Med Chem Res
DOI 10.1007/s00044-017-2069-5

MEDICINAL CHEMISTRY RESEARCH

ORIGINAL RESEARCH

QSAR studies of indoyl aryl sulfides and sulfones as reverse transcriptase inhibitors

Pablo R. Duchowicz[1] · Daniel E. Bacelo[2] · Silvina E. Fioressi[2] · Valeria Palermo[3] · Nnenna E. Ibezim[4] · Gustavo P. Romanelli[3,5]

Received: 21 April 2017 / Accepted: 6 September 2017

Abstract The inhibitory HIV reverse transcriptase activity of 172 non-nucleoside indoyl aryl sulfones and sulfides is studied with a QSAR analysis, in order to identify the molecular characteristics influencing the interaction with the reverse transcriptase enzyme. This work increases the available QSAR studies of indoyl aryl sulfones and sulfides using the reported experimental EC_{50} values against HIV-1 wild type (IIIB) in human T-lymphocyte (CEM) cells. Different approaches are proposed, involving 0D, 1D and 2D molecular descriptors from PaDEL freeware, and also based on flexible descriptors from CORAL freeware. Three models are finally presented, which correlate the inhibitory HIV reverse transcriptase activity with good accuracy. It is demonstrated that the established models are predictive in the validation process. The novelty of the present work relies on the development of structure-inhibitory HIV activity relationships, through a computational technique that does not require the knowledge of the molecular conformation during the structural representation. The obtained results would contribute to guide the design of more effective compounds for HIV treatment.

Keywords Indoyl aryl sulfides and sulfones · HIV reverse transcriptase inhibitors · QSAR · PaDEL · CORAL

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00044-017-2069-5) contains supplementary material, which is available to authorized users.

✉ Pablo R. Duchowicz
pabloducho@gmail.com

✉ Daniel E. Bacelo
dbacelo@aol.com

[1] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, Diag. 113y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina

[2] Departamento de Química, Facultad de Ciencias Exactas y Naturales, Universidad de Belgrano, Villanueva 1324, CP 1426 Buenos Aires, Argentina

[3] Centro de Investigación y Desarrollo en Ciencias Aplicadas "Dr. J. J. Ronco" (CINDECA), Departamento de Química, Facultad de Ciencias Exactas, CONICET, UNLP, Calle 47 No. 257, B1900AJK La Plata, Argentina

[4] Department of Computer Education, University of Nigeria, Nsukka, Nigeria

[5] Cátedra de Química Orgánica, Centro de Investigación en Sanidad Vegetal (CISaV), Facultad de Ciencias Agrarias y Forestales, Universidad Nacional de La Plata, Calles 60y 119s/n, B1904AAN La Plata, Argentina

Introduction

Human immunodeficiency virus (HIV) is a pathogenic lentivirus of the Retroviridae family, a causative factor for the acquired immune deficiency syndrome (AIDS). Since this virus has been identified in the early 1980s, more than 25 milion people have died due to this disease (Defant et al. 2015; Ashok et al. 2015a). The antiretroviral therapy (ART) for people with HIV has caused an important decrease in mortality in the last decade (UNAIDS gap report 2015). However, AIDS-related diseases are still one of the leading causes of death and are expected to cause a significant premature mortality in the coming decades (La Regina et al. 2011; Piscitelli et al. 2009; WHO 2014).

The therapeutic strategies against the HIV involve enzymes that act in the viral replication cycle: virus adsorption, virus–cell fusion, virus uncoating, and enzyme

Published online: 23 September 2017

inhibition (Nguyen Van Nhien et al. 2005). The main implicated enzymes are: HIV reverse transcriptase (RT), which is essencial for the replication of the virus; HIV integrase (IN), which catalyzes the integration of viral DNA into the host DNA before replication (Pannecouque et al. 2010); and HIV protease, which is required in the final maturation of newly formed viruses to make up an infectious virion (Roy and Leonard 2004; Broder and Fauci 1988). The RT inhibitors have been the first agents approved for the treatment of HIV-1, and, together with protease inhibitors, are the most attractive targets for the anti-HIV drug development process, meanwhile IN inhibitors are still in clinical trials. Other widely used anti-HIV agents are: fusion inhibitors (FIs), co-receptor inhibitors (CRIs), and the viral entry inhibitor Enfuvirtide (Piscitelli et al. 2009; Roy and Leonard 2004; Oversteegen et al. 2007; De Clercq 2009; De Martino et al. 2006; Bonini et al. 2014; Williams 2003).

HIV is a single stranded (ss) RNA virus, which produces a double stranded (ds) DNA provirus, the RT enzyme catalyzes this process in the cytoplasm. The generated ds DNA is then integrated into the host cell genome (Ashok et al. 2015b; Musah 2004). There are three classes of RT inhibitors: nucleoside RT inhibitors (NRTIs), nucleotide RT inhibitors (NtRTIs), and non-nucleoside RT inhibitors (NNRTIs). NNRTIs do not require preliminary phosphorylation and are less toxic than nucleoside analogs, because they do not affect the activity of cellular polymerases. They lock themselves into inactive conformations by fitting into an allosteric site (a flexible hydrophobic pocket), ~10 Å from the polymerase active site (the NNRTI binding site (NNBS)). This causes several conformational changes and induces restrictions in the dynamics of the enzyme and its inactivity (Defant et al. 2015; Ashok et al. 2015b; Hunter et al. 2008; Ribone et al. 2012; Tantillo et al. 1994). Despite the potent antiviral activity of NNRTIs, they produce the rapid emergence of drug resistance caused by mutations of some amino acids in the NNBS (Silvestri et al. 2003; Samuele et al. 2011).

There are NNRTIs of diverse scaffolds: multicyclic, benzo-fused heterocycles, six-membered and five-membered heterocycles, amide or thioamide linker, and diphenyl among others. X-ray crystallographic studies of NNRTIs in complex with RT have shown that most of them contain a common pharmacophore model, with a "butterflylike" shape with one "body" and two hydrophobic "wings". More than 50 structurally diverse NNRTIs have been identified (Nguyen Van Nhien et al. 2005; Ashok et al. 2015b; Silvestri et al. 2003; Witvrouw et al. 2000). They can be classified into first and second generation classes. The first generation NNRTIs are inflexible hydrophobic compounds, such as efavirenz (Sustiva®), nevirapine (Viramune®), and delavirdine (Rescriptor®), that could effectively inhibit proliferation of the wild-type (WT) HIV. The rapid emergence of resistance forced the development of new NNRTI, the second generation NNRTIs, as a result of strategies involving computational chemistry (molecular modeling), structure-based rational drug design and synthesis, along with biological and pharmacokinetic assays. Generally, second generation NNRTIs tend to be more active against WT and mutant viruses, have minimal side effects, are more flexible, and can inhibit HIV strains resistant to first generation NNRTIs (Piscitelli et al. 2009; De Martino et al. 2006; Zhan et al. 2013; Samuele et al. 2009; Singh et al. 2012; Ragno et al. 2005). There are many second-generation NNRTIs reported in the literature, with many different scaffolds such as diarylpyrimidines (Ribone et al. 2012), heteroaryl esters (Cesarini et al. 2010), amides and thioamides (Hunter et al. 2008), multicyclics (Pannecouque et al. 2010; Witvrouw et al. 2000), benzophenones, and benzimidazoles (Pan et al. 2015; Di Santo and Costi 2005; Artico et al. 2000; Guendel et al. 2014). Among the diarylsulfones, the 2-nitrophenyl phenyl sulfone (NPPS) presents the best antiviral activity of this family (McMahon et al. 1993). Another important kind of compounds with activity against RT are the 3-pyrrolyl aryl sulfones (PASs), first described in 1995. They are more potent and selective than NNPS due to the presence of a *p*-chloroanilino moiety (Silvestri et al. 2003; Di Santo and Costi 2005; Artico et al. 1995).

The first indoyl aryl sulfone (IAS) NNRTI has been reported by Williams et al. in 1993, the L-737,126 (5-chloro-3-(phenylsulfonyl)indole-2-carboxamide) (Williams et al. 1993). Despite its potent activity, it has not been suitable for clinical trials because of its poor water solubility (Williams et al. 1994). The Silvestri group has explored chemical modifications to L-737,126 IAS at the indole-2-carboxamide function in order to improve the activity against mutants and solubility. Changing the position of benzenesulfonyl moiety from position 3 to 1 (De Martino et al. 2006; Silvestri et al. 2003), modification of the carboxyamide side chain and shift from position 2 to 3 (Samuele et al. 2009), and replacement of carboxyamide by a carboxyhydrazide chain (Ragno et al. 2005) did not result in more potent inhibitors (Artico et al. 2000; La Regina et al. 2007; Ragno et al. 2006; Silvestri et al. 2004). The replacement of the 3-phenylsulfonyl with a pyrrolidin-1-ylsulfonyl moiety or the introduction of an additional heterocycle at position 2 of the indole has also been studied (Zhao et al. 2008). The introduction of an amino acid to the carboxamide chain produced more active derivatives than L-737,126 against the mutant strains HIV-1 WT (Piscitelli et al. 2009; Silvestri et al. 2004; Young et al. 1995). Related esters and hydrazides were always less potent, whereas the carboxylic acids were completely inactive (Silvestri et al. 2003). Cancio et al. (2005) have studied the mechanism of

inhibition of HIV-1 RT by IAS. It was found that IAS could be made selective for different enzyme–substrate complexes, depending on the substitutions on the IAS (Silvestri et al. 2003, 2004; Samuele et al. 2009). Halo and nitro IASs have also been studied, and it was found that derivatives bearing two halogen atoms at positions 4 and 5 of the indole were an optimal substitution pattern for the antiviral activitiy of IASs (La Regina et al. 2007, 2011; Samuele et al. 2009). In general, derivatives bearing an amide terminus were more cytotoxic than hydrazide. On the other hand, amides were more potent than hydrazides against the HIV-1 WT and the NNRTI-resistant strains (Young et al. 1995). With respect to indoyl aryl sulfides, sulfone derivatives were less cytotoxic and more potent (Silvestri et al. 2003).

Quantitative structure–activity relationships (QSARs) have proven to be a useful tool in compound design, since they represent a cheaper and faster alternative to in vivo and in vitro assays (Consonni et al. 2009; Duchowicz et al. 2009; Weaver and Gleeson 2008). For example, an extensive 3D-QSAR study and Docking simulation has been conducted on IASs, comprising NNRTIs involving 83 molecules, 75 of them taking part as the training set and eight of them as the test set (Ragno et al. 2005). The models found present correlation coefficients in the range (0.79, 0.93), and cross-validated standard deviations of prediction falling in the range (0.69, 1.07). Based on such models, 24 new IAS derivatives were proposed.

The aim of the present study is to perform a QSAR analysis for the RT inhibitory activity of reported indoyl aryl sulfones and sulfides, in order to identify the molecular characteristics influencing the interaction with NNRT. In this way, we expect to contribute to the design of more effective compounds for HIV treatment.

Models involving quantum-chemical descriptors imply difficult calculations of the optimum molecular geometries, high computational costs, and long times. In contrast, conformation independent 0D-2D-QSAR methods emerge as an alternative approach for developing simple models based on constitutional and topological molecular features (Aranda et al. 2016). The exclusion of 3D-structural aspects also avoids problems associated with ambiguities resulting from the existence of various conformational molecular states. For this reason, we use three different QSAR approaches for establishing predictive models. In the first one, the freely available descriptor generator PaDEL program (PaDEL 2016) is used to generate 1444 0D-2D molecular descriptors and 16,092 fingerprints. In the second approach, the CORALSEA freeware (CORALSEA 2016) is used to generate conformation independent flexible descriptors. Finally, we also explore models combining both descriptor types.

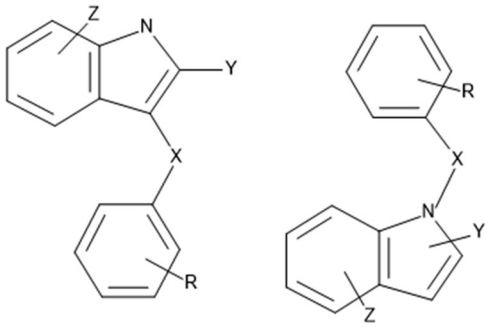

Fig. 1 Base structures of the indoyl aryl sulfones and sulfides used in this study

Methodology

Experimental dataset

The QSAR analysis is performed on 172 indoyl aryl sulfones and sulfides (see Fig. 1), which are found to be active against HIV-1 wild type (IIIB). IAS have been evaluated against the HIV-WT in human T-lymphocyte cells, and the inhibitory activity is quantified through the EC_{50} (concentration required to protect CEM cells against the cythopathogenicity of HIV by 50%). For the QSAR analysis, the $-\log_{10} EC_{50}$ (pEC_{50}) is used as a measure for the inhibitory activity. The experimental data are collected from the literature (La Regina et al. 2011; Piscitelli et al. 2009; Silvestri et al. 2003, 2004; Ragno et al. 2005; Young et al. 1995; Prajapati and Doshi 2011) and presented in Table 1S of Supplementary Materials, along with the molecular structural details.

Structural representation and molecular descriptors calculation

The structures of the compounds are generated in SMILES notation and verified for their correctness, and afterwards 2D structures are drawn with ACDLabs ChemSketch freeware (ChemSketch 2015). The descriptors are calculated using two different methodologies:

a. Theoretical conformation-independent molecular descriptors and fingerprints are calculated using the freely available PaDEL software (version 2.20) (PaDEL 2016). Before calculating the descriptors, a conversion to MDL-mol format (the recommended format in PaDEL) is performed using Open-Babel version 2.3.2 (The Open Babel Package 2015). Constant values and descriptors found to be pairwise correlated are excluded from the original matrix of variables to minimize redundant information. In total, 1444 1D and 2D descriptors and 12 fingerprint types (16092) are obtained. These are calculated using The

Chemistry Development Kit, and additional descriptors are added, such as atom type electrotopological state descriptors, Crippen's logP and MR, extended topochemical atom (ETA) descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, count of chemical substructures identified by Laggner, and binary fingerprints and count of chemical substructures identified by Klekota and Roth (Yap 2011) resulting in a total of 17,536 non-conformational descriptors explored.

b. Flexible molecular descriptors are calculated with the CORAL freeware (CORALSEA 2016). At first, the SMILES notations of the compounds are provided as input to the CORAL program, together with the studied experimental activity values (pEC_{50}). Three different structural representation (SR) approaches are available in CORAL: i. a chemical graph, like hydrogen-suppressed graph (HSG), hydrogen-filled graph (HFG), and graph of atomic orbitals (GAO); ii. SMILES; and iii. a hybrid representation which includes both graph and SMILES (CORALSEA 2016). The most appropriate combination of structural attributes (local descriptors, SA) are chosen for the modeling process. The selected SR, i.e. graph-based or SMILES-based, defines the number and types of local descriptors that participate in the QSAR analysis.

In the graph approach of the HSG, HFG, or GAO type, the structural attributes that can be used are the Morgan's extended connectivity indices of kth order for vertex (atom) Z ($^{k}EC_{Z}$, $k = 0$–3). The index of zero-th order $^{0}EC_{Z}$ represents the vertex degree for atom Z (number of neighbor atoms to Z in HSG), while the higher order indices $^{k}EC_{Z}$ are obtained through a recursive formula based on $^{0}EC_{Z}$ (see in Table 2S). In the SMILES approach, the one-element, two-element, and three-element SMILES attributes $^{1}s_{k}$, $^{2}s_{k}$, $^{3}s_{k}$, respectively, can be calculated.

In order to achieve the best statistical quality of the final model, the most relevant structural attributes for a specific structural representation are found in a stepwise manner. First, the single best attribute is searched; after that, a second attribute that best combines with the first is searched, and then the following attributes are continuously added in the same way.

In the CORAL framework, the QSAR model is obtained through a one-variable linear correlation between pEC_{50} and a properly defined flexible descriptor (DCW, descriptor of correlation weights). The DCW descriptor is a linear combination of special coefficients called correlation weights (CW). A CW value is calculated for each SA type of the training set. The CW values for all the structural attributes are calculated through the Monte Carlo (MC) simulation method, by searching for the highest correlation coefficient (r) between pEC_{50} and the DCW descriptor (Table 2S).

The DCW flexible descriptor depends upon the threshold value (T) and the number of epochs or iterations (N_{epochs}) used. These parameters are positive integers from the MC method that should be correctly specified in order to calculate the DCW values. The T defines rare (noise) SMILES attributes that do not contribute to the predicted inhibitory activity, so that all SMILES attributes that take place in less than T SMILES notations of the training set are classified as rare instead of as active. N_{epochs} is the number of iterations used during the numerical optimization procedure. In current study, T ranges from 0 to 5 and the maximum number of iterations used is 50.

Model development

Molecular descriptors selection in MLR

We employ the replacement method (RM) technique (Duchowicz et al. 2006) in order to generate MLR models on the training set, by searching in a pool having $D =$ 17,538 descriptors for optimal subsets having $d \ll D$ ones with smallest values for the standard deviation (S_{train}) or the root mean square deviation (RMS_{train}). Table 2S includes a list of mathematical equations involved in the present study. All the Matlab (Matlab 2016) programmed algorithms used in our calculations are available upon request.

Model validation

In order to verify the predictive capability of the proposed QSAR models, the dataset is splited into a training set (58 compounds) used for model development, a validation set (57 compounds) used for partial model validation, and a test set (57 compounds) used for model external validation. The split of the dataset is performed with the balanced subsets method (BSM), which is based on k-means cluster analysis (k-MCA) (Rojas et al. 2015a, 2015b). The BSM ensures that the training set is representative of both the validation and test sets, and thus similar structure–property relationships are found in the three molecular subsets.

Results and discusion

The 172 selected indoyl aryl sulfones and sulfides are potent inhibitors of HIV-1 WT replication in CEM cells, and show inhibitory concentrations in the low nanomolar range. Statistical parameters for the three explored models are provided in Table 3S–11S.

QSAR models based on PaDEL descriptors

The initial pool of 1444 PaDEL descriptors is first reduced to 1016 descriptors due to its linear dependency. Table 1

Table 1 Descriptors identified for modeling the inhibitory HIV reverse transcriptase activity along with the squared correlation coefficient and the standard deviation for training, validation, and test sets.

d	Descriptor symbols	R^2_{train}	S_{train}	R^2_{val}	S_{val}	R^2_{test}	S_{test}
1	*NaasC*	0.51	1.37	0.58	1.04	0.48	1.02
2	*naasC, maxHBint5*	0.71	1.06	0.67	0.94	0.67	0.85
3	*AATSC7v, MATS4e, maxHBint5*	0.79	0.91	0.69	0.90	0.67	0.84
4	*GATS8m, SpMin6_Bhv, maxHBint4, MDEO-12*	0.81	0.87	0.74	0.85	0.62	0.89
5	*ATSC8v, BCUTc-1h, ndsN, minHaaNH, JGI4*	0.83	0.82	0.83	0.72	0.76	0.72
6	*AATS7i, MATS7p, GATS4s, maxHBint5, maxHaaNH, SRW9*	**0.86**	**0.77**	**0.82**	**0.73**	**0.81**	**0.68**
7	*ALogp2, nAromBond, MATS7p, GATS8v, GATS4s, maxHBint5, maxaaNH*	0.87	0.75	0.79	0.82	0.72	0.87

The best model appears in bold

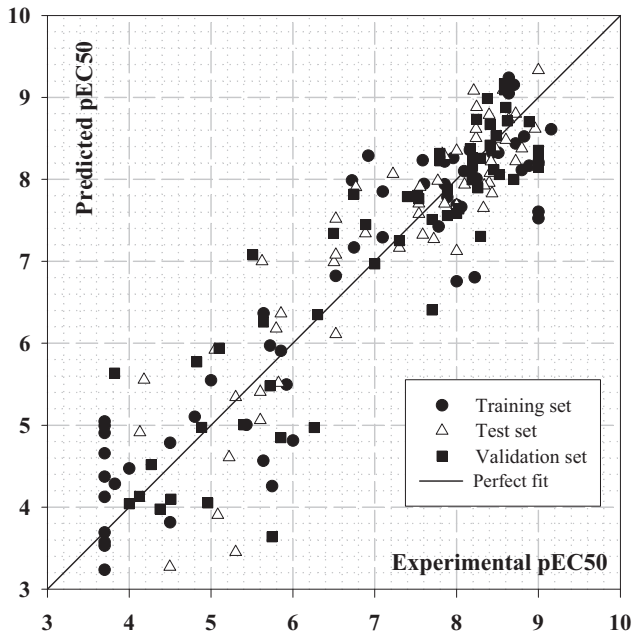

Fig. 2 Predicted and experimental pEC_{50} values for the training, validation and test sets (Eq. 1)

includes the QSAR models obtained, revealing that the best predictive performance is achieved for six molecular descriptors. Figure 2 plots the calculated pEC_{50} as a function of the experimental values for the equation:

$$pEC_{50} = -0.26\,AATS7i + 7.27\,MATS7p - 2.81\,GATS4s$$
$$+0.25\,maxHBint5 + 3.91\,maxHaaNH + 1.34\,SRW9 + 40.79 \quad (1)$$

$$N_{train} = 58,\ R^2_{train} = 0.86,\ S_{train} = 0.77,\ F = 51,\ o2.5 = 0$$
$$N_{val} = 57,\ R^2_{val} = 0.82,\ S_{val} = 0.73$$
$$N_{test} = 57,\ R^2_{test} = 0.81,\ S_{test} = 0.68$$

Here, F is the Fisher parameter and $o2.5$ indicates the number of outlier compounds in the training set having a residual (difference between experimental and calculated pEC_{50}) greater than 2.5-times S_{train} and lower than 3-times S_{train}.

The descriptors *MATS7p*, *AATS7i*, and *GATS4s* are 2D-autocorrelation descriptors of the topological structure of Broto–Moreau (ATS), Moran (MATS), and Geary (GATS). As a result, these descriptors consider the topology of the structure or parts of it associated with a selected physicochemical atomic property. The two indices following the descriptor symbol represent the topological distance between pairs of atoms or lag, and the physicochemical property considered in the weighting component for its computation. For example, the *MATS7p* descriptor represents a Moran autocorrelation with lag 7 weighted by polarizability. In the above model the *MATS7p* descriptor correlates positively to the inhibitory activity, whereas the *AATS7i* (average Broto–Moreau autocorrelation, lag 7, weighted by first ionization potential) and *GATS4s* (Geary autocorrelation, lag 4, weighted by I-state) contribute negatively to the activity. The electrotopological state atom type descriptors *maxHBint5* (maximum E-state descriptors of strength for potential hydrogen bonds of path length 5) and *maxHaaNH* (maximum atom-type H E-state::NH:) correlate positively to the inhibitory activity, as well as the topological path descriptor *SRW9* (self-returning walk count of order 9). Equation 1 also satisfies the external validation conditions (Golbraikh and Tropsha 2002):

- $1 - R^2_0/R^2_{test} < 0.1$ (0.03) and $1 - R'^2_0/R^2_{test} < 0.1$ (0.00) and,
- $0.85 \le k \le 1.15$ (0.99) and $0.85 \le k' \le 1.15$ (1.00)
- $R^2_m > 0.5$ (0.69)

The addition of molecular fingerprints provides 1568 independent descriptors and fingerprints to the calculations. For this case, the best results are achieved with five descriptors (see Table 2). Figure 3 plots the predicted and experimental values for the training, validation, and test sets using three descriptors and fingerprints. The resultant

Table 2 Descriptors and fingerprints identified for modeling the inhibitory H IV reverse transcriptase activity along with the squared correlation coefficient and the standard deviation for training, validation, and test sets.

d	Descriptor symbols	R^2_{train}	S_{train}	R^2_{val}	S_{val}	R^2_{test}	S_{test}
1	*SubFP200*	0.63	1.19	0.71	0.90	0.66	0.84
2	*MACCSFP54,SubFP200*	0.78	0.92	0.77	0.80	0.75	0.73
3	*ATSC8v, maxHBint4, MACCSFP79*	0.80	0.88	0.78	0.79	0.69	0.81
4	*GATS7m, SpMax2_Bhi, PubchemFP643, SubFP28*	0.86	0.75	0.78	0.76	0.73	0.77
5	*AATS7e, ASP.7, MACCSFP54, PubchemFP533, APC2D4CS*	**0.85**	**0.77**	**0.83**	**0.69**	**0.74**	**0.74**
6	*GATS7m, ASP.7, ETABetaPnsd, MACCSFP79, PubchemFP643, KRFP820*	0.88	0.70	0.84	0.68	0.69	0.86

The best model appears in bold

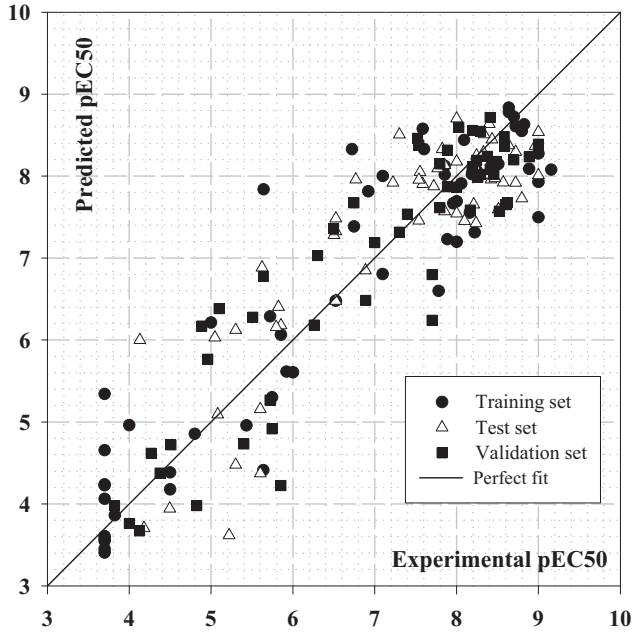

Fig. 3 Predicted and experimental pEC_{50} values for the training, validation and test sets (Eq. 2)

equation for this model is

$$pEC_{50} = -2.09\,AATS7e - 362.10\,ASP.7 + 2.02\,MACCSFP54 + 1.50\,PubchemFP533 + 0.31\,APC2D4CS + 28.89 \quad (2)$$

$N_{train} = 58$, $R^2_{train} = 0.85$, $S_{train} = 0.77$, $F = 61$, $o2.5 = 1$
$N_{val} = 57$, $R^2_{val} = 0.83$, $S_{val} = 0.69$
$N_{test} = 57$, $R^2_{test} = 0.74$, $S_{test} = 0.74$

The descriptors *AATS7e* (average Broto–Moreau autocorrelation - lag 7/weighted by Sanderson electronegativities) and *ASP.7* (average simple path, order 7) are 2D descriptors originated as autocorrelations of the topological structure of Broto–Moreau (ATS), and PaDEL Chi Paths, respectively. Equation 2 also includes three fingerprints, the MACCS key QHAAQH *MACCSFP54*, the Simple SMARTS patterns S-C:C-N *PubchemFP533*, and the Count of C-S at topological distance 4 *APC2D4CS*.

Equation 2 (see Fig. 3) has one outlier compound with a residual higher than 2.5 standard deviations, **138**. Young et al. (1995) have pointed out that this particular compound has a difference of three orders of magnitude in the EC_{50} with compounds of similar structures. They justify this diference with the fact that **138** is not capable of co-planarity with the receptor. We assume that since we are only using 1D and 2D descriptors, the three-dimensional arrangement of the molecule is not fully represented, resulting in a poor prediction for **138**. Equation 2 also satisfies the external validation criteria (Golbraikh and Tropsha 2002):

- $1 - R^2_0/R^2_{test} < 0.1$ (0.01) and $1 - R'^2_0/R^2_{test} < 0.1$ (0.02) and,
- $0.85 \le k \le 1.15$ (1.00) and $0.85 \le k' \le 1.15$ (0.99)
- $R^2_m > 0.5$ (0.73)

QSAR models based on flexible molecular descriptors

In order to decide which structural attributes are the most efficient for each SR during the flexible descriptor design, the DCW flexible descriptor is optimized by increasing R^2_{train}, until the model starts to loose predictive capability in the validation set. The classical scheme is adopted: the test set is not involved during the model development. Table 3 contains a summary for the statistical quality of the best QSAR models found by trying different possible CORAL methods. It reveals that the best choice is a hybrid approach that includes both graph and SMILES representations. We follow the common practice of keeping the model's size as small as possible (Ockham's razor), in order to avoid any fortuitous correlation. Therefore, no more than three attributes are considered in the DCW calculation, because no further improvement can be obtained beyond that value (Toropov et al. 2015). The model with two attributes is chosen due to its good performance and simplicity. Figure 4 shows the predicted and experimental values for the training, validation, and test sets using DCW and the following

Table 3 The stepwise search for the best QSAR model in the hybrid approach as structural representation.

Structural attributes	R^2_{train}	S_{train}	R^2_{val}	S_{val}	R^2_{test}	S_{test}
^3S_k	0.85	0.51	0.85	0.53	0.68	0.75
^2S_k, ^3S_k	0.84	0.53	0.83	0.55	0.71	0.70
^1S_k, ^2S_k, ^3S_k	0.85	0.52	0.84	0.54	0.71	0.70
$Pt2_k$	0.82	0.83	0.82	0.70	0.83	0.59
$Pt2_k$, ^1EC_j	0.83	0.79	0.83	0.70	0.80	0.64
^3S_k, $Pt2_k$	**0.91**	**0.60**	**0.82**	**0.71**	**0.83**	**0.63**
^1S_k, ^3S_k, $Pt2_k$	0.90	0.61	0.82	0.71	0.84	0.61

The best model appears in bold

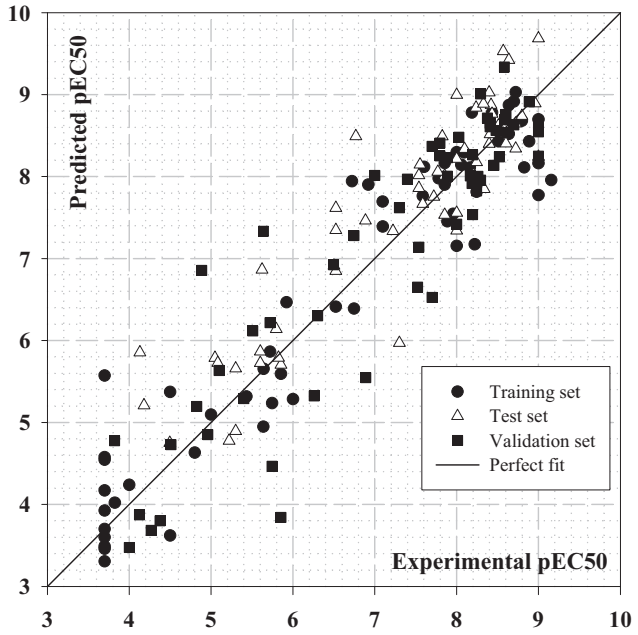

Fig. 4 Predicted and experimental pEC_{50} values for the training, validation and test sets (Eq. 3)

linear model:

$$pEC_{50} = 0.12\,DCW + 0.27 \quad (3)$$

$$N_{train} = 58,\ R^2_{train} = 0.91,\ S_{train} = 0.60,\ F = 537,\ o2.5 = 1$$
$$N_{val} = 57,\ R^2_{val} = 0.82,\ S_{val} = 0.71$$
$$N_{test} = 57,\ R^2_{test} = 0.83,\ S_{test} = 0.73$$

The parameters used during model building are $T = 5$ and $N_{epochs} = 50$. Table 11S includes an example for calculating the DCW for **1**. Figure 4 shows that the predicted vs. experimental inhibitory activity obtained with Eq. 3 follows a straight line. The flexible descriptor of Eq. 3 considers $Pt2_k$ as local graph invariant, and the structural attributes that contribute to the DCW calculation are listed in Table 10S. Furthermore, higher positive CW values tend to predict higher activity values.

Equation 3 (see Fig. 4) presents only one outlier compound in the training set that has a residual higher than 2.5 standard deviations, **27**. Such compound has an experimental EC_{50} value not well-defined (>200 μM) (Ragno et al. 2005). Hence, we assume that this irregular behavior may be attributed to the intrinsic ambiguity in the experimental value reported for this molecule.

Equation 3 also satisfies the external validation conditions (Golbraikh and Tropsha 2002):

- $1 - R^2_0/R^2_{test} < 0.1$ (0.03) and $1 - R'^2_0/R^2_{test} < 0.1$ (0.00) and,
- $0.85 \leq k \leq 1.15$ (0.99) and $0.85 \leq k' \leq 1.15$ (1.00)
- $R^2_m > 0.5$ (0.70)

The QSAR models given by Eqs. 1–3 predict with good accuracy the inhibitory activity of 172 structurally diverse indoyl aryl sulfones and sulfides, and compare favorably to previous published results. For instance, the models presented in this study analyze a greater number of compounds than a previous reported study (Ragno et al. 2005), and only involve simpler 1D and 2D descriptors when compared to the 3D QSAR used in such research. Furthermore, our developed models are properly internally and externally validated.

Finally, calculations combining both PaDEL and CORAL descriptors and fingerprints show no significant improvement in the results. The combinations of various flexible descriptors or flexible descriptors with traditional molecular descriptors do not create models having a better prediction quality, and, moreover, increase significantly the complexity of the models.

Conclusions

In this work, the inhibitory HIV reverse transcriptase activity is modeled with a good accuracy through a computational technique, that does not require the knowledge of the molecular conformation during the structural representation. We succeeded in proposing molecular structure-based models that correlate well with the experimental pEC_{50} values, and demonstrate that such models are predictive in the validation process. The QSAR models established by using traditional molecular descriptors are accurate, while similar predictive performance is obtained with CORAL using only one flexible descriptor. The simplicity of the proposed models makes their applicability easy and new results will be published shortly elsewhere.

Acknowledgements PRD acknowledges the financial support from the National Research Council of Argentina (CONICET) PIP11220130100311 project and to Ministerio de Ciencia, Tecnología e Innovación Productiva for the electronic library facilities.

The authors are members of the scientific researcher career of the National Research Council of Argentina (CONICET).

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

Aranda JF, Garro Martinez JC, Castro EA, Duchowicz PR (2016) Conformation-independent QSPR approach for the soil sorption coefficient of heterogeneous compounds. Int J Mol Sci 17:1247–1255

Artico M, Silvestri R, Stefancich G, Massa S, Pagnozzi E, Musu D, Scintu F, Pinna E, Tinti E, La Colla P (1995) Synthesis of pyrryl aryl sulfones targeted at the HIV-1 reverse transcriptase. Arch Pharm 328:223–229

Artico M, Silvestri R, Pagnozzi E, Bruno B, Novellino B, Greco G, Masaa S, Ettorre A, Loi AG, Scintu F, La Colla P (2000) Structure-based design, synthesis, and biological evaluation of novel pyrrolyl aryl sulfones: HIV-1 non-nucleoside reverse transcriptase inhibitors active at nanomolar concentrations. J Med Chem 43:1886–1891

Ashok P, Chander S, Balzarini J, Pannecouque C, Murugesan S (2015a) Design, synthesis of new β-carboline derivatives and their selective anti-HIV-2 activity. Bioorg Med Chem Lett 25:1232–1235

Ashok P, Sharma H, Lathiya H, Chander S, Murugesan S (2015b) In-silico design and study of novel piperazinyl β-carbolines as inhibitor of HIV-1 reverse transcriptase. Med Chem Res 24:513–522

Bonini C, Chiummiento L, Di Blasio N, Funicello M, Lupattelli P, Tramutola F, Berti F, Ostric A, Miertus S, Frecer V, Kong DX (2014) Synthesis and biological evaluation of new simple indolic non peptidic HIV Protease inhibitors: the effect of different substitution patterns. Bioorg Med Chem 22:4792–4802

Broder SA, Fauci AS (1988) Progress in drug therapies for HIV infection. Public Health Rep 103:224

Cancio R, Silvestri R, Ragno R, Artico M, De Martino G, La Regina G, Crespan E, Zanoli S, Hübscher U, Spadari S, Maga G (2005) High potency of indolyl aryl sulfone nonnucleoside inhibitors towards drug-resistant human immunodeficiency virus type 1 reverse transcriptase mutants is due to selective targeting of different mechanistic forms of the enzyme. Antimicrob Agents Chemother 49:4546–4554

Cesarini S, Spallarossa A, Ranise A, Schenone S, La Colla P, Collu G, Sanna G, Loddo R (2010) (Hetero) aroyl esters of 2-(N-phthalimido) ethanol and analogues: parallel synthesis, anti-HIV-1 activity and cytotoxicity. Med Chem Res 19:311–336

ChemSketch. http://www.acdlabs.com. Accessed 1 Sep 2015

Consonni V, Ballabio D, Todeschini R (2009) Comments on the definition of the Q 2 parameter for QSAR validation. J Chem Inf Model 49:1669–1678

CORALSEA. http://www.insilico.eu/coral. Accessed 2 May 2016

De Clercq E (2009) Highlights in the discovery of antiviral drugs: a personal retrospective. J Med Chem 53:1438–1450

Defant A, Mancini I, Tomazzolli R, Balzarini J (2015) Design, synthesis, and biological evaluation of novel 2H-pyran-2-one derivatives as potential HIV-1 reverse transcriptase inhibitors. Arch Pharm 348:23–33

De Martino G, La Regina G, Ragno R, Coluccia A, Bergamini A, Ciaprini C, Sinistro A, Maga G, Crespan E, Artico M, Silvestri R (2006) Indolyl aryl sulphones as HIV-1 non-nucleoside reverse transcriptase inhibitors: synthesis, biological evaluation and binding mode studies of new derivatives at indole-2-carboxamide. Antivir Chem Chemother 17:59–77

Di Santo R, Costi R (2005) 2H-pyrrolo [3, 4-b][1, 5] benzothiazepine derivatives as potential inhibitors of HIV-1 reverse transcriptase. Il Farmaco 60:385–392

Duchowicz PR, Castro EA, Fernández FM (2006) Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. MATCH Commun Math Comput Chem 55:179–192

Duchowicz PR, Goodarzi M, Ocsachoque MA, Romanelli GP, Ortiz ED, Autino JC, Bennardi DO, Ruiz DM, Castro EA (2009) QSAR analysis on Spodoptera litura antifeedant activities for flavone derivatives. Sci Total Environ 408:277–285

Golbraikh A, Tropsha A (2002) A beware of q2! J Mol Graph Model 20:269–276

Guendel I, Iordanskiy S, Van Duyne R, Kehn-Hall K, Saifuddin M, Das R, Jaworski E, Sampey GC, Senina S, Shultz L, Narayanan A (2014) Novel neuroprotective GSK-3β inhibitor restricts Tat-mediated HIV-1 replication. J Virol 88:1189–1208

Hunter R, Younis Y, Muhanji CI, Curtin TL, Naidoo KJ, Petersen M, Bailey CM, Basavapathruni A, Anderson KS (2008) C-2-Aryl O-substituted HI-236 derivatives as non-nucleoside HIV-1 reverse-transcriptase inhibitors. Bioorg Med Chem 16:10270–10280

La Regina G, Coluccia A, Piscitelli F, Bergamini A, Sinistro A, Cavazza A, Maga G, Samuele A, Zanoli S, Novellino E, Artico MG, Silvestri R (2007) Indolyl aryl sulfones as HIV-1 non-nucleoside reverse transcriptase inhibitors: role of two halogen atoms at the indole ring in developing new analogues with improved antiviral activity. J Med Chem 50:5034–5038

La Regina G, Coluccia A, Brancale A, Piscitelli F, Gatti V, Maga G, Samuele A, Pannecouque C, Schols D, Balzarini J, Novellino E, Silvestri R (2011) Indolylarylsulfones as HIV-1 non-nucleoside reverse transcriptase inhibitors: new cyclic substituents at indole-2-carboxamide. J Med Chem 54:1587–1598

Matlab 7.0. http://www.mathworks.com. Accessed 29 July 2016

McMahon JB, Gulakowski RJ, Weialow OS, Shultz RJ, Narayanan VL, Clanton DJ, Pedemonte R, Wassmundt FW, Buckheit Jr RW, Decker WD, White EL, Bader JP, Boyd MR (1993) Diarylsulphones, a new chemical class of nonnucleoside antiviral inhibitors of immunodeficiency virus type 1 reverse transcriptase. Antimicrob Agents Chemother 37:754–760

Musah RA (2004) The HIV-1 nucleocapsid zinc finger protein as a target of antireviral therapy. Curr Top Med Chem 4:1605–1622

Nguyen Van Nhien A, Tomassi C, Len C, Marco-Contelles JL, Balzarini J, Pannecouque C, De Clercq E, Postel DA (2005) First synthesis and evaluation of the inhibitory effects of aza analogues of TSAO on HIV-1 replication. J Med Chem 48:4276–4284

Oversteegen L, Shah M, Rovini H (2007) HIV combination products. Nat Rev Drug Discov 6:951–952

PaDEL. http://www.yapcwsoft.com. Accessed 2 May 2016

Pan T, He X, Chen B, Chen H, Geng G, Luo H, Zhang H, Bai C (2015) Development of benzimidazole derivatives to inhibit HIV-1 replication through protecting APOBEC3G protein. Eur J Med Chem 95:500–513

Pannecouque C, Szafarowicz B, Volkova N, Bakulev V, Dehaen W, Mély Y, Daelemans D (2010) Inhibition of HIV-1 replication by a bis-thiadiazolbenzene-1, 2-diamine that chelates zinc ions from retroviral nucleocapsid zinc fingers. Antimicrob Agents Chemother 54:1461–1468

Piscitelli F, Coluccia A, Brancale A, La Regina G, Sansone A, Giordano C, Balzarini J, Maga G, Zanoli S, Samuele A, Cirilli R, La Torre F, Lavecchia A, Novellino E, Silvestri R (2009) Indolylarylsulfones bearing natural and unnatural amino acids discovery of potent inhibitors of HIV-1 non-nucleoside wild type

and resistant mutant strains reverse transcriptase and coxsackie B4 virus. J Med Chem 52:1922–1934

Prajapati HR, Doshi AV (2011) Mixed mesomorphism-I: determination of Latent Transition Temperatures (LTTs). Der Pharma Chem 3:123–133

Ragno R, Artico R, De Martino G, La Regina G, Coluccia A, Di Pasquali A, Silvestri R (2005) Docking and 3-D QSAR studies on indolyl aryl sulfones Binding mode exploration at the HIV-1 reverse transcriptase non-nucleoside binding site and design of highly active N-(2-hydroxyethyl) carboxamide and N-(2-hydroxyethyl) carbohydrazide derivatives. J Med Chem 48:213–223

Ragno R, Coluccia A, La Regina G, De Martino G, Piscitelli F, Lavecchia A, Novellino E, Bergamini A, Ciaprini C, Sinistro A, Maga G, Crespan E, Artico MG, Silvestri R (2006) Design, molecular modeling, synthesis and anti-HIV-1 activity of new indolyl aryl sulfones Novel derivatives of the indole-2-carboxamide. J Med Chem 49:3172–3184

Ribone SR, Leen V, Madrid M, Dehaen W, Daelemans D, Pannecouque C, Briñón MC (2012) Synthesis, biological evaluation and molecular modeling of 4, 6-diarylpyrimidines and diarylbenzenes as novel non-nucleosides HIV-1 reverse transcriptase inhibitors. Eur J Med Chem 58:485–492

Rojas C, Duchowicz PR, Tripaldi P, Pis Diez R (2015a) QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. Chemom Intel Lab Syst 140:126–132

Rojas C, Duchowicz PR, Tripaldi P, Pis Diez R (2015b) Quantitative structure-property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase. J Chromatogr A 1422:277–288

Roy K, Leonard JT (2004) QSAR modeling of HIV-1 reverse transcriptase inhibitor 2-amino-6-arylsulfonylbenzonitriles and congeners using molecular connectivity and E-state parameters. Bioorg Med Chem 12:745–754

Samuele A, Bisi S, Kataropoulou A, La Regina G, Piscitelli F, Gatti V, Silvestri R, Maga G (2011) Mechanism of interaction of novel indolylarylsulfone derivatives with K103N and Y181I mutant HIV-1 reverse transcriptase in complex with its substrates. Antivir Chem Chemother 22:107–118

Samuele A, Kataropoulou A, Viola M, Zanoli S, La Regina G, Piscitelli F, Silvestri R, Maga G (2009) Non-nucleoside HIV-1 reverse transcriptase inhibitors di-halo-indolyl aryl sulfones achieve tight binding to drug-resistant mutants by targeting the enzyme–substrate complex. Antivir Res 81:47–55

Silvestri R, De Martino G, La Regina G, Artico M, Massa S, Vargiu L, Mura M, Loi AG, Marceddu T, La Colla P (2003) Novel indolyl aryl sulfones active against HIV-1 carrying NNRTI resistance mutations: synthesis and SAR studies. J Med Chem 46:2482–2493

Silvestri R, Artico M, De Martino G, La Regina G, Loddo R, La Colla M, La Colla P (2004) Simple, short peptide derivatives of a sulfonylindolecarboxamide (L-737,126) active in vitro against HIV-1 wild type and variants carrying non-nucleoside reverse transcriptase inhibitor resistance mutations. J Med Chem 47:3892–3896

Singh K, Marchand B, Rai DK, Sharma B, Michailidis E, Ryan EM, Matzek KB, Leslie MD, Hagedorn AN, Li Z, Norden PR (2012) Biochemical mechanism of HIV-1 resistance to rilpivirine. J Biol Chem 287:38110–38123

Tantillo C, Ding J, Jacobo-Molina A, Nanni RG, Boyer PL, Hughes SH, Pauwels R, Andries K, Janssen PA, Arnold EC (1994) Locations of anti-AIDS drug binding sites and resistance mutations in the three-dimensional structure of HIV-1 reverse transcriptase Implications for mechanisms of drug inhibition and resistance. J Mol Biol 243:369–387

The Open Babel package, version 231. http://openbabel.org. Accessed 1 Sep 2015

Toropov AA, Toropova AP, Benfenati E, Nicolotti O, Carotti A, Nesmerak K, Veselinović AM, Veselinović JB, Duchowicz PR, Bacelo DE, Castro EA, Rasulev BF, Leszczynska D, Leszczynski J (2015) Quantitative structure-activity relationships in drug design, predictive toxicology, and risk assessment. In: Roy K (ed) QSPR/QSAR analyses by means of the CORAL software: results, challenges, perspectives. IGI Global, Hershey, pp 560–585

UNAIDS gap report (2015) United Nations Programme on HIV/AIDS (UNAIDS). The GAP report 2014, UNAIDS, Geneva

Weaver S, Gleeson MP (2008) The importance of the domain of applicability in QSAR modeling. J Mol Graph Model 26: 1315–1326

WHO (9789240692671_eng) (2014) World Health Organization world health statistics. http://appswhoint/iris/bitstream/10665/112738/1/9789240692671_engpdf?ua=1. Accessed 15 Sept 2015

Williams TM, Ciccarone TM, Saari WS, Wai JS, Greenlee WJ, Balani SK, Goldman ME, Hoffman JM Jr, Lumma Jr WC, Huff J, Rooney C, Sanderson PE, Theoharides AD (1994) Indole derivatives as inhibitors of HIV reverse transcriptase. Patent PCT Int Appl WO 9419321, USA

Williams IG (2003) Enfuvirtide (Fuzeon): the first fusion inhibitor. Int J Clin Pract 57:890–897

Williams TM, Ciccarone TM, MacTough SC, Rooney CS, Balani SK, Condra JH, Emini EA, Goldman ME, Greenlee WJ, Kauffman LR, O'Brien JA, Sardana VV, Schleif WA, Theoharides AD, Anderson PS (1993) 5-Chloro-3-(phenylsulfonyl) indole-2-carboxamide: a novel, non-nucleoside inhibitor of HIV-1 reverse transcriptase. J Med Chem 36:1291–1294

Witvrouw M, Fikkert V, Pluymers W, Matthews B, Mardel K, Schols D, Raff J, Debyser Z, De Clercq E, Holan G, Pannecouque C (2000) Polyanionic (ie polysulfonate) dendrimers can inhibit the replication of human immunodeficiency virus by interfering with both virus adsorption and later steps (reverse transcriptase/integrase) in the virus replicative cycle. Mol Pharmacol 58:1100–1108

Yap CW (2011) PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474

Young SD, Amblard MC, Britcher SF, Grey VE, Tran LO, Lumma WC, Huff JR, Schleif WA, Emini EE, O'Brien JA, Pettibone DJ (1995) 2-Heterocyclic indole-3-sulfones as inhibitors of HIV-1 reverse transcriptase. Bioorg Med Chem Lett 5:491–496

Zhan P, Chen X, Li D, Fang Z, Clercq E, Liu X (2013) HIV-1 NNRTIs: structural diversity, pharmacophore similarity, and implications for drug design. Med Res Rev 33:E1–E72

Zhao Z, Wolkenberg SE, Lu M, Munshi V, Moyer G, Feng M, Carella AV, Ecto LT, Gabryelski LJ, Lai MT, Prasad SG (2008) Novel indole-3-sulfonamides as potent HIV non-nucleoside reverse transcriptase inhibitors (NNRTIs). Bioorg Med Chem Lett 18:554–559