Supporting Information for

Using the Variable Nearest Neighbor Method to Identify P-Glycoprotein Substrates and Inhibitors

Patric Schyman[1]*, *Ruifeng Liu*[1], *and Anders Wallqvist*[1]*

[1]DoD Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, 2405 Whittier Drive, Frederick, MD 21702, USA

Table of Contents

Table S1. Optimizing Tanimoto-Distance Threshold Value (*d0*) Using 10-Fold Cross Validation for High Accuracy (HA) Parameters. Training Set for Pgp Substrates from Li et al. (*Mol. Pharmaceutics* 2014, 11, 716-726).

	Smoothing Factor	Tanimoto Distance	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	kappa	Coverage
	0.6	1.00	293	105	204	20	0.64	0.94	0.34	0.28	1.00
	0.6	0.90	279	118	191	34	0.64	0.89	0.38	0.27	1.00
	0.6	0.80	219	208	88	90	0.71	0.71	0.70	0.41	0.97
	0.6	0.70	183	163	61	60	0.74	0.75	0.73	0.48	0.75
	0.6	0.65	169	139	48	52	0.75	0.76	0.74	0.51	0.66
HA	0.6	0.60	163	124	42	45	0.77	0.78	0.75	0.53	0.60
	0.6	0.55	154	105	40	42	0.76	0.79	0.72	0.51	0.55
	0.6	0.50	144	90	36	38	0.76	0.79	0.71	0.50	0.50
	0.6	0.40	117	61	26	27	0.77	0.81	0.70	0.51	0.37
	0.6	0.30	99	27	13	13	0.83	0.88	0.68	0.56	0.24
	0.6	0.20	47	6	2	1	0.95	0.98	0.75	0.77	0.09

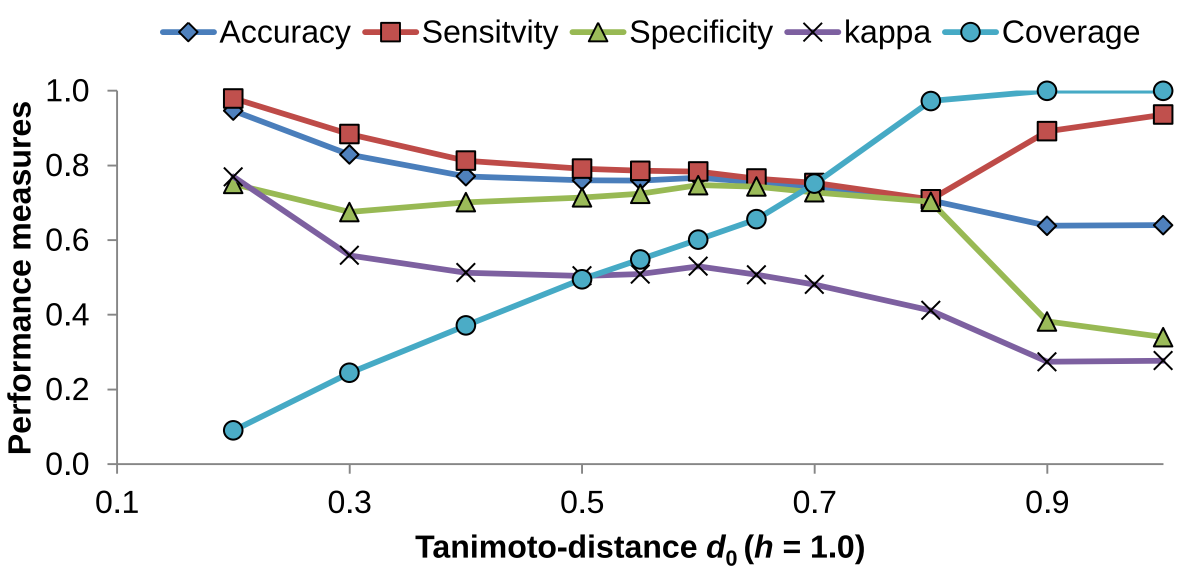

Figure S1. Performance measures of the v-NN Pgp substrate model as a function of Tanimoto-distance threshold *d0* at a constant smoothing factor *h* of 0.6.

Table S2. Optimizing the Smoothing Factor using 10-Fold Cross Validation for High Coverage (HC) Parameters. Training Set for Pgp Substrates from Li et al. (*Mol. Pharmaceutics* 2014, 11, 716-726)

Smoothing Factor		Tanimoto Distance	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	kappa	Coverage
	1	1	311	29	280	2	0.55	0.99	0.09	0.09	1.00
	0.9	1	309	46	263	4	0.57	0.99	0.15	0.14	1.00
	0.8	1	305	64	245	8	0.59	0.97	0.21	0.18	1.00
	0.7	1	301	84	225	12	0.62	0.96	0.27	0.23	1.00
	0.6	1	293	105	204	20	0.64	0.94	0.34	0.28	1.00
	0.5	1	281	132	177	32	0.66	0.90	0.43	0.33	1.00
	0.4	1	267	165	144	46	0.69	0.85	0.53	0.39	1.00
HC	0.3	1	243	200	109	70	0.71	0.78	0.65	0.42	1.00
	0.2	1	223	220	89	90	0.71	0.71	0.71	0.42	1.00
	0.1	1	206	222	87	107	0.69	0.66	0.72	0.38	1.00

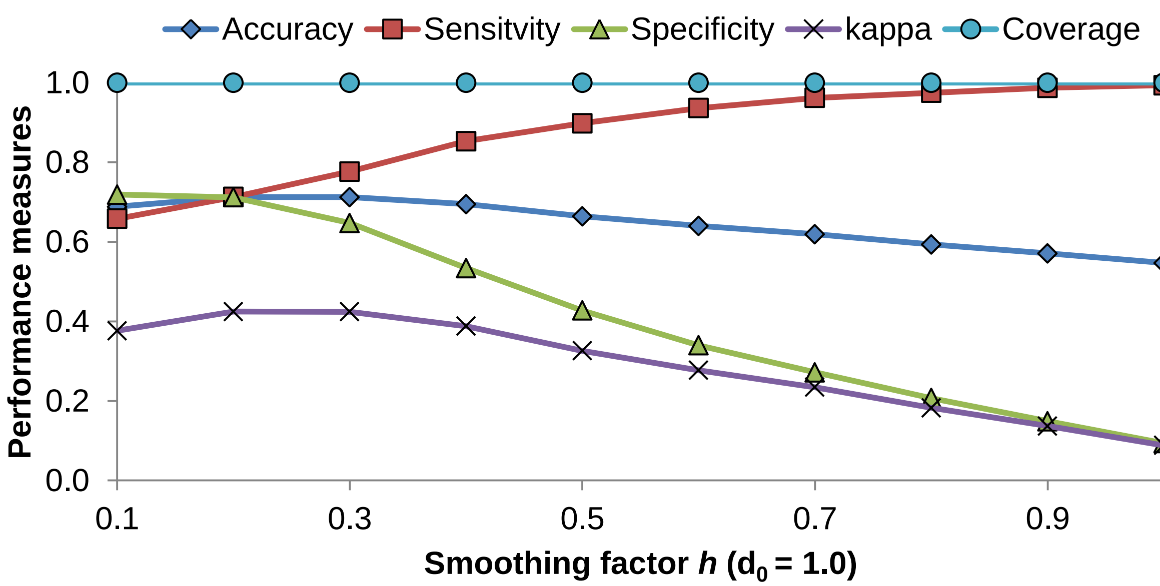

Figure S2. Performance measures of the v-NN Pgp substrate model as a function of smoothing factor *h* at constant Tanimoto-distance threshold *d0* of 1.0.

Table S3. Optimizing the *d0* Value using 10-Fold Cross Validation for High Coverage (HC) and High Accuracy (HA) Parameters. Training Set for Pgp inhibitors from Broccatelli et al. (*J. Med Chem*. 2011, 54, 1740)

Smoothing Factor		Tanimoto Distance	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	kappa	Coverage
HC	0.2	1	286	370	68	48	0.85	0.86	0.84	0.70	1.00
	0.2	0.9	286	369	68	48	0.85	0.86	0.84	0.70	1.00
	0.2	0.8	287	356	62	47	0.86	0.86	0.85	0.71	0.97
	0.2	0.7	272	270	43	36	0.87	0.88	0.86	0.75	0.80
HA	0.2	0.6	254	215	28	20	0.91	0.93	0.88	0.81	0.67
	0.2	0.5	234	164	22	17	0.91	0.93	0.88	0.82	0.57
	0.2	0.4	218	100	19	15	0.90	0.94	0.84	0.78	0.46
	0.2	0.3	195	48	11	9	0.92	0.96	0.81	0.78	0.34
	0.2	0.2	136	14	3	1	0.97	0.99	0.82	0.86	0.20

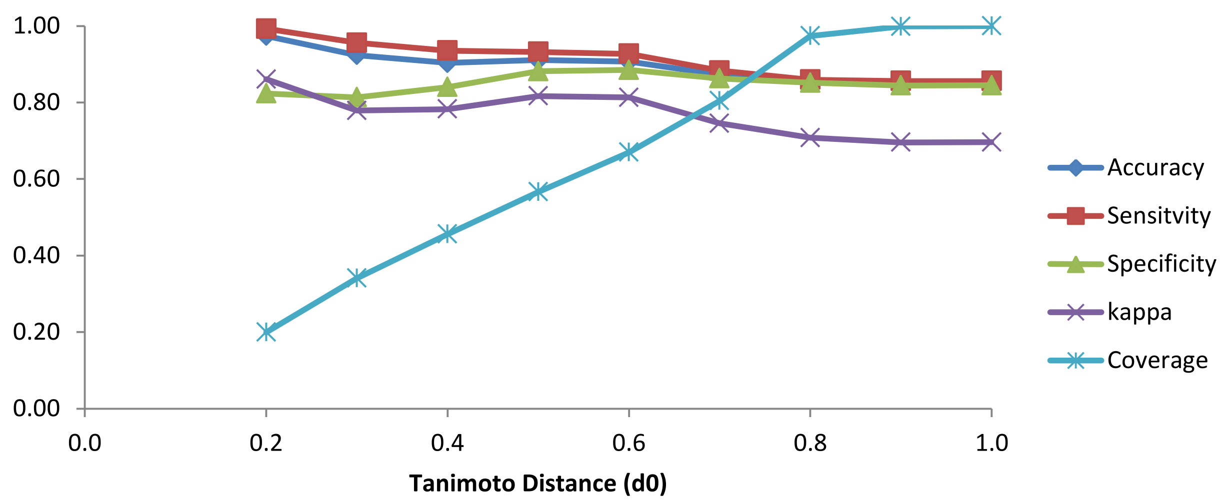

Figure S3. Performance measures of the v-NN Pgp inhibitor model as a function of Tanimoto-distance threshold *d0* at a constant smoothing factor *h* of 0.6.

Table S4. Optimizing the *d0* Value using Leave-One-Out (LOO) Cross Validation for High Accuracy (HA) Parameters. Training Set for Pgp from Chen et al. (*Mol. Pharmaceutics* 2011, 8, 889)

	Smoothing factor	Tanimoto Distance	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	kappa	Coverage
	0.4	0.2	167	9	12	10	0.89	0.94	0.43	0.39	0.20
	0.4	0.3	292	38	40	16	0.85	0.95	0.49	0.49	0.39
	0.4	0.4	369	80	60	37	0.82	0.91	0.57	0.51	0.54
HA	0.4	0.5	436	127	76	41	0.83	0.91	0.63	0.57	0.68
	0.4	0.6	477	163	102	50	0.81	0.91	0.62	0.55	0.79
	0.4	0.7	524	195	139	51	0.79	0.91	0.58	0.52	0.91
	0.4	0.8	566	170	213	43	0.74	0.93	0.44	0.41	0.99
	0.4	0.9	606	43	349	3	0.65	1.00	0.11	0.12	1.00
	0.4	1	609	8	385	0	0.62	1.00	0.02	0.02	1.00

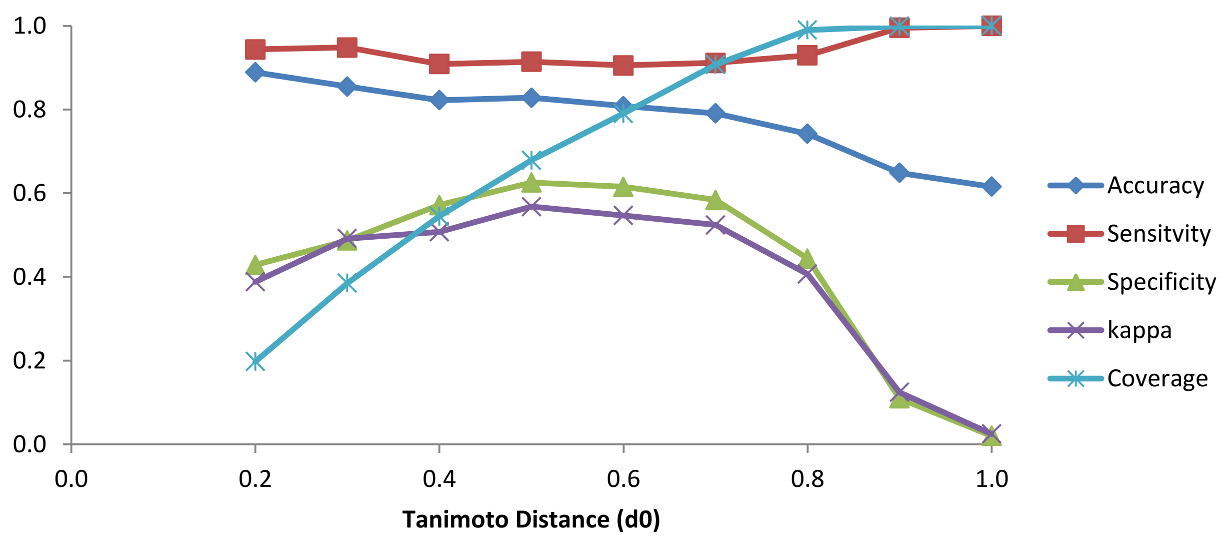

Figure S4. Performance measures of the v-NN Pgp inhibitor model as a function of Tanimoto-distance threshold *d0* at a constant smoothing factor *h* of 0.4.

Table S5. Optimizing the Smoothing Factor Value using Leave-One-Out (LOO) Cross Validation for High Coverage (HC) Parameters. Training Set for Pgp from Chen et al. (*Mol. Pharmaceutics* 2011, 8, 889)

Smoothing Factor		Tanimoto Distance	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	kappa	Coverage
HC	0.1	1	511	253	140	98	0.76	0.84	0.64	0.49	1.00
	0.2	1	533	228	165	76	0.76	0.88	0.58	0.47	1.00
	0.3	1	573	132	261	36	0.70	0.94	0.34	0.31	1.00
	0.4	1	609	8	385	0	0.62	1.00	0.02	0.02	1.00

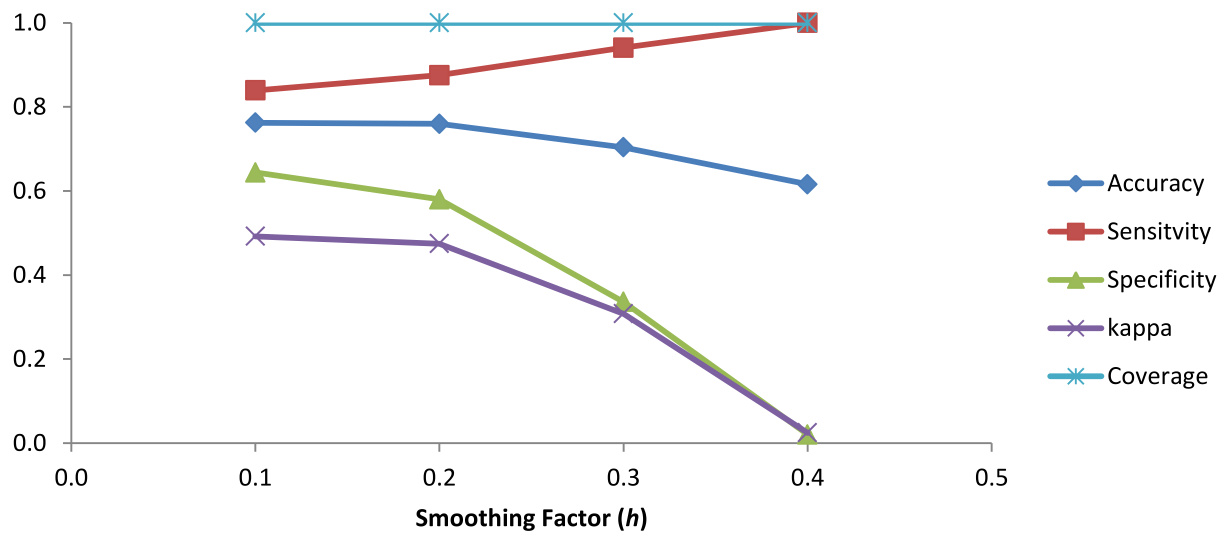

Figure S5. Performance measures of the v-NN Pgp inhibitor model as a function of smoothing factor *h* at constant Tanimoto-distance threshold *d0* of 1.0.

Table S6. Optimizing the *d0* Value using Leave-One-Out (LOO) Cross Validation for High Accuracy (HA) Parameters. Training Set for All Pgp Inhibitors

	Smoothing Factor	Tanimoto Distance	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	kappa	Coverage
	0.3	0.2	516	84	35	29	0.90	0.95	0.71	0.67	0.32
	0.3	0.3	764	164	67	52	0.89	0.94	0.71	0.66	0.51
	0.3	0.4	889	246	98	66	0.87	0.93	0.72	0.67	0.63
HA	0.3	0.5	984	348	125	75	0.87	0.93	0.74	0.68	0.75
	0.3	0.6	1040	424	159	90	0.85	0.92	0.73	0.67	0.83
	0.3	0.7	1086	499	214	106	0.83	0.91	0.70	0.63	0.93
	0.3	0.8	1125	541	286	94	0.81	0.92	0.65	0.60	1.00
	0.3	0.9	1162	402	435	57	0.76	0.95	0.48	0.47	1.00
	0.3	1	1167	341	496	52	0.73	0.96	0.41	0.40	1.00

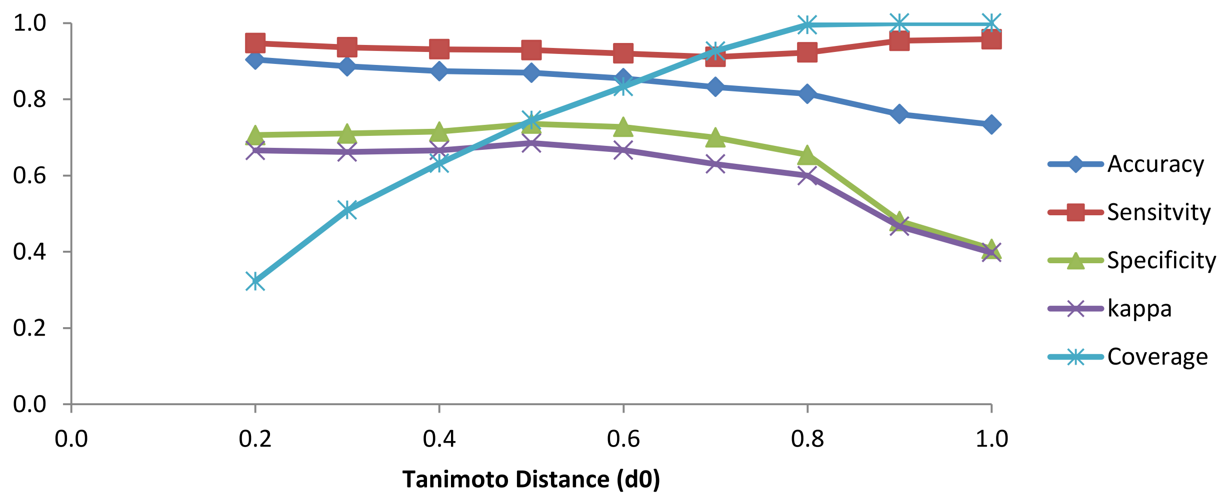

Figure S6. Performance measures of the v-NN Pgp inhibitor model as a function of Tanimoto-distance threshold *d0* at a constant smoothing factor *h* of 0.3.

Table S7. Optimizing the Smoothing Factor Value using Leave-One-Out (LOO) Cross Validation for High Coverage (HC) Parameters. Training Set All Pgp Inhibitors.

Smoothing Factor	Tanimoto Distance	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	kappa	Coverage
0.1	1	1065	643	194	154	0.83	0.87	0.77	0.65	1.00
0.2	1	1101	589	248	118	0.82	0.90	0.70	0.62	1.00
0.3	1	1167	341	496	52	0.73	0.96	0.41	0.40	1.00

Table S8. Performance Measures of the Pgp Substrate Model Using Different Fingerprints.

Data	Fingerprints	Smoothing Factor	Tanimoto Distance	Accuracy	Sensitivity	Specificity	kappa	Coverage
			High Coverage Predictions					
Training	FCFP4	0.3	1	0.73	0.77	0.69	0.46	100%
	MDLKeys	0.2	1	0.74	0.88	0.60	0.47	100%
	ECFP4	0.3	1	0.72	0.78	0.65	0.43	100%
Test	FCFP4	0.3	1	0.78	0.84	0.69	0.54	100%
	MDLKeys	0.2	1	0.75	0.85	0.62	0.48	100%
	ECFP4	0.3	1	0.76	0.80	0.71	0.51	100%
			High Accuracy Predictions					
Training	FCFP4	0.6	0.50	0.76	0.77	0.74	0.51	62%
	MDLKeys	0.2	0.25	0.77	0.81	0.72	0.52	63%
	ECFP4	0.6	0.60	0.77	0.78	0.75	0.53	60%
Test	FCFP4	0.6	0.50	0.82	0.85	0.79	0.64	71%
	MDLKeys	0.2	0.25	0.80	0.87	0.72	0.59	72%
	ECFP4	0.6	0.60	0.81	0.82	0.78	0.60	70%

Table S9. Performance Measures of the Pgp Inhibitor Model Using Different Fingerprints.

Data	Fingerprints	Smoothing Factor	Tanimoto Distance	Accuracy	Sensitivity	Specificity	kappa	Coverage
			High Coverage Predictions					
Broccatelli Training	FCFP4	0.2	1	0.86	0.87	0.85	0.72	100%
	MDLKeys	0.1	1	0.85	0.89	0.82	0.69	100%
	ECFP4	0.2	1	0.85	0.86	0.84	0.69	100%
Ext. Test	FCFP4	0.2	1	0.79	0.89	0.60	0.52	100%
	MDLKeys	0.1	1	0.81	0.92	0.59	0.55	100%
	ECFP4	0.2	1	0.76	0.81	0.67	0.48	100%
Internal Test	FCFP4	0.2	1	0.86	0.89	0.83	0.72	100%
	MDLKeys	0.1	1	0.86	0.89	0.83	0.72	100%
	ECFP4	0.2	1	0.84	0.84	0.83	0.67	100%
Chen Training	FCFP4	0.1	1	0.78	0.86	0.66	0.53	100%
	MDLKeys	0.1	1	0.75	0.86	0.56	0.45	100%
	ECFP4	0.1	1	0.76	0.84	0.64	0.50	100%
Chen Test	FCFP4	0.1	1	0.78	0.87	0.63	0.51	100%
	MDLKeys	0.1	1	0.75	0.85	0.59	0.45	100%
	ECFP4	0.1	1	0.76	0.87	0.59	0.48	100%
			High Accuracy Predictions					
Broccatelli Training	FCFP4	0.2	0.50	0.89	0.91	0.87	0.78	68%
	MDLKeys	0.1	0.25	0.90	0.93	0.85	0.79	69%
	ECFP4	0.2	0.60	0.91	0.93	0.88	0.81	67%
Ext. Test	FCFP4	0.2	0.50	0.86	0.92	0.70	0.65	55%
	MDLKeys	0.1	0.25	0.83	0.93	0.59	0.56	58%
	ECFP4	0.2	0.60	0.88	0.91	0.80	0.71	53%
Internal Test	FCFP4	0.2	0.50	0.89	0.91	0.86	0.77	64%
	MDLKeys	0.1	0.25	0.91	0.97	0.81	0.80	64%
	ECFP4	0.2	0.60	0.89	0.88	0.91	0.78	66%
Chen Training	FCFP4	0.1	0.4	0.83	0.90	0.66	0.58	69%
	MDLKeys	0.1	0.2	0.79	0.88	0.57	0.48	71%
	ECFP4	0.6	0.5	0.83	0.91	0.64	0.58	68%
Chen Test	FCFP4	0.1	0.4	0.82	0.90	0.62	0.53	72%
	MDLKeys	0.1	0.2	0.78	0.86	0.59	0.45	73%
	ECFP4	0.6	0.5	0.79	0.90	0.50	0.43	72%