

Developing an Integrated Model Management Solution to Assure Quality of Predicted Data at the US EPA's National Center of Computational Toxicology

Christopher Grulke¹

Antony Williams¹

Amar Singh¹

Jeff Edwards¹



1. National Center of Computational Toxicology, U.S. EPA

American Chemical Society Meeting, Spring 2019
1 April 2019, Orlando FL

The Comptox Chemicals Dashboard Overview

A **publicly accessible website** delivering access:

- ~875,000 chemicals with related property data
- Experimental Human and Ecological hazard data
- Integration to “biological assay data” for 1000s of chemicals
- Information regarding consumer products containing chemicals
- Links to other agency websites and public data resources
- “Literature” searches for chemicals using public resources
- “Batch searching” for thousands of chemicals
- Experimental and predicted physicochemical property data
- Real time prediction of physchem and toxicity endpoints

QSAR Predictions in the Dashboard

Download Predicted Data

Source	Result	Calculation Details	QMRF
EPISUITE	114	Not Available	Not Available
NICEATM	153	Not Available	Available
TEST	154	TEST Report	Not Available
OPERA	181	OPERA Model Report [Inside AD]	Available

Prediction results (colors defined in table below)

Chemicals	MAE*
Entire set	11.46
Similarity coefficient ≥ 0.5	6.68

*Mean absolute error in Å°C

CAS	Structure	Similarity Coefficient	Experimental value Å°C	Predicted value Å°C
NOCAS_871126 (test chemical)	<chem>CC(C)(O)C</chem>		175.00	179.54
107-41-5	<chem>CC(C)(O)C</chem>	0.88	198.00	199.20
684-84-4	<chem>CC(C)(O)C</chem>	0.87	200.00	196.96

Predicted average

2.00
151
323
4.48e-4
3.49e-6
176
53.8
1.60
58.5
23.2
1.25
170
158
1.25e-8
8.38

HAZARD

- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COM
- GENRA (BETA
- RELATED SUB

Model Results

Predicted value: 186 °C
Global applicability domain: Inside
Local applicability domain index: 0.925
Confidence level: 0.830

Model Performance

Weighted KNN model

5-fold CV (75%)		Training (75%)		Test
Q2	RMSE	R2	RMSE	R2
0.930	22.5	0.930	22.1	0.930

Nearest Neighbors from the Training Set

CC(C)(O)C
(3R)-2-Methylbutane-2,3-diol
Measured: 175
Predicted: 185.944

CC(C)(O)C
2,4-Pentanediol
Measured: 199
Predicted: 199.144

CC(C)(O)C
1,3-Butanediol, 3-methyl-
Measured: 202.5
Predicted: 200.052

New Prediction



Property	Experimental Value	Consensus
96 hour fathead minnow LC50	4.158 -Log10(mol/L) 14.993 mg/L	4.001 -Log10(mol/L) 21.524 mg/L
48 hour D. magna LC50	3.601 -Log10(mol/L) 54.062 mg/L	3.854 -Log10(mol/L) 30.189 mg/L
48 hour T. pyriformis IGC50		
Oral rat LD50	2.506 -Log10(mol/kg) 672.807 mg/kg	2.197 -Log10(mol/kg) 1370.798 mg/kg
Bioaccumulation factor	0.557 Log10 3.606	0.950 Log10 8.910
Developmental toxicity		true
Ames mutagenicity	false	false
Estrogen Receptor RBA		
Estrogen Receptor Binding	false	false
Normal boiling point		312.2 °C
Melting point	173.0 °C	163.6 °C
Flash point	176.7 °C	175.0 °C
Vapor pressure	-6.539 Log10(mmHg) 2.891*10 ⁻⁷ mmHg	-6.369 Log10(mmHg) 4.275*10 ⁻⁷ mmHg
Density	1.269 g/cm ³	1.248 g/cm ³
Surface tension		
Thermal conductivity		158.367 mW/mK



Select properties to predict

H

C

N

O

S

P

F

Cl

Br

I

PT

T.E.S.T.

- ☒ Toxicological properties
 - ☒ 96 hour fathead minnow LC50
 - ☒ 48 hour D. magna LC50
 - ☒ 48 hour T. pyriformis IGC50
 - ☒ Oral rat LD50
 - ☒ Bioaccumulation factor
 - ☒ Developmental toxicity
 - ☒ Ames mutagenicity
 - ☒ Estrogen Receptor RBA
 - ☒ Estrogen Receptor Binding
- ☒ Physical properties
 - ☒ Normal boiling point
 - ☒ Melting point
 - ☒ Flash point
 - ☒ Vapor pressure
 - ☒ Density
 - ☒ Surface tension
 - ☒ Thermal conductivity
 - ☒ Viscosity
 - ☒ Water solubility

Calculate

Chiral

5

Known Issues: Prediction of Tautomers in T.E.S.T.

DETAILS

EXECUTIVE SUMMARY

PROPERTIES

ENV. FATE/TRANSPORT

HAZARD

ADME

EXPOSURE

BIOACTIVITY

SIMILAR COMPOUNDS

GENRA (BETA)

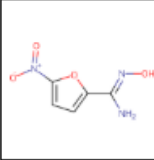
RELATED SUBSTANCES

SYNONYMS

LITERATURE

LINKS

COMMENTS



5-Nitro-2-furamidoxime
772-43-0 | DTXSID00209
Searched by DSSTox Compound Id.

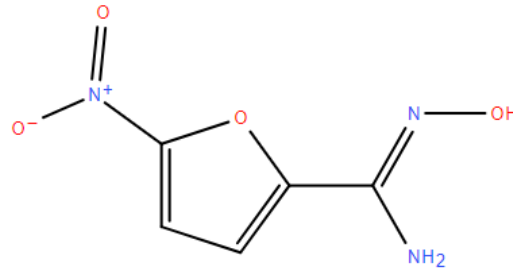
Property
Flash Point

Download Summary

Type	Average
Experimental	-
Predicted	84.4

Download Predicted Data

Source	Result
ACD/Labs	86.1
TEST	82.7



Provider: T.E.S.T.
Download Summary

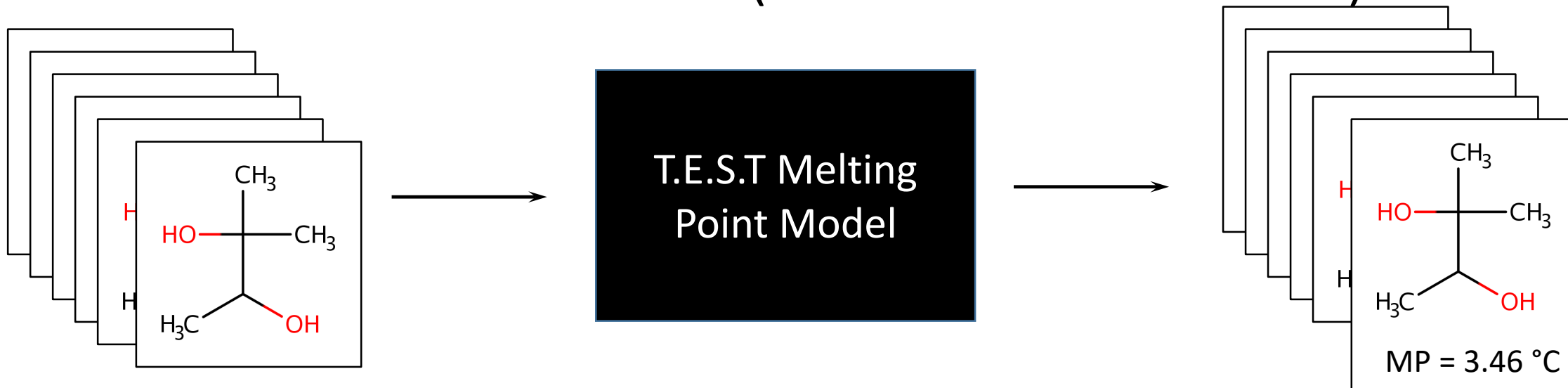
Property	Experimental Value	Consensus	Hierarchical clustering	Single model
96 hour fathead minnow LC50		3.061 -Log10(mol/L) 148.865 mg/L	3.303 -Log10(mol/L) 85.145 mg/L	3.231 -Log10(mol/L) 100.608 mg/L
48 hour D. magna LC50				
48 hour T. pyriformis IGC50				
Oral rat LD50				
Bioaccumulation factor		0.042 Log10 1.102	0.042 Log10 1.102	0.042 Log10 1.102
Developmental toxicity		false	false	true
Ames mutagenicity		false	false	
Estrogen Receptor RBA				
Estrogen Receptor Binding		false	false	false
Normal boiling point		259.4 °C		
Melting point		180.2 °C		
Flash point		140.3 °C		

Certain Point of View



“So, what I told you was true... from a certain point of view” - Obi-Wan Kenobi

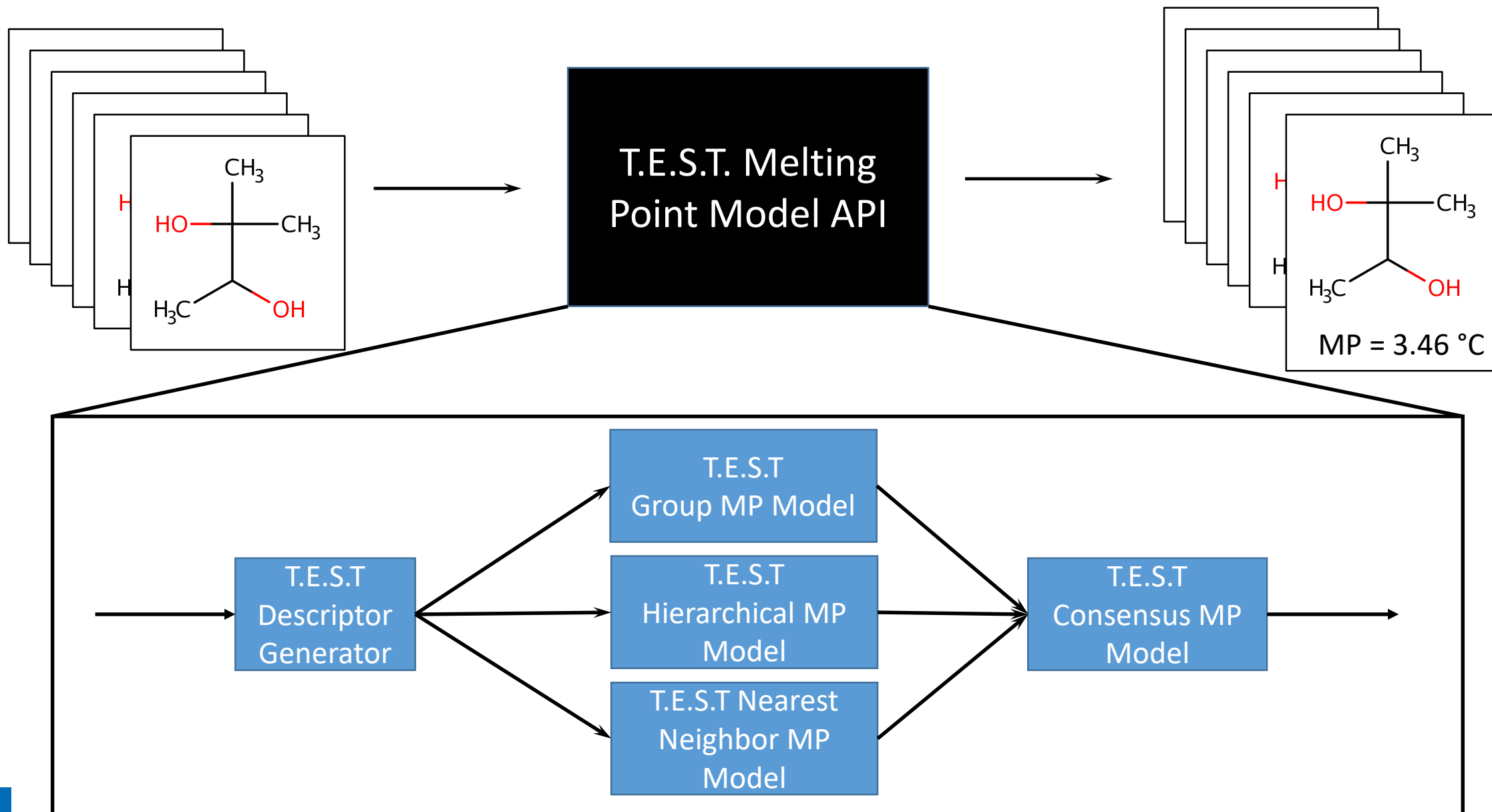
Model Prediction (for non-modelers)



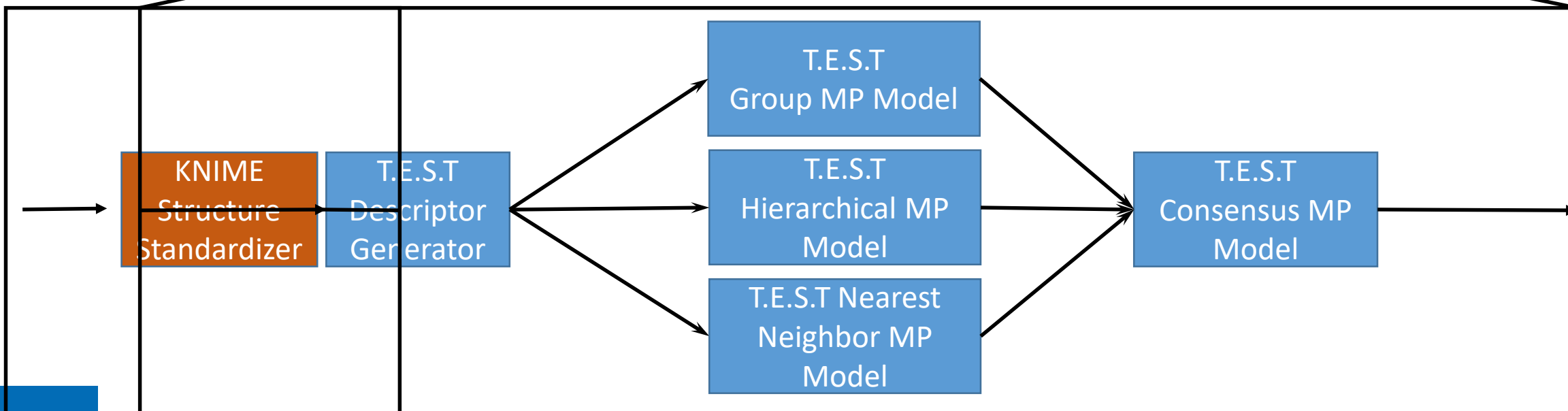
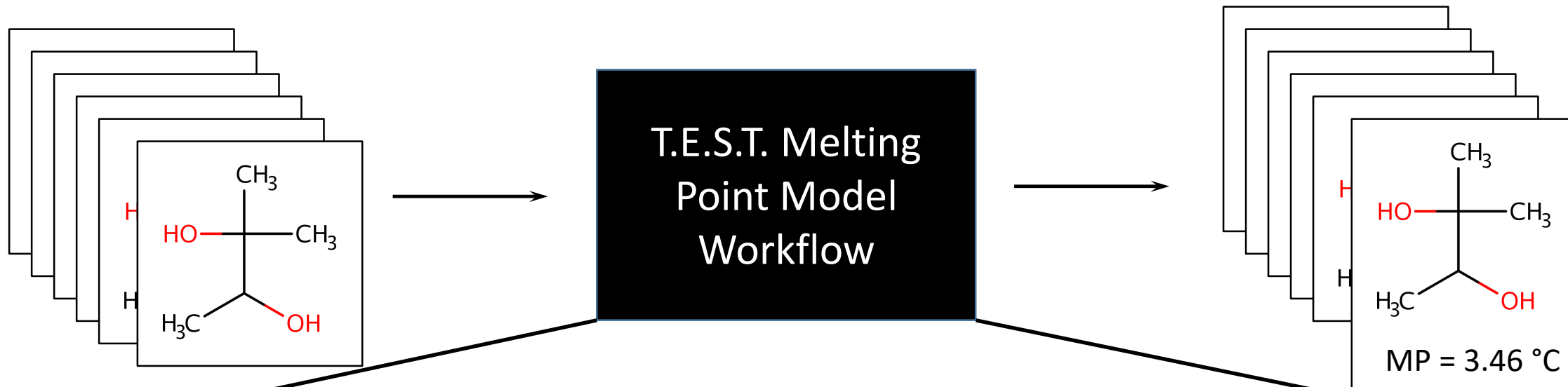
EXPECTATIONS

It works
If it fails, it tells you why
Reproducible
Same Chemical, Same Result
Changes to the model are clearly conveyed

Model Prediction (T.E.S.T.)



Stored Model Predictions (T.E.S.T.)



11

Chemical reaction scheme showing the conversion of 5-nitro-2-(hydroxyimino)phenylhydrazide to 5-amino-2-nitrophenylhydrazide.

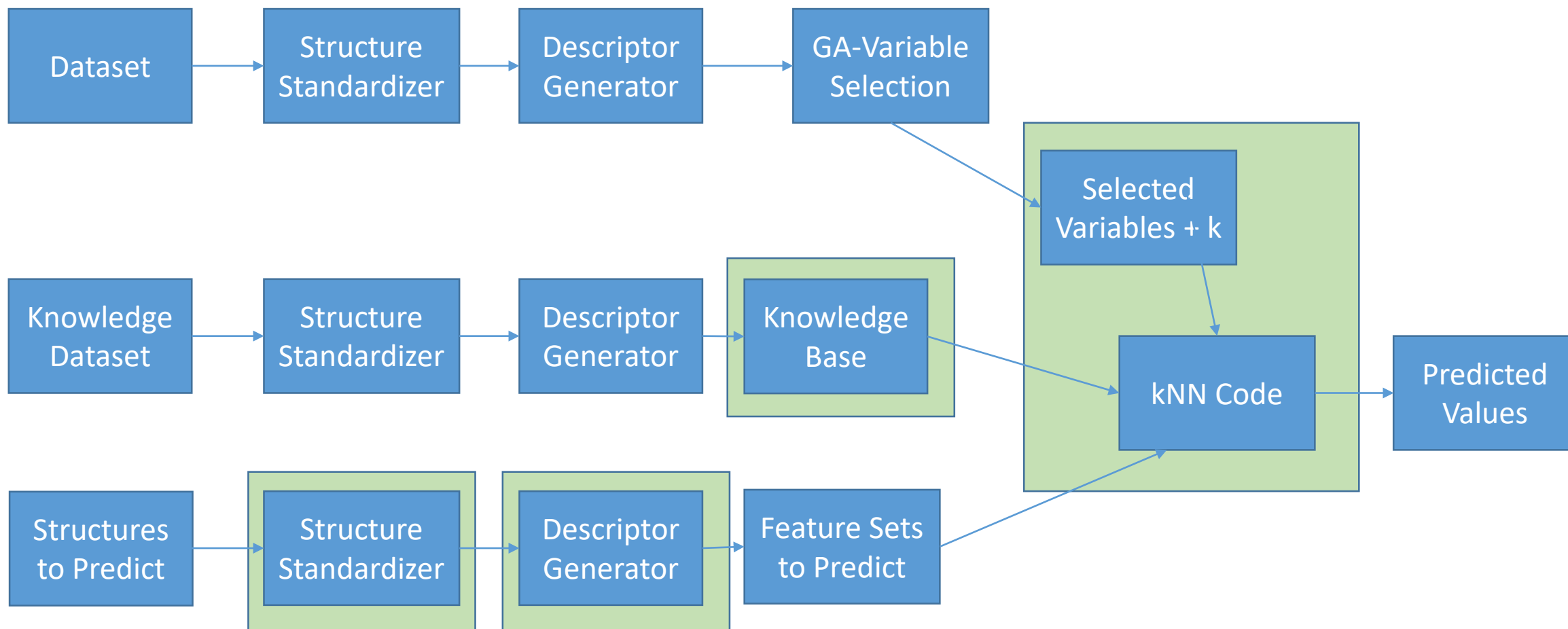
Reactant: 5-nitro-2-(hydroxyimino)phenylhydrazide (Structure 1)

Product: 5-amino-2-nitrophenylhydrazide (Structure 2)

The reaction involves the reduction of the nitro group to an amino group and the conversion of the hydroxyimino group to a nitro group.

What is the Model?

Genericized GA-kNN Modeling Workflow



Known Issues: EpiSuite Tautomers

EPI Suite

File Edit Functions Batch Mode Show Structure Output Fugacity SIP Help

EPI Suite - Welcome Screen

KOWWIN v1.69

File Edit Functions BatchMode ShowStructure Zwitterions Help

Draw Previous Get User Save User CAS Input ExpValAdj Calculate

Enter SMILES: NC(N=O)C1=CC=C(O1)[N+][O-]

Enter NAME:

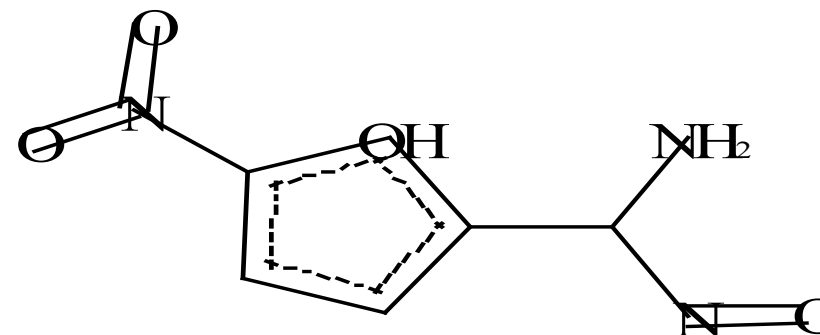
NameLookup

AOPWIN
 KOWWIN
 BIOWIN
 MPBPVP
 WSKOW
 WATERNT
 HENRYWIN
 KOAWIN
 KOCWIN
 BCFBAF
 HYDROWIN
 BioHCwin
 DERMWIN
 ECOSAR
 EPI Links

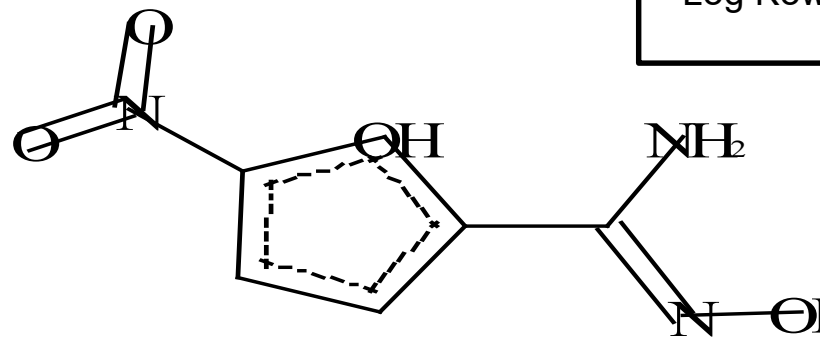
The Estimation Program
 and Toxics and Syrac
 chemicals for release
 (measured) values are

 EPI Suite™ cannot
 chemicals generally

 Important information
 found under the Help
 programs except Bio



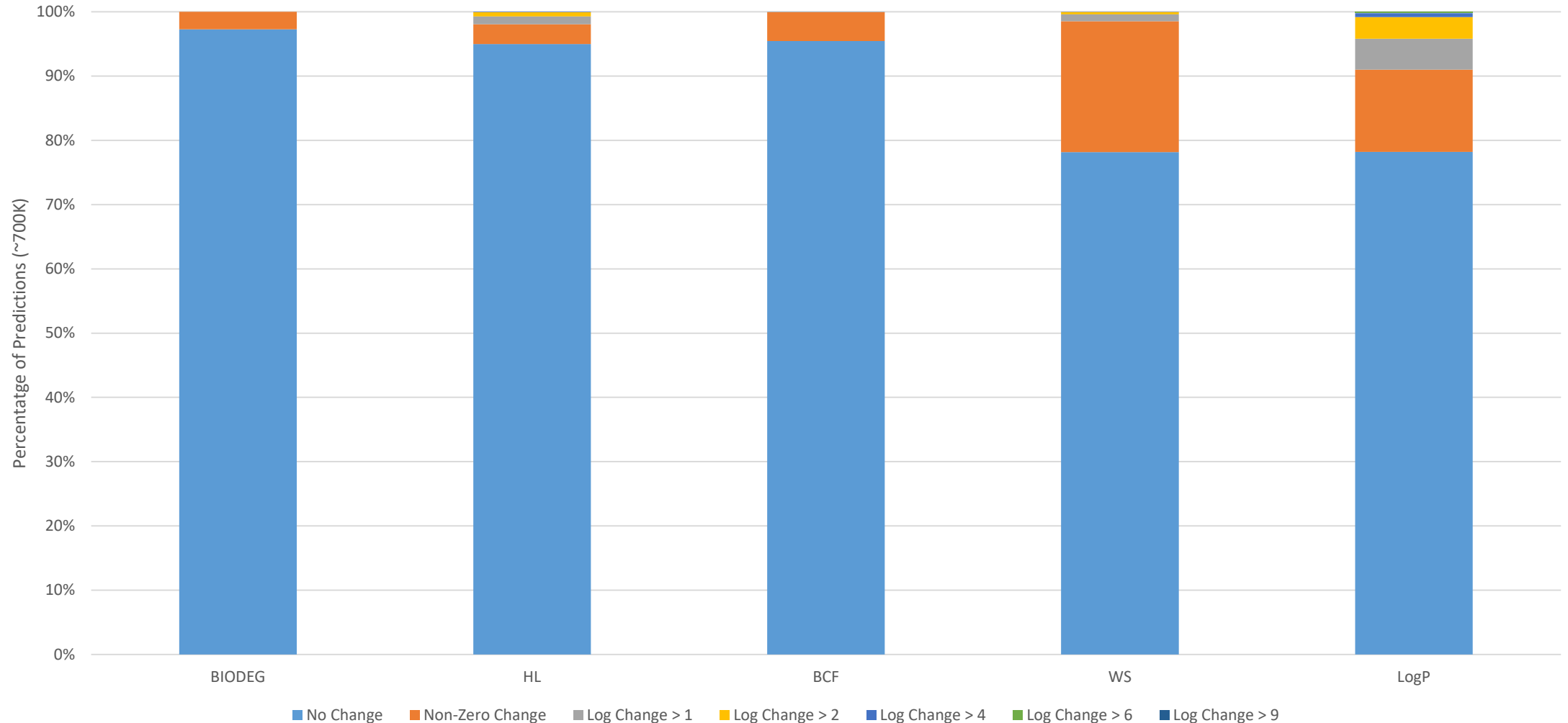
Log Kow (estimated): -0.00



Log Kow (estimated): -0.36

Known Issues: OPERA 2.0 Released

Prediction Differences between OPERA versions



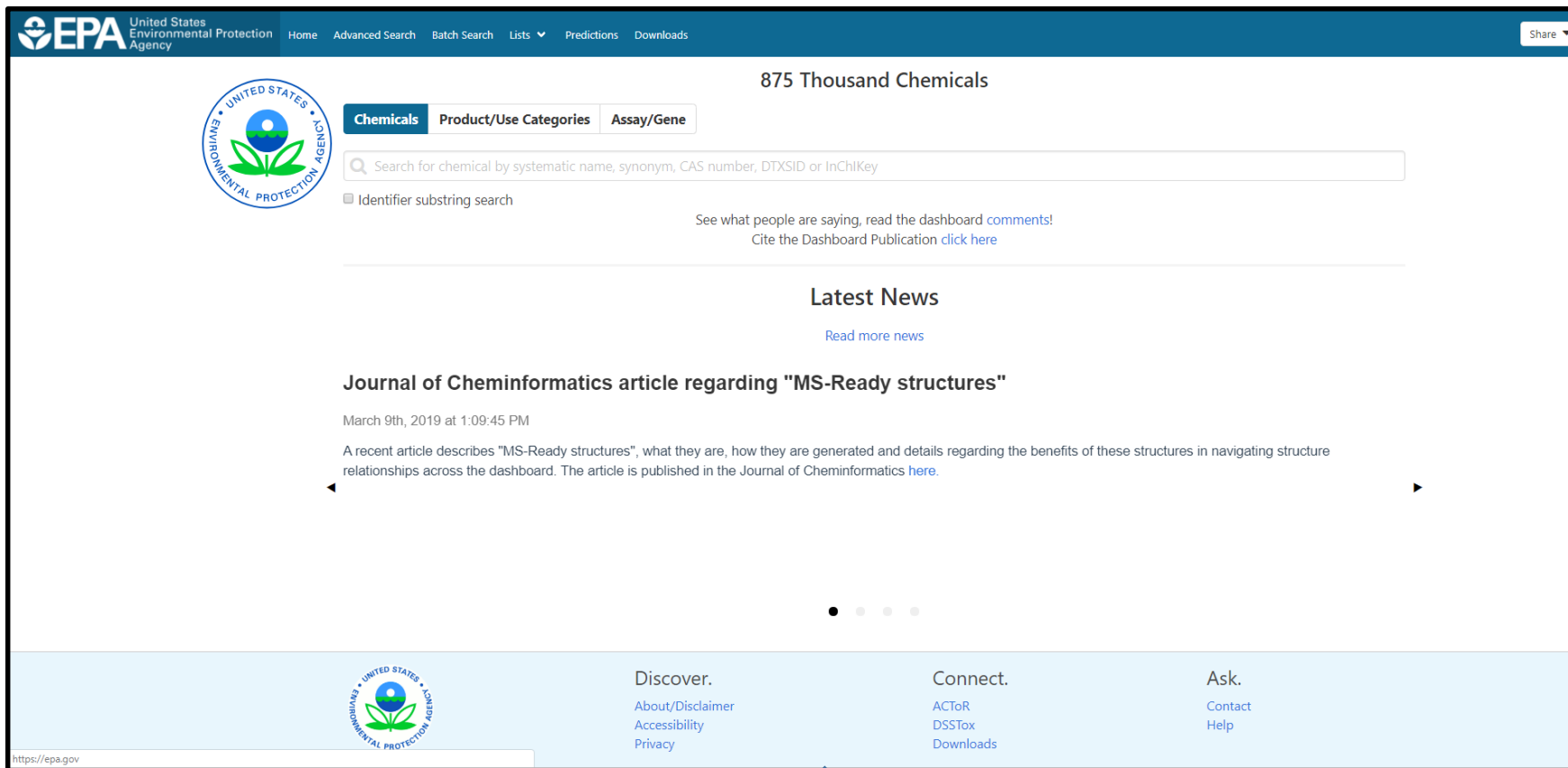
Standardization Improvements

Knowledge Base Updated

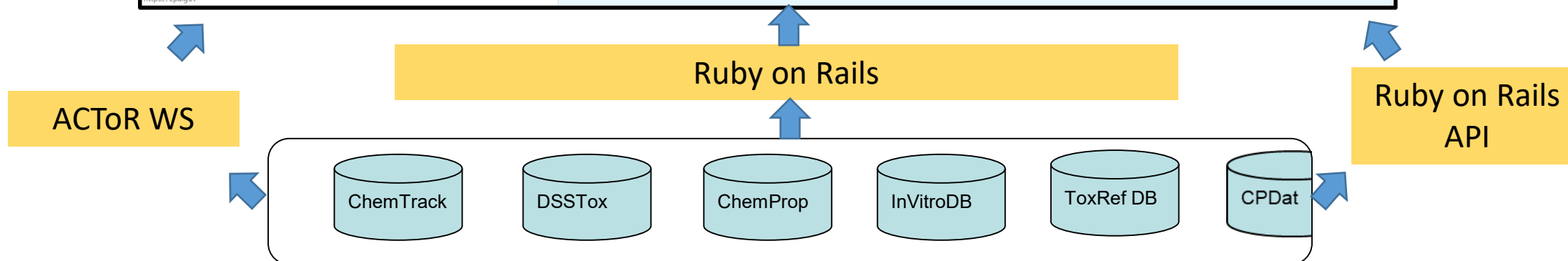


This is not the web-
based application
talk I was looking
for?

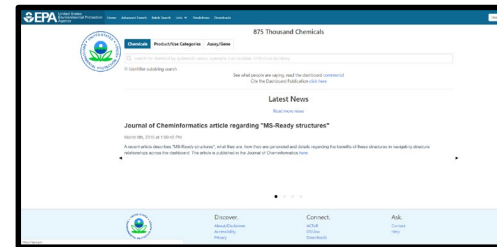
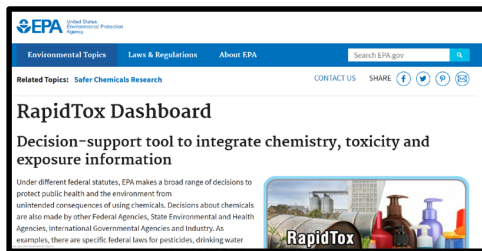
Current Dashboard Structure



The screenshot shows the EPA Chemicals Dashboard. At the top, the EPA logo and navigation links (Home, Advanced Search, Batch Search, Lists, Predictions, Downloads) are visible. The main header displays "875 Thousand Chemicals". Below this, there are tabs for "Chemicals", "Product/Use Categories", and "Assay/Gene". A search bar prompts users to "Search for chemical by systematic name, synonym, CAS number, DTXSID or InChIKey". A section titled "Latest News" features an article from the "Journal of Cheminformatics" dated March 9th, 2019, about "MS-Ready structures". The footer contains links for "Discover.", "Connect.", and "Ask.".



Proposed EPA-ORD Data Structure



Production Application Layer

Virtual or Physical Application Databases
– Data Structured to efficiently support
our external production applications

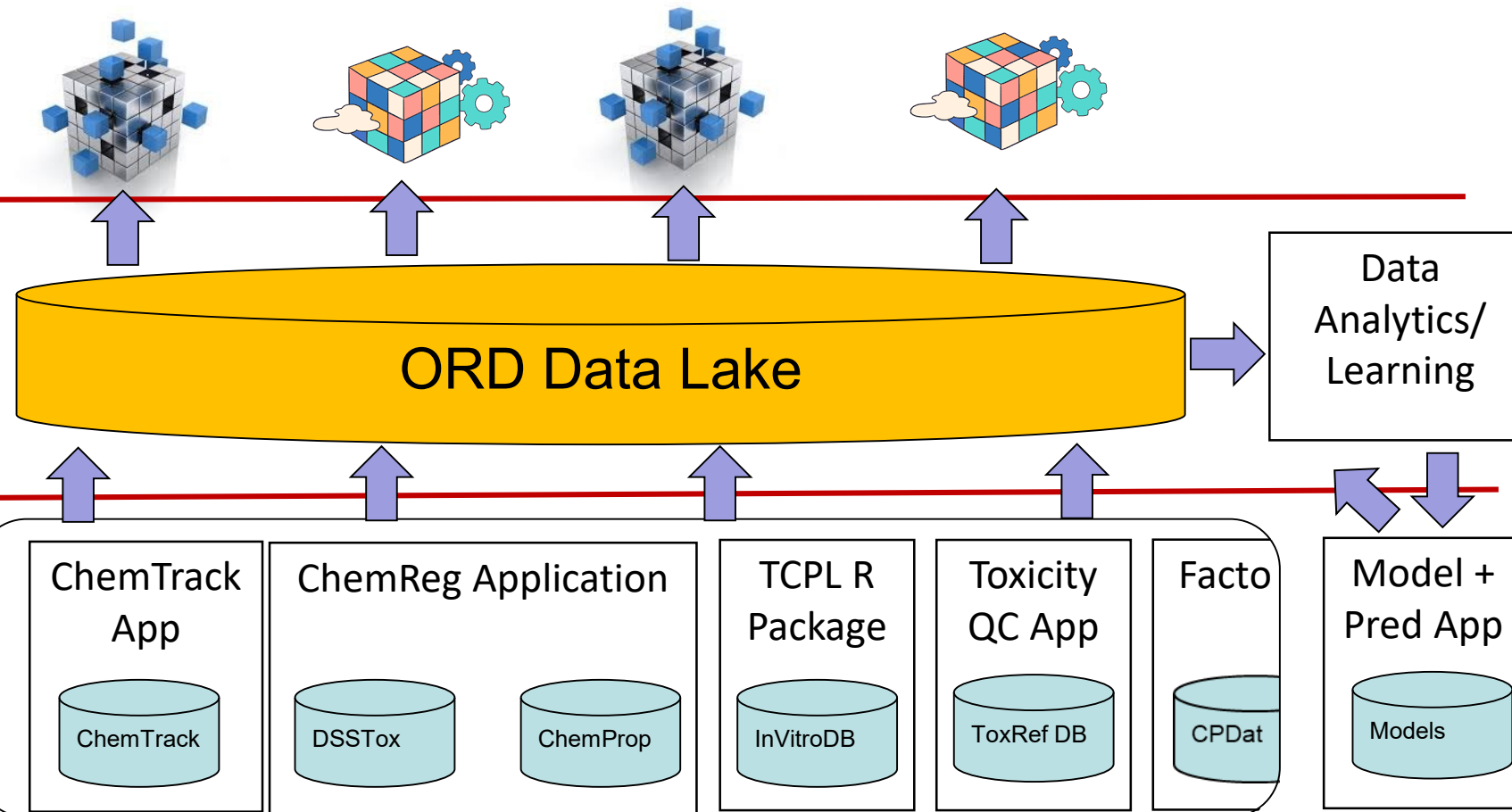
Presentation Layer (Fit-for-Purpose)

Generalized data model
providing integration
across the domains for
scientific exploration

Integration Layer

Transactional Layer

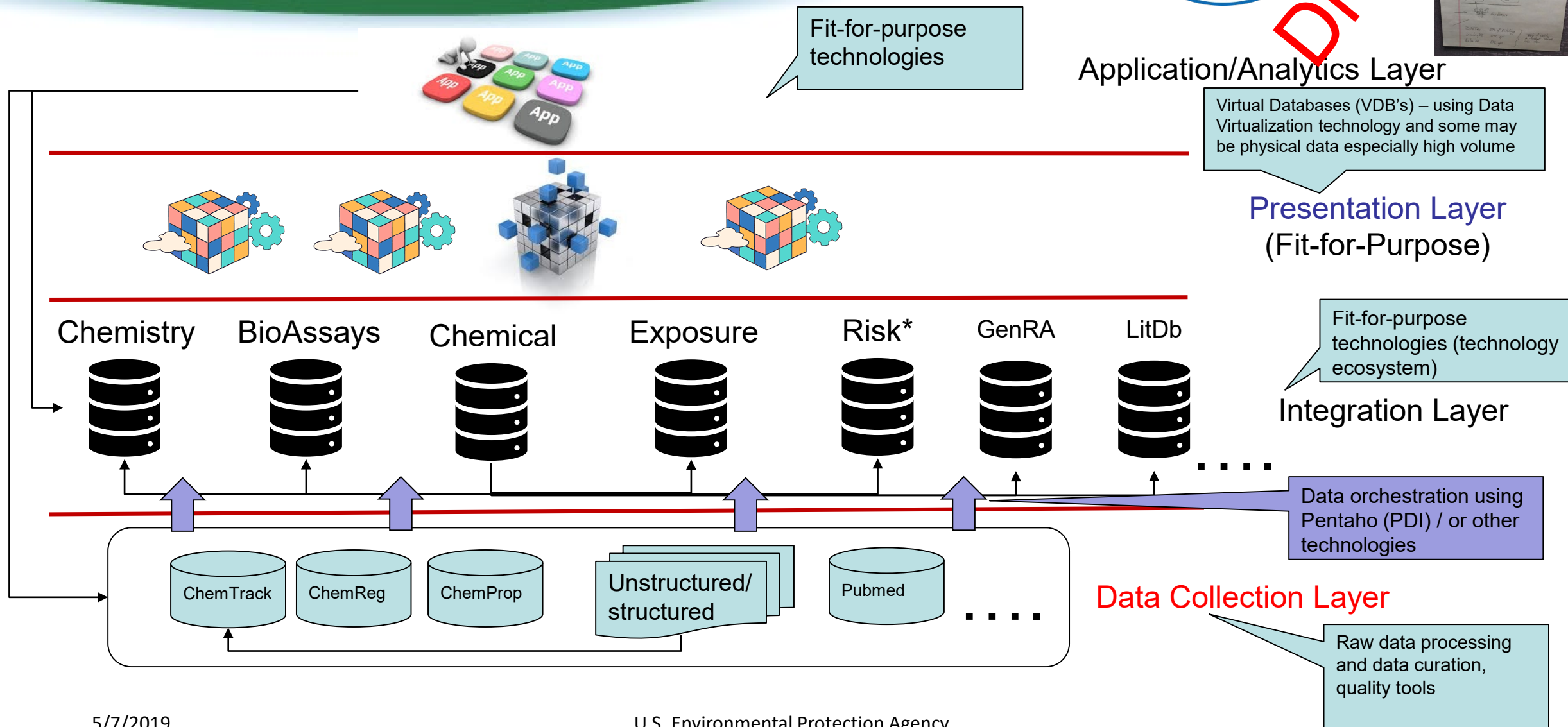
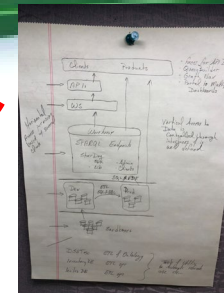
Raw data processing
and data curation,
quality tools



High level Information/Integration Architecture (30K feet)



Draft



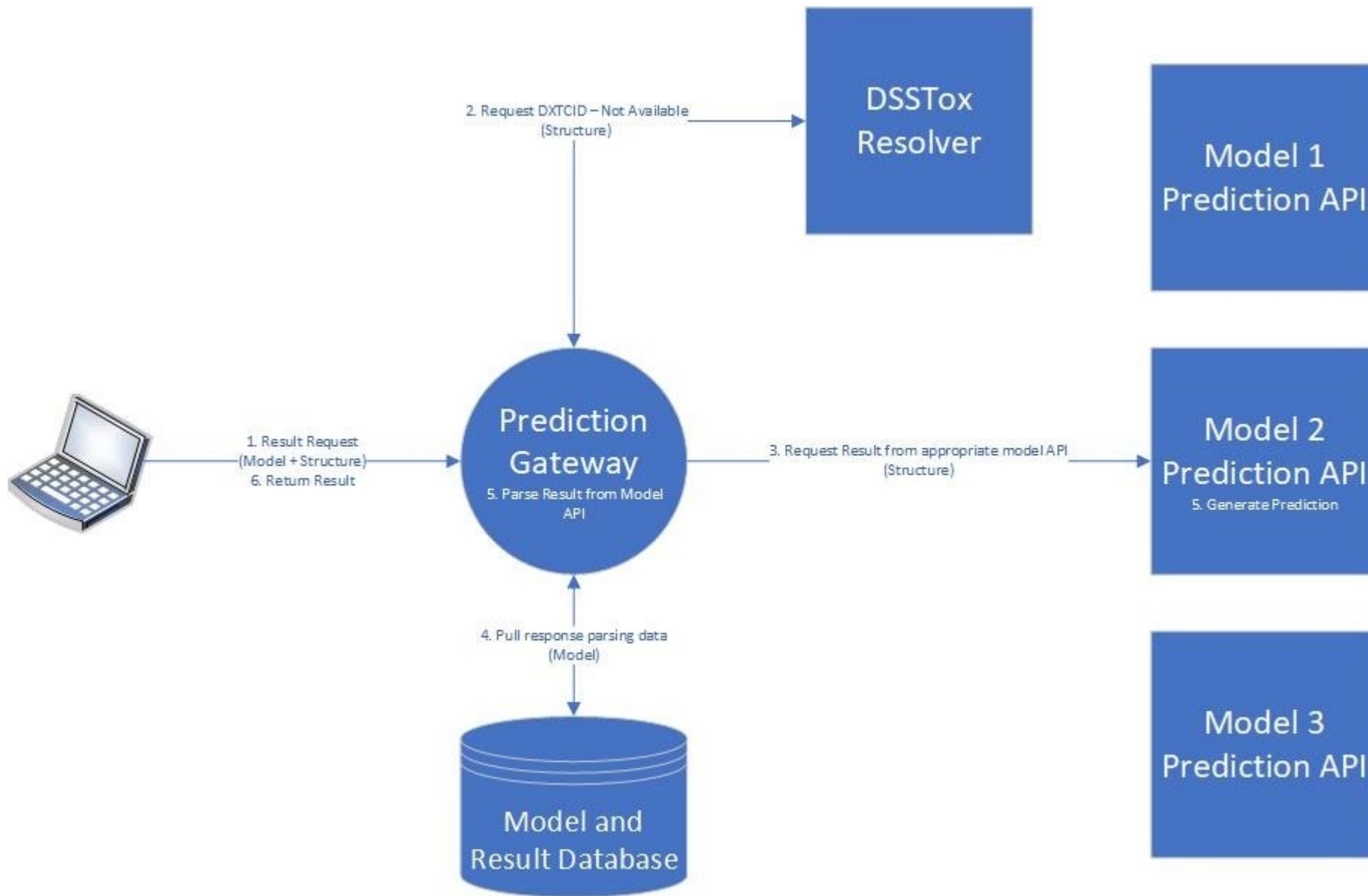
5/7/2019

U.S. Environmental Protection Agency

Contribution: Jeff Edwards and Amar Singh

* Decision that can be stored with consultations with key stakeholders

Model Prediction Gateway



Model registration interface

ACToR-DSSTox Chemical Registration

View/Edit a Single Record Structure Search Browse/Curate Records Export DSSTox Chemotypes Manage Chemical Lists Manage Property Data Add Deleted Casrns **Manage Models**

Welcome cgrulke

Search for a Qsar Model

OPERA

Search

Source:

OPERA

Name:

OPERA_VP

Model Class:

☒ Disconnected ☐ External Service ☐ Local Model

Label:

OPERA Vapor Pressure

Short Description

The vapor pressure model from the OPERA software application v2.0

Long Description

The vapor pressure model from the v2.0

Software

OPERA_v2.0

Implementation Notes

OPERA vapor pressure prediction pr 2/26/2019. This set of data was st standardization workflow with desc prediction completed using the OPE

Next

Cancel

Model Name: OPERA_VP

New Prediction (in a tsv format):

Save New Data

Model Classes

Disconnected

- Models not accessible via API
- Data parsed and stored in prediction gateway database
- Not available in prediction interface

e.g., all current models

External Service

- Models accessible via API
- API JSON schema not “controlled”
- Available in prediction interface
- May require structural pre-processing

e.g., T.E.S.T Models

Local Model

- Models accessible via API
- API JSON schema “controlled” (pre-defined parsing)
- Available in prediction interface
- Likely uses support services

e.g., None

Model Result Parsing

Search for a Qsar Model

Source:
Name:
Model Class: ☒ Dis ☐ Con
Label:
Short Description:
Long Description:
Software:
Implementation Notes:

Model Name: OPERA_VP
New Prediction (in a tsv format):

MoleculeID	BP_pred	AD_BP	AD_index_BP	Conf_index_BP	BP_DSSTOXMPID_neighbor_1	BP_DSSTOXMPID_neighbor_2
BP_DSSTOXMPID_neighbor_3			BP_DSSTOXMPID_neighbor_4		BP_DSSTOXMPID_neighbor_5	
DTXCID60883654	312.4264288	1	0.364508111	0.495081851	19506	19526
DTXCID60893157	300.9209418	1	0.466389508	0.651009516	20975	19560
DTXCID70893163	331.7435009	0	0.364907678	0.439346717	24646	22987
DTXCID90893165	244.5177002	1	0.616416613	0.636233575	22620	19798
DTXCID20893173	327.84903	0	0.30368404	0.377789572	19468	19132
DTXCID60872568	274.2450955	1	0.562789584	0.694340407	24549	23136
DTXCID70893688	301.2280978	1	0.368082174	0.602620494	20975	21001
DTXCID90883667	247.8285442	1	0.951535784	0.884089592	24056	21356
DTXCID60893192	337.6601623	0	0.182064063	0.44478338	22394	22573
DTXCID00893191	295.4580244	1	0.27809264	0.393890342	22396	24603
DTXCID50893186	276.5252297	1	0.256688427	0.51175549	23198	20484
DTXCID50894274	222.0173729	1	0.389031834	0.522414704	19830	23912
DTXCID10894275	210.8582755	1	0.678463756	0.690978191	23836	24240
DTXCID20893198	288.1721205	1	0.2637774	0.466554246	23281	22107
DTXCID60893197	279.4804796	1	0.218004333	0.351941404	21686	20484
DTXCID20893875	231.0121354	0	0.253011705	0.372803467	19799	20511
DTXCID40889295	470.438083	0	0.100184203	0.307075581	23443	24628
DTXCID20893951	214.8296955	1	0.262994668	0.542646808	22630	23457
DTXCID20881217	318.2570382	0	0.313377812	0.412573358	19633	22677
DTXCID70888978	324.5091213	0	0.069590918	0.280110689	19782	19159
DTXCID7029586	343.2158127	1	0.40004618	0.50747105	22395	22949
DTXCID50884756	301.3213089	1	0.26353058	0.496579182	23198	22593
DTXCID50894335	278.0202743	1	0.227134213	0.354563758	20484	21686
DTXCID60872780	413.8526075	0	0.028028877	0.175557587	23922	19159
DTXCID10889025	338.8526517	1	0.273167398	0.470516955	19763	20369
DTXCID80893432	300.9529916	0	0.159013265	0.321981288	20367	19797
DTXCID60893470	316.6093532	0	0.175705584	0.358082268	19390	23152
DTXCID00885314	318.1740396	0	0.157491668	0.317578918	21234	19390
DTXCID30893523	346.1157953	1	0.551526981	0.575580557	19169	19168
DTXCID90894394	292.3420884	1	0.290951072	0.472955893	22593	23198
DTXCID60894402	282.0595866	0	0.156681965	0.325758436	20367	19797
DTXCID20894403	373.6647139	0	0.118062916	0.254699026	19159	19167
DTXCID50894411	374.7050022	1	0.449536473	0.600414317	21774	21573
DTXCID00883456	331.7667474	0	0.206846456	0.340776469	21234	19390
DTXCID90873902	421.3331183	0	0.133077034	0.37093668	24628	21113

Model Classes

Disconnected

- Models not accessible via API
- Data parsed and stored in prediction gateway database
- Not available in prediction interface

e.g., all current models

External Service

- Models accessible via API
- API JSON schema not “controlled”
- Available in prediction interface
- May require structural pre-processing

e.g., T.E.S.T Models

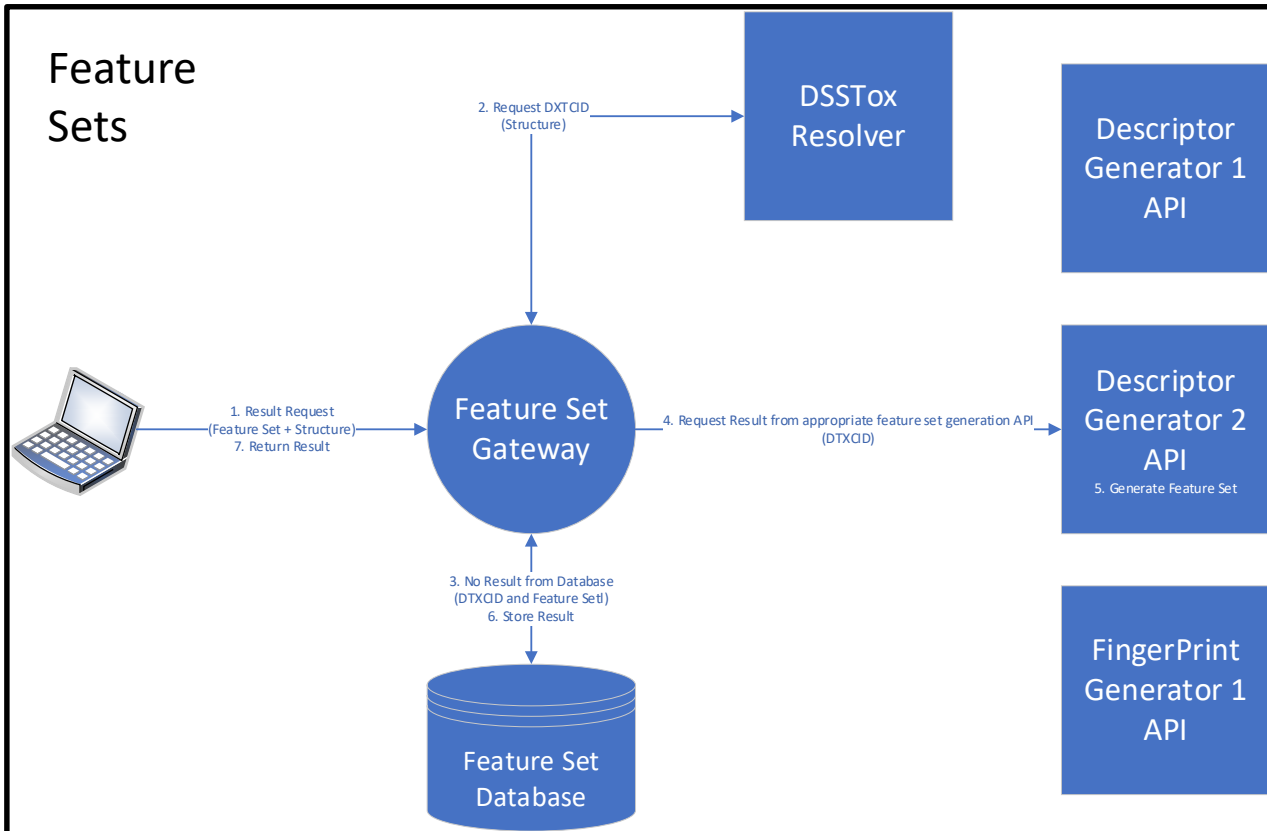
Local Model

- Models accessible via API
- API JSON schema “controlled” (pre-defined parsing)
- Available in prediction interface
- Likely uses support services

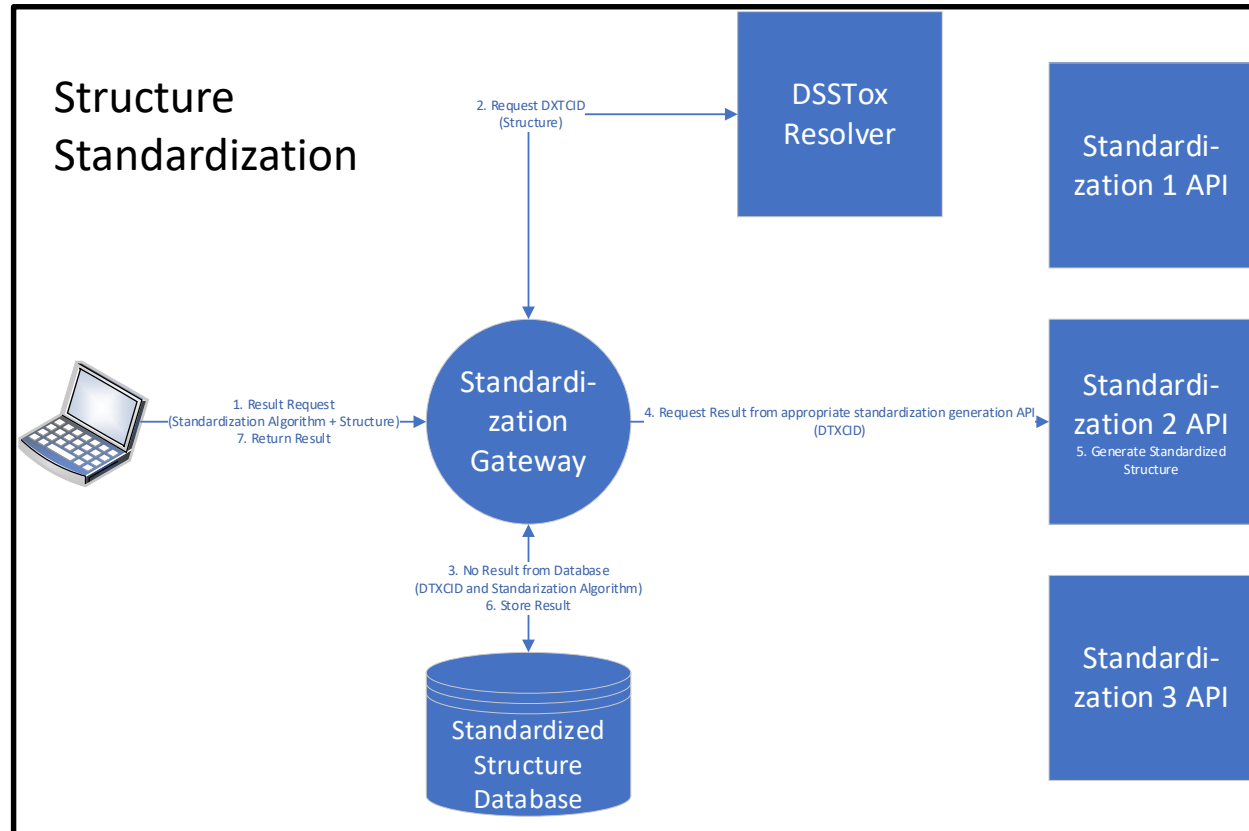
e.g., None

Supporting Service Gateways

Feature Sets



Structure Standardization



Descriptor and Fingerprint Availability in Dashboard

Batch Search?



Step Four: Select Output Format

Please enter one identifier per line

Select Input Type(s)

- ☐ Identifiers
 - ☐ Chemical Name
 - ☐ CASRN
 - ☐ InChiKey
 - ☐ DSSTox Substance ID
 - ☒ DSSTox Compound ID
 - ☐ InChiKey Skeleton
 - ☐ MS-Ready Formula(e)
 - ☐ Exact Formula(e)
 - ☐ Monoisotopic Mass

Display All Chemicals

Download Chemical Data

Enter Identifiers to Search (searches should be limited to <5000 identifiers)

DTXCID101

Select Output Format:

Download as...

Download

- ☐ PubChem Data Sources
- ☐ CPDat Product Occurrence Count
- ☐ IRIS
- ☐ PPRTV
- ☐ QC Notes
- ☐ Include links to ACToR reports - SLOW! (BETA)

Enhanced Data Sheets

- ☐ MetFrag Input File (Beta)
- ☒ ToxPrint single fingerprints
- ☐ Abstract Sifter Input File (Beta)
- ☐ Synonyms and Identifiers

- [EPA: Consumer Products Suspect Screening Result](#)
- [EPA: High Production Volume List](#)
- [EPA: IRIS Chemicals](#)
- [EPA: Mechanism of Action \(MoA\) for aquatic toxicity](#)
- [EPA: National-Scale Air Toxics Assessment \(NATA\)](#)
- [EPA: PPRTV Chemical Report](#)
- [EPA: Safer Choice Chemical List](#)
- [EPA: Toxicity Values Version 5 \(Aug 2018\)](#)
- [EPA: Toxics Release Inventory](#)
- [EPA|ECOTOX: Fathead Minnow Acute Toxicity](#)
- [EPA|TSCA: TSCA Work Plan Chemicals](#)

Standardization in the Dashboard

Batch Search



Step Four: Select Output Format

Please enter one identifier per line

Select Input Type(s)

- ☐ Identifiers
 - ☐ Chemical Name
 - ☐ CASRN
 - ☐ InChIKey
 - ☐ DSSTox Substance ID
 - ☒ DSSTox Compound ID
 - ☐ InChIKey Skeleton
 - ☐ MS-Ready Formula(e)
 - ☐ Exact Formula(e)
 - ☐ Monoisotopic Mass

Display All Chemicals Download Chemical Data

Enter Identifiers to Search (searches should be limited to <5000 identifiers)

DTXCID101

Select Output Format:

Download as... Download

Same Connectivity: 3 records (based on first layer of InChI)

InChIKey

IUPAC Name

Structures

Mol File

SMILES

InChI String

MS-Ready SMILES

QSAR-Ready SMILES

Intrinsic And Predicted Properties

Molecular Formula

Average Mass

Monoisotopic Mass

TEST Model Predictions

OPERA Model Predictions

California Office of Environmental Health Hazard Assessment

CERAPP: Collaborative Estrogen Receptor Activity Prediction Project

CHEMINV: ToxCast/Tox21 Chemical inventory available as DMSO solutions (20181123)

CHEMINV: EPA Chemical Inventory for ToxCast

CHEMINV: EPA ToxCast CHEMINV list of volatiles

CHEMINV: EPA ToxCast Cheminventory chemicals with stability problems

CHEMINV: EPA ToxCast Cheminventory DMSO Insolubles

CHEMINV: EPA ToxCast Cheminventory List of Reactives

DRUGS: DrugBank database from the University of Alberta

DRUGS: ITNANTIBIOTIC list of antibiotics

DRUGS: Pharmaceutical List with EU, Swiss, US Consumption Data

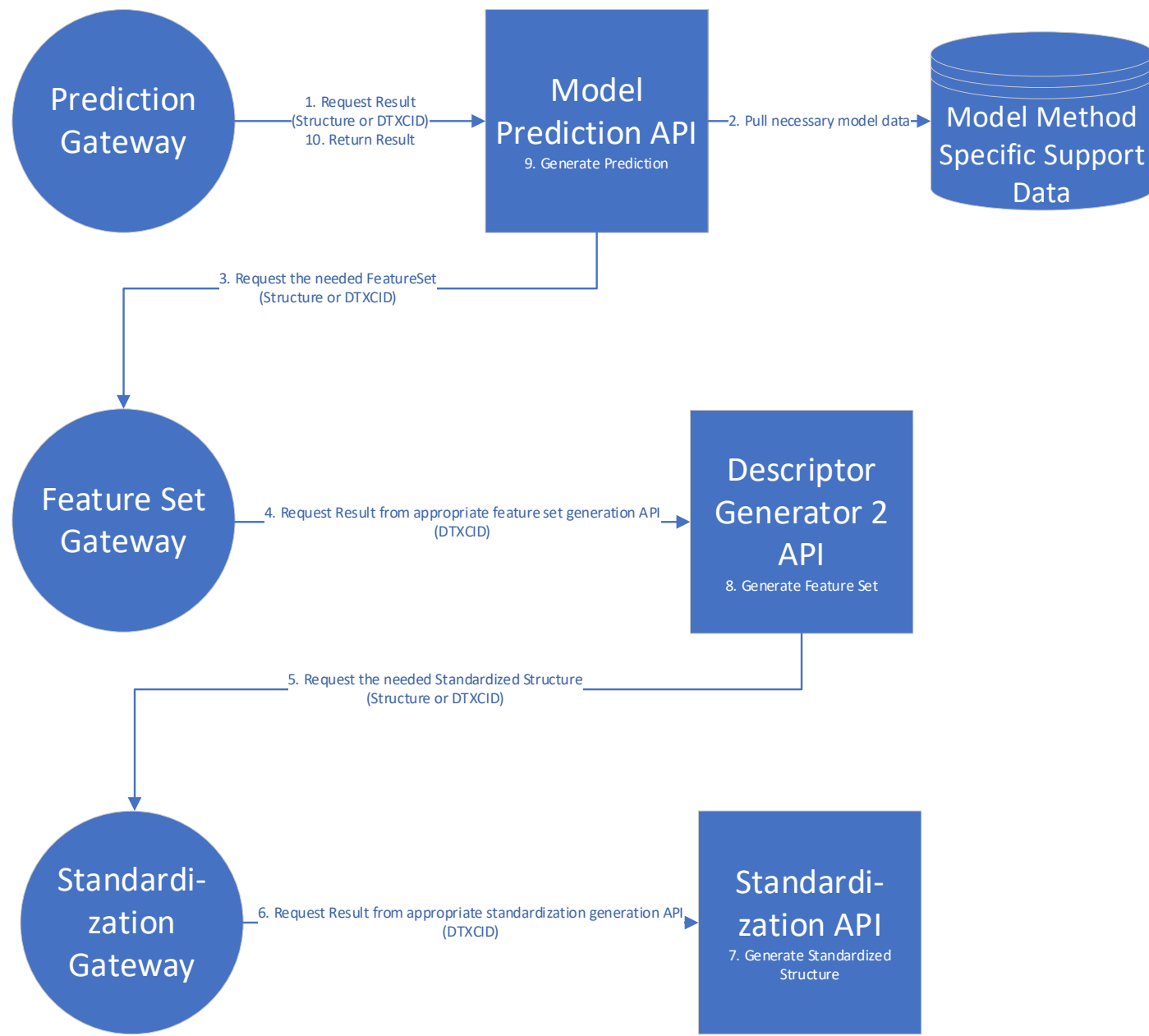
DRUGS: Statin drugs

ECOTOX: Ecotoxicology knowledgebase

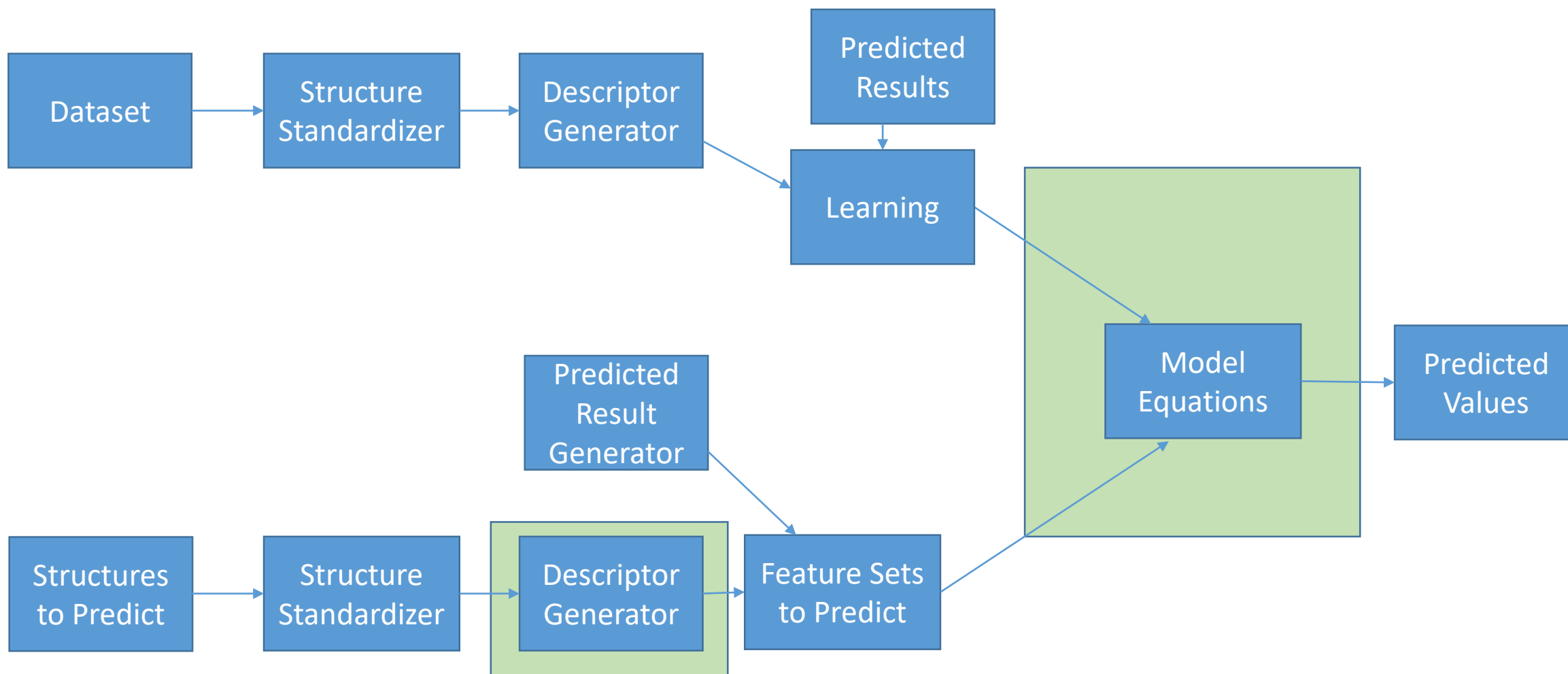
Endocrine Disruptor Screening Program (EDSP) University of Chemicals

id as a compone

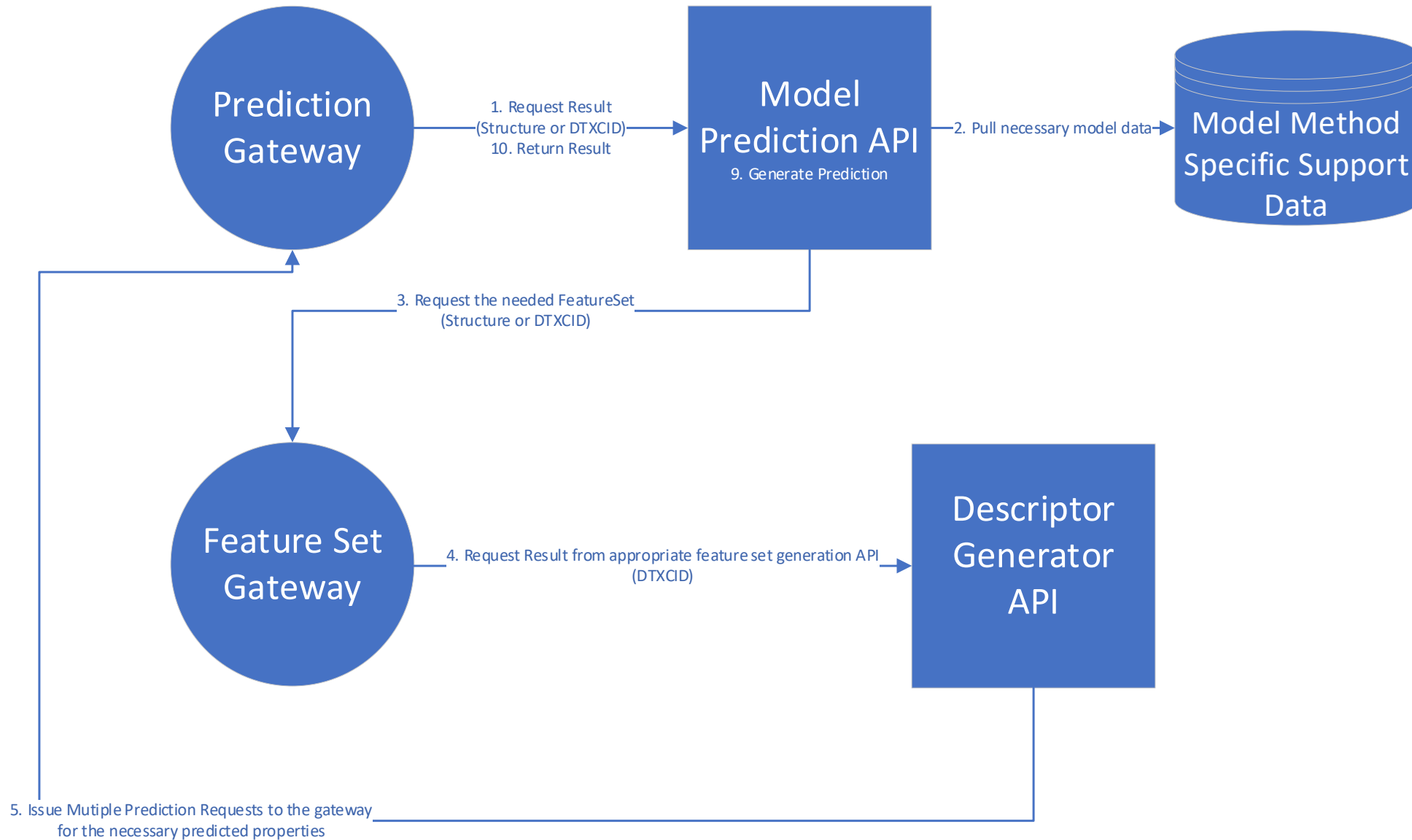
Local Model Services



More Complexity: “Dependent” Models



The Circle of Modeling!!!



Conclusions

- The Comptox Chemicals Dashboard provides lots of predicted data
- There are issues in our “on the fly” prediction tools
- Versioning, history and clear documentation of the entire “modeling workflow” are necessary to prevent confusion
- Web tools and technologies are the answer to providing robust prediction services through the Dashboard

Questions?