Table S1. PubMed identifiers of the publications used in this study

Training set	NCBI (PMC) PubMed ID
Relevant publications	1282569; 1375293; 7538590; 7689109; 8523406; 8558522; 8809165; 9240358; 9622549; 9685236; 9719591; 9748354; 10052969; 10579849; 10821714; 26345647; 11384233; 11472208; 11844664; 12642128; 16107158; 16162014; 16220981; 18506577; 23217210; 19058881; 19411176; 20728367; 21193314; 25522204; 28859311
Irrelevant publications	8667357; 11078020; 11170624; 11212098; 11229762; 11371163; 11575933; 14717471; 15183338; 15183348; 15634005; 15916427; 16279773; 16527484; 16782042; 16884295; 17081812; 17488516; 18155520; 18155520; 18313992; 19046616; 19386130; 19626612; 19665597; 20060625; 20449621; 21713384; 23416260; 25462235; 26045359

Table S2. Results of automatic categorizing publications into relevant and irrelevant

Data sets	Based on abstracts			Based on full texts		
	Sens	Spec	BA	Sens	Spec	BA
Data set 1 5-fold CV	0.79	0.87	0.83	0.77	0.87	0.81
Data set 1 (training)/ Data set 2 (test)	0.82	0.84	0.83	0.66	0.92	0.79
Random sampling with replacement (Data set 1 + Data set 2)	0.82 ± 0.12*	0.80 ± 0.11	0.81 ± 0.06	0.79 ± 0.11	0.84 ± 0.09	0.82 ± 0.05

* The Mean ± (Standard Deviation) is given.

Table S3. Occurrence of MeSH terms and in the publications (a) manually divided into relevant and irrelevant and (b) classified to be relevant and irrelevant.

(a)

Data set 1 (manual classification)			
Keyword	A (Frequency in Relevant), %	B (Frequency in irrelevant), %	R (A/B)*
Human immunodeficiency virus	16	-	-
Reverse transcriptase	16	-	-
(Q)SAR analysis	-	3	-

MeSH terms	A (Frequency in Relevant), %	B (Frequency in irrelevant), %	R (A/B)*
HIV Reverse Transcriptase	77	56	1.38
Reverse Transcriptase Inhibitors	74	70	1.05
HIV-1	74	36	2.05
Anti-HIV Agents	51	46	1.1
Humans	54	46	1.17

(b)

PubMed data set (automatic classification relevant/irrelevant)			
Keyword	A (Frequency in Relevant), %	B (Frequency in irrelevant), %	R (A/B)*
HIV-1	8.99	1.33	6.75
HIV	2.5	3.5	0.71
Reverse transcriptase	6	0.75	8
NNRTI	2.6	0.6	4
Anti-HIV activity	1.3	0.007	185
Anti-HIV-1 activity	1.75	0.0035	500
MeSH terms	A (Frequency in Relevant), %	B (Frequency in irrelevant), %	R (A/B)*
Reverse transcriptase inhibitors	50	45	1.1
HIV-1	43.5	51.5	0.81
Anti HIV Agents	41	50	0.69
HIV Reverse Transcriptase	41	21	1.95
Drug Resistance, Viral	11	20	0.55

* Ratio of occurrence of the term in the set of relevant publications to that in the set of irrelevant ones.

Table S4. Results of categorizing the FoTs into classes according to bioassay characteristics.

Category	N_{tr}^{pos}*	N_{tr}^{neg}**	N_{test}^{pos}***	N_{test}^{neg}****	Sens	Spec	BA
Cell-based	149	102	137	111	0.96	0.98	0.97
RT-based	102	149	111	137	0.98	0.96	0.97
I	63	188	64	184	0.71	0.96	0.84
II	86	165	73	175	0.76	0.90	0.83
III	65	186	61	191	0.63	0.96	0.79
IV	24	227	21	231	0.59	0.96	0.78
V	41	210	30	221	0.64	0.96	0.80

* N_{tr}^{pos} is the number of positive samples (fragments belonging to the particular category) in the training set
** N_{tr}^{neg} is the number of negative samples (fragments not belonging to the particular category) in the training set
*** N_{test}^{pos} is the number of positive samples in the test set
**** N_{test}^{neg} is the number of negative samples in the test set

Table S5. Results of categorizing the publications into classes according to bioassay characteristics without preliminary extracting of the FoTs with bioassay description.

Category	Based on abstracts			Based on full texts		
	Sens	Spec	BA	Sens	Spec	BA
Cell-based	0.37	0.48	0.43	0.62	0.41	0.52
RT-based	0.48	0.37	0.43	0.41	0.62	0.52
I	0.99	0.12	0.56	0.94	0.12	0.53
II	0.98	0.05	0.51	0.95	0.07	0.51
III	0.60	0.41	0.51	0.62	0.49	0.56
IV	0.94	0.05	0.50	0.95	0.07	0.51
V	0.99	0.05	0.52	0.94	0.06	0.47