Gadaleta et al.:

Automated Integration of Structural, Biological and Metabolic Similarities to Sustain Read-Across

Supplementary Data

Tab. S1: List of SMARTS codifying functional groups
SMARTS are retrieved and readapted from RDKit's Functional Group Filter KNIME node

NAME	SMARTS	NAME	SMARTS
Aromatic Acid Chloride	[$(C-!@[a])](=O)(Cl)	Aromatic Amine	[N;!H0;$(N-c);!$(N-[!#6;!#1]);!$(N-C=[O,N,S])]
Aliphatic Acid Chloride	[$(C-!@[A;!Cl])](=O)(Cl)	Aliphatic Amine	[N;!H0;!$(N-c);$(N-C);!$(N-[!#6;!#1]);!$(N-C=[O,N,S])]
Aromatic Carboxylic Acid	[$(C-!@[a])](=O)([O;H,-])	Cyclic Amine	[N;!H0;R;$(N-[#6]);!$(N-[!#6;!#1]);!$(N-C=[O,N,S])]
Aliphatic Carboxylic Acid	[$(C-!@[A;!O])](=O)([O;H,-])	Aromatic Boronic Acid	[$(B-!@c)](O)(O)
Alpha Amino Acid	[$(C-[C;!$(C=[!#6])]-[N;!H0;!$(N-[!#6;!#1]);!$(N-C=[O,N,S])])](=O)([O;H,-])	Aliphatic Boronic Acid	[$(B-!@C)](O)(O)
Aromatic Sulfonyl Chloride	[$(S-!@c)](=O)(=O)(Cl)	Aromatic Isocyanate	[$(N-!@c)](=!@C=!@O)
Aliphatic Sulfonyl Chloride	[$(S-!@C)](=O)(=O)(Cl)	Aliphatic Isocyanate	[$(N-!@C)](=!@C=!@O)
Aliphatic Alcohol	[O;H1;$(O-!@[C;!$(C=!@[O,N,S])])]	Aromatic Alcohol	[O;H1;$(O-!@c)]
Aromatic Aldehyde	[CH;D2;$(C-!@[a])](=O)	Aliphatic Halogen	[$([F,Cl,Br,I]-!@C);!$([F,Cl,Br,I]-!@C-!@[F,Cl,Br,I])]
Aliphatic Aldehyde	[CH;D2;$(C-!@C)](=O)	Aromatic Azide	[N;H0;$(N-c);D2]=[N;D2]=[N;D1]
Aromatic Halogen	[F,Cl,Br,I;$(*-!@c)]	Aliphatic Azide	[N;H0;$(N-C);D2]=[N;D2]=[N;D1]

Tab. S2: Information on the chemical space covered by source datasets implemented in the RAX workflow
For each activity category in the two datasets, the range of (and the mean) values of relevant physico-chemical properties for the two datasets are reported. Properties were calculated with the Chemistry Developmental Kit (CDK) "Molecular Properties" node available in KNIME (https://cdk.github.io/).

DATASET	CATEGORIES	LOGP	TPSA	MW
DILIRank	No-DILI	-2.06 to 14.44 (3.02)	0.0 to 3115.35 (169.41)	60.02 to 7049.04 (530.52)
	Less-DILI	-1.73 to 5.53 (2.67)	3.24 to 872.52 (109.86)	123.04 to 1619.71 (384.42)
	Most-DILI	0.91 to 5.75 (2.69)	3.24 to 518.52 (90.4)	76.03 to 1269.44 (360.03)
ToxRef	Non-Hepatotox	0.03 to 4.76 (2.42)	0.0 to 474.9 (58.67)	60.03 to 972.32 (265.28)
	Hepatotox	1.02 to 4.76 (2.27)	0.0 to 165.84 (59.4)	42.02 to 756.36 (288.4)

doi:10.14573/altex.2002281s1

Tab. S3: Prediction statistics of the RAX integrated approach applied to the ToxRef binary classification
For each combination of similarity lists (i.e., number of lists including a single analogue) and pre-filtering method, the number of true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), the sensitivity (SEN), the specificity (SPE), the Balanced Accuracy (BA) and the number and the percentage (%) of predictions returned are reported. The first row of the table refers to the benchmark performance related to the sole use of the closest structural neighbour to infer the prediction.

Group	Filters	TP	FP	TN	FN	SEN	SPE	BA	#preds[1]	%
-	-	*117*	*77*	*363*	*88*	*0.571*	*0.825*	***0.698***	*645*	*1.000*
1	none	68	24	416	137	0.332	0.945	**0.639**	645	1.000
2	none	93	57	302	63	0.596	0.841	**0.719**	515	0.798
3	none	23	7	51	10	0.697	0.879	**0.788**	91	0.141
1	FG+MCS	63	31	292	81	0.438	0.904	**0.671**	467	0.724
2	FG+MCS	69	31	269	62	0.527	0.897	**0.712**	431	0.668
3	FG+MCS	51	27	134	39	0.567	0.832	**0.699**	251	0.389

[1] 58 out of 691 SMILES in the dataset were unread from the RAX workflow and were not considered for statistical calculation.

Tab. S4: Prediction statistics of the RAX integrated approach applied to the DILIRank three-class classification

For each combination of similarity lists (i.e., lists including a single analogue) and pre-filtering method, the number of true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), the sensitivity (SEN), the specificity (SPE) are reported for each classification category. The overall number and the percentage (%) of predictions returned are also indicated. SEN_{avg} and SPE_{avg} are the average of values computed separately for each class, while BA_{avg} is the arithmetic mean of the average SEN and SPE. The first row of the table refers to the benchmark performance related to the sole use of the closest structural neighbour to infer the prediction.

Group	Filters	No-DILI						Less-DILI						Most-DILI						SEN_{avg}	SPE_{avg}	BA_{avg}	#preds[1]	%
		TP	FP	TN	FN	SEN	SPE	TP	FP	TN	FN	SEN	SPE	TP	FP	TN	FN	SEN	SPE					
-	-	*134*	*94*	*318*	*86*	*0.609*	*0.772*	*70*	*77*	*383*	*102*	*0.407*	*0.833*	*123*	*134*	*258*	*117*	*0.513*	*0.658*	*0.510*	*0.754*	***0.632***	*633*	*1.000*
1	none	125	90	322	96	0.566	0.782	157	158	235	83	0.654	0.598	60	43	418	112	0.349	0.907	0.523	0.762	**0.642**	633	1.000
2	none	111	67	282	62	0.642	0.808	129	111	205	77	0.626	0.649	53	51	328	90	0.371	0.865	0.546	0.774	**0.660**	522	0.825
3	none	22	8	61	8	0.733	0.884	30	21	31	17	0.638	0.596	5	13	64	17	0.227	0.831	0.533	0.770	**0.652**	99	0.156
1	FG+MCS	103	65	193	50	0.673	0.748	84	70	180	77	0.522	0.720	50	39	275	47	0.515	0.876	0.570	0.781	**0.676**	411	0.649
2	FG+MCS	88	57	188	46	0.657	0.767	83	70	156	70	0.542	0.690	45	36	251	47	0.489	0.875	0.563	0.777	**0.670**	379	0.599
3	FG+MCS	49	34	118	34	0.590	0.776	52	49	94	40	0.565	0.657	27	24	151	33	0.450	0.863	0.535	0.766	**0.650**	235	0.371

[1]18 out of 663 SMILES in the dataset were unread from the RAX workflow and were not considered for statistical calculation.

Tab. S5: Prediction statistics of the RAX integrated approach for specific chemical categories with the application of MCS and FG filters

The total number of compounds (#) and those included in each activity class, the sensitivity (SEN), the specificity (SPE) and the balanced accuracy (BA) are reported for each chemical category and for each dataset. SEN_{avg} and SPE_{avg} are the average of values computed separately for each class of the DILIRank dataset, while BA_{avg} is the arithmetic mean of the average SEN and SPE. Chemical categories were defined based on SMARTS described in the RDKit Functional Group Filter KNIME node (see Table S1). Only chemical categories including at least 10 compounds were analyzed for each dataset.

Chemical category	DILIRank							ToxRef					
	#	Most_DILI	Less_DILI	No_DILI	SEN_{avg}	SPE_{avg}	BA_{avg}	#	HT	nHT	SEN	SPE	BA
Aliphatic Alcohol	114	20	33	61	0.58	0.79	0.68	46	9	37	0.67	0.97	0.82
Aliphatic Amine	87	14	40	33	0.42	0.72	0.57	-	-	-	-	-	-
Aliphatic Carboxylic Acid	78	17	39	22	0.73	0.87	0.80	29	4	25	0.75	1.00	0.88
Aliphatic Halogen	-	-	-	-	-	-	-	15	12	3	0.75	0.67	0.71
Aromatic Alcohol	55	11	18	26	0.57	0.80	0.68	66	10	56	0.10	0.93	0.51
Aromatic Amine	37	13	18	6	0.60	0.78	0.69	40	13	27	0.23	0.93	0.58
Aromatic Carboxylic Acid	14	5	7	2	0.58	0.68	0.63	13	0	13	N/A	0.92	0.92
Aromatic Halogen	70	23	29	18	0.46	0.72	0.59	74	45	29	0.60	0.52	0.56

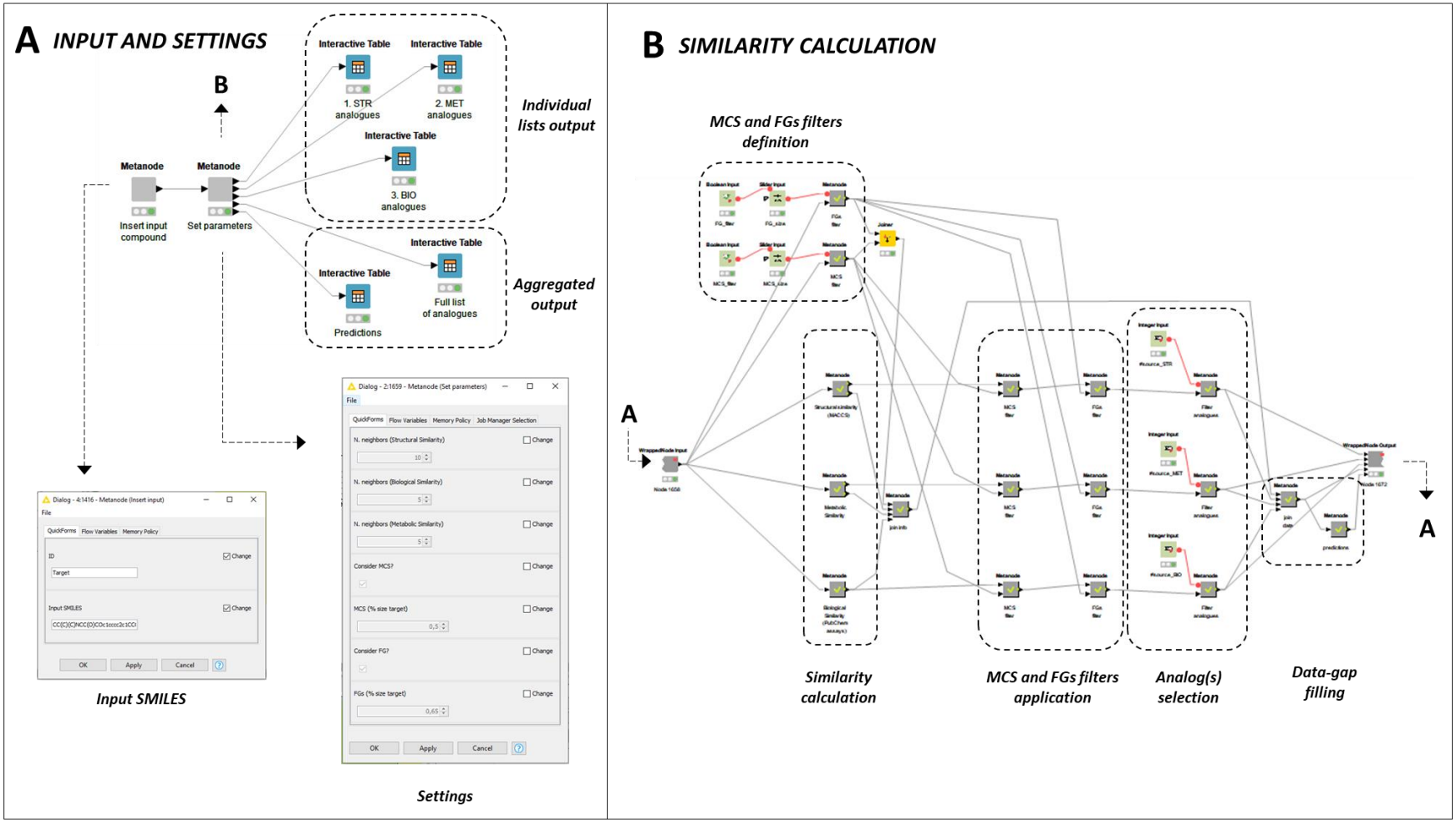

Fig. S1: KNIME implementation of the RAX workflow
The left-hand panel shows the input window where user can insert target SMILES and modify the settings (e.g., number of analogue(s) selected for each similarity, application of maximum-common-substructure and functional groups based filters), whereas the right-hand panel shows the nodes responsible of similarities calculations and data-gap-filling.

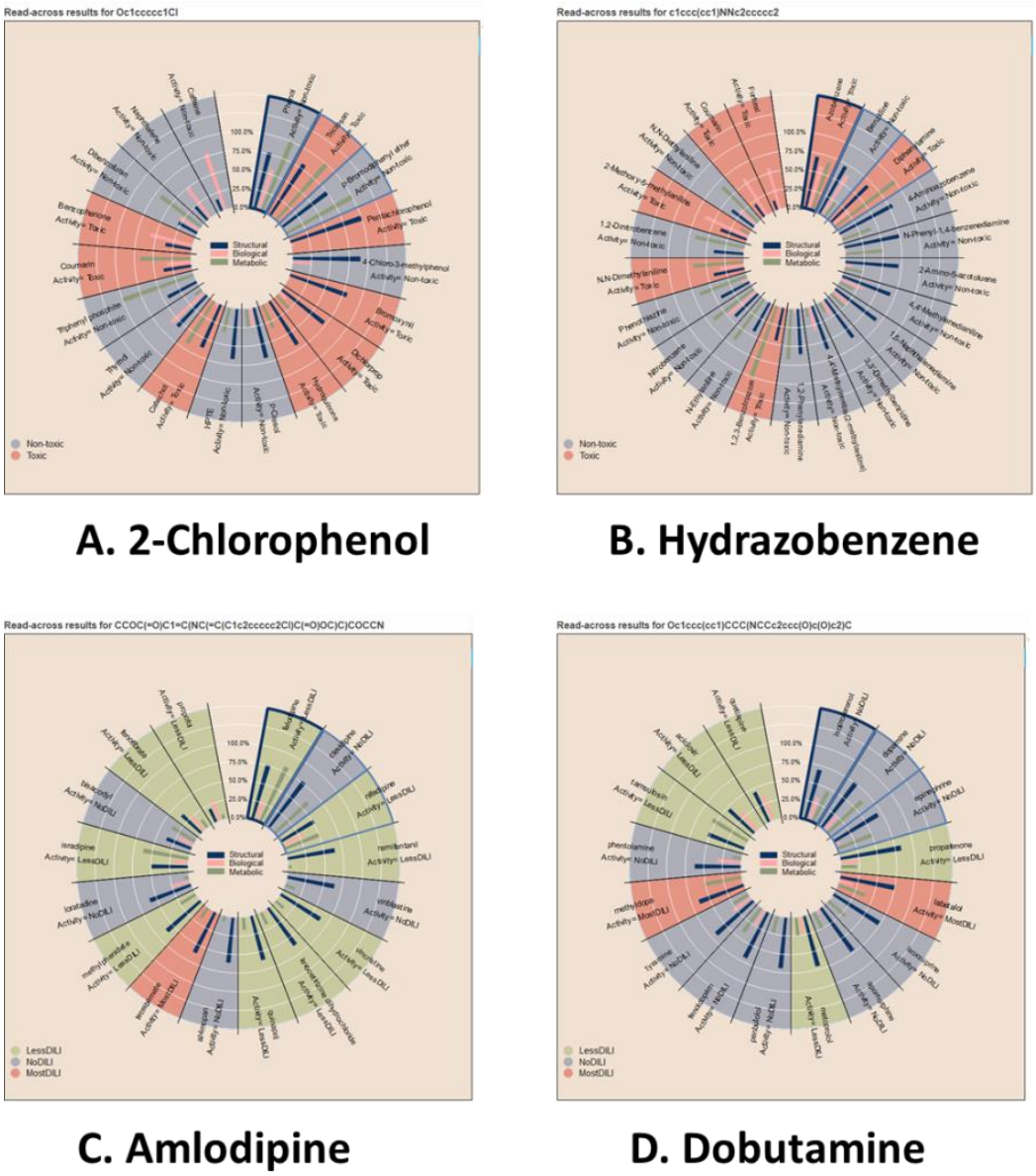

Figure S2. Example of graphical output of the RAX workflow. Graphical output for A. 2-Chlorphenol, B. Hydrazobenzene, C. Amlodipine and D. Dobutamine are showed as a pie chart. Each slice of the graph represents a different analogue, with the background color that is indicative of the analog's activity and histogram bars that give estimation of the four similarities with respect of the target. Outlined slices indicate analogues included in multiple similarity lists (i.e., dark blue for three lists and light blue for two lists).