CERAPP - Collaborative Estrogen Receptor Activity Prediction Project

Consensus modeling

Background and Goals

- U.S. Congress mandated that the EPA screen chemicals for their potential to be endocrine disruptors
- Led to development of the Endocrine Disruptor Screening Program (EDSP)
- Initial focus was on environmental estrogens, but program expanded to include androgens and thyroid pathway disruptors

EDSP Chemicals

- EDSP Legislation contained in:
  - FIFRA: Federal Insecticide, Fungicide, Rodenticide Act
  - SDWA: Safe Drinking Water Act
- Chemicals:
  - All pesticide ingredients (actives and inerts)
  - Chemicals likely to be found in drinking water to which a significant population can be exposed
- Total EDSP Chemical universe is ~10,000
- Subsequent filters brings this to about 5,000 to be tested

The Problem with EDSP

- EDSP Consists of Tier 1 and Tier 2 tests
- Tier 1 is a battery of 11 in vitro and in vivo assays
- Cost ~$1,000,000 per chemical
- Throughput is ~50 chemicals / year
- Total cost of Tier 1 is billions of dollars and will take 100 years at the current rate
- Need pre-tier 1 filter
- Use combination of structure modeling tools and high-throughput screening “EDSP21”

CERAPP Goals

- Use structure-based models to predict ER activity for all of EDSP Universe and aid in prioritization for EDSP Tier 1

- Because models are relatively easy to run on large numbers of chemicals, extend to all chemicals with likely human exposure

- Chemicals with significant evidence of ER activity can be queued further testing

Thinking about need for accuracy

- Goal is prioritizing chemicals for further testing
  - Sensitivity more important than specificity
  - Better to leave in “funny” structures than to discard
  - OK predictions today are better than perfect predictions tomorrow
- There will be errors in:
  - Chemical structures
  - Chemical identities
  - Model predictions
  - Experimental data
- Structure library can improve / expand going forward
  - Will be used for other prediction projects

Participants:

- **DTU/food:** Technical University of Denmark/ National Food Institute
- **EPA/NCCT:** U.S. Environmental Protection Agency / National Center for Computational Toxicology
- **FDA/NCTR/DBB:** U.S. Food and Drug Administration/ National Center for Toxicological Research/Division of Bioinformatics and Biostatistics
- **FDA/NCTR/DSB:** U.S. Food and Drug Administration/ National Center for Toxicological Research/Division of Systems Biology
- **Helmholtz/ISB:** Helmholtz Zentrum Muenchen/Institute of Structural Biology
- **ILS&EPA/NCCT:** ILS Inc & EPA/NCCT
- **IRCSS:** Istituto di Ricerche Farmacologiche “Mario Negri”
- **JRC_Ispra:** Joint Research Centre of the European Commission, Ispra.
- **LockheedMartin&EPA:** Lockheed Martin IS&GS/ High Performance Computing
- **NIH/NCATS:** National Institutes of Health/ National Center for Advancing Translational Sciences
- **NIH/NCI:** National Institutes of Health/ National Cancer Institute
- **RIFM:** Research Institute for Fragrance Materials, Inc
- **UMEA/Chemistry:** University of UMEA/ Chemistry department
- **UNC/MML:** University of North Carolina/ Laboratory for Molecular Modeling
- **UniBA/Pharma:** University of Bari/ Department of Pharmacy
- **UNIMIB/Michem:** University of Milano-Bicocca/ Milano Chemometrics and QSAR Research Group
- **UNISTRA/Infochim:** University of Strasbourg/ ChemoInformatique

Plan of the project

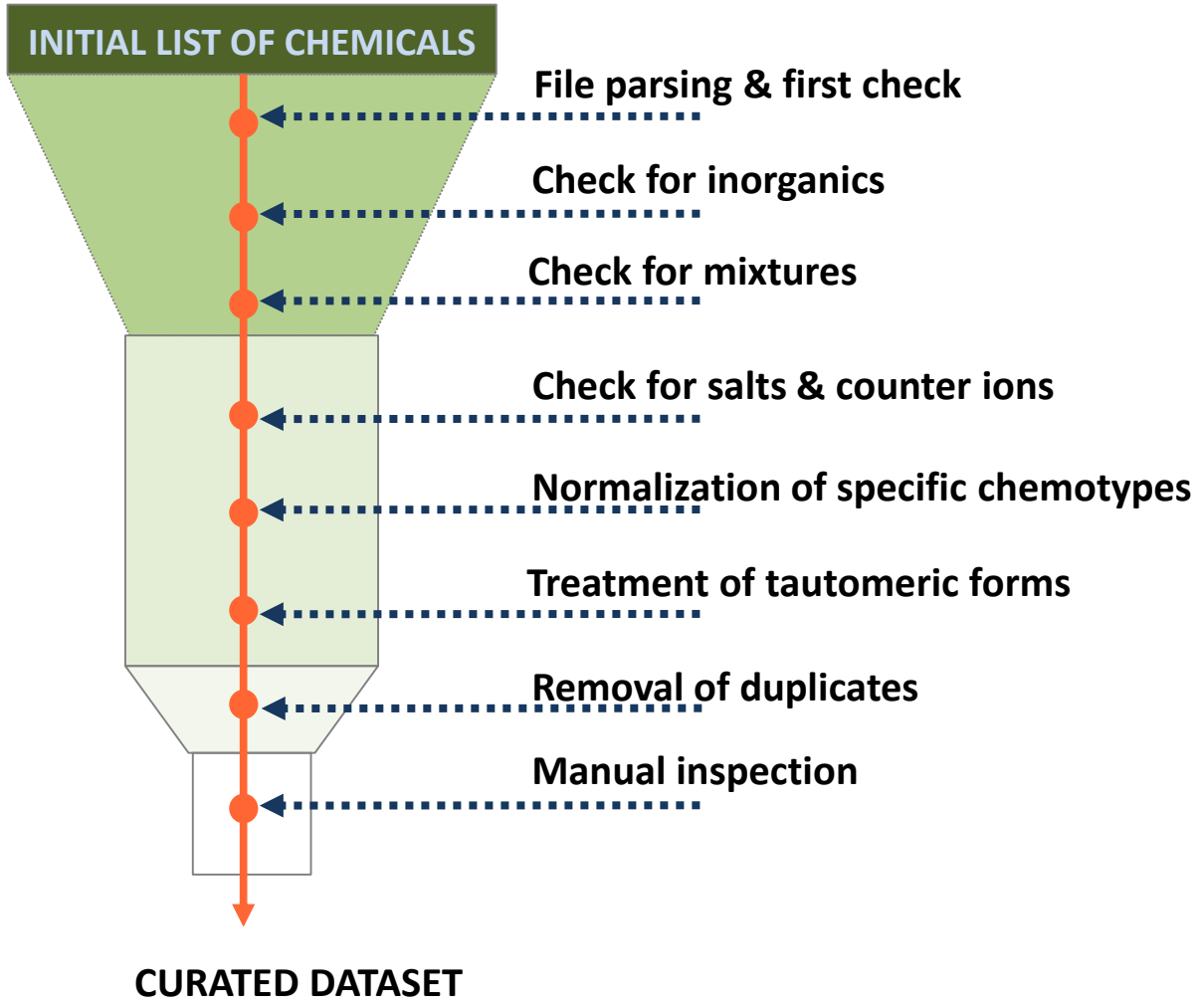
1: Structures curation	- Collect chemical structures from different sources - Design and document a workflow for structure cleaning - Deliver the QSAR-ready training set and prediction set
2: Experimental data preparation	- Collect and clean experimental data for the evaluation set - Define a strategy to evaluate the models separately
3: Modeling & predictions	- Train/refine the models based on the training set - Deliver predictions and applicability domains for evaluation
4: Model evaluation	- Analyze the training and evaluation datasets - Evaluate the predictions of each model separately
5: Consensus strategy	- Define a score for each model based on the evaluation step - Define a weighting scheme from the scores
6: Consensus modeling & validation	- Combine the predictions based on the weighting scheme - Validate the consensus model using an external dataset.

Chemical structures curation (standardization)

Subgroup:

- U.S. EPA-NCCT
- University of North Carolina
- Danish Technical University-DTU Food

Scheme of the curation workflow UNC, DTU, EPA Consensus

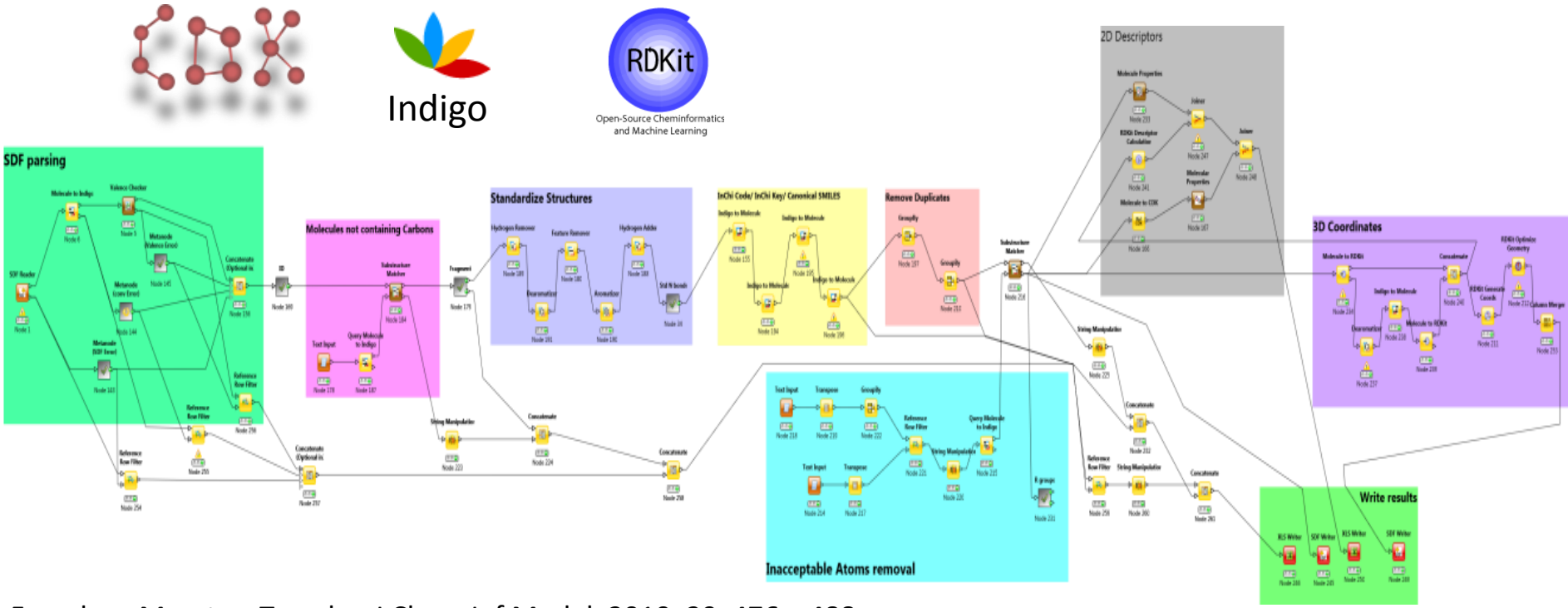


Fourches, Muratov, Tropsha. J Chem Inf Model, 2010, 29, 476 – 488

KNIME workflow

Aim of the workflow:

- Combine (<u>not reproduce</u>) different procedures and ideas
- Minimize the differences between the structures used for prediction by different groups
- Produce a flexible free and open source workflow to be shared



Fourches, Muratov, Tropsha. J Chem Inf Model, 2010, 29, 476 – 488

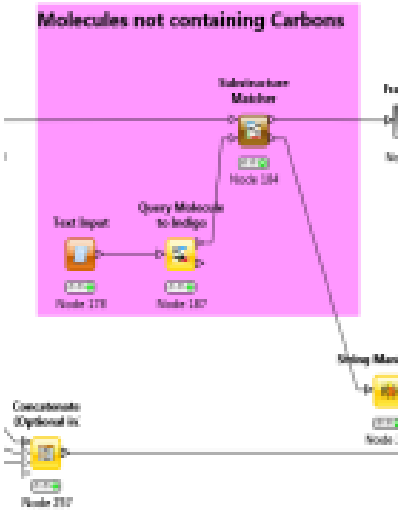
Wedebye, Niemelä, Nikolov, Dybdahl, Danish EPA Environmental Project No. 1503, 2013

Parsing and 1st filter

SDF Parser: 40125 initial compounds

(Webservices: Pubchem, Chemspider)

40117 parsed compounds
Unique IDs

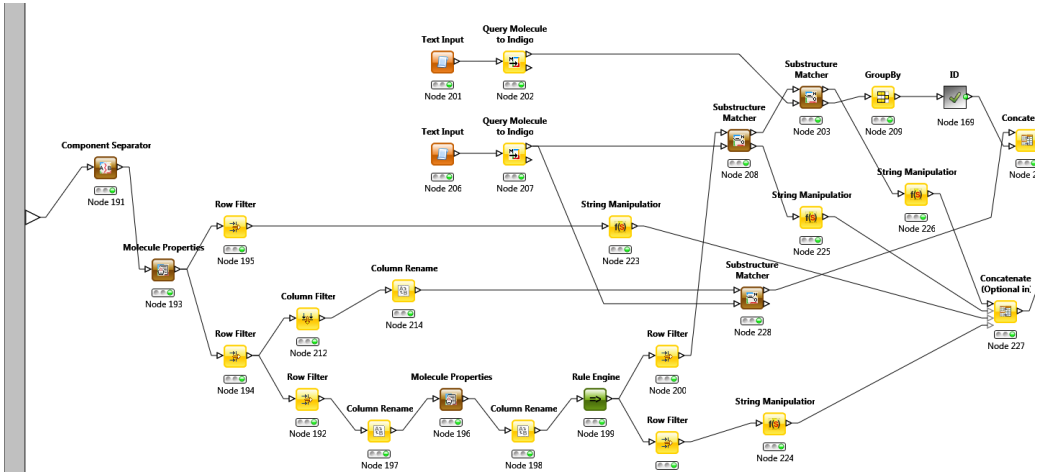


Errors reported

Unconnected structures

1. Separate unconnected fragments
2. MW filter on biggest Cpd

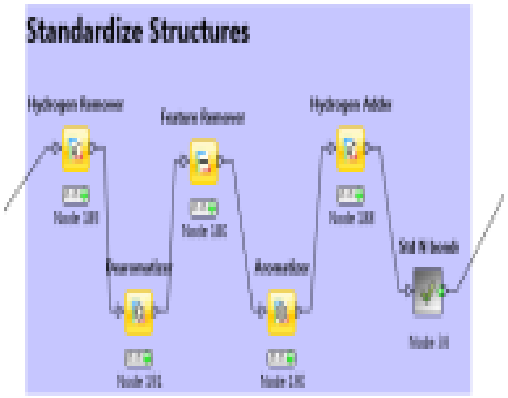
(497 compounds removed)



1. 2^{nd} biggest is removed if:

- It was the same/stereo as the biggest component
- Not containing carbons
- It was a salt/solvent from the defined list of accepted salts and solvents.

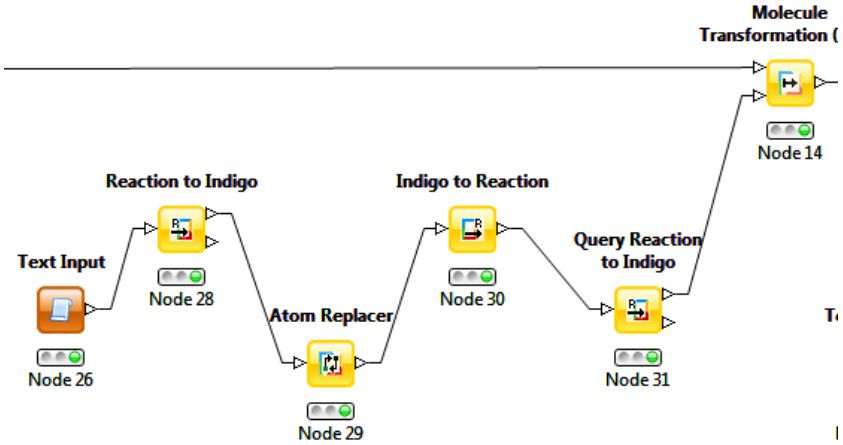
Standardization of structures

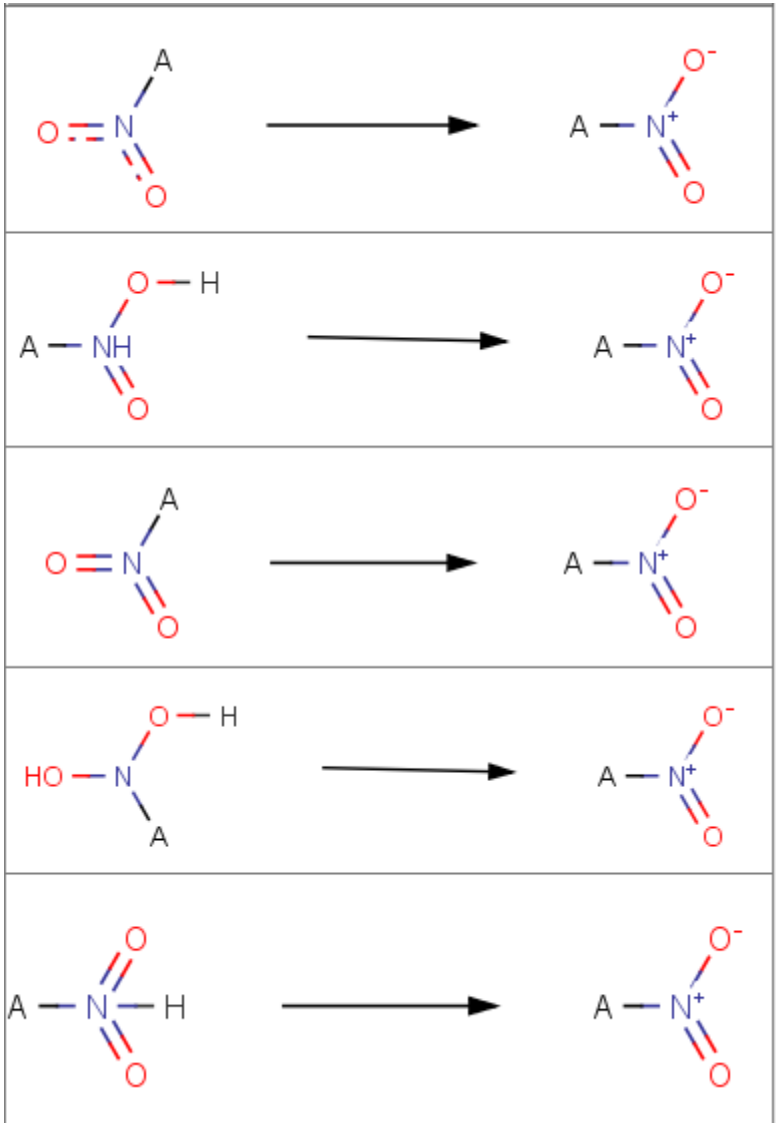


- Explicit hydrogen removed
- Dearomtization
- Removal of chirality/stereochemistry info, isotopes and pseudo-atoms
- Aromatization + add explicit hydrogen atoms
- Standardize Nitro groups
- Other: tautomerize/mesomerize
- Neutralize (when possible)

Standardize Nitro mesomers

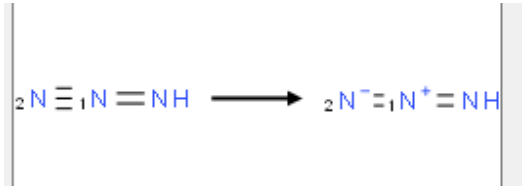
SMARTS query to reaction

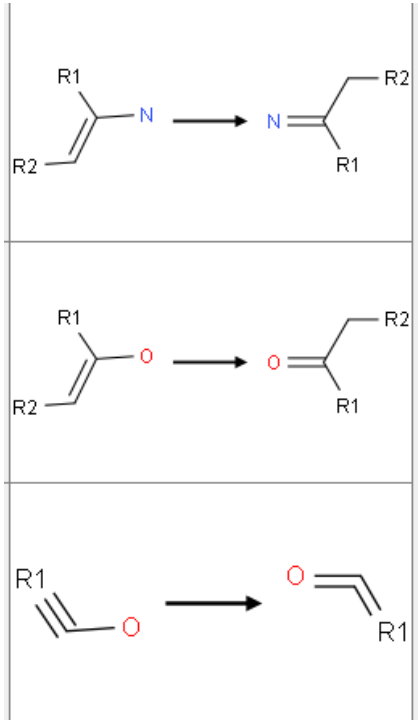


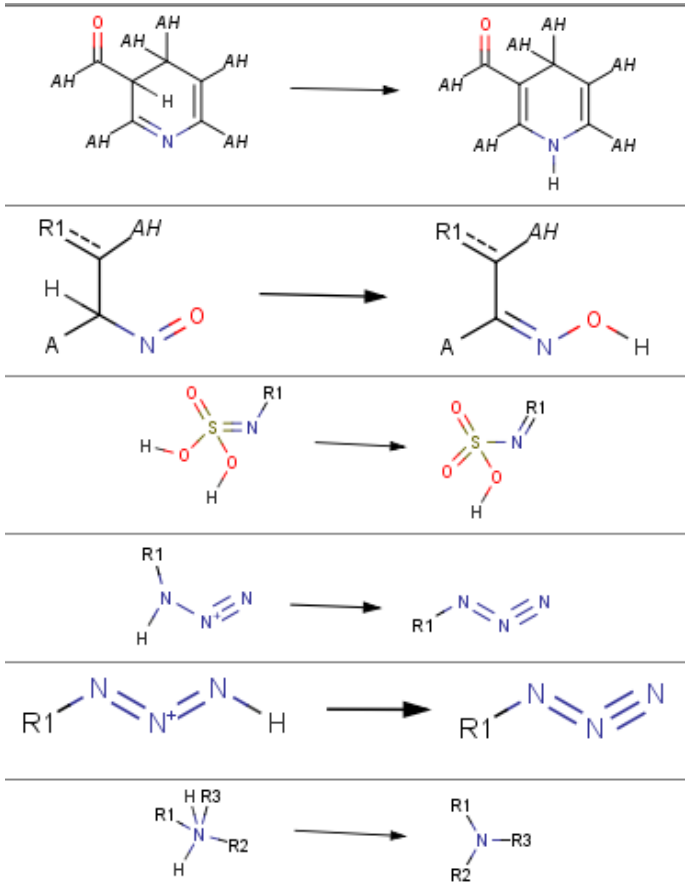


Mesomerization/tautomerization

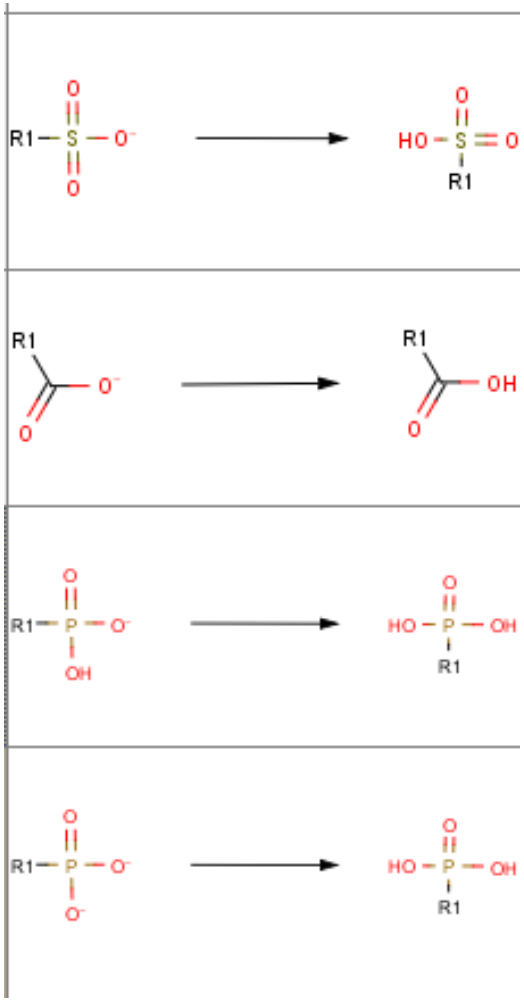
- Azide mesomers
- Exo-enol tautomers
- Enamine-Imine tautomers
- Ynol-ketene tautomers
- ….

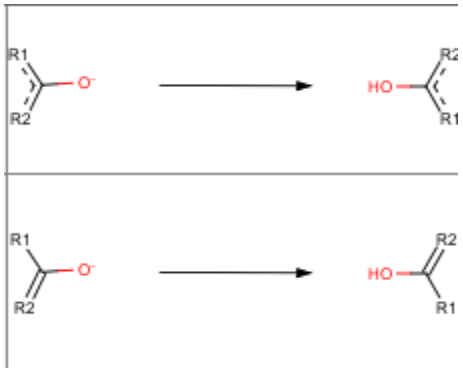


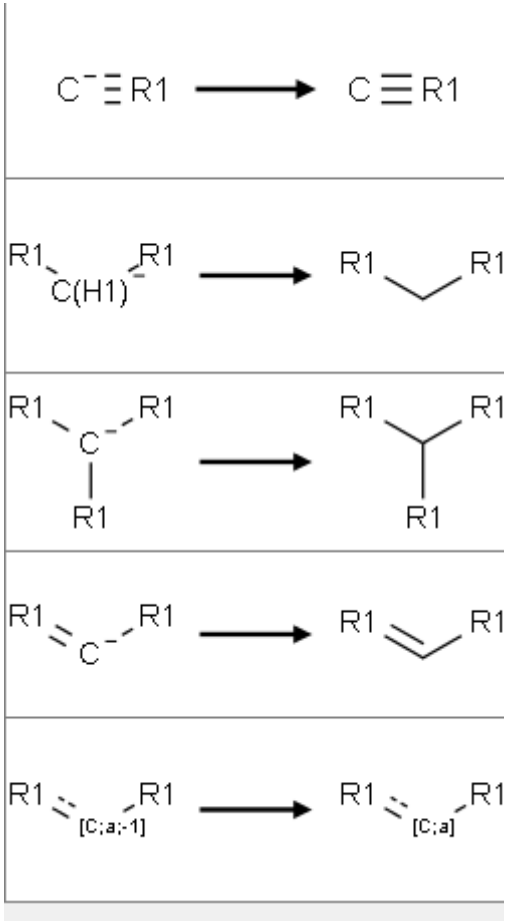




Neutralize Structures







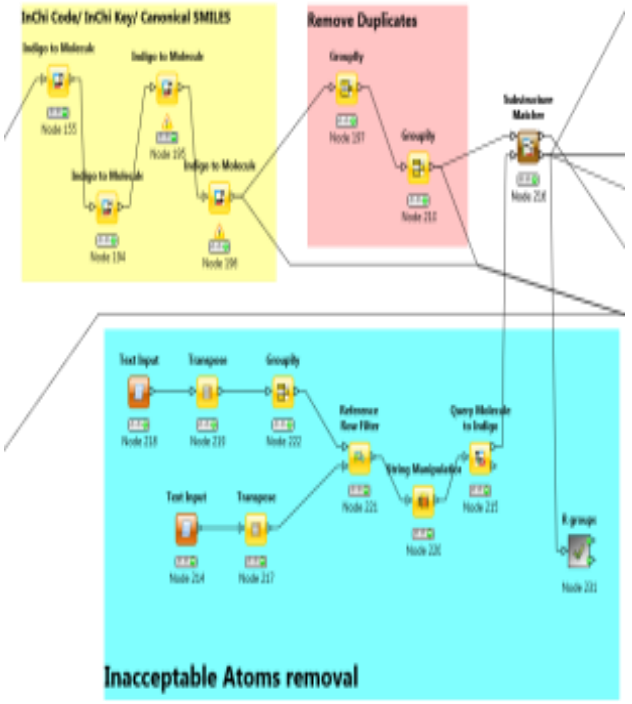
Filter inacceptable atoms

- Generate InChi, InChi Key and Canonical Smiles.

- Remove duplicates (InChis & canonical SMI)

- Remove molecules with inacceptable atoms. Other then:

  H, C, N, O, P, S, Se, F, Cl, Br, I, Li, Na, K, B, Si

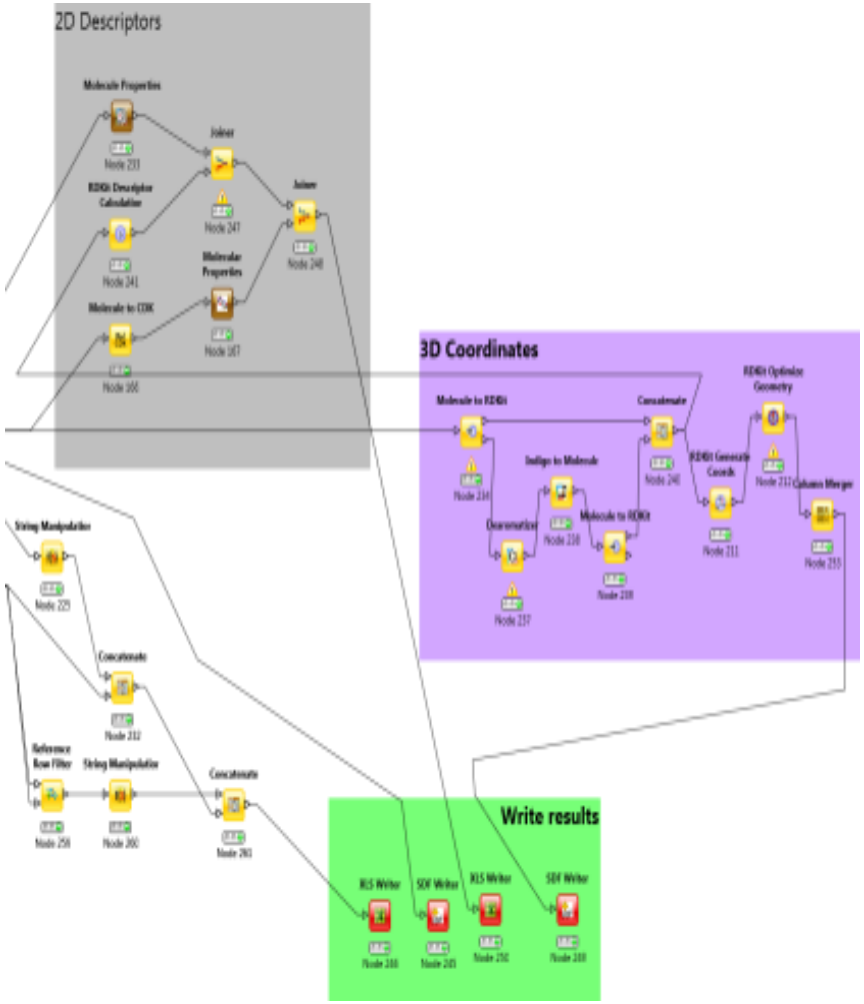


Write results

- Calculate 2D descriptors (Indigo, CDK, RDKit)
- Generate 3D conformers
- Optimize geometry (MMFF94S)

<u>Generated files:</u>

- Sdf file containing the 2D structures
- Excel file containing 2D descriptors
- Sdf file containing the 3D structures
- Excel file for error messages



Chemicals for Prediction:
The Human Exposure Universe

- EDSP Universe (10K)
- Chemicals with known use (40K) (CPCat & ACToR)
- Canadian Domestic Substances List (DSL) (23K)
- EPA DSSTox – structures of EPA/FDA interest (15K)
- ToxCast and Tox21 (In vitro ER data) (8K)

⟹ **~55k to ~32K unique set of structures**

- Training set (ToxCast): 1677 Chemicals
- Prediction Set: 32464 Chemicals

Experimental data for evaluation

Subgroup:

- U.S.EPA/NCCT: Kamel Mansouri, Jayaram Kancherla, Ann Richard, Richard Judson
- UMEA/Chem: Aleksandra Rybacka, Patrik Andersson
- FDA/NCTR/DBB: Huixiao Hong
- NIH/NCATS: Ruili Huang
- Helmholtz/ISB: Igor Tetko

Tasks to fulfill

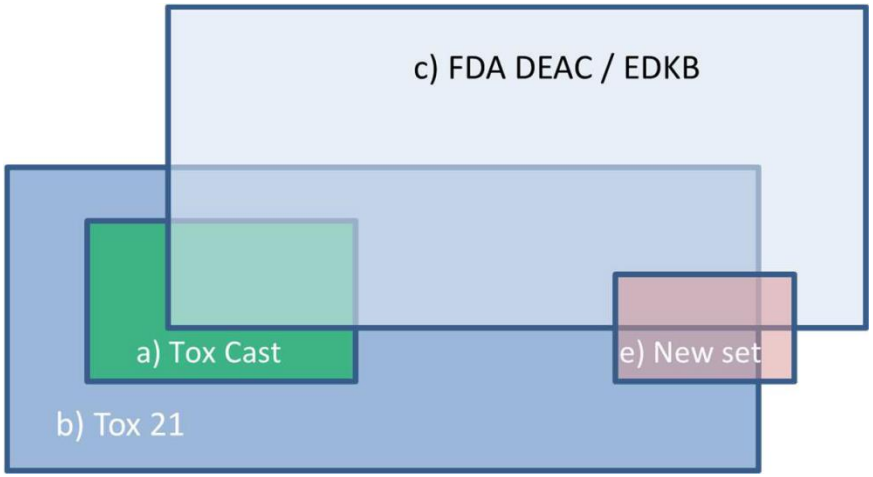
- Collect the experimental data for the evaluation step.
- Combine the different sources of literature.
- Define a strategy to evaluate the models separately.

Experimental data for evaluation set

EPA/NCCT, UMEA/Chem, FDA/NCTR/DBB, NIH/NCATS, Helmholtz/ISB

a) Tox21, ~8000 chemicals in 4 assays;
b) FDA EDKB database of ~8000 chemicals from the literature;
c) METI database, ~2000 chemicals;
d) ChEMBL database, ~2000 chemicals.

60,000 entries for ~15,000 chemicals



Cleaning procedure

- Knime workflow for structure cleaning
- INChi code for chemical matching
- **7,600 chemicals with CERAPP IDs**
- Remove: *in-vivo*, cytotoxicity, ambiguous, missing values, non-defined endpoints/units
- Categorize assays: binding, reporter gene or cell proliferation
- Normalize units
- Use of reference chemicals to categorize into 5 classes.

7547 CERAPP compounds from 44641 entries

Categorize chemicals

- Merge entries with AC50, PC50, IC50, GI50 and EC50.
- Use of 36 reference categorized chemicals
- 5 classes created:
  - Strong : 0-0.09 => score =1
  - Moderate: 0.09-0.18 => score = 0.25
  - Weak: 0.18-20 => score = 0.5
  - Very Weak: 20-800 => score = 0.75
  - Inactive: 800> => score = 0

Evaluation set

Evaluation set for binary classification models

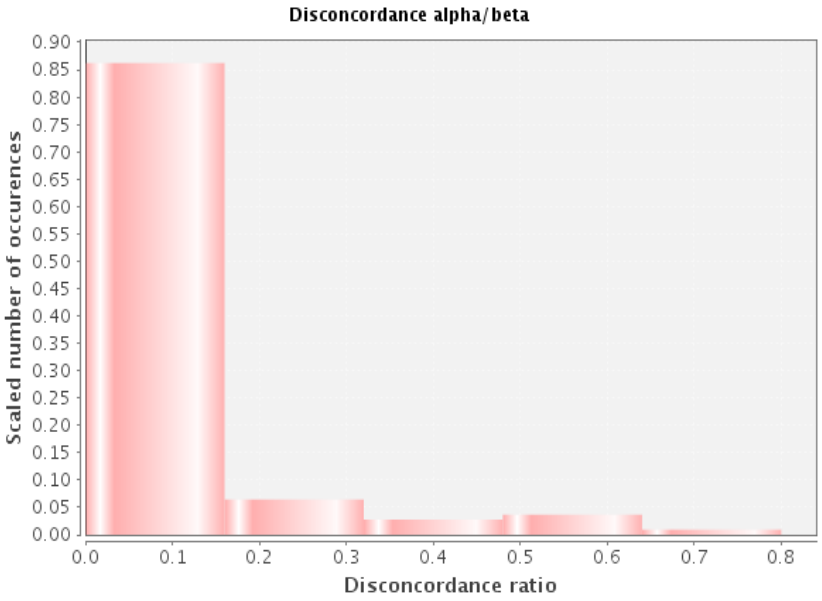
	Active	Inactive	Total
Binding	1982	5301	7283
Agonist	350	5969	6319
Antagonist	284	6255	6539
Total	2616	17525	20141

Evaluation set for quantitative models

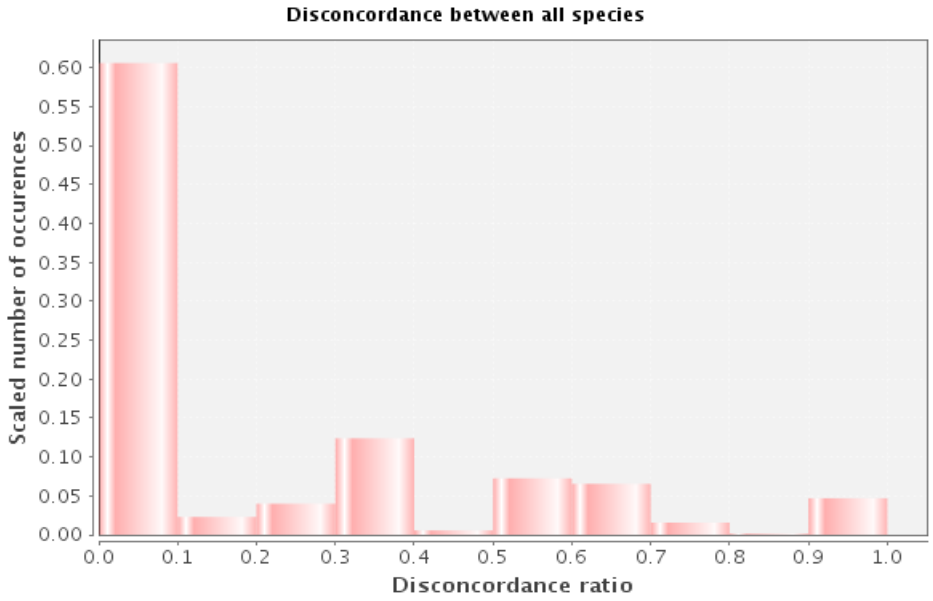
	Inactive	V. Weak	Weak	Moderate	Strong	Total
Binding	5042	685	894	72	77	6770
Agonist	5892	19	179	31	42	6163
Antagonist	6221	76	188	10	10	6505
Total	17155	780	1261	113	129	19438

Consistency of the data

Consistency alpha/beta



Consistency between species



Consistency training set/ evaluation set

	data	N. Chem	B. Acc	Sn	Sp
All orig.	all	1659	78.57	89.29	67.86
No VW	all	1424	84.57	88.16	80.99
All orig.	Multi Src	1410	84.42	88.46	80.38
No VW	Multi Src	1306	87.79	87.67	87.9

Evaluation & consensus

(consensus subgroup: most of participants)

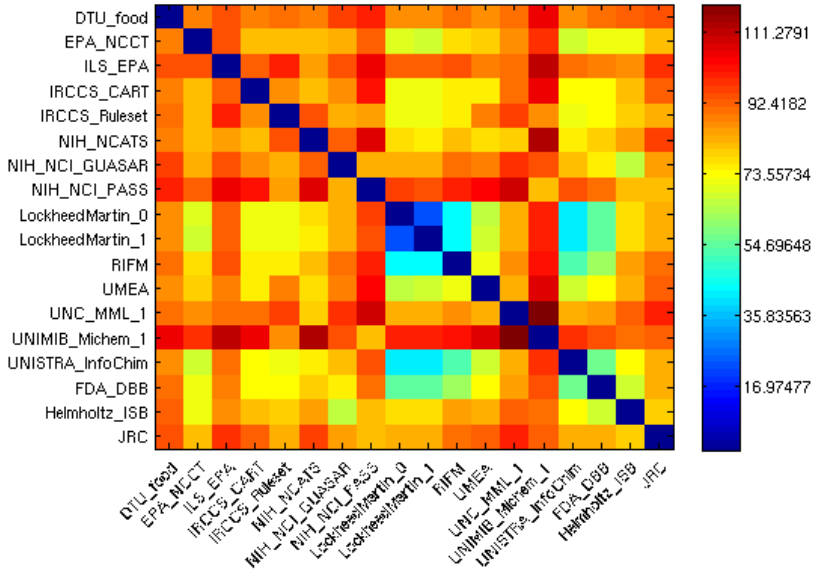
Models received:

- Classification / Qualitative:
  - Binding: **22 models**
  - Agonists: **11 models**
  - Antagonists: **9 models**
- Regression / Quantitative:
  - Binding: **3 models**
  - Agonists: **3 models**
  - Antagonists: **2 models**

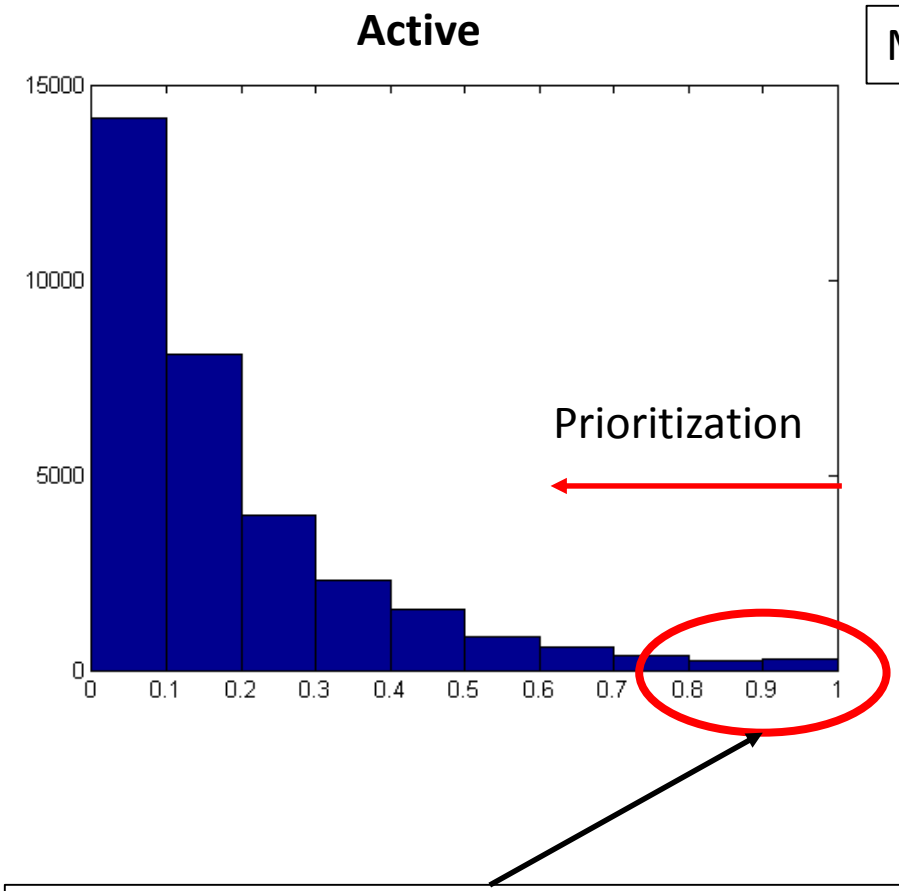
Preliminary Results

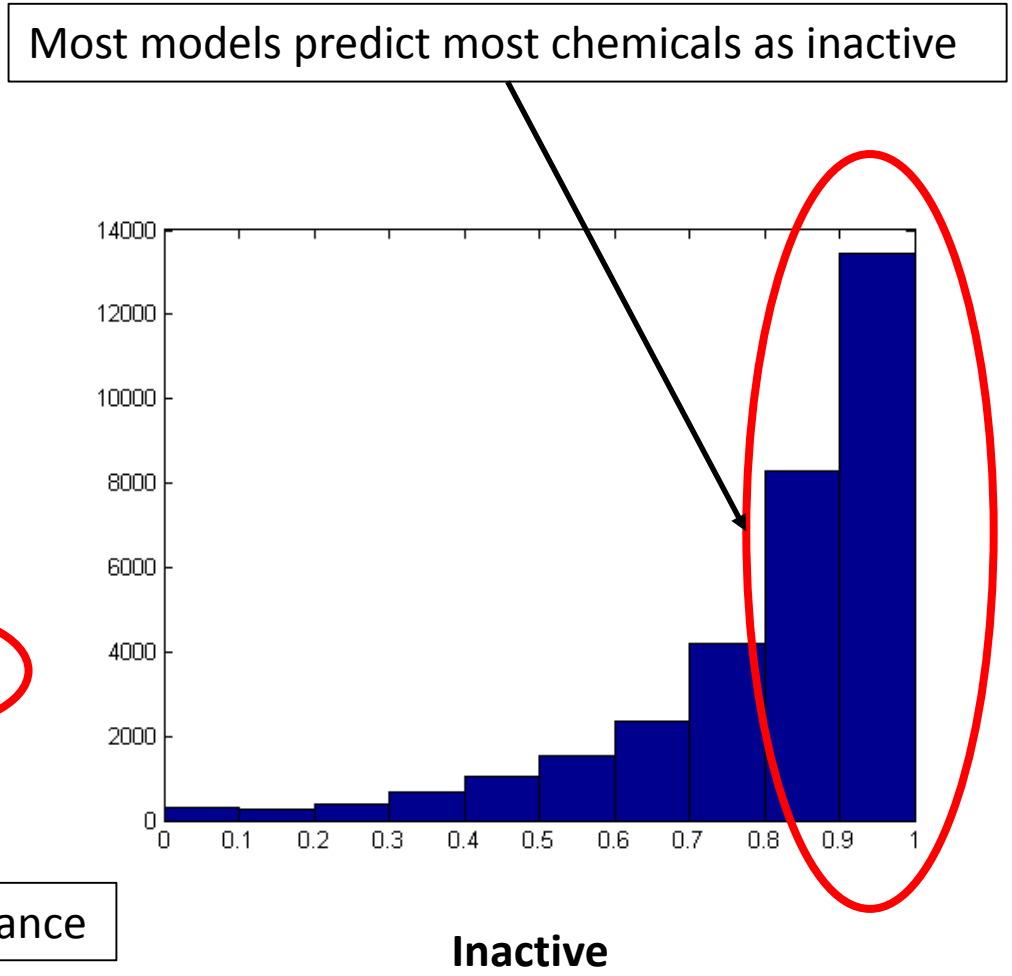
18 binding models with most chemicals predicted

Euclidean distance



Concordance of the 22 classification models for binding





Evaluation procedure:

- On the EPA training set (1677)
- On the full evaluation set (~7k)
- Evaluation set with multi-sources
- Remove “VeryWeak”
- Remove single source
- Remove chemicals outside the AD

Score functions & weights for consensus predictions

$$g\_score = {}^{1}/_{3}\left(\frac{NER_{ToxCast} * N\_pred_{ToxCast}}{N\_tot_{ToxCast}} + \frac{N\_pred}{N\_tot} + {}^{1}/_{N_{filter}} \sum_{i=1}^{N_{filter}} \frac{NER_i * N\_pred_i}{N\_tot_i}\right)$$

$$opt\_score = {}^{1}/_{2}\left(NER_{ToxCast} + NER_{all\_filters}\right)$$

Evaluation of binding models

Models	Training set	B. Ac. Training	Evaluation set	B. Ac. Eval	Unambiguous	Accu Unambig	All predicted	g_score	opt_score
DTU_1	873	0.82	3840	0.64	2695	0.78	16063	0.43	0.80
DTU_2	737	0.79	3268	0.61	2383	0.71	13442	0.36	0.75
EPA_NCCT	1529	0.87	7283	0.57	5275	0.69	32463	0.82	0.78
FDA_NCTR_DBB	1529	0.99	7283	0.60	5991	0.68	32464	0.87	0.84
FDA_NCTR_DSB	0	0.00	534	0.53	431	0.53	2008	0.03	0.53
Helmholtz_ISB	1512	0.89	7123	0.62	5860	0.72	31629	0.83	0.80
ILS_EPA	1506	0.84	7068	0.66	5814	0.75	31318	0.82	0.79
IRCCS_CART	1529	0.80	7280	0.61	3620	0.75	32442	0.78	0.77
IRCCS_Ruleset	1383	0.91	6603	0.56	5416	0.62	28958	0.75	0.77
JRC_Ispra	1465	0.82	6900	0.58	5672	0.67	30801	0.77	0.74
LockheedMartin_EPA_1	1529	0.83	7283	0.55	1539	0.66	32464	0.75	0.75
LockheedMartin_EPA_2	1529	0.76	7283	0.54	1539	0.64	32464	0.72	0.70
NIH_NCATS	1528	0.69	7271	0.59	5981	0.65	32184	0.77	0.67
NIH_NCI_GUASAR	1529	0.99	7283	0.61	5951	0.69	32455	0.88	0.84
NIH_NCI_PASS	1465	0.86	6900	0.58	5672	0.66	30800	0.78	0.76
RIFM	1529	0.73	7283	0.58	5991	0.65	32463	0.78	0.69
UMEA	1529	0.82	7280	0.61	5989	0.70	32430	0.82	0.76
UNC_MML_1	1529	0.80	7283	0.59	5991	0.65	32464	0.80	0.73
UNC_MML_2	1529	0.49	7283	0.55	5991	0.60	32464	0.69	0.55
UNIBA	750	0.86	3259	0.62	2753	0.73	15178	0.40	0.80
UNIMIB_Michem_1	1529	0.76	7283	0.55	5991	0.59	32464	0.77	0.68
UNIMIB_Michem_2	531	0.98	2780	0.62	2241	0.71	11832	0.32	0.85
UNISTRA_InfoChim	1529	0.86	7283	0.57	4755	0.60	32464	0.80	0.73

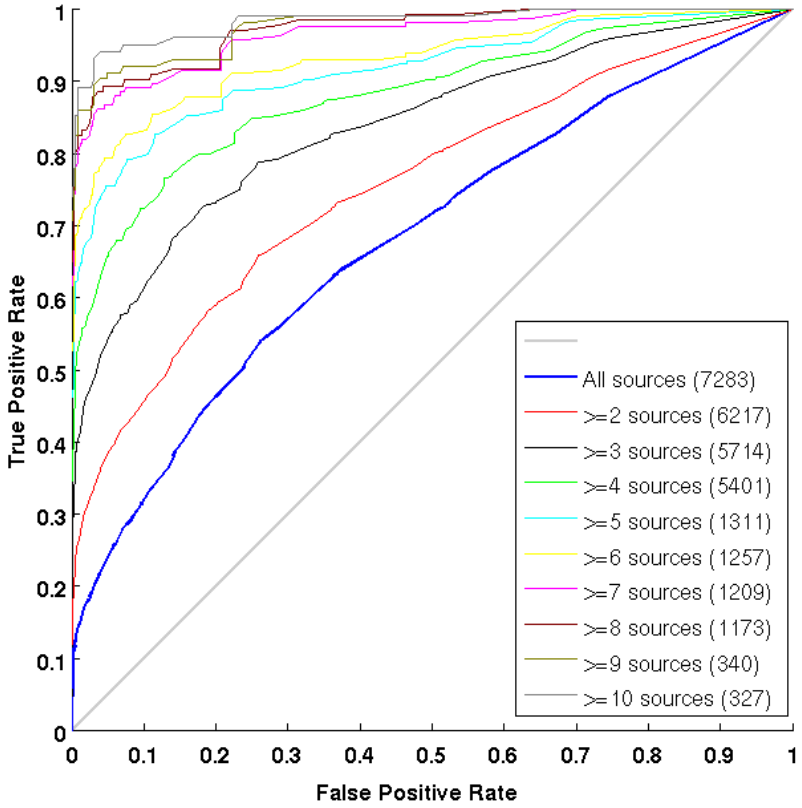
Consensus_1 predictions

	Binding								Agonist				Antagonist			
CERAPP ID	n_act	score	n_no	score	cons	act_conc	inact_c	Potency	cons	act_c	inact_c	Potency	cons	act_c	inact_c	Potency
10001	1	0.05	15	0.71	0	0.06	0.94	Inactive	0	0.00	1.00	Inactive	0	0.00	1.00	Inactive
10005	3	0.11	17	0.65	0	0.15	0.85	Inactive	0	0.11	0.89	Inactive	0	0.14	0.86	Inactive
10007	4	0.18	12	0.58	0	0.25	0.75	Inactive	0	0.00	1.00	Inactive	0	0.00	1.00	Inactive
10008	0	0.00	18	0.76	0	0.00	1.00	Inactive	0	0.00	1.00	Inactive	0	0.00	1.00	Inactive
10009	1	0.04	17	0.71	0	0.06	0.94	Inactive	0	0.00	1.00	Inactive	0	0.00	1.00	Inactive
10016	21	0.76	0	0.00	1	1.00	0.00	Strong	1	1.00	0.00	Strong	1	1.00	0.00	Inactive
10017	21	0.76	0	0.00	1	1.00	0.00	Strong	1	1.00	0.00	Strong	1	1.00	0.00	Inactive
10018	16	0.61	4	0.15	1	0.80	0.20	VeryWeak	1	0.89	0.11	VeryWeak	1	0.86	0.14	Inactive
10027	19	0.72	1	0.04	1	0.95	0.05	Moderate	0	0.10	0.90	Inactive	0	0.13	0.88	Moderate
10033	4	0.17	13	0.58	0	0.24	0.76	Inactive	0	0.00	1.00	Inactive	0	0.00	1.00	Inactive
10034	21	0.75	0	0.00	1	1.00	0.00	Moderate	1	0.89	0.11	Moderate	1	0.86	0.14	Inactive
10088	11	0.42	9	0.34	1	0.55	0.45	VeryWeak	1	0.78	0.22	VeryWeak	1	0.86	0.14	Inactive
10089	1	0.04	19	0.72	0	0.05	0.95	Inactive	0	0.00	1.00	Inactive	0	0.00	1.00	Inactive
10099	2	0.09	15	0.66	0	0.12	0.88	Inactive	0	0.11	0.89	Inactive	0	0.13	0.88	Inactive
10100	6	0.24	12	0.50	0	0.33	0.67	Inactive	0	0.00	1.00	Inactive	0	0.00	1.00	Inactive
10101	3	0.12	16	0.64	0	0.16	0.84	Inactive	0	0.11	0.89	Inactive	0	0.14	0.86	Inactive
10102	12	0.43	9	0.32	1	0.57	0.43	VeryWeak	1	0.78	0.22	VeryWeak	1	0.71	0.29	Inactive
10111	3	0.12	16	0.64	0	0.16	0.84	Inactive	0	0.00	1.00	Inactive	0	0.00	1.00	Inactive
10112	22	0.75	0	0.00	1	1.00	0.00	Weak	1	1.00	0.00	Weak	1	1.00	0.00	Inactive
10113	21	0.75	0	0.00	1	1.00	0.00	Weak	1	1.00	0.00	Weak	1	1.00	0.00	Inactive
10119	12	0.46	8	0.30	1	0.60	0.40	VeryWeak	1	0.78	0.22	VeryWeak	1	0.71	0.29	Inactive
10120	11	0.39	10	0.36	1	0.52	0.48	VeryWeak	1	0.78	0.22	VeryWeak	1	0.71	0.29	Inactive

Consensus_1 evaluation

Total binders: 2576
Agonists: 2312
Antagonists: 2779

	ToxCast data	Evaluation set
Sensitivity	0.85	0.23
Specificity	0.98	0.95
Balanced accuracy	0.92	0.59



Rules for consensus_2

Agonist and antagonist consensus models first, then on binding consensus:

1) If chemical *i* is active in classification consensus_1
   - active in Potency_class consensus_2
2) If chemical *i* is active in regression & >= 3 positive classification models
   - active in classification consensus_2
3) If chemical *i* is active in regression & < 3 positive classification models
   - Inactive in Potency_class consensus_2

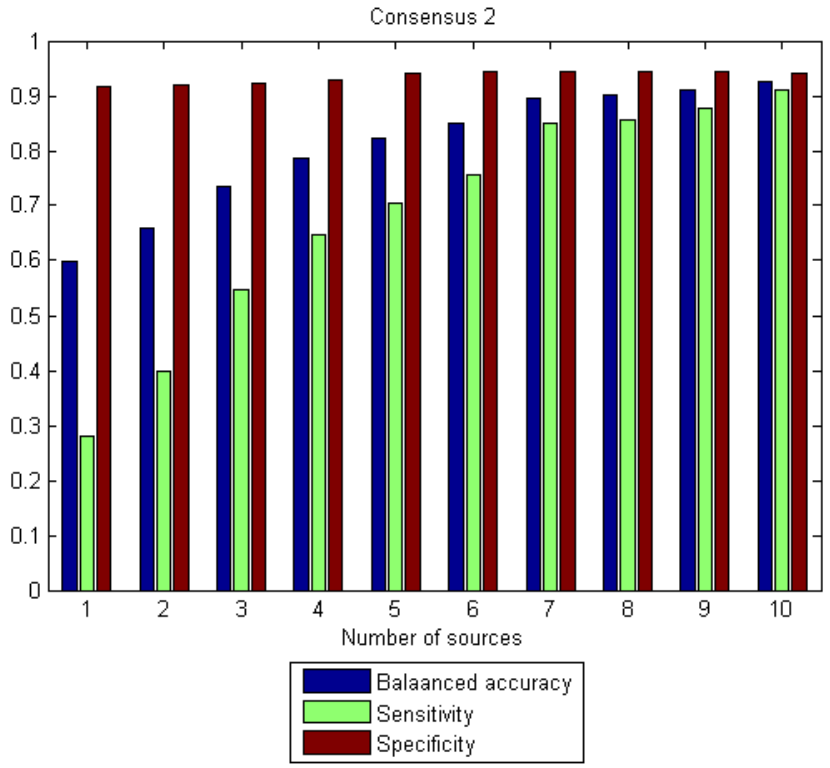
Binding consensus:

4) If chemical *i* is active agonist or active antagonist

- Active in classification consensus_2
- Potency_class consensus_2 = Potency_class agonist/antagonist

Consensus_2 evaluation

Total binders: 3961
Agonists: 2494
Antagonists: 2793

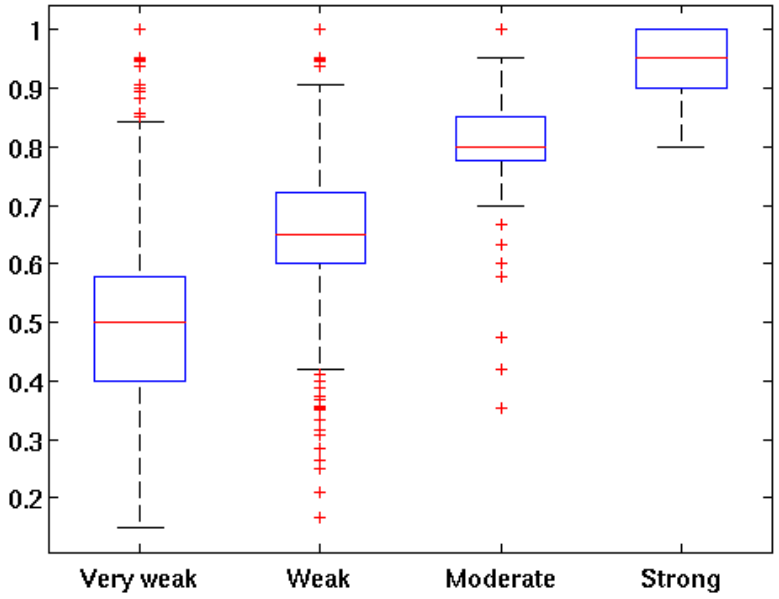


	ToxCast data		Literature data	
Observed\Predicted	**Actives**	**Inactives**	**Actives**	**Inactives**
Actives	83	6	597	1385
Inactives	40	1400	463	4838

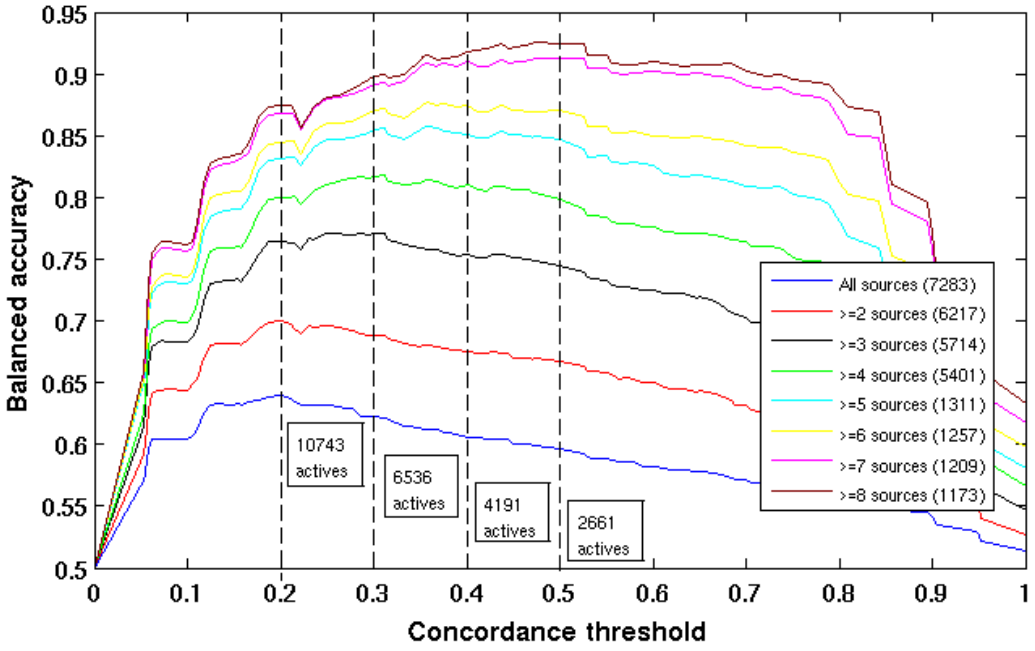
	ToxCast data	Literature data (All: 7283)	Literature data (>6 sources: 1209)
Sensitivity	0.93	0.30	0.87
Specificity	0.97	0.91	0.94
Balanced accuracy	0.95	0.61	0.91

Positive concordance & Potency level

- positive concordance < 0.6 => Potency class= Very weak
- 0.6=<positive concordance<0.75 => Potency class= Weak
- 0.75=<positive concordance<0.9 => Potency class= Moderate
- positive concordance>=0.9 => Potency class= Strong



Box plot of the positive classes of the consensus model.



Variation of the balanced accuracy with positive concordance thresholds

New External validation set

ToxCast phIII+ Tox21 agonist assays

All matching chemicals: 620

Observed\Predicted	Actives	Inactives
Actives	19	23
Inactives	17	561

Specificity: 0.97 Sensitivity: 0.45 Balanced accuracy: 0.71

Only chemicals in agreement with other literature sources: 584

Observed\Predicted	Actives	Inactives
Actives	13	3
Inactives	17	551

Specificity: 0.97 Sensitivity: 0.81 Balanced accuracy: 0.89

Conclusions

- High quality training set (1677 chemicals)
- Free & open-source structure curation workflow
- Curated structures with potential exposure (32k)
- QSAR-ready dataset from the literature (~7k)
- Consensus models for binding, agonist & antagonist
- 32k list predicted for prioritization.
- EDSP dashboard: http://actor.epa.gov/edsp21/

future work

- Validate binding consensus with the new external set
- Clean literature data from cytotoxicity. Use it as QSAR ready set.