadjura, C. D'Arienzo, B. Eggers, and K. Riccardi. Inhibitors of hiv-1 attachment. part 9: An assessment of oral prodrug approaches to improve the plasma exposure of a tetrazole-containing derivative. *Bioorganic & Medicinal Chemistry Letters*, 23(1):209–212, 2013.

[52] K. Yeung, Z. Qiu, Q. Xue, H. Fang, Z. Yang, L. Zadjura, C. D'Arienzo, B. Eggers, and K. Riccardi. Inhibitors of hiv-1 attachment. part 8: The effect of c7-heteroaryl substitution on the potency, and in vitro and in vivo profiles of indole-based inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 23(1): 203–208, 2013.

[53] K. Yeung, Z. Qiu, Q. Xue, H. Fang, Z. Yang, L. Zadjura, C. D'Arienzo, B. Eggers, K. Riccardi, and P. Shi. Inhibitors of hiv-1 attachment. part 7: Indole-7-carboxamides as potent and orally bioavailable antiviral agents. *Bioorganic & Medicinal Chemistry Letters*, 23(1):198–202, 2013.

[54] A. Zakharov, M. Peach, M. Sitzmann, and M. Nicklaus. QSAR modeling of imbalanced high-throughput screening data in PubChem. *Chemical Information and Modeling*, 54:705–712, 2009.

[55] Q. Zhao, L. Ma, S. Jiang, H. Lu, and S.Liu. Identification of n-phenyl-n'-(2,2,6,6-tetramethyl-piperidin-4-yl)-oxalamides as a new class of hiv-1 entry inhibitors that prevent gp120 binding to cd4. *Virology*, 339(2):213–225, 2005.