Enhancing Automated Curation of QSAR Datasets

How can we best associate source-provided identifiers to structure data?

100% human curation

100% human curation

Single-identifier mapping

100% human curation

“Smart” multi-identifier mapping

Single-identifier mapping

100% human curation

Progressive human curation

“Smart” multi-identifier mapping

Single-identifier mapping

“Smart” multi-identifier mapping

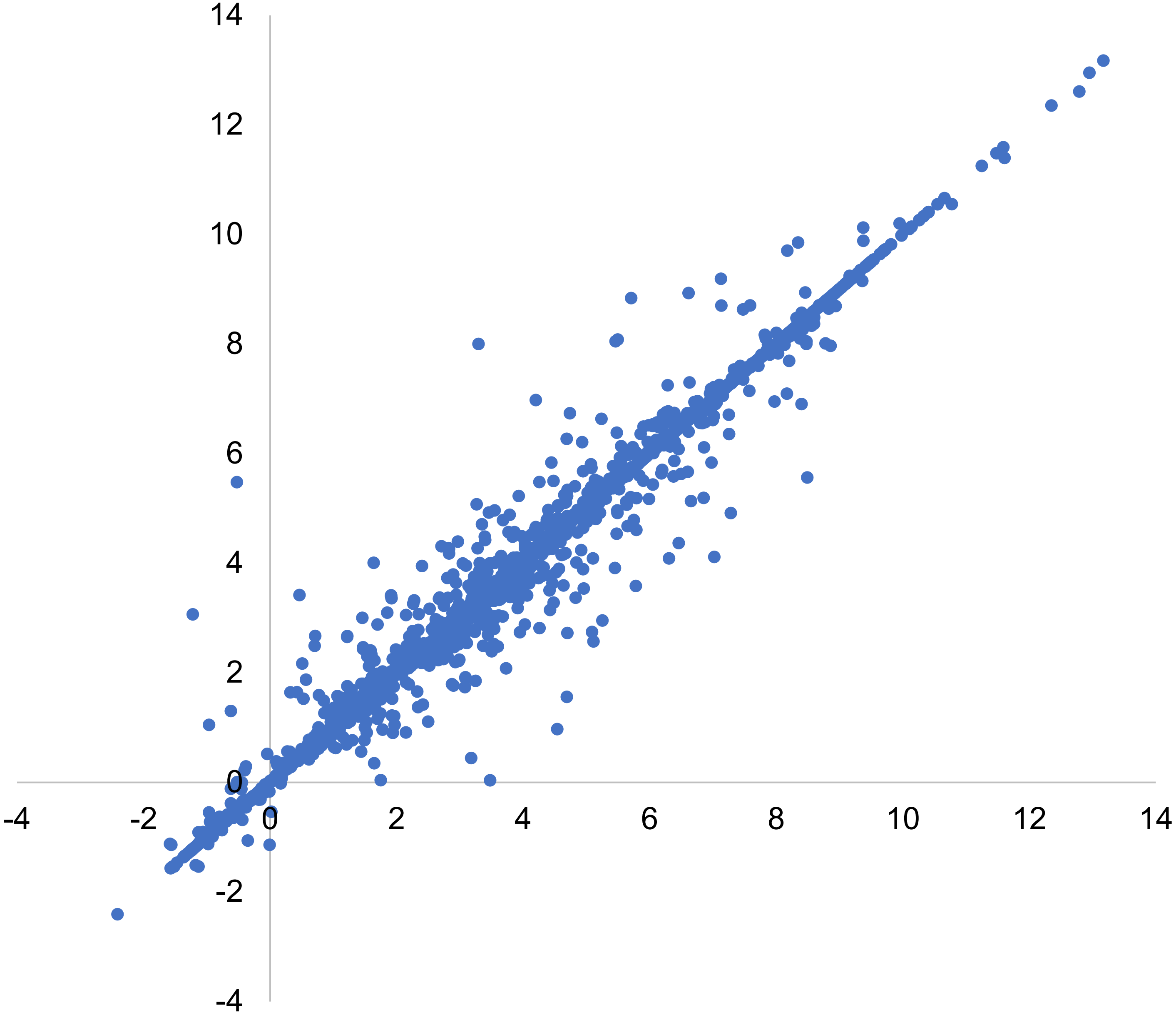
Based on EPA DSSTox database

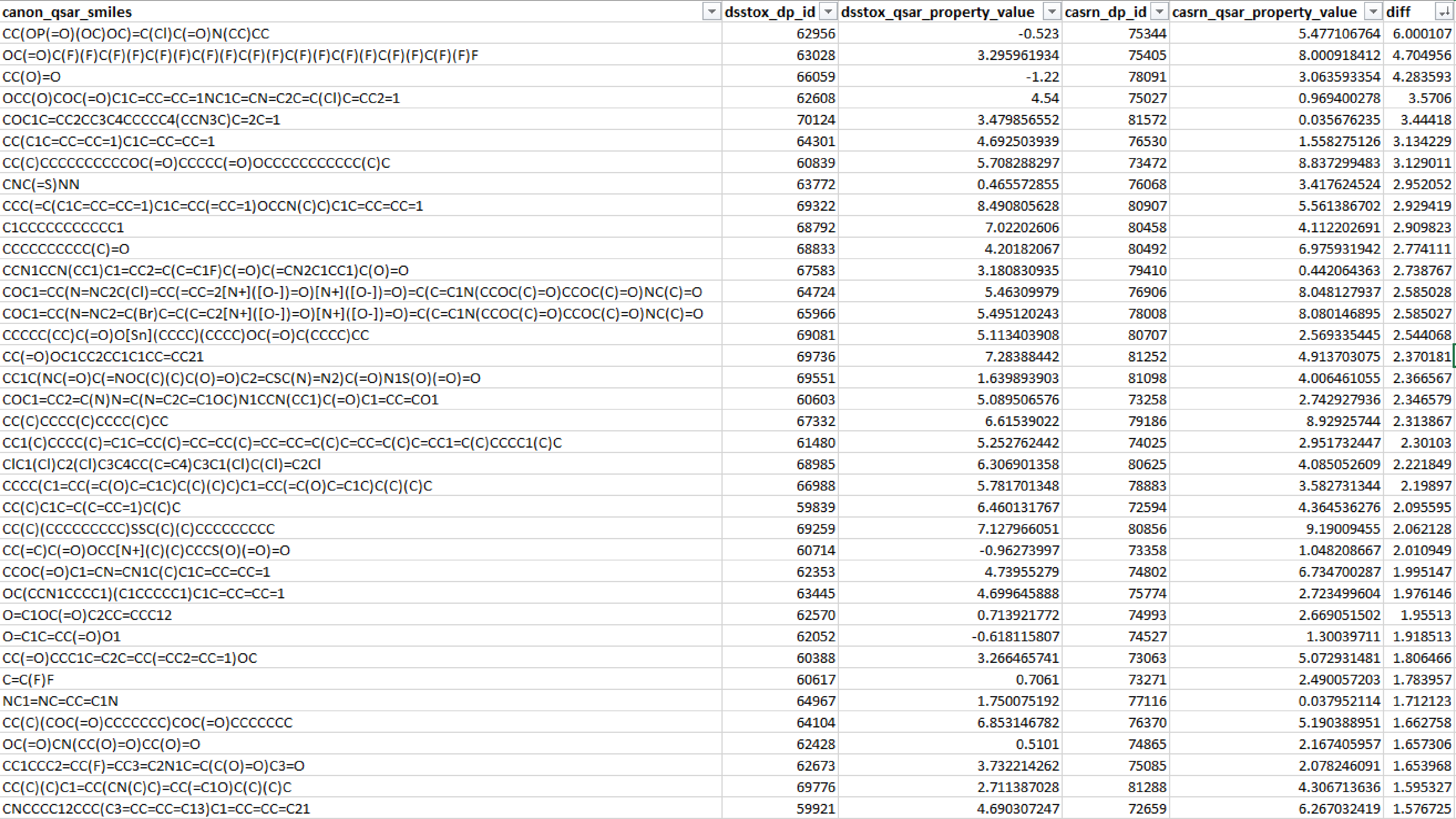
Find and score matches based on CASRN, deprecated CASRN, name, synonym, structure, etc.

> *“Bin” scoring system developed in consultation with human curation team (thanks, guys!)*

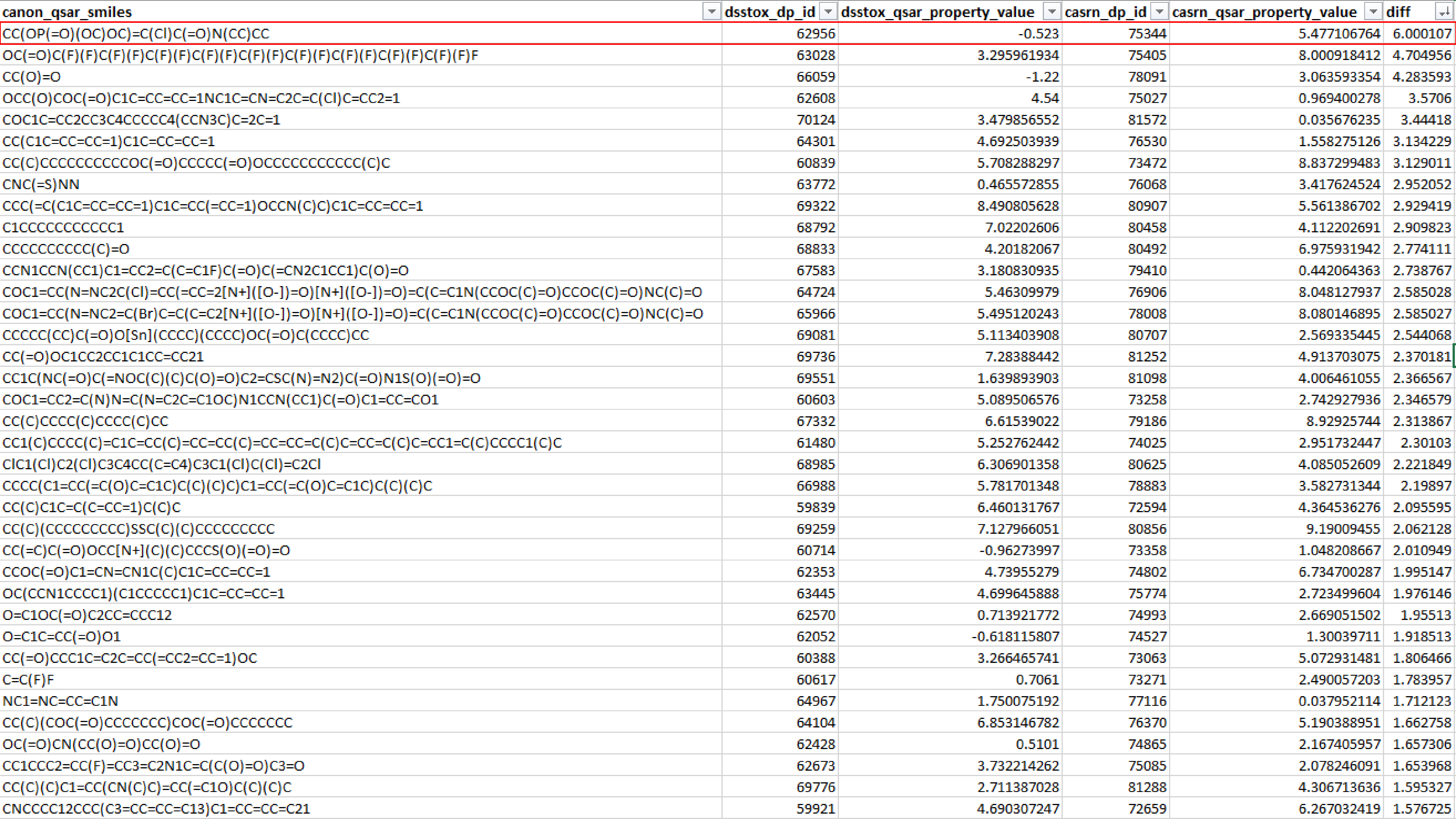
Resolve conflicts based on QSAR-ready SMILES structure

Property Values by CASRN vs. by DSSTox Mapping





canon_qsar_smiles	dsstox_dp_id	dsstox_qsar_property_value	casrn_dp_id	casrn_qsar_property_value	diff
CC(OP(=O)(OC)OC)=C(Cl)C(=O)N(CC)CC	62956	-0.523	75344	5.477106764	6.000107
OC(=O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F	63028	3.295961934	75405	8.000918412	4.704956
CC(O)=O	66059	-1.22	78091	3.063593354	4.283593
OCC(O)COC(=O)C1C=CC=CC=1NC1C=CN=C2C=C(Cl)C=CC2=1	62608	4.54	75027	0.969400278	3.5706
COC1C=CC2CC3C4CCCCC4(CCN3C)C=2C=1	70124	3.479856552	81572	0.035676235	3.44418
CC(C1C=CC=CC=1)C1C=CC=CC=1	64301	4.692503939	76530	1.558275126	3.134229
CC(C)CCCCCCCCCCOC(=O)CCCCC(=O)OCCCCCCCCCCC(C)C	60839	5.708288297	73472	8.837299483	3.129011
CNC(=S)NN	63772	0.465572855	76068	3.417624524	2.952052
CCC(=C(C1C=CC=CC=1)C1C=CC(=CC=1)OCCN(C)C)C1C=CC=CC=1	69322	8.490805628	80907	5.561386702	2.929419
C1CCCCCCCCCCC1	68792	7.02202606	80458	4.112202691	2.909823
CCCCCCCCCC(C)=O	68833	4.20182067	80492	6.975931942	2.774111
CCN1CCN(CC1)C1=CC2=C(C=C1F)C(=O)C(=CN2C1CC1)C(O)=O	67583	3.180830935	79410	0.442064363	2.738767
COC1=CC(N=NC2C(Cl)=CC(=CC=2[N+]([O-])=O)[N+]([O-])=O)=C(C=C1N(CCOC(C)=O)CCOC(C)=O)NC(C)=O	64724	5.46309979	76906	8.048127937	2.585028
COC1=CC(N=NC2=C(Br)C=C(C=C2[N+]([O-])=O)[N+]([O-])=O)=C(C=C1N(CCOC(C)=O)CCOC(C)=O)NC(C)=O	65966	5.495120243	78008	8.080146895	2.585027
CCCCC(CC)C(=O)O[Sn](CCCC)(CCCC)OC(=O)C(CCCC)CC	69081	5.113403908	80707	2.569335445	2.544068
CC(=O)OC1CC2CC1C1CC=CC21	69736	7.28388442	81252	4.913703075	2.370181
CC1C(NC(=O)C(=NOC(C)(C)C(O)=O)C2=CSC(N)=N2)C(=O)N1S(O)(=O)=O	69551	1.639893903	81098	4.006461055	2.366567
COC1=CC2=C(N)N=C(N=C2C=C1OC)N1CCN(CC1)C(=O)C1=CC=CO1	60603	5.089506576	73258	2.742927936	2.346579
CC(C)CCCC(C)CCCC(C)CC	67332	6.61539022	79186	8.92925744	2.313867
CC1(C)CCCC(C)=C1C=CC(C)=CC=CC(C)=CC=CC=C(C)C=CC=C(C)C=CC1=C(C)CCCC1(C)C	61480	5.252762442	74025	2.951732447	2.30103
ClC1(Cl)C2(Cl)C3C4CC(C=C4)C3C1(Cl)C(Cl)=C2Cl	68985	6.306901358	80625	4.085052609	2.221849
CCCC(C1=CC(=C(O)C=C1C)C(C)(C)C)C1=CC(=C(O)C=C1C)C(C)(C)C	66988	5.781701348	78883	3.582731344	2.19897
CC(C)C1C=C(C=CC=1)C(C)C	59839	6.460131767	72594	4.364536276	2.095595
CC(C)(CCCCCCCCC)SSC(C)(C)CCCCCCCCC	69259	7.127966051	80856	9.19009455	2.062128
CC(=C)C(=O)OCC[N+](C)(C)CCCS(O)(=O)=O	60714	-0.96273997	73358	1.048208667	2.010949
CCOC(=O)C1=CN=CN1C(C)C1C=CC=CC=1	62353	4.73955279	74802	6.734700287	1.995147
OC(CCN1CCCC1)(C1CCCCC1)C1C=CC=CC=1	63445	4.699645888	75774	2.723499604	1.976146
O=C1OC(=O)C2CC=CCC12	62570	0.713921772	74993	2.669051502	1.95513
O=C1C=CC(=O)O1	62052	-0.618115807	74527	1.30039711	1.918513
CC(=O)CCC1C=C2C=CC(=CC2=CC=1)OC	60388	3.266465741	73063	5.072931481	1.806466
C=C(F)F	60617	0.7061	73271	2.490057203	1.783957
NC1=NC=CC=C1N	64967	1.750075192	77116	0.037952114	1.712123
CC(C)(COC(=O)CCCCCCC)COC(=O)CCCCCCC	64104	6.853146782	76370	5.190388951	1.662758
OC(=O)CN(CC(O)=O)CC(O)=O	62428	0.5101	74865	2.167405957	1.657306
CC1CCC2=CC(F)=CC3=C2N1C=C(C(O)=O)C3=O	62673	3.732214262	75085	2.078246091	1.653968
CC(C)(C)C1=CC(CN(C)C)=CC(=C1O)C(C)(C)C	69776	2.711387028	81288	4.306713636	1.595327
CNCCCC12CCC(C3=CC=CC=C13)C1=CC=CC=C21	59921	4.690307247	72659	6.267032419	1.576725



canon_qsar_smiles	dsstox_dp_id	dsstox_qsar_property_value	casrn_dp_id	casrn_qsar_property_value	diff
CC(OP(=O)(OC)OC)=C(Cl)C(=O)N(CC)CC	62956	-0.523	75344	5.477106764	6.000107
OC(=O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F	63028	3.295961934	75405	8.000918412	4.704956
CC(O)=O	66059	-1.22	78091	3.063593354	4.283593
OCC(O)COC(=O)C1C=CC=CC=1NC1C=CN=C2C=C(Cl)C=CC2=1	62608	4.54	75027	0.969400278	3.5706
COC1C=CC2CC3C4CCCCC4(CCN3C)C=2C=1	70124	3.479856552	81572	0.035676235	3.44418
CC(C1C=CC=CC=1)C1C=CC=CC=1	64301	4.692503939	76530	1.558275126	3.134229
CC(C)CCCCCCCCCCOC(=O)CCCCC(=O)OCCCCCCCCCCC(C)C	60839	5.708288297	73472	8.837299483	3.129011
CNC(=S)NN	63772	0.465572855	76068	3.417624524	2.952052
CCC(=C(C1C=CC=CC=1)C1C=CC(=CC=1)OCCN(C)C)C1C=CC=CC=1	69322	8.490805628	80907	5.561386702	2.929419
C1CCCCCCCCCCC1	68792	7.02202606	80458	4.112202691	2.909823
CCCCCCCCCC(C)=O	68833	4.20182067	80492	6.975931942	2.774111
CCN1CCN(CC1)C1=CC2=C(C=C1F)C(=O)C(=CN2C1CC1)C(O)=O	67583	3.180830935	79410	0.442064363	2.738767
COC1=CC(N=NC2C(Cl)=CC(=CC=2[N+]([O-])=O)[N+]([O-])=O)=C(C=C1N(CCOC(C)=O)CCOC(C)=O)NC(C)=O	64724	5.46309979	76906	8.048127937	2.585028
COC1=CC(N=NC2=C(Br)C=C(C=C2[N+]([O-])=O)[N+]([O-])=O)=C(C=C1N(CCOC(C)=O)CCOC(C)=O)NC(C)=O	65966	5.495120243	78008	8.080146895	2.585027
CCCCC(CC)C(=O)O[Sn](CCCC)(CCCC)OC(=O)C(CCCC)CC	69081	5.113403908	80707	2.569335445	2.544068
CC(=O)OC1CC2CC1C1CC=CC21	69736	7.28388442	81252	4.913703075	2.370181
CC1C(NC(=O)C(=NOC(C)(C)C(O)=O)C2=CSC(N)=N2)C(=O)N1S(O)(=O)=O	69551	1.639893903	81098	4.006461055	2.366567
COC1=CC2=C(N)N=C(N=C2C=C1OC)N1CCN(CC1)C(=O)C1=CC=CO1	60603	5.089506576	73258	2.742927936	2.346579
CC(C)CCCC(C)CCCC(C)CC	67332	6.61539022	79186	8.92925744	2.313867
CC1(C)CCCC(C)=C1C=CC(C)=CC=CC(C)=CC=CC=C(C)C=CC=C(C)C=CC1=C(C)CCCC1(C)C	61480	5.252762442	74025	2.951732447	2.30103
ClC1(Cl)C2(Cl)C3C4CC(C=C4)C3C1(Cl)C(Cl)=C2Cl	68985	6.306901358	80625	4.085052609	2.221849
CCCC(C1=CC(=C(O)C=C1C)C(C)(C)C)C1=CC(=C(O)C=C1C)C(C)(C)C	66988	5.781701348	78883	3.582731344	2.19897
CC(C)C1C=C(C=CC=1)C(C)C	59839	6.460131767	72594	4.364536276	2.095595
CC(C)(CCCCCCCCC)SSC(C)(C)CCCCCCCCC	69259	7.127966051	80856	9.19009455	2.062128
CC(=C)C(=O)OCC[N+](C)(C)CCCS(O)(=O)=O	60714	-0.96273997	73358	1.048208667	2.010949
CCOC(=O)C1=CN=CN1C(C)C1C=CC=CC=1	62353	4.73955279	74802	6.734700287	1.995147
OC(CCN1CCCC1)(C1CCCCC1)C1C=CC=CC=1	63445	4.699645888	75774	2.723499604	1.976146
O=C1OC(=O)C2CC=CCC12	62570	0.713921772	74993	2.669051502	1.95513
O=C1C=CC(=O)O1	62052	-0.618115807	74527	1.30039711	1.918513
CC(=O)CCC1C=C2C=CC(=CC2=CC=1)OC	60388	3.266465741	73063	5.072931481	1.806466
C=C(F)F	60617	0.7061	73271	2.490057203	1.783957
NC1=NC=CC=C1N	64967	1.750075192	77116	0.037952114	1.712123
CC(C)(COC(=O)CCCCCCC)COC(=O)CCCCCCC	64104	6.853146782	76370	5.190388951	1.662758
OC(=O)CN(CC(O)=O)CC(O)=O	62428	0.5101	74865	2.167405957	1.657306
CC1CCC2=CC(F)=CC3=C2N1C=C(C(O)=O)C3=O	62673	3.732214262	75085	2.078246091	1.653968
CC(C)(C)C1=CC(CN(C)C)=CC(=C1O)C(C)(C)C	69776	2.711387028	81288	4.306713636	1.595327
CNCCCC12CCC(C3=CC=CC=C13)C1=CC=CC=C21	59921	4.690307247	72659	6.267032419	1.576725

Property Value (by DSSTox) =

–0.523 –(log10(M))

CASRN	Name	SMILES	Value	Unit	Source
	phosphamidon	CCN(CC)C(=O)/C(Cl)=C(/C)O[P](=O)(OC)OC	3.336568151385261	M	AqSolDB
	phosphamidon	CCN(CC)C(=O)\C(Cl)=C(/C)O[P](=O)(OC)OC	3.336568151385261	M	AqSolDB
	Dimecron	CCN(CC)C(=O)C(=CCOP(=O)(OC)OC)Cl	3.3342641276323497	M	AqSolDB
	phosphamidon	ClC(=CCOP(=O)(OC)OC)C(=O)N(CC)CC	3.311311214826	M	Bradley
23783-98-4		CCN(CC)C(=O)C(Cl)=C(C)OP(=O)(OC)OC	0.000999	g/L	OChem

Property Value (by CASRN) =

5.477 –(log10(M))

CASRN	Name	SMILES	Value	Unit	Source
23783-98-4		CCN(CC)C(=O)C(Cl)=C(C)OP(=O)(OC)OC	0.000999	g/L	OChem

Progressive human curation

Versioned database for dataset storage

Detailed parameters for mapping strategy in each dataset version

Backwards linkage to source data points

“Guarantee” existing data points or make individualized corrections where appropriate

More to think about…

Detecting experimental differences by modality of distribution within structure and dataset

Detecting duplicate data by distribution and covariance across structures and datasets

*How similar is **too** similar?*

Accounting for data spread and sample size when unifying property values