SUPPLEMENTARY MATTERIAL

A 4D structure-activity study of HIV-1 integrase strand transfer inhibitors through a LQTA-QSAR approach

Eduardo B. de Melo[a], Márcia M. C. Ferreira[b]*.

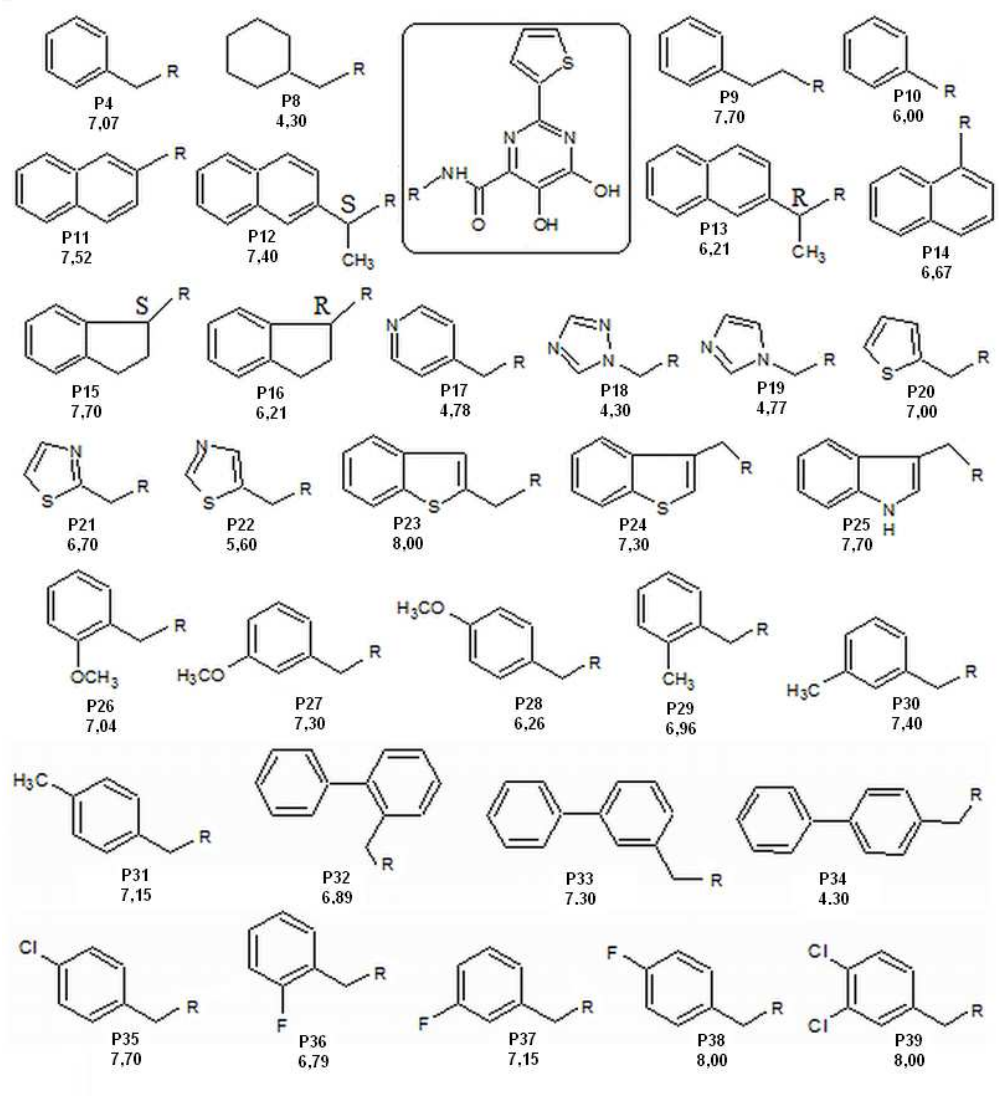
[a]Theoretical Medicinal and Environmental Chemistry Laboratory (LQMAT), Department of Pharmacy, Western Parana State University – Unioeste, 2069 Universitaria St, Cascavel, PR, 85819-110, Brazil.

[b]Laboratory for Theoretical and Applied Chemometrics (LQTA), Institute of Chemistry, University of Campinas – UNICAMP, P.O. Box 6154, Campinas, SP, 13084-971, Brazil.

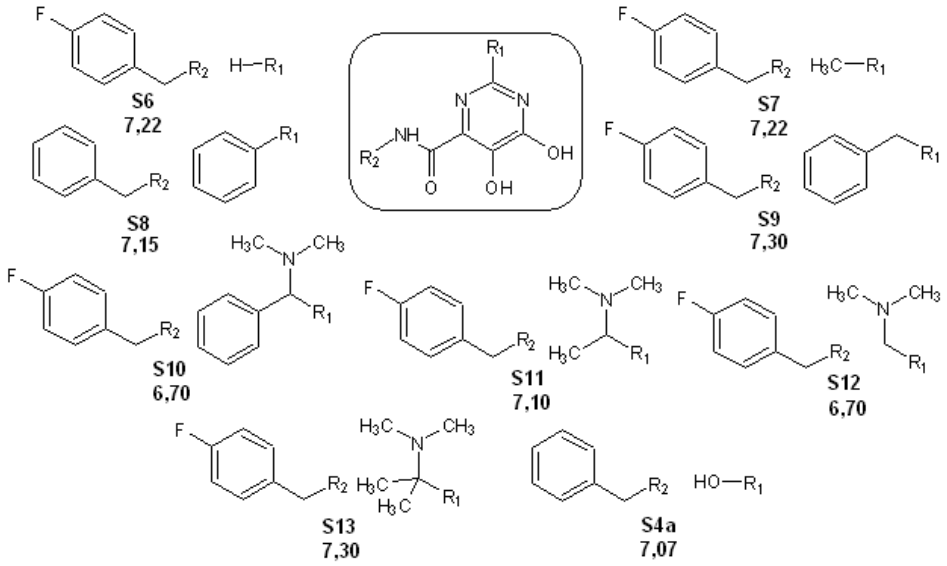
*Corresponding author. Email: marcia@iqm.unicamp.br

INFORMATION S1: STRUCTURE OF THE STUDIED COMPOUNDS.

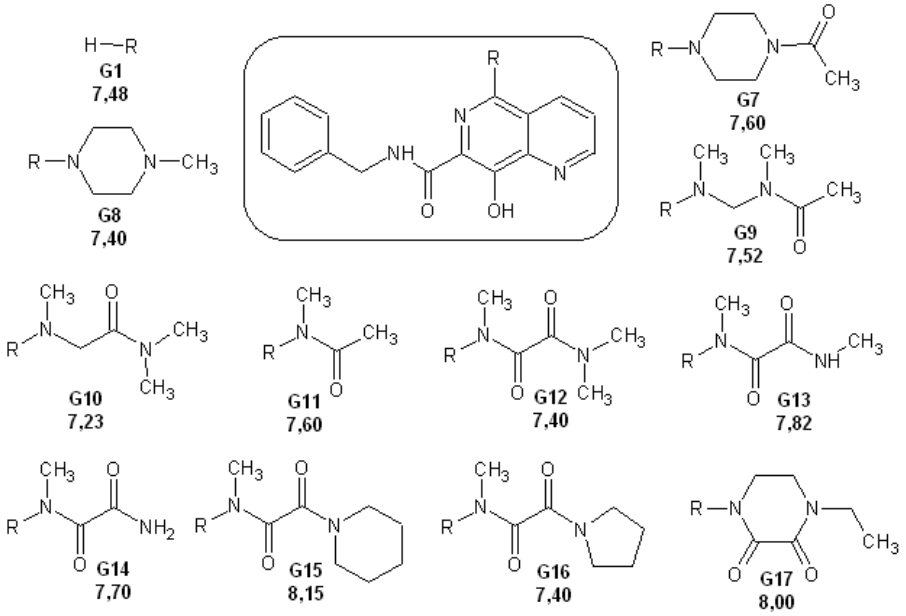
Note: the numbers are the pIC_{50} ($-\log IC_{50}$).



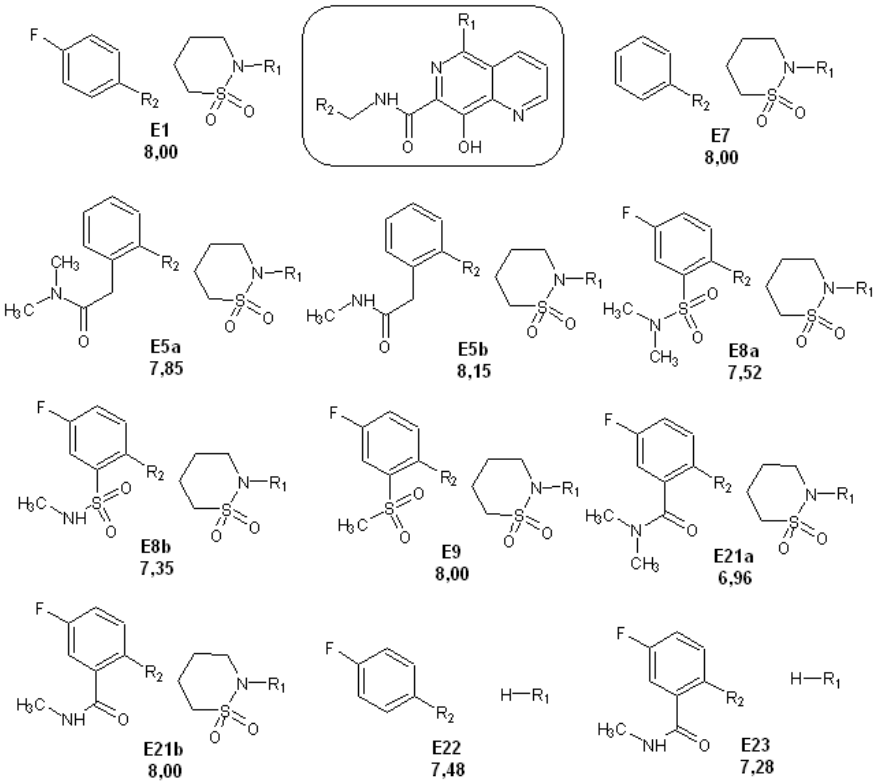
From ref. 12



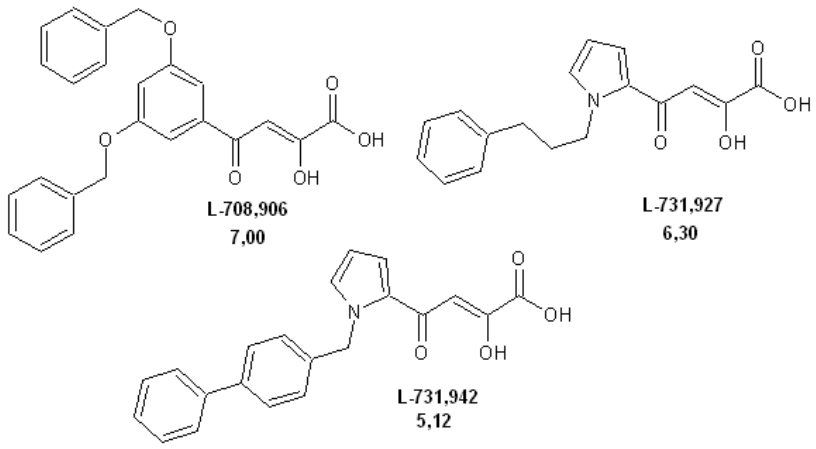
From ref. 18



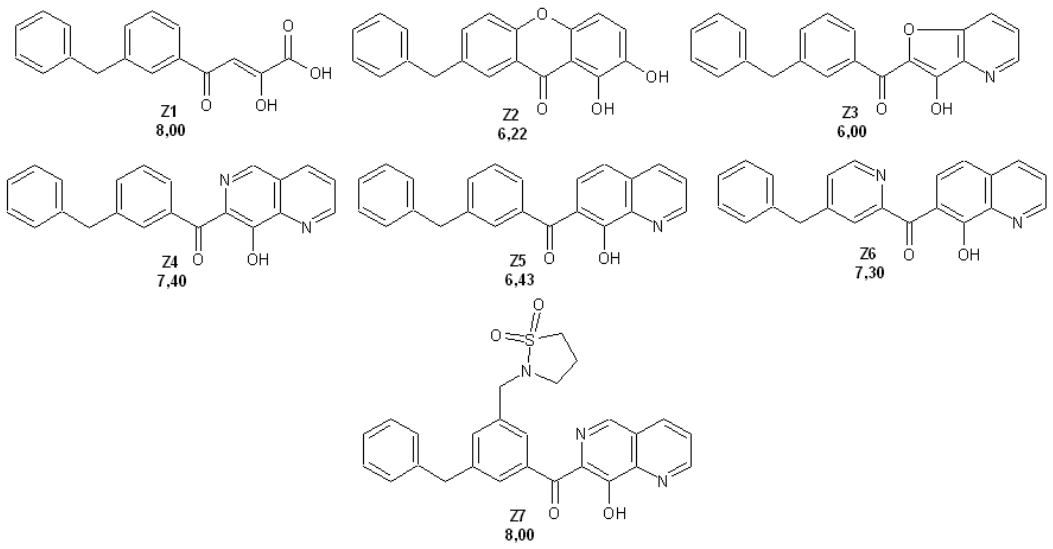
From ref. 19



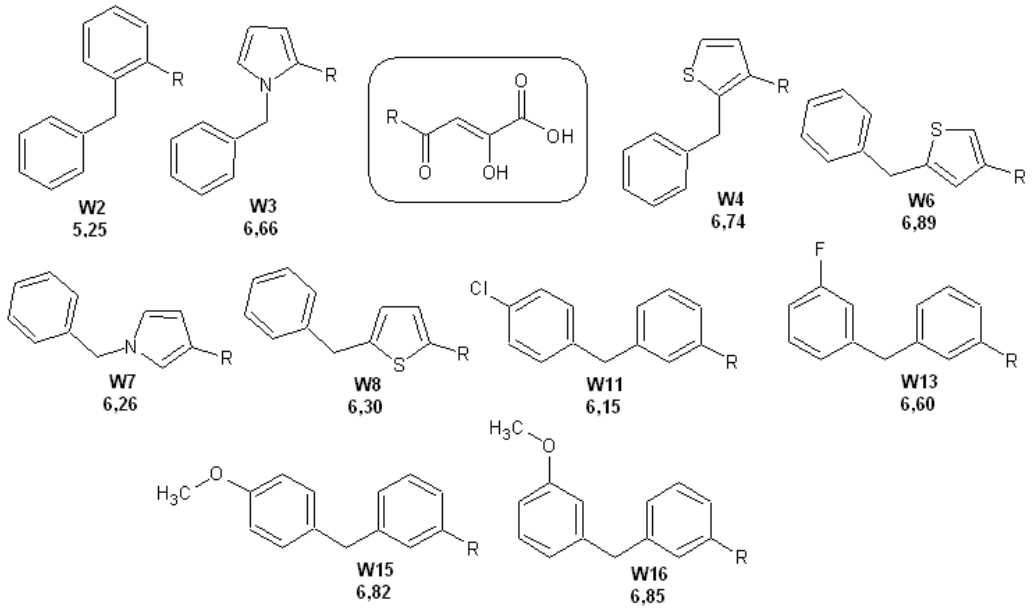
From ref. 20



From ref. 21

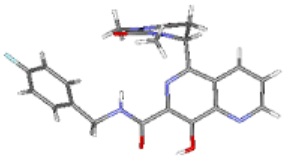


From ref. 22

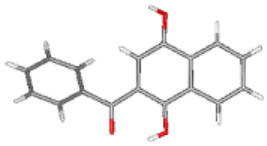


From ref. 23

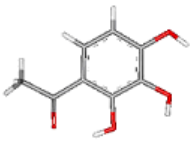
INFORMATION S2: CRYSTALLOGRAPHIC STRUCTURES SELECTED FROM THE CAMBRIDGE STRUCTURAL DATABASE (CSD) THAT WERE USED AS BASIS FOR THE CONSTRUCTION OF THE TRAINING SET. THE NAMES OF EACH STRUCTURE CORRESPOND TO YOUR SPECIFIC CODE IN CSD.



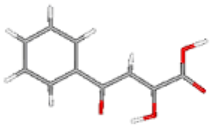
GEDKUW
R-factor: 6.79%
Average sigma: 0.001-0.005 Å



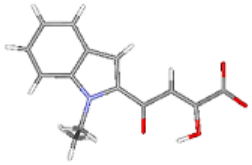
LUCFAQ
R-factor: 7.35%
Average sigma: 0.001-0.005 Å



DOTRUZ
R-factor: 4.40%
Average sigma: 0.001-0.005 Å



XEPJIL
R-factor: 4.50%
Average sigma: 0.001-0.005 Å



RALTUU
R-factor: 3.98%
Average sigma: 0.001-0.005 Å

INFORMATION S3: VALIDATION TOOLS

Table S1. Statistics parameters, respective equations, and adopted limits for the assessment of the internal quality of the model.

Parameter	Symbol	Equation	Limits
Coefficient of multiple determination of calibration[a]	R^2	$1-\frac{\sum_i (y_i-\hat{y}_{ci})^2}{\sum_i (y_i-\bar{y})^2}$	> 0.6
Standard deviation of calibration model[a]	SEC	$\sqrt{\frac{\sum_i (y_i-\hat{y}_{ci})^2}{n-p-1}}$	As much lower as possible
F-test (with 95% confidence interval)[a]	$F_{(p,n-p-1)}$	$\frac{\sqrt{\frac{\sum_i (y_i-\hat{y}_{ci})^2}{k}}}{\sqrt{\frac{\sum_i (y_i-\bar{y})^2}{n-p-1}}}$	Higher than the tabulated critical value
Coefficient of determination of leave-one-out cross validation[b]	Q^2_{LOO}	$1-\frac{\sum_i (y_i-\hat{y}_{vi})^2}{\sum_i (y_i-\bar{y})^2}$	> 0.5
Standard error of cross validation[b]	SEV	$\sqrt{\frac{\sum_i (y_i-\hat{y}_{vi})^2}{n}}$	As much lower as possible
Predictive Residual Sum of Squares of Validation[b]	$PRESS_{val}$	$\sum_i (y_i-\hat{y}_{vi})^2$	Higher than Ssy

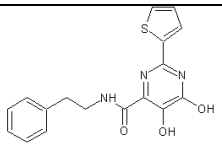
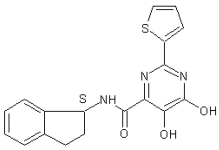
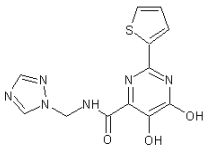
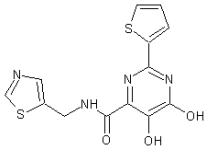
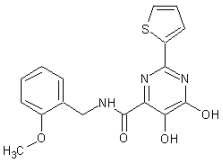
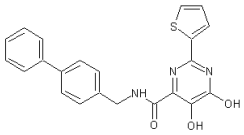
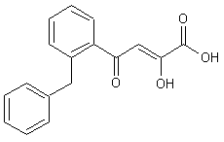
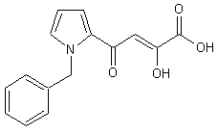
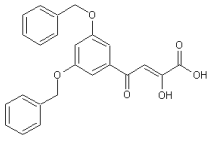
[a]Data fit; [b]cross validation; y_i: observed pIC_{50}; $\bar{y}$: average observed pIC_{50} for the training set; $\hat{y}_{ci}$: estimated pIC_{50} in the calibration model; $\hat{y}_{vi}$: estimated pIC_{50} in the cross-validation; n: number of samples in the training set; p: number of latent variables in the model.

Table S2. Statistics parameters, corresponding equations, and adopted limits for the evaluation of the external quality of the model.

Parameter	Symbol	Equation	Limits
Coefficient of multiple determination of prediction[a]	R^2_{pred}	$1-\frac{\sum_i (y_i - \hat{y}_{ei})^2}{\sum_i (y_i - \overline{y}_{ev})^2}$	□ 0.5
Standard error of prediction	SEP	$\sqrt{\frac{\sum_i (y_i - \hat{y}_{ei})^2}{n_{ev}}}$	As much lower as possible
Average relative error of prediction	ARE_{pred}	$\frac{\frac{\sum_i \lvert y_i - \hat{y}_{ei}\rvert^2}{y_i}100}{n}$	As much lower as possible
Slopes of the linear regression lines	k and k'	$\frac{\sum_i (y_i - \hat{y}_{ei})}{\sum_i y_{ei}}$ and $\frac{\sum_i (y_i - \hat{y}_{ei})}{\sum_i y_i}$	$0.85 \le x \le 1.15$ (x = k or k')
The absolute value of the difference between the coefficient of determination between y_{obsi} and y_{evi} and the coefficient of determination between y_{evi} and y_{obsi}		$\lvert R_0^2 - R'^2_0 \rvert$	< 0.3

[a]For R^2_{pred}, $\overline{y}_{ev}$ is the average value of observed pIC_{50} for the training set without the test set; y_i: observed pIC_{50}; $\hat{y}_{ei}$: estimated pIC_{50} in the external validation; n: number of samples in the training set; n_{ev}: number of samples in the test set.

INFORMATION S4: OUTLIERS

Outliers	pIC_{50}
 **P9**	7.700
 **P15**	7.700
 **P18**	4.300
 **P22**	5.600
 **P26**	7.040
 **P34**	4.300
 **W2**	5.250
 **W3**	6.660
 **L708906**	7.000

INFORMATION S5: DENDROGRAM (AUTOSCALED DATA; LINKAGE METHOD FLEXIBLE) OF THE COMPLETE TRAINING SET. THE TEST SET IS SHOWN DETACHED (BLACK DOTS).