Supplemental data

Multiple Machine Learning Comparisons of HIV Cell-Based and Reverse Transcriptase Datasets

Kimberley M. Zorn[†, ‡], Thomas R. Lane[†, ‡], Daniel P. Russo[†, $], Alex M. Clark[§], Vadim Makarov[#] and Sean Ekins[†, *]

[†]Collaborations Pharmaceuticals, Inc., Main Campus Drive, Lab 3510, Raleigh, NC 27606, USA.

[$]The Rutgers Center for Computational and Integrative Biology, Camden, NJ, 08102, USA.

[§]Molecular Materials Informatics, Inc., 2234 Duvernay St, Montreal, Quebec, Canada, H3J2Y3, Canada.

[#]Bach Institute of Biochemistry, Research Center of Biotechnology of the Russian Academy of Sciences, Leninsky Prospekt 33-2, Moscow, 119071, Russia

[‡] authors contributed equally

*To whom correspondence should be addressed.

E-mail: sean@collaborationspharma.com

Phone: 215-687-1320

Supplemental Tables

Table S1. Correlation of cell-based and RT data from the NIAID ChemDB.

Spearman Correlation							
Filter Applied	**Num. of Molecules**	**Units**	R	**95% CI**	**R^2**	**P value (two-tailed)**	**Significant**
RT ≤ 10μM	1647	μM	0.5013	0.4631-0.5376	N/A	<0.0001	Y
Pearson Correlation							
RT ≤ 10μM	1647	μM	0.1890	0.1420-02352	0.03574	< 0.0001	Y
RT ≤ 10μM	1647	-logM	0.5009	0.4638-0.5361	0.2509	< 0.0001	Y
RT ≤ 1μM	1137	-logM	0.4389	0.3908-0.4847	0.1927	<0.0001	Y
RT ≤ 0.1μM	633	-logM	0.2378	0.1629-0.3100	0.05657	< 0.0001	Y
RT ≤ 0.01μM	118	-logM	0.2469	0.06922-0.4094	0.06095	0.0070	Y

Table S2. Individual dataset diversity. Number of Fingerprint Features: The number of unique fingerprints normalized by the number of ligands. Number of Assemblies: The number of unique assemblies normalized by the number of ligands. The type of assembly used is determined by the fragments to generate parameter. Fingerprint Distances: The minimum, maximum, and average distances between all pairs of fingerprints. Property Distances: The average Euclidean distance between all pairs of properties. Diversity_NumAssemblies is defined as the total number of assemblies divided by the number of molecules. Diversity_NumFPFeatures is defined as the total number of fingerprint features divided by the number of molecules. Fingerprint distance is defined as 1 - similarity for every pair of molecules. Property distance is defined as the Euclidean distance of the specified numerical properties for every pair of molecules.

	HIV WC CHEMBL full test set	**HIV WC CHEMBL 500MW test set**	**HIV WC Literature test set**	**Training Set (WC, nonspecific, 500WM cutoff)**	**HIV RTase CHEMBL full test set**	**HIV Rtase CHEMBL 500MW test set**	**HIV RTase Literature test set**	**Training Set (RTase, nonspecific, 500WM cutoff)**
Diversity: Number of Assemblies	0.258	0.255	0.297	0.295	0.393	0.387	0.359	0.285
Diversity: Fingerprint Features	6.749	7.530	8.158	5.798	12.213	12.340	9.542	7.151

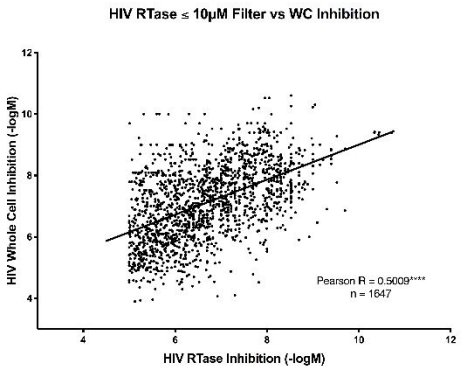
Diversity: Fingerprint Distances								
Average Fingerprint Distance	0.869	0.874	0.775	0.900	0.898	0.895	0.772	0.895
Minimum Fingerprint Distance	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum Fingerprint Distance	0.991	0.990	0.955	1.000	1.000	1.000	0.948	1.000
Diversity: Property Distances								

Average Property Distance	1.304	1.333	1.351	1.317	1.206	1.307	1.353	1.280
Minimum Property Distance	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum Property Distance	3.254	3.981	2.823	6.008	8.675	4.952	2.972	7.351

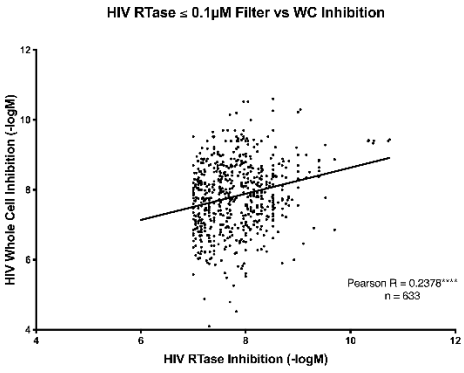
Supplemental Figures

Figure S1. Correlation of whole-cell HIV and reverse transcriptase inhibition using data from the NIAID ChemDB with an RT cut off at A. 10 μM, B. 0.1 μM, C. 0.01μM.

A.



B



C.

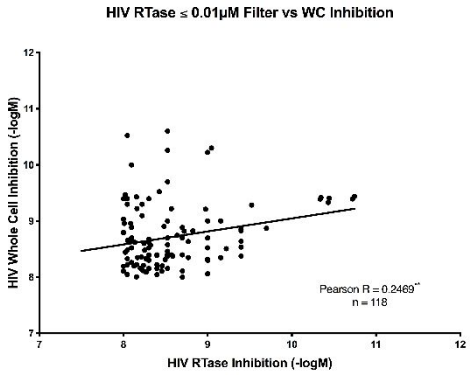
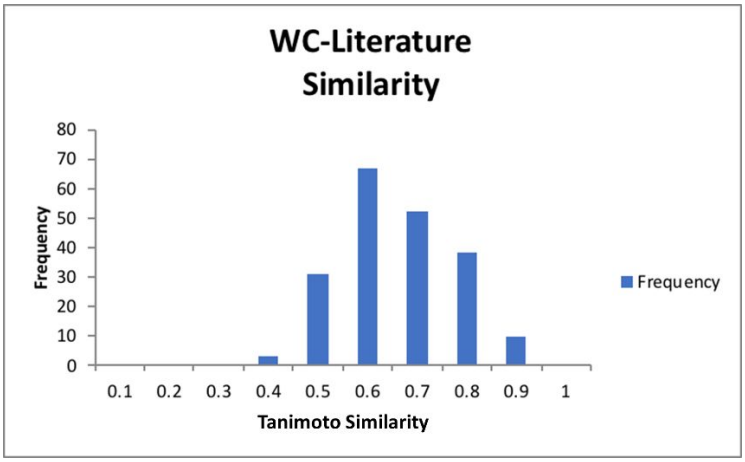
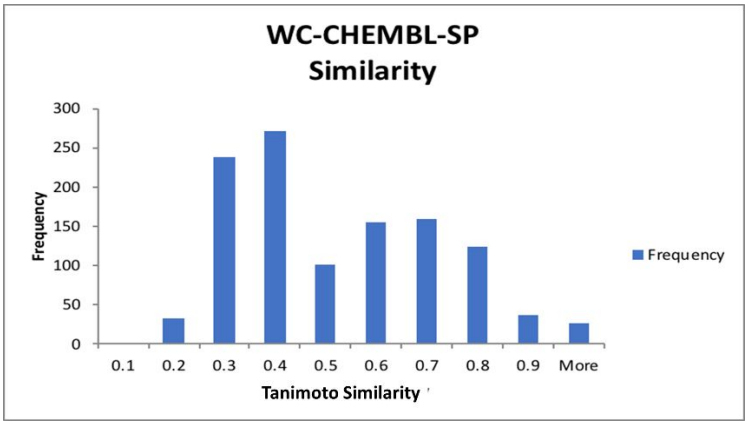
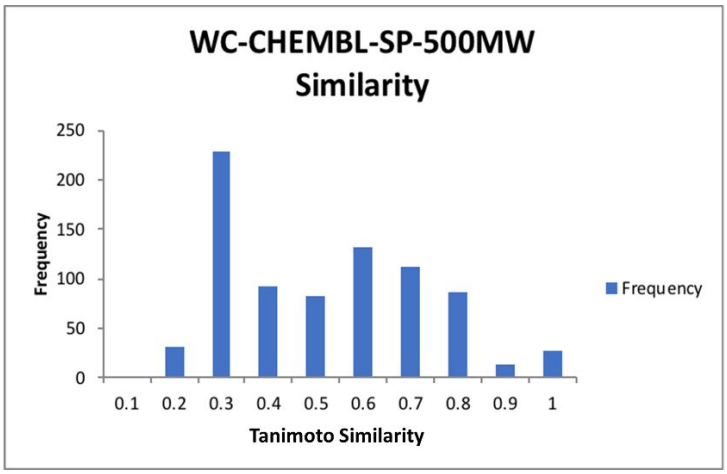
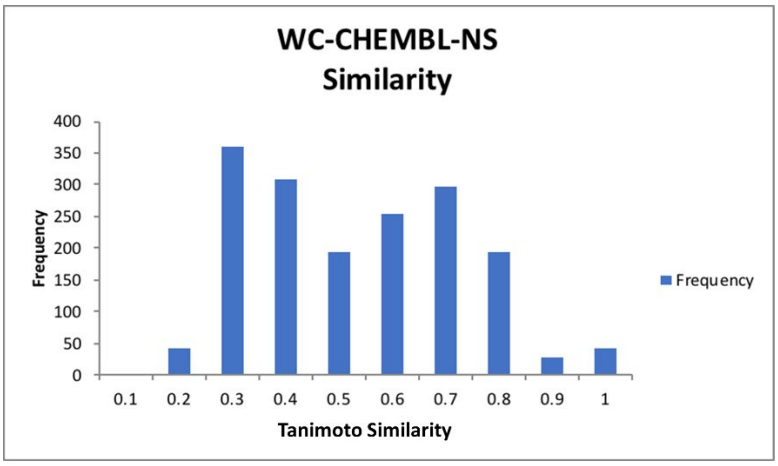


Figure S2. Whole-cell dataset diversity metrics: Training set (WC nonspecific, 500MW cutoff) vs various test sets Tanimoto Similarity (ECFP6, 2D only).








WC-CHEMBL-NS
Similarity
400
350
300
250
200
150
100
50
0
Frequency
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Tanimoto Similarity
Frequency

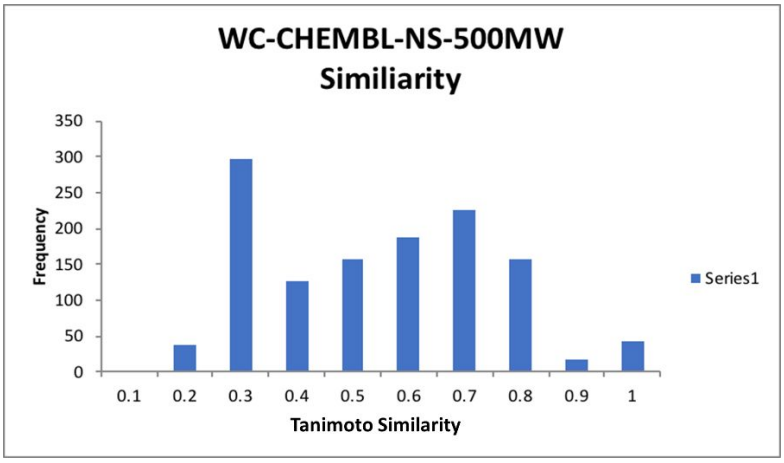
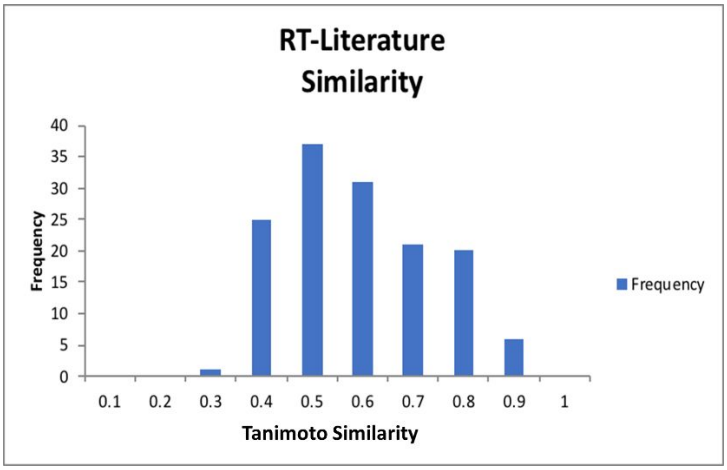
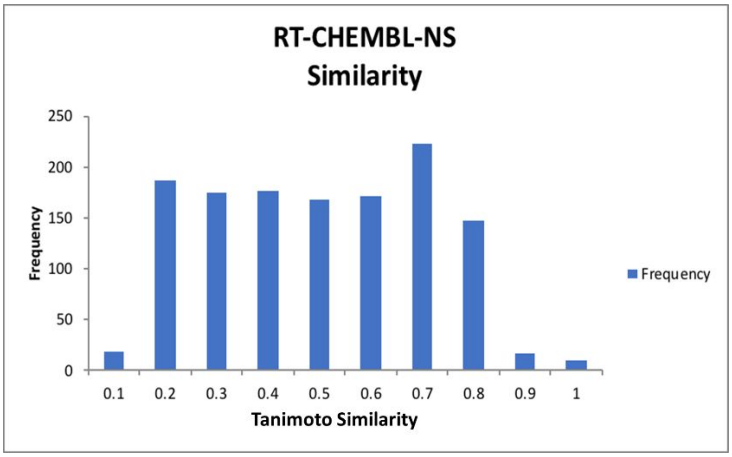

WC-CHEMBL-NS-500MW
Similiarity
350
300
250
200
150
100
50
0
Frequency
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Tanimoto Similarity
Series1

Figure S3. Diversity metrics: Training set (RTase nonspecific, 500MW cutoff) vs various test sets. Tanimoto Similarity (ECFP6, 2D only).





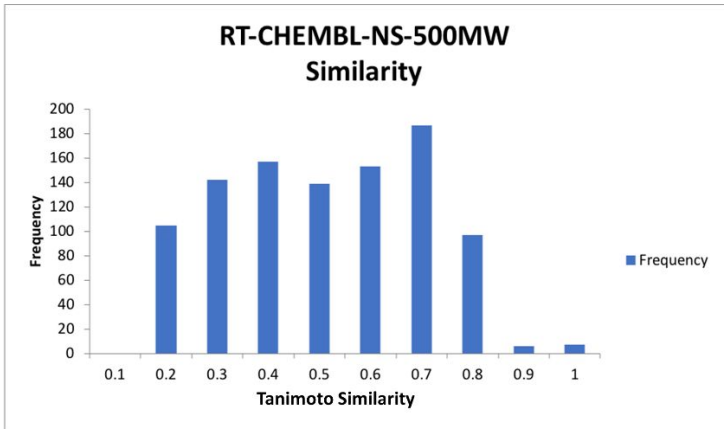
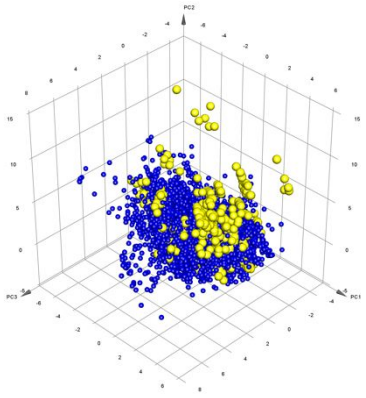


Figure S4. PCA analysis using AlogP, Molecular weight, Number of Hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, number of rings, number of aromatic rings, molecular fractional polar surface area. A. Whole-cell vs ChEMBL non-specific test set in which 3 principal components explains 78.2% of the variance, B. Whole-cell vs literature test set in which 3 principal components explains 77.2% of the variance. (yellow = test set molecules).

A.



B.

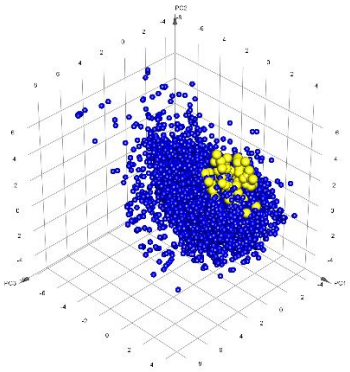
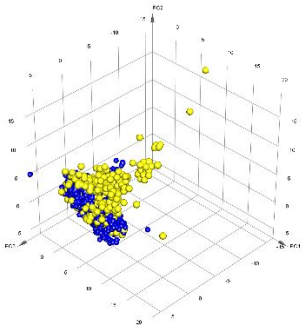
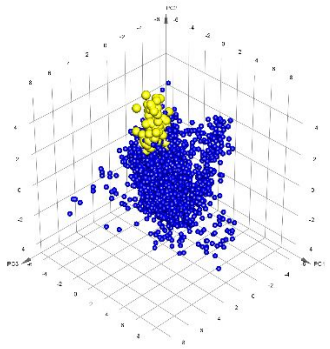


Figure S5. PCA analysis using AlogP, Molecular weight, Number of Hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, number of rings, number of aromatic rings, molecular fractional polar surface area. A. RT vs ChEMBL test set in which 3 principal components explains 81.7% of the variance, B. RT vs literature test set in which 3 principal components explains 81.9% of the variance. (yellow = test set molecules)



A.



B.