SUPPORTING INFORMATION

Predictive structure-based toxicology approaches to assess the androgenic potential of chemicals

Daniela Trisciuzzi[a], Domenico Alberga[a,b], Kamel Mansouri[c,d,e], Richard Judson[d], Ettore Novellino[f], Giuseppe Felice Mangiatordi[a,b]* and Orazio Nicolotti[a,b]*

a. *Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro," Via E. Orabona 4, I-70126 Bari, Italy*

b. *Centro Ricerche TIRES, Università degli Studi di Bari "Aldo Moro", Via Amendola 173, I-70126 Bari, Italy*

c. *Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA*

d. *National Center for Computational Toxicology, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA*

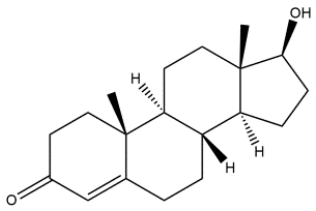
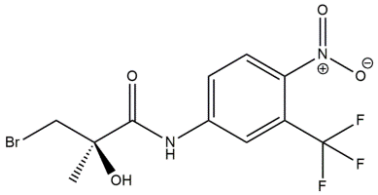
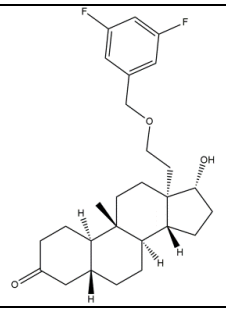
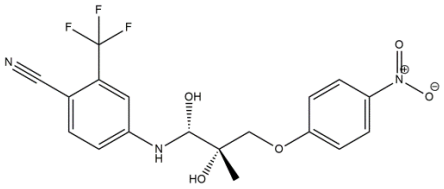
e. *ScitoVation LLC, 6 Davis Dr, Research Triangle Park, NC 27709*

f. *Dipartimento di Farmacia – Università degli Studi di Napoli "Federico II", Via D. Montesano 49, 80131 Napoli, Italy*

* Authors to whom correspondence should be addressed; e-mails: giuseppe.mangiatordi@uniba.it and orazio.nicolotti@uniba.it; telephone: +39-080-544-2551/2765; fax: +39-080-5442230

Tables

Table S1. Chemical structures of X-ray solved cognate ligands of PDB entries 2AM9, 2AX9, 2PNU, 3B66, 3L3X, 3RLJ, 5CJ6, 4QL8 and 2HVC. Hydrogen bonds, hydrophobic, π –π and π-cation interactions of each cognate ligands in the corresponding binding site residues are also reported.

PDB CODE	X-ray solved cognate ligand	Hydrogen bond	Hydrophobic interactions	π -π π-cation interactions.
2AM9		Asn705 Thr877	Leu873 Met895	-
2AX9		Leu704 Asn705 Gln711	Met745	-
2PNU		Asn705 Thr877	Trp741 Met742 Met745 Leu873	Trp741
3B66		Asn705	Trp741 Met745 Thr877 Met895	Trp741

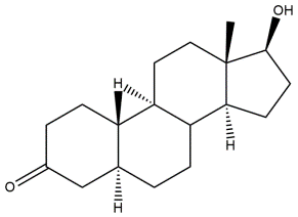
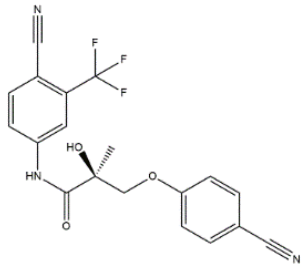
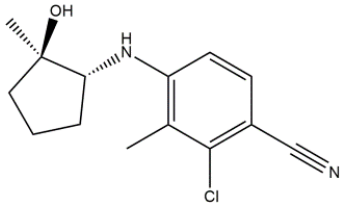
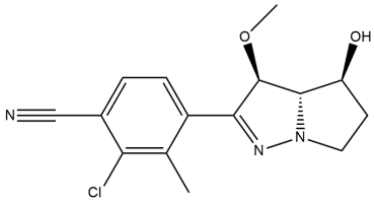
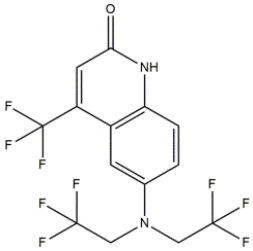
3L3X		Asn705 Thr877	Met745 Met780 Met895 Leu873	-
3RLJ		Leu704 Asn705	Met742 Met745 Thr877 Met895	Trp741
5CJ6		Asn705	Leu704 Met745 Thr877	-
4QL8		Asn705 Thr877	Gly708	-
2HVC		Gln711 Arg752	Leu704	-

Table S2. RMSD values (Å) computed comparing the Cartesian coordinates of the heavy atoms of the docking poses with those available in the X-ray cognate ligands.

PDB CODE	RMSD
2AM9	0.495
2AX9	0.436
2PNU	0.890
3B66	0.574
3L3X	0.654
3RLJ	2.069
5CJ6	1.906
4QL8	0.544
2HVC	0.441

Table S3. Maximum and minimum values of the 162 molecular descriptors.

Molecular Descriptor	Minimum	Maximum
MW	42.04	1701.19
AlogP	-13.97	18.10
HBA	0	46
HBD	0	25
RB	0	32
HeavyAtomCount	3	122
ChiralCenterCount	0	40
ChiralCenterCountAllPossible	0	40
RingCount	0	11
PSA	0	777.98
Estate	6.5	390.01
MR	6.92	381.42
Polar	3.97	182.39
Atom Count	3.00	122.00
Atoms in Ring System	0.00	66.00
Bond Count	2.00	132.00
Bonds in Ring System	0.00	66.00
Centralization	0.00	77312.00
Cyclomatic number	0.00	11.00
Eccentric connectivity	6.00	5688.00
Eccentricity	5.00	2668.00
First Zagreb	6.00	658.00
Gutman Molecular Topological	6.00	435922.00
Number of ring systems	0.00	11.00
Polarity	0.00	228.00
Quadratic	0.00	88.00
Ramification	0.00	55.00
Ring Count 10	0.00	1.00
Ring Count 12	0.00	1.00
Ring Count 14	0.00	1.00
Ring Count 3	0.00	3.00
Ring Count 4	0.00	2.00
Ring Count 5	0.00	4.00
Ring Count 6	0.00	11.00
Ring Count 7	0.00	1.00
Ring Count 8	0.00	1.00
Ring bridge count	0.00	24.00
Ring perimeter	0.00	66.00
Schultz Molecular Topological	16.00	412910.00
Second Zagreb	4.00	787.00
Sum of topological distances between Br..Br	0.00	117.00
Sum of topological distances between Cl..Br	0.00	20.00
Sum of topological distances between Cl..Cl	0.00	262.00

Sum of topological distances between F..Br	0.00	12.00
Sum of topological distances between F..Cl	0.00	150.00
Sum of topological distances between F..F	0.00	1170.00
Sum of topological distances between F..I	0.00	87.00
Sum of topological distances between I..I	0.00	36.00
Sum of topological distances between N..Br	0.00	42.00
Sum of topological distances between N..Cl	0.00	258.00
Sum of topological distances between N..F	0.00	268.00
Sum of topological distances between N..I	0.00	41.00
Sum of topological distances between N..N	0.00	351.00
Sum of topological distances between N..O	0.00	505.00
Sum of topological distances between N..P	0.00	13.00
Sum of topological distances between N..S	0.00	85.00
Sum of topological distances between O..Br	0.00	124.00
Sum of topological distances between O..Cl	0.00	140.00
Sum of topological distances between O..F	0.00	342.00
Sum of topological distances between O..I	0.00	104.00
Sum of topological distances between O..O	0.00	15037.00
Sum of topological distances between O..P	0.00	80.00
Sum of topological distances between O..S	0.00	337.00
Sum of topological distances between P..Br	0.00	27.00
Sum of topological distances between P..Cl	0.00	24.00
Sum of topological distances between P..F	0.00	9.00
Sum of topological distances between P..P	0.00	12.00
Sum of topological distances between S..Br	0.00	24.00
Sum of topological distances between S..Cl	0.00	34.00
Sum of topological distances between S..F	0.00	97.00
Sum of topological distances between S..I	0.00	4.00
Sum of topological distances between S..P	0.00	40.00
Sum of topological distances between S..S	0.00	46.00
Topological diameter	2.00	28.00
Topological radius	1.00	20.00
Unipolarity	2.00	963.00
Variation	0.00	1288.00
Wiener	4.00	97399.00
ALOGP1	-2.10	328.90
ALOGP10	-1.21	844.18
ALOGP2	-22.35	414.63
ALOGP3	0.00	760.39
ALOGP4	-3.84	87.52
ALOGP5	-4.32	177.07
ALOGP6	0.00	161.96
ALOGP7	0.00	315.19
ALOGP8	0.00	177.83
ALOGP9	0.00	67.88
Average connectivity index chi-0	0.66	0.90

Average connectivity index chi-1	0.37	0.71
Average eccentricity	1.67	21.87
Average valence connectivity index chi-0	0.42	1.34
Average valence connectivity index chi-1	0.21	1.31
Average vertex distance degree	2.67	1596.70
Balaban-type index from Z weighted distance matrix - Barysz matrix	0.94	11.59
Balaban-type index from mass weighted distance matrix	0.94	11.61
Balaban distance connectivity index	0.83	7.08
Connectivity chi-1 [Randic connectivity]	1.41	57.35
Connectivity index chi-0	2.71	89.38
E-state topological parameter	0.16	776.74
Eccentric	0.00	3.36
First Mohar	-122.29	5679.05
First Zagreb index by valence vertex degrees	8.13	2410.00
Global topological charge	0.00	1.00
Gutman MTI by valence vertex degrees	7.31	1837350.00
Harary	2.50	900.32
Hyper-distance-path index	5.00	827956.00
Kier Hall electronegativity	-0.38	66.00
Kier benzene-likeliness index	0.63	3.94
Log of product of row sums	1.26	389.55
MR1	-7.99	93.01
MR8	41.07	1493.91
Mean Distance Degree Deviation	0.00	285.90
Mean Square Distance Balaban	1.41	14.53
Mean Wiener	1.33	13.20
Mean topological charge index of order 1	0.00	0.75
Mean topological charge index of order 10	0.00	0.02
Mean topological charge index of order 2	0.00	0.22
Mean topological charge index of order 3	0.00	0.19
Mean topological charge index of order 4	0.00	0.15
Mean topological charge index of order 5	0.00	0.12
Mean topological charge index of order 6	0.00	0.06
Mean topological charge index of order 7	0.00	0.04
Mean topological charge index of order 8	0.00	0.04
Mean topological charge index of order 9	0.00	0.03
Modified Randic connectivity	9.19	373.39
Molecular electrotopological variation	0.00	411.11
Molecule cyclized degree	0.00	1.00
Narumi Geometric Topological	1.26	2.35
Narumi Harmonic Topological	1.18	2.31
Narumi Simple Topological	0.69	82.60
Petitjean 2D shape	0.00	1.00
Pogliani	6.50	290.00
Quasi Wiener	4.00	82992.83

Radial centric	0.00	3.82
Reciprocal hyper-distance-path index	2.33	353.23
Schultz Molecular Topological by valence vertex degrees	27.25	848506.00
Second Mohar	0.40	10.97
Second Zagreb index by valence vertex degrees	5.04	2222.00
Solvation connectivity index chi-0	2.71	89.38
Solvation connectivity index chi-1	1.41	57.35
Spanning tree number	0.00	19.71
Square reciprocal distance sum	2.25	273.75
Topological charge index of order 1	0.00	33.00
Topological charge index of order 10	0.00	2.57
Topological charge index of order 2	0.00	20.22
Topological charge index of order 3	0.00	14.14
Topological charge index of order 4	0.00	11.88
Topological charge index of order 5	0.00	6.19
Topological charge index of order 6	0.00	5.36
Topological charge index of order 7	0.00	4.06
Topological charge index of order 8	0.00	3.41
Topological charge index of order 9	0.00	3.22
Total structure connectivity	0.00	0.71
Valence connectivity index chi-0	1.52	59.89
Valence connectivity index chi-1	0.51	33.18
Van der Waals surface area	61.59	1493.91
Wiener-type index from Z weighted distance matrix - Barysz matrix	1.71	76755.33
Wiener-type index from mass weighted distance matrix	1.70	76784.78
Xu	1.65	81.71
Reciprocal distance Randic-type index	1.15	8.84
Reciprocal distance square Randic-type index	3.46	2020.05

Table S4. Synoptic view of confusion matrix.

		EXPERIMENTAL CLASS	
		P	N
PREDICTED CLASS	P	True Positive	False Positive
	N	False Negative	True Negative

Table S5. Number (percentage) of excluded compounds undocked or returning unrealistic (positive) values of docking score for all the considered crystal structures (PDB entries: 2AM9, 2AX9, 2PNU, 3B66, 3L3X, 3RLJ, 5CJ6, 4QL8 and 2HVC).

PDB CODE	Total EPA-ARDB	Excluded compounds (%)
2AM9	1592	97 (6.09%)
2AX9	1599	90 (5.62%)
2PNU	1643	46 (2.79%)
3B66	1636	53 (3.23%)
3L3X	1595	94 (5.89%)
3RLJ	1635	54 (3.30%)
5CJ6	1607	82 (5.10%)
4QL8	1612	77 (4.77%)
2HVC	1622	67 (4.13%)

Table S6. Number (percentage) of excluded compounds after the application of the first filter (VS – Bounding box) and after the application of both filters (VS – Bounding box/Convex hull) for all the three docked VS.

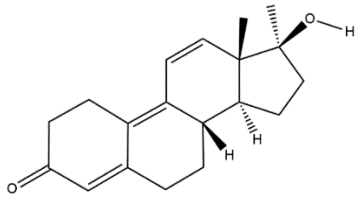
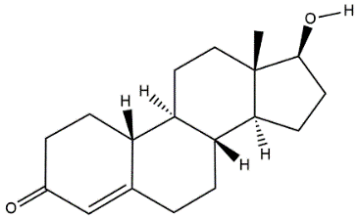
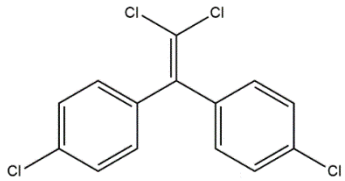
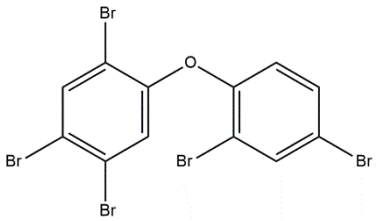
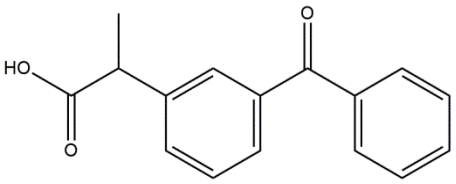
	VS – Bounding box (%)	VS – Bounding box/Convex hull (%)
VS1	102 (3.93%)	355 (13.70%)
VS2	115 (4.44%)	374 (14.44%)
VS3	104 (4.01%)	361 (13.93%)

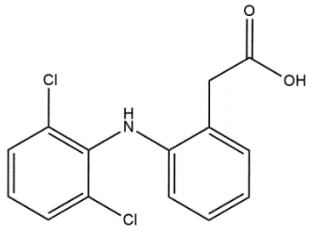
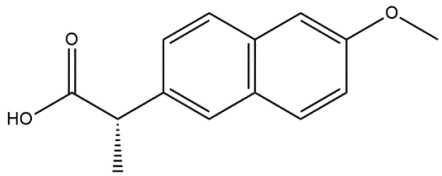
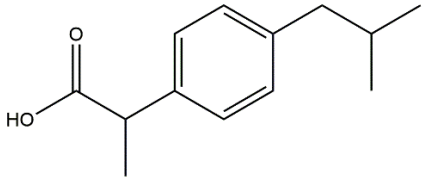
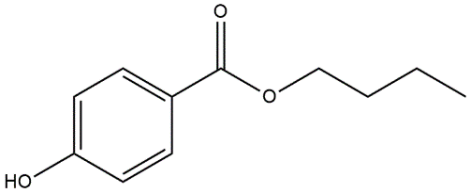
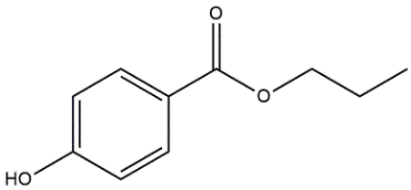
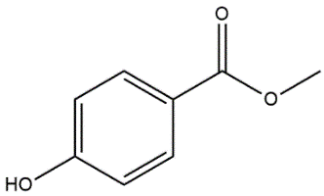
Table S7. Percentage of binders found at SE >0.75 in the class of predicted non-binders for all the considered methods on the three VS (DES: Molecular descriptors; ECFP: Extended Connectivity Fingerprint; FCFP: Functional Connectivity Fingerprint; DAY: Daylight Fingerprint; PCA: Principal component analysis; t-SNE: t-Distributed Stochastic Neighbour Embedding).

	VS1	VS2	VS3
Bounding box/Convex hull DES (PCA)	4.42%	4.59%	5.09%
Bounding box/Convex hull ECFP (PCA)	6.61%	6.98%	7.53%
Bounding box/Convex hull FCFP (PCA)	7.60%	7.96%	8.85%
Bounding box/Density ECFP (PCA)	4.57%	5.50%	5.92%
Bounding box/Density FCFP (PCA)	8.16%	8.03%	9.26%
Bounding box/Convex hull ECFP (t-SNE*)	7.17%	7.65%	8.19%
Bounding box/Convex hull FCFP (t-SNE*)	7.15%	7.52%	8.10%

* t-SNE was employed over a set of 2048 bit Morgan circular fingerprints (equivalent to ECFP or FCFP computed using the RDkit module in Python[1]), with a radius of 2.

Table S8. Docking score relative to twelve representative substances predicted by the 2PNU best performing classification model.

Referend compound	Chemical structure	Docking score (kJ/mol)
R1881		**-41.99**
Nandrolone		**-39.43**
p,p'-DDE		**-34.40**
Bisphenol-A		**-33.09**
PBDE-99		**-31.99**
Ketoprofen		**-30.08**

Diclofenac		**-29.77**
Naproxen		**-27.57**
Ibuprofen		**-26.44**
Butylparaben		**-25.13**
Propylparaben		**-22.17**
Methylparaben		**-19.73**

Figures

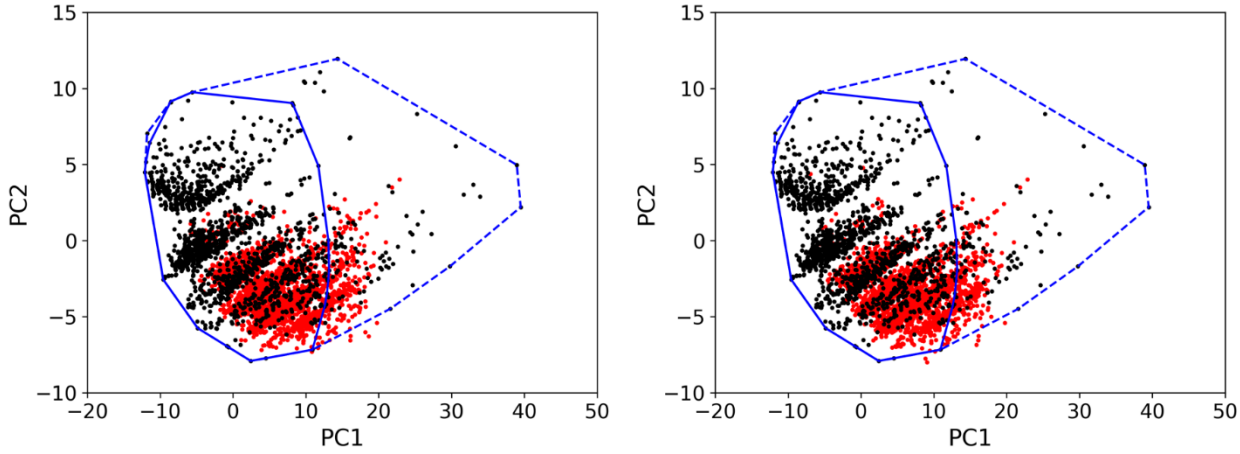


Figure S1. The convex-hull was defined on the top two PCs obtained from the 162 descriptors computed for each compound of the EPA-ARDB. The outer polygon (dashed line) takes into accounts all the chemicals in the EPA-ARDB (black circles), while the inner polygon (solid line) retains the 95% of them based on a user-dependent inclusive threshold. Chemicals (red circles) of the VS2 (on the left side) and VS3 (on the right side) outside the inner 95% polygon are flagged as outside AD.

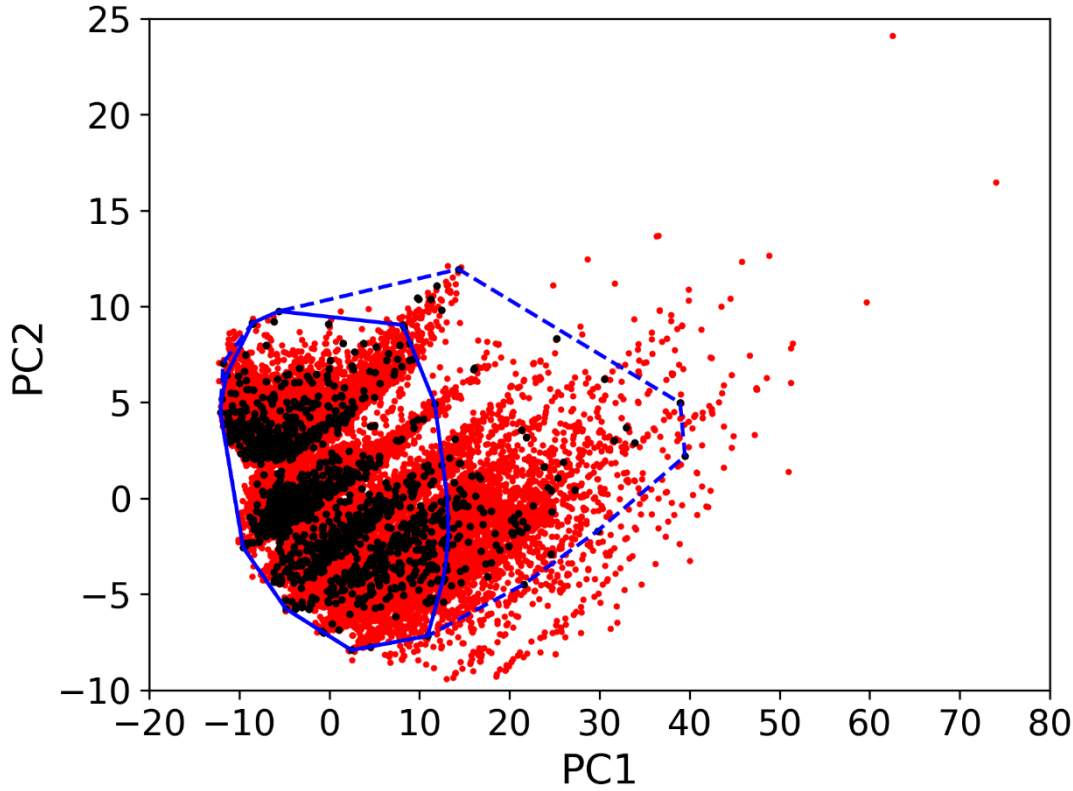


Figure S2. The convex-hull was defined on the top two PCs obtained from the 162 descriptors computed for each compound of the EPA-ARDB. The outer polygon (dashed line) takes into accounts all the chemicals in the EPA- ARDB (black circles), while the inner polygon (solid line) retains the 95% of them based on a user-dependent inclusive threshold. Chemicals of the external blind dataset (red circles) outside the inner 95% polygon are flagged as outside AD.

Supplementary methodological details

The definition of AD is given by the subsequent application of two-step approach: the former is the bounding box method; the latter is applied by means of an interpolation space based on:

- the Cartesian coordinates of the top two principal components (PCs) obtained from the initial 162 descriptors (see Bounding box/Convex hull DES (PCA) in Table S7);
- the Cartesian coordinates of the top two PCs obtained from the Extended-connectivity fingerprints (ECFPs) (see Bounding box/Convex hull ECFP (PCA) in Table S7);
- the Cartesian coordinates of the top two PCs obtained from the Functional-connectivity fingerprints (FCFPs) (see Bounding box/Convex hull FCFP (PCA) in Table S7);
- the probability density distribution of the Cartesian coordinates of the top two PCs obtained from the ECFPs following the Jouan-Rimbaud et al. protocol[2] (see Bounding box/Density ECFP (PCA) in Table S7);
- the probability density distribution of the Cartesian coordinates of the top two PCs obtained from the FCFPs following the Jouan-Rimbaud et al. protocol[2] (see Bounding box/Density FCFP (PCA) in Table S7);
- ECFPs calculated for each compound of EPA-ARDB. The chemicals were projected into a 2D map using t-SNE technique.[3] This map was employed to define a convex hull filter (see Bounding box/Convex hull ECFP (t-SNE) in Table S7);
- FCFPs calculated for each compound of EPA-ARDB. The chemicals were projected also into a 2D map using t-SNE technique.[3] This map was employed to define a convex hull filter (see Bounding box/Convex hull FCFP (t-SNE) in Table S7).

REFERENCES

(S1) RDKit: Open-Source Cheminformatics; Http://Www.rdkit.org ; Last Accessed 12/09/2017.

(S2) Jouan-Rimbaud, D.; Bouveresse, E.; Massart, D. L.; de Noord, O. E. Detection of Prediction Outliers and Inliers in Multivariate Calibration. *Anal. Chim. Acta* **1999**, *388*, 283–301.

(S3) Maaten, L. van der; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.