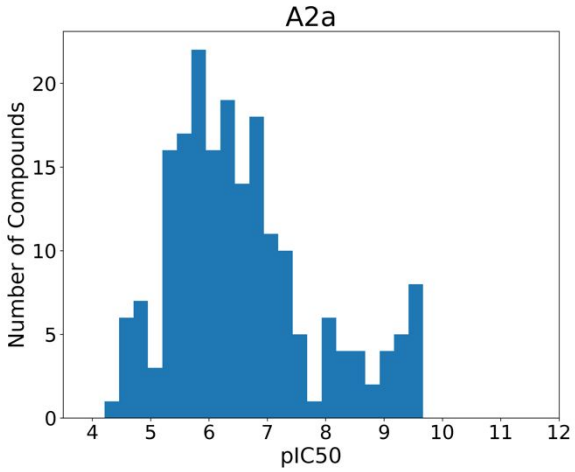
Supporting Information:

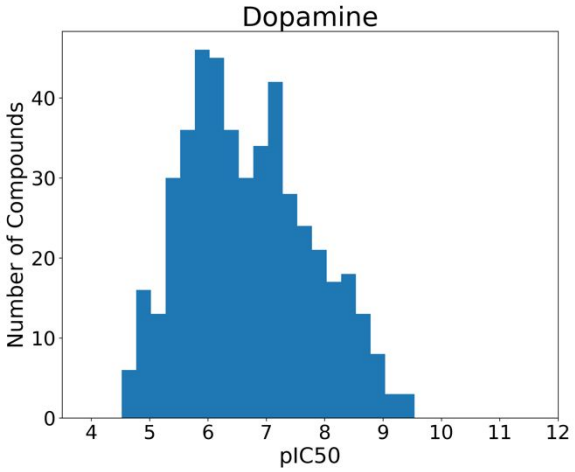
SMILES Pair Encoding: A Data-Driven Substructure Tokenization Algorithm for Deep Learning

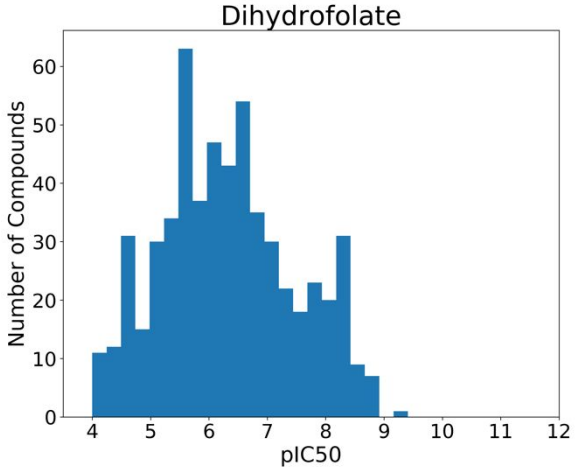
Xinhao Li & Denis Fourches*

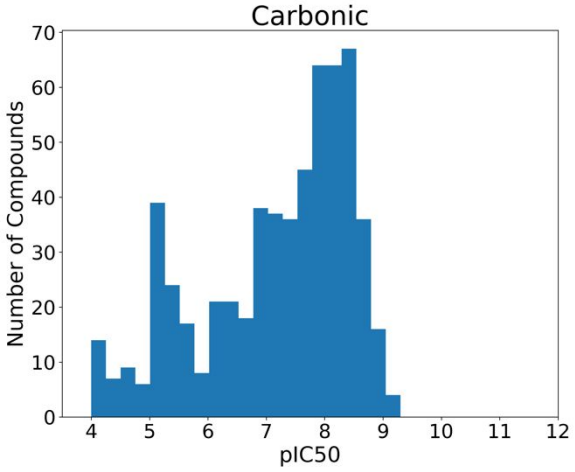
Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, United States.

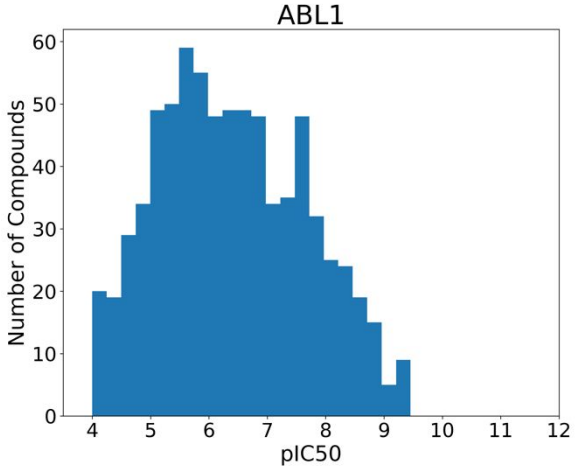
** To whom correspondence should be sent. Email: dfourch@ncsu.edu*

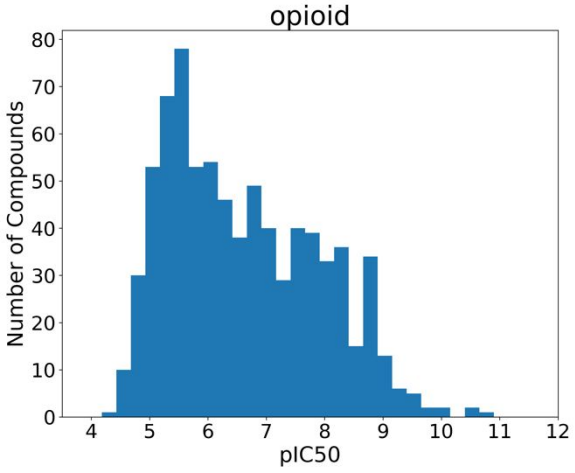

A2a
Number of Compounds
pIC50

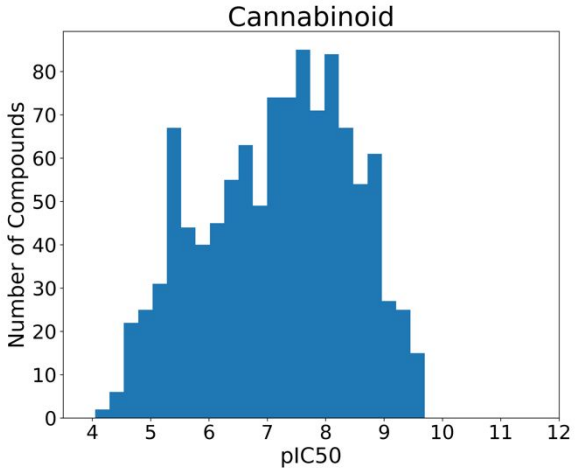

Dopamine
Number of Compounds
pIC50

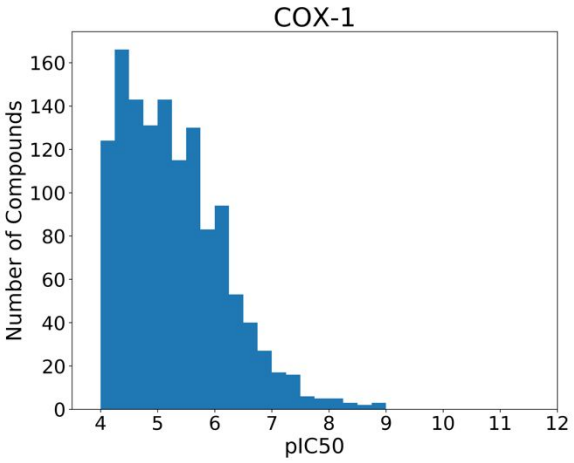

Dihydrofolate
Number of Compounds
pIC50

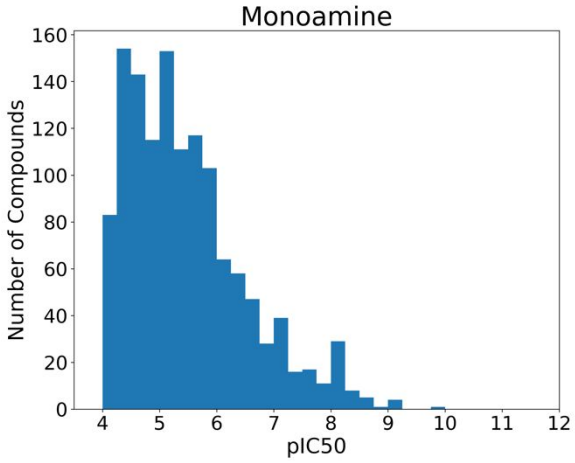

Carbonic
Number of Compounds
pIC50

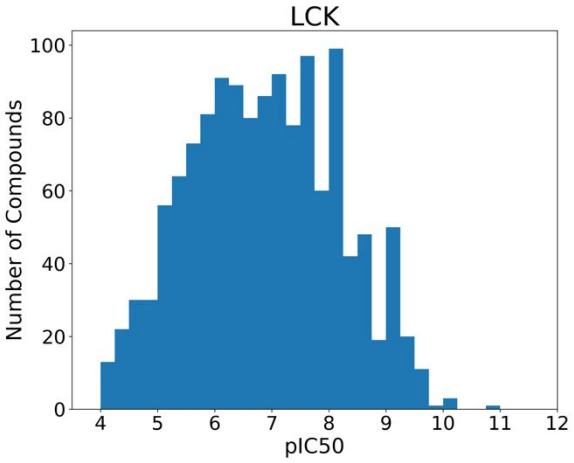

ABL1
Number of Compounds
pIC50

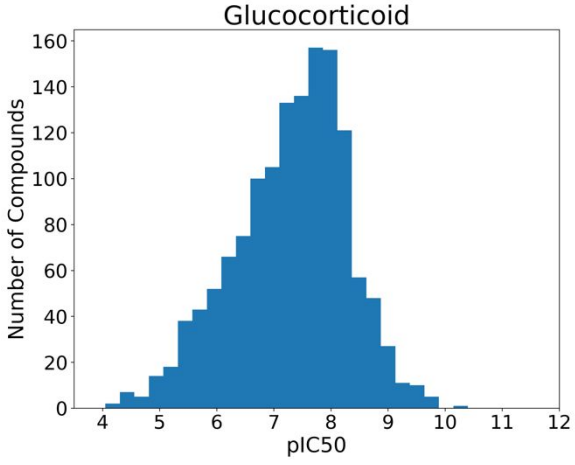

opioid
Number of Compounds
pIC50

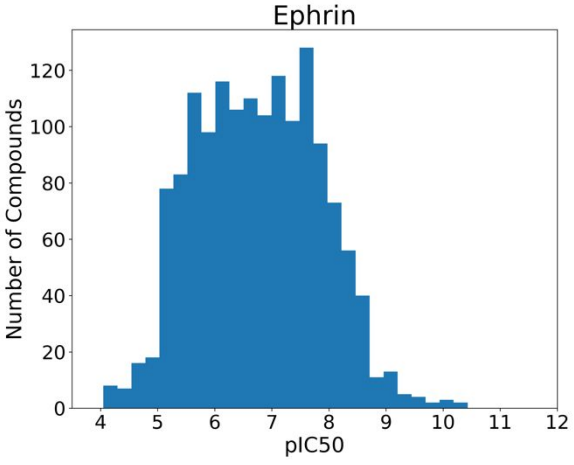

Cannabinoid
Number of Compounds
pIC50

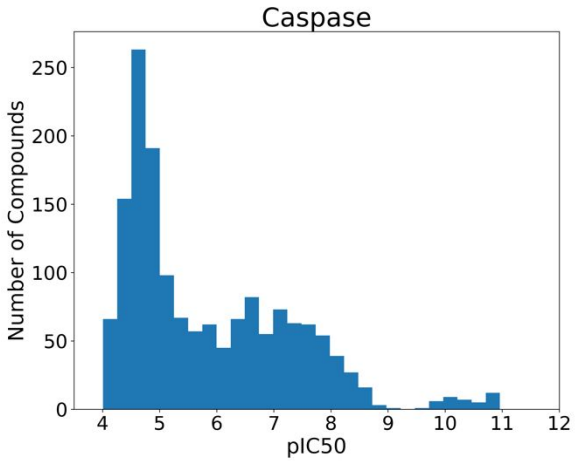

COX-1
Number of Compounds
pIC50

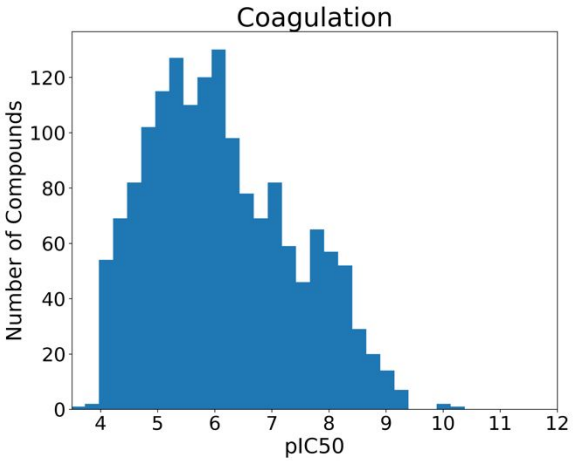

Monoamine
Number of Compounds
pIC50

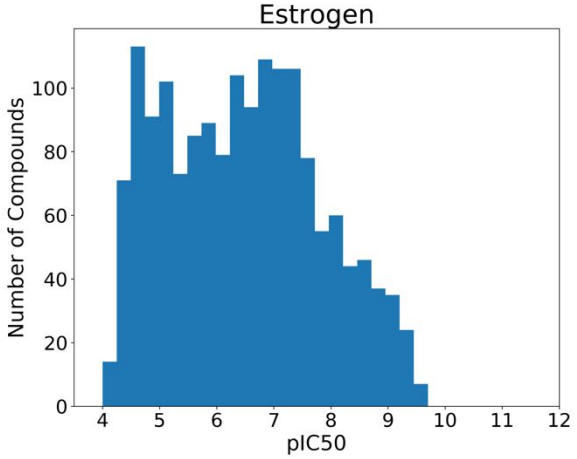

LCK
Number of Compounds
pIC50

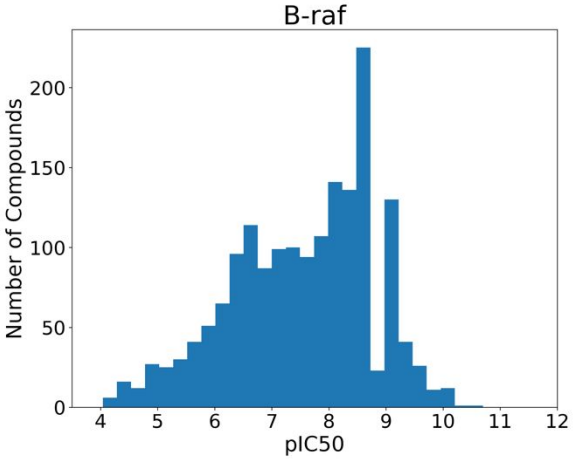

Glucocorticoid
Number of Compounds
pIC50

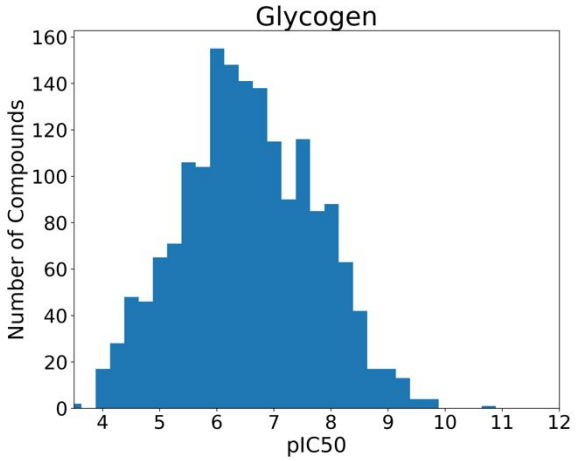

Ephrin
Number of Compounds
pIC50

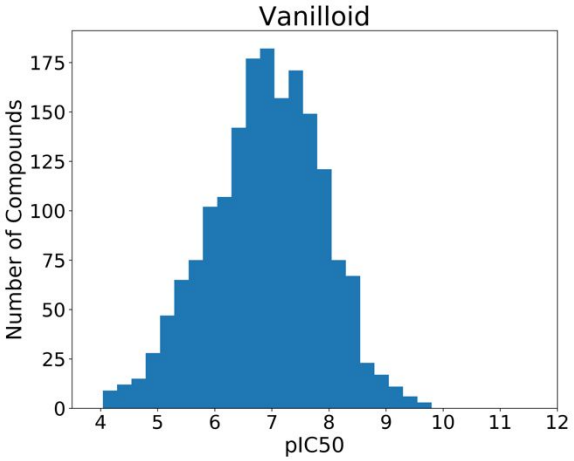

Caspase
Number of Compounds
pIC50


Coagulation
Number of Compounds
pIC50


Estrogen
Number of Compounds
pIC50


B-raf
Number of Compounds
pIC50


Glycogen
Number of Compounds
pIC50


Vanilloid
Number of Compounds
pIC50

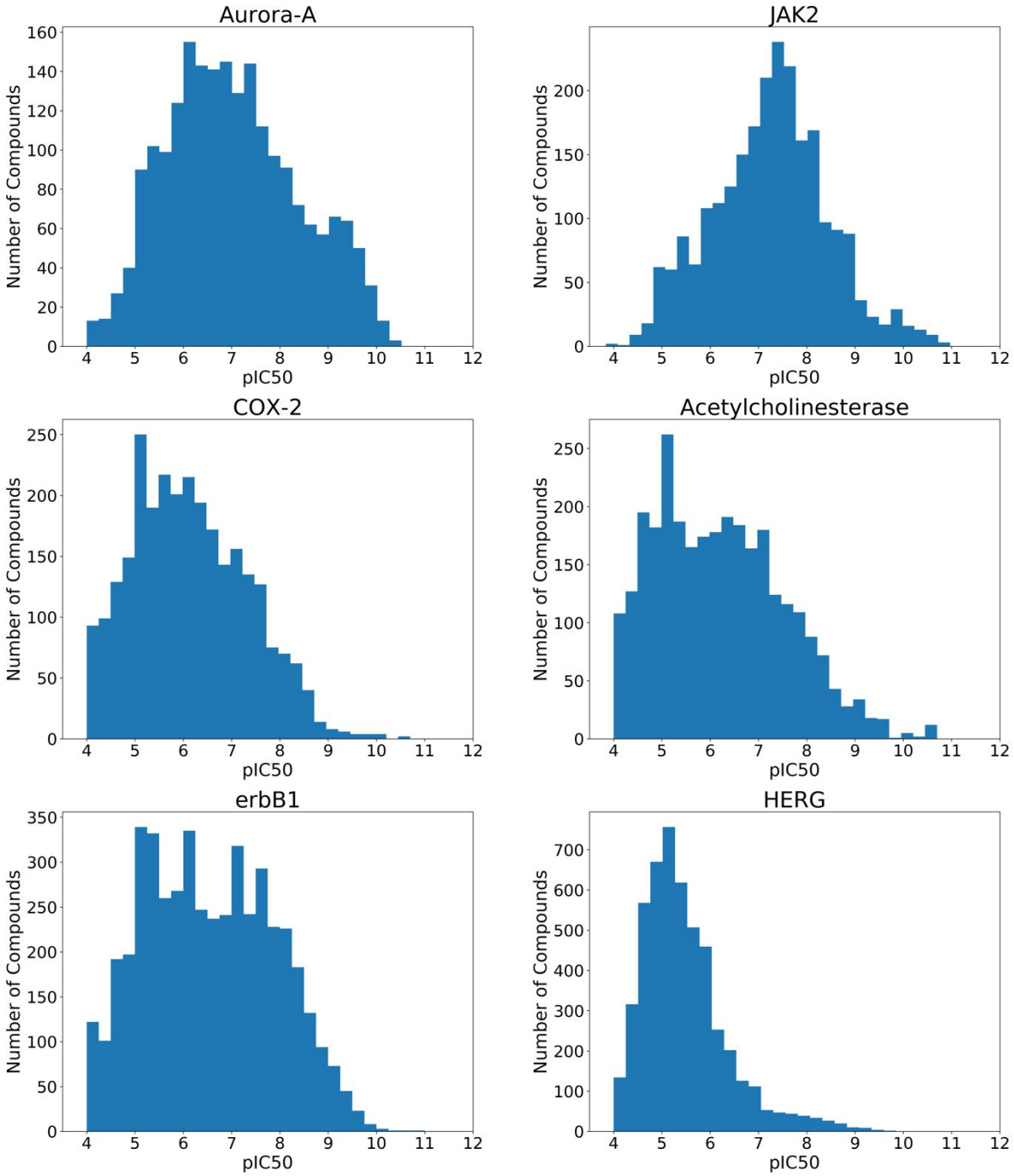


Figure S1. Distributions of QSAR Benchmark Datasets.

Table S1. Number of Extracted Substructures.

	Training Set	**SPE**	**Atom-level**
# of BRICS Fragments	300,315	388,839	407,865
# of Functional Group	14,500	28,859	28,156
# of Scaffolds	474,646	601,502	649,974
# of Ring Systems	35,107	83,891	85,265

Table S2. Performance of QSAR models trained with SPE tokenization.

Targets	RMSE	R^2	MAE
A2a	0.669±0.111	0.724±0.116	0.523±0.078
DRD2	0.719±0.055	0.522±0.080	0.551±0.045
Dihydrofolate	0.771±0.053	0.545±0.049	0.586±0.042
Carbonic	0.582±0.035	0.781±0.046	0.427±0.032
ABL1	0.746±0.050	0.637±0.067	0.566±0.045
opioid	0.636±0.046	0.742±0.042	0.477±0.032
Cannabinoid	0.671±0.041	0.715±0.036	0.500±0.021
COX-1	0.655±0.064	0.486±0.064	0.489±0.046
Monoamine	0.604±0.054	0.645±0.046	0.454±0.037
LCK	0.758±0.035	0.673±0.048	0.589±0.028
Glucocorticoid	0.541±0.042	0.692±0.057	0.426±0.030
Ephrin	0.652±0.051	0.647±0.037	0.494±0.026
Caspase	0.586±0.098	0.835±0.079	0.440±0.093
Coagulation	0.771±0.037	0.622±0.039	0.580±0.031
Estrogen	0.630±0.037	0.780±0.024	0.459±0.032
B-raf	0.599±0.057	0.756±0.046	0.452±0.036
Glycogen	0.731±0.028	0.593±0.041	0.548±0.023
Vanilloid	0.669±0.055	0.543±0.080	0.522±0.032
Aurora-A	0.711±0.049	0.723±0.033	0.530±0.035
JAK2	0.623±0.046	0.725±0.041	0.467±0.026
COX-2	0.728±0.043	0.605±0.050	0.537±0.032
Acetylcholinesterase	0.675±0.0049	0.749±0.033	0.495±0.033
erbB1	0.658±0.023	0.757±0.011	0.492±0.019
HERG	0.536±0.019	0.625±0.033	0.395±0.019

Table S3. Performance of QSAR models trained with atom-level tokenization.

Targets	RMSE	R^2	MAE
A2a	0.776±0.224	0.612±0.215	0.550±0.151
DRD2	0.748±0.097	0.479±0.140	0.576±0.071
Dihydrofolate	0.794±0.101	0.525±0.101	0.592±0.072
Carbonic	0.578±0.069	0.792±0.046	0.421±0.052
ABL1	0.750 0.046	0.635±0.046	0.574±0.034
opioid	0.642±0.072	0.735±0.066	0.485±0.049
Cannabinoid	0.717±0.055	0.679±0.034	0.552±0.033
COX-1	0.665±0.094	0.484±0.089	0.478±0.058
Monoamine	0.624±0.061	0.633±0.053	0.467±0.048
LCK	0.835±0.165	0.591±0.181	0.617±0.047
Glucocorticoid	0.535±0.058	0.695±0.074	0.411±0.042
Ephrin	0.664±0.055	0.636±0.043	0.508±0.027
Caspase	0.587±0.061	0.837±0.050	0.444±0.046
Coagulation	0.770±0.037	0.622±0.045	0.582±0.023
Estrogen	0.655±0.044	0.761±0.028	0.474±0.031
B-raf	0.599±0.067	0.762±0.046	0.443±0.043
Glycogen	0.744±0.045	0.579±0.052	0.555±0.040
Vanilloid	0.670±0.065	0.542±0.082	0.515±0.040
Aurora-A	0.744±0.073	0.698±0.044	0.547±0.040
JAK2	0.642±0.035	0.708±0.042	0.481±0.023
COX-2	0.736±0.064	0.596±0.074	0.543±0.048
Acetylcholinesterase	0.679±0.060	0.745±0.044	0.485±0.037
erbB1	0.661±0.019	0.754±0.013	0.492±0.016
HERG	0.531±0.025	0.637±0.030	0.391±0.020

Table S4. Performance of QSAR models trained with k-mer tokenization.

Targets	RMSE	R^2	MAE
A2a	0.746 ±0.197	0.648 ±0.177	0.532 ±0.125
DRD2	0.754 ±0.074	0.468 ±0.126	0.564 ±0.049
Dihydrofolate	0.783 ±0.078	0.540 ±0.065	0.571 ±0.054
Carbonic	0.588 ±0.062	0.783 ±0.050	0.431 ±0.054
ABL1	0.764 ±0.062	0.620 ±0.063	0.572 ±0.050
opioid	0.651 ±0.073	0.726 ±0.070	0.487 ±0.059
Cannabinoid	0.722 ±0.069	0.674 ±0.047	0.535 ±0.045
COX-1	0.667 ±0.100	0.47 ±0.010	0.480 ±0.062
Monoamine	0.630 ±0.067	0.625 ±0.059	0.464 ±0.046
LCK	0.768 ±0.045	0.663 ±0.053	0.590 ±0.026
Glucocorticoid	0.531 ±0.057	0.699 ±0.068	0.408 ±0.040
Ephrin	0.642 ±0.057	0.699 ±0.068	0.490 ±0.040
Caspase	0.604 ±0.066	0.828 ±0.050	0.458 ±0.067
Coagulation	0.771 ±0.044	0.621 ±0.033	0.581 ±0.026
Estrogen	0.647 ±0.052	0.766 ±0.029	0.470 ±0.046
B-raf	0.584 ±0.056	0.773 ±0.043	0.421 ±0.037
Glycogen	0.727 ±0.041	0.598 ±0.047	0.539 ±0.037
Vanilloid	0.671 ±0.070	0.538 ±0.105	0.510 ±0.042
Aurora-A	0.722 ±0.071	0.715 ±0.050	0.526 ±0.049
JAK2	0.620 ±0.043	0.726 ±0.044	0.459 ±0.023
COX-2	0.718 ±0.053	0.616 ±0.056	0.522 ±0.036
Acetylcholinesterase	0.686 ±0.047	0.740 ±0.035	0.492 ±0.032
erbB1	0.655 ±0.031	0.758 ±0.018	0.483 ±0.020
HERG	0.533 ±0.028	0.633 ±0.033	0.390 ±0.020

Table S5. Performance of random forest models trained ECFP6.

Targets	RMSE	R^2	MAE
A2a	0.738 ±0.238	0.647 ±0.224	0.541 ±0.120
DRD2	0.737 ±0.070	0.494 ±0.110	0.557 ±0.052
Dihydrofolate	0.790 ±0.098	0.527 ±0.111	0.584 ±0.068
Carbonic	0.619 ±0.062	0.763 ±0.036	0.435 ±0.043
ABL1	0.753 ±0.073	0.631 ±0.069	0.572 ±0.051
opioid	0.691 ±0.067	0.692 ±0.007	0.511 ±0.047
Cannabinoid	0.707 ±0.057	0.688 ±0.029	0.532 ±0.042
COX-1	0.719 ±0.092	0.397 ±0.084	0.522 ±0.058
Monoamine	0.663 ±0.037	0.583 ±0.058	0.487 ±0.022
LCK	0.785 ±0.040	0.649 ±0.047	0.609 ±0.035
Glucocorticoid	0.544 ±0.050	0.685 ±0.064	0.411 ±0.032
Ephrin	0.697 ±0.073	0.600 ±0.052	0.521 ±0.037
Caspase	0.540 ±0.056	0.863 ±0.040	0.395 ±0.032
Coagulation	0.811 ±0.049	0.580 ±0.049	0.608 ±0.035
Estrogen	0.702 ±0.037	0.725 ±0.029	0.507 ±0.024
B-raf	0.576 ±0.056	0.780 ±0.041	0.419 ±0.035
Glycogen	0.736 ±0.052	0.589 ±0.049	0.552 ±0.042
Vanilloid	0.708 ±0.044	0.489 ±0.073	0.552 ±0.032
Aurora-A	0.740 ±0.068	0.701 ±0.044	0.534 ±0.044
JAK2	0.624 ±0.033	0.724 ±0.035	0.459 ±0.022
COX-2	0.786 ±0.045	0.541 ±0.054	0.581 ±0.035
Acetylcholinesterase	0.738 ±0.052	0.699 ±0.037	0.539 ±0.043
erbB1	0.691 ±0.026	0.731 ±0.017	0.516 ±0.018
HERG	0.605 ±0.026	0.527 ±0.032	0.447 ±0.018

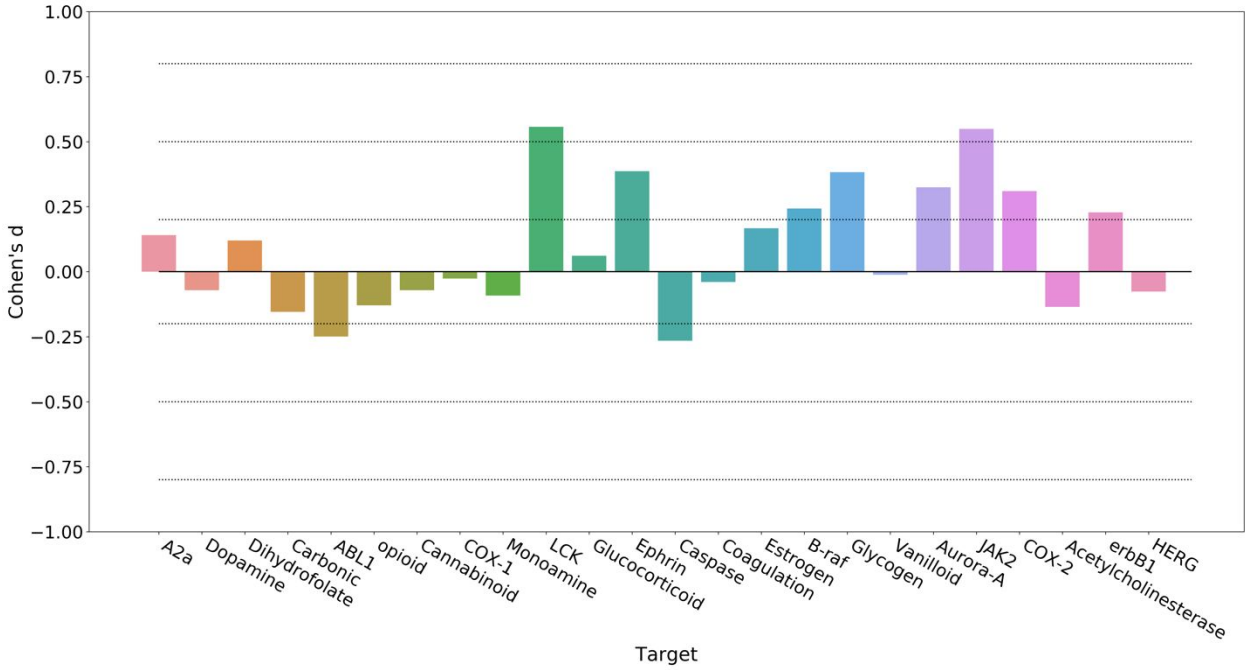


Figure S2. Cohen's d: Atom-level vs. k-mer

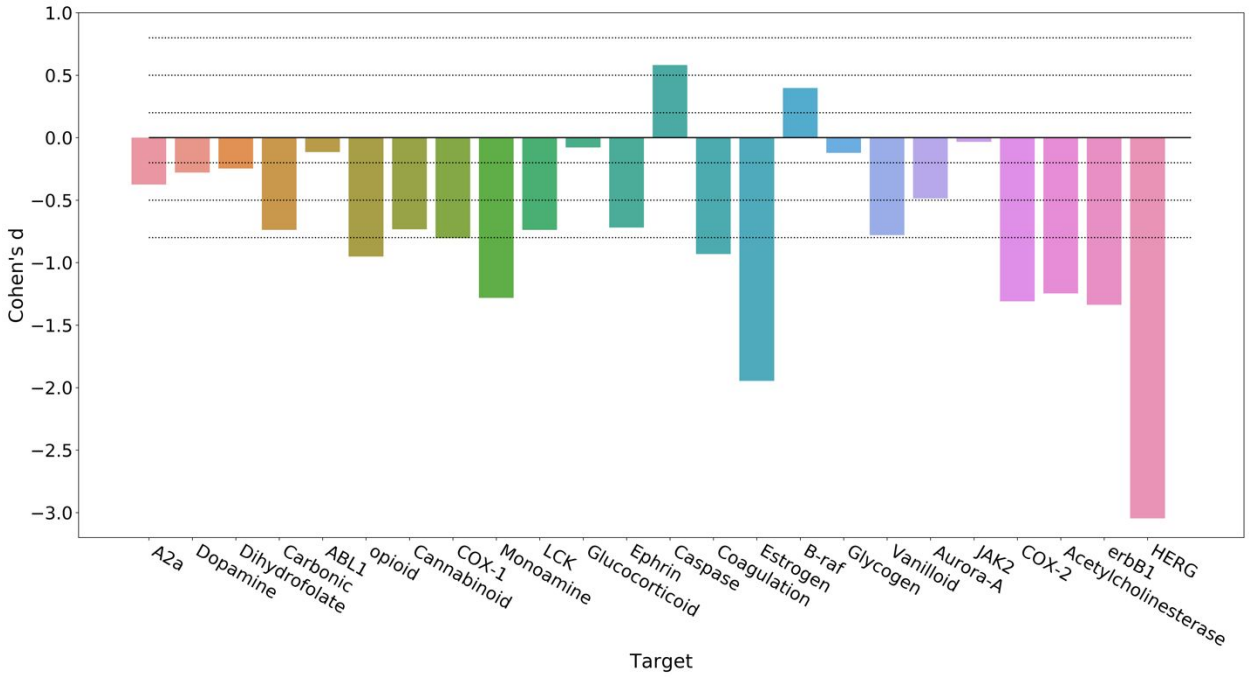


Figure S3. Cohen's d: SPE vs. ECFP6 (Random Forest Model)

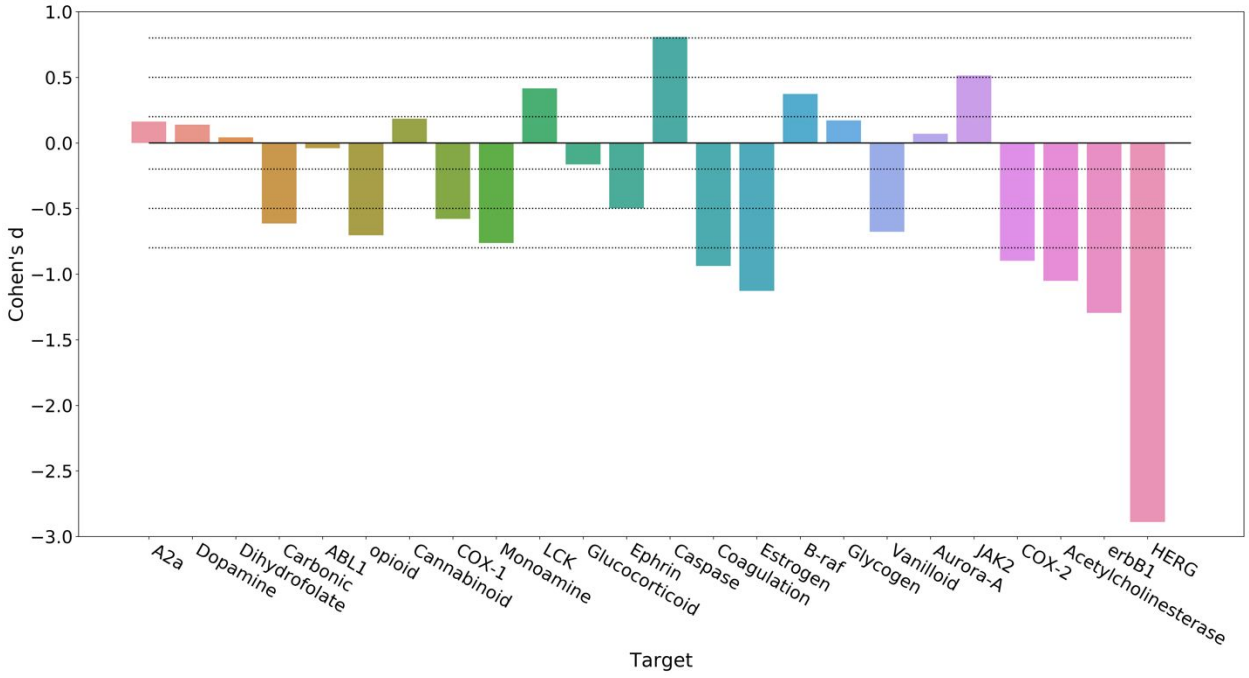


Figure S4. Cohen's d: Atom-level vs. ECFP6 (Random Forest Model)

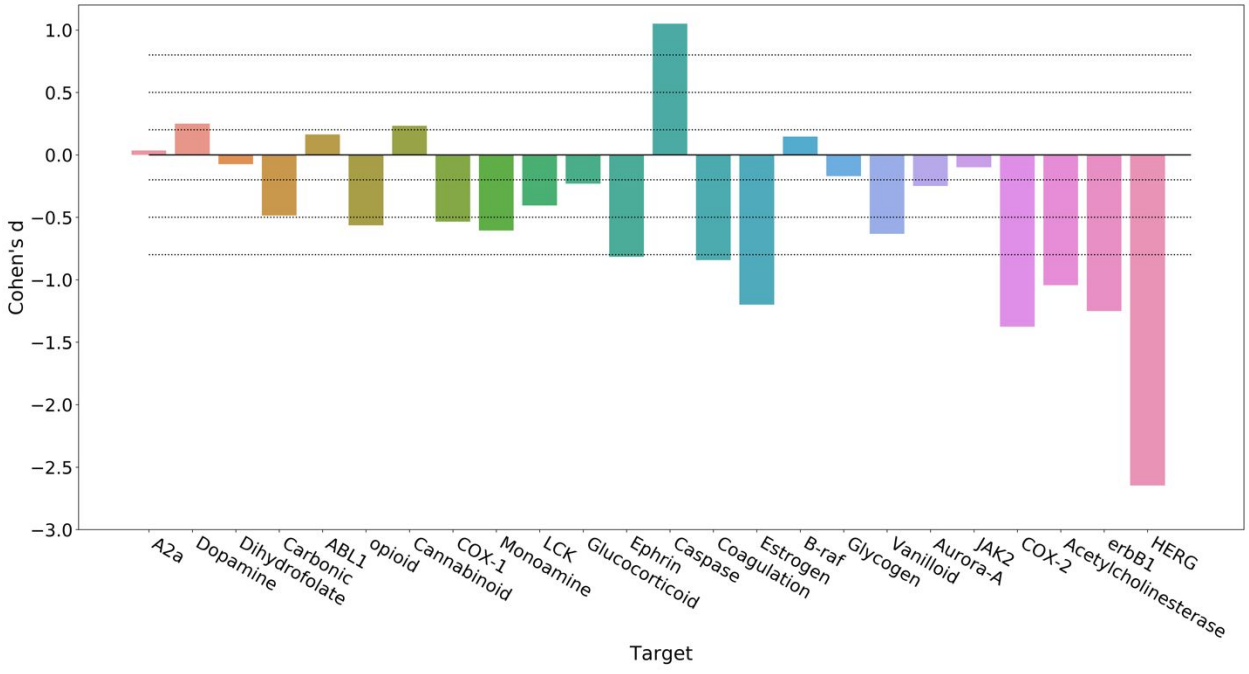


Figure S5. Cohen's d: k-mer vs. ECFP6 (Random Forest Model)