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

Christopher Grulke[1]
Antony Williams[1]
Amar Singh[1]
Jeff Edwards[1]

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

Predicted

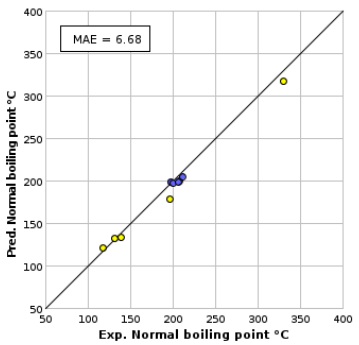
Download Predicted Data

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

DETAILS
EXECUTIVE SUM
PROPERTIES
ENV
HAZ
ADM
EXP
BIO
SIM
GEN
REL
SYN
LITE
LIN
CON

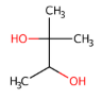
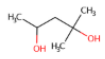
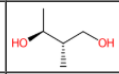
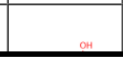
12

Prediction results (colors defined in table below)



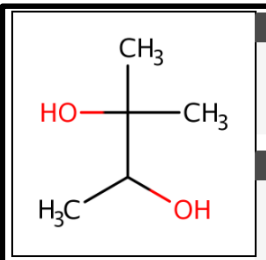
Chemicals	MAE*
Entire set	11.46
Similarity coefficient ≥ 0.5	6.68

*Mean absolute error in Â°C

CAS	Structure	Similarity Coefficient	Experimental value Â°C	Predicted value Â°C
NOCAS_871126 (test chemical)			175.00	179.54
107-41-5		0.88	198.00	199.20
684-84-4		0.87	200.00	196.96
				

	Predicted average
	2.66
	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

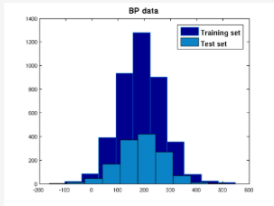
HAZARD
ADME
EXPOSURE
BIOACTIVITY
SIMILAR COM
GENRA (BETA
RELATED SUB

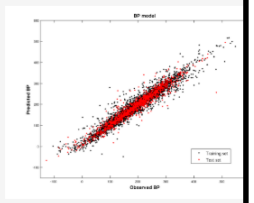


Model Results

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

Model Performance



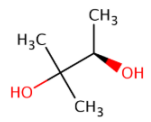


QMRF

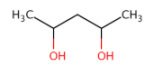
Weighted KNN model

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

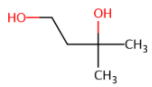
Nearest Neighbors from the Training Set



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



2,4-Pentanediol
Measured:199
Predicted: 199.144



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/cm³	1.248 g/cm³
Surface tension		
Thermal conductivity		158.367 mW/mK

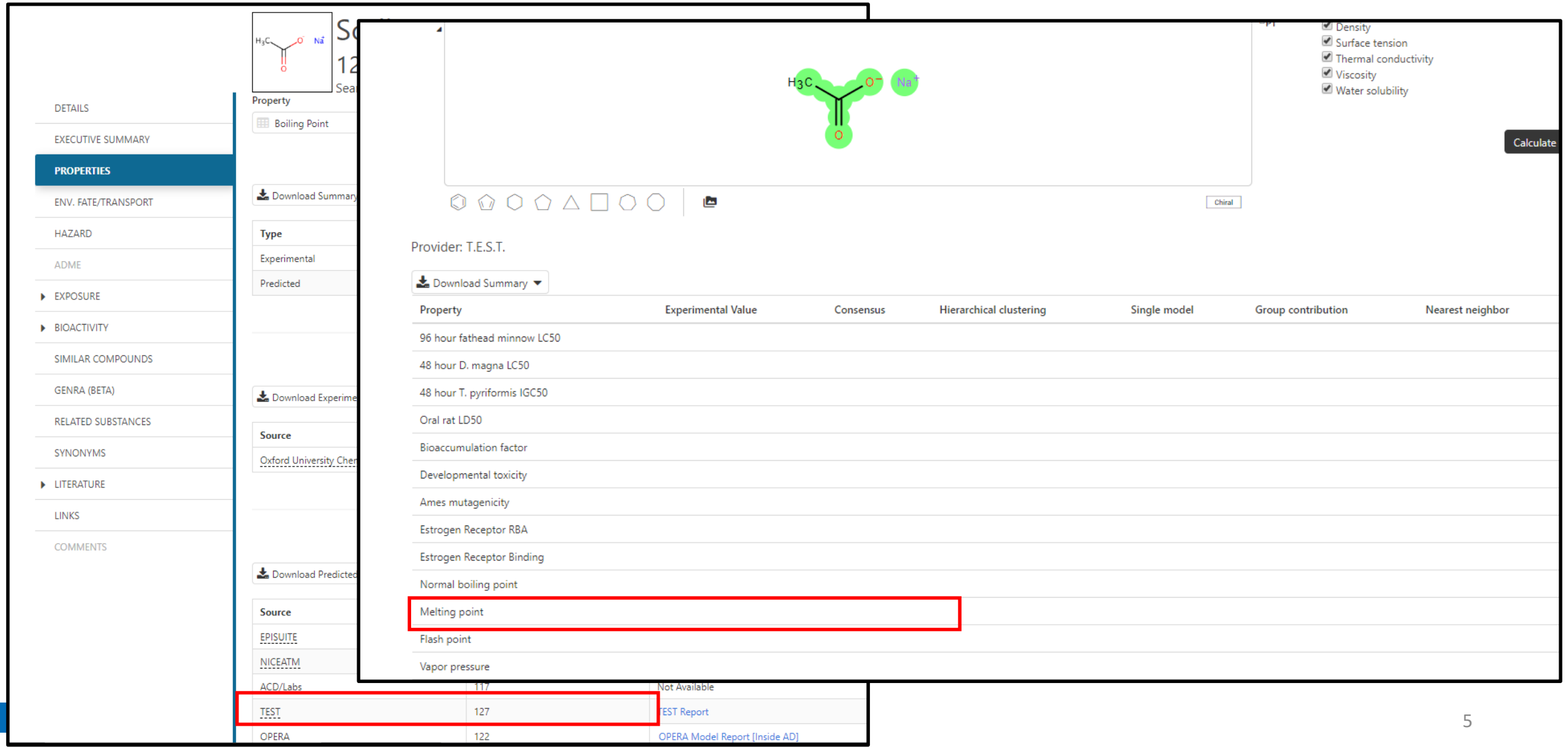
Select properties to predict

T.E.S.T.

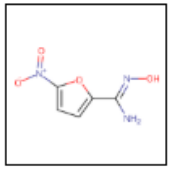
- Toxicological properties
  - 96 hour fathead minnow LC50
  - 48 hour D. magna LC50
  - 48 hour T. pyriformis IGC50
  - Oral rat LD50
  - Bioaccumulation factor
  - Developmental toxicity
  - Ames mutagenicity
  - Estrogen Receptor RBA
  - Estrogen Receptor Binding
- Physical properties
  - Normal boiling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

Calculate

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



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



5-Nitro-2-furamidox

772-43-0 | DTXSID00209

Searched by DSSTox Compound Id.

- DETAILS
- EXECUTIVE SUMMARY
- **PROPERTIES**
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS
- COMMENTS

Property

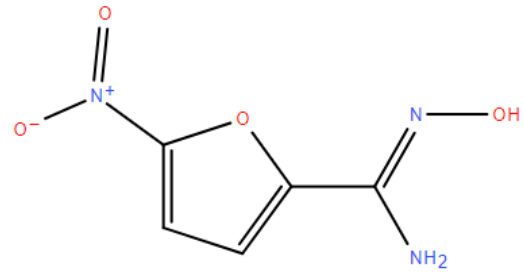
Flash Point

Download Summary

Type	Aver
Experimental	-
Predicted	84.4

Download Predicted Data

Source	Result
ACD/Labs	86.1
TEST	82.7



Provider: T.E.S.T.

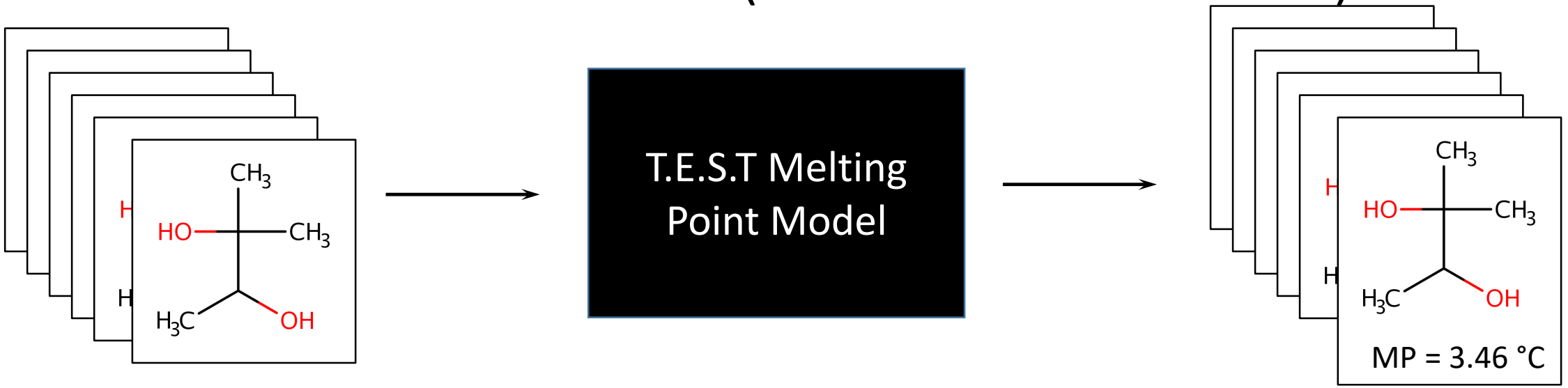
Download Summary

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

R. Marquand *et al*. 1986. *Return of the Jedi*. United States: Twentieth Century-Fox Film Corporation

Model Prediction (for non-modelers)



EXPECTATIONS

It works

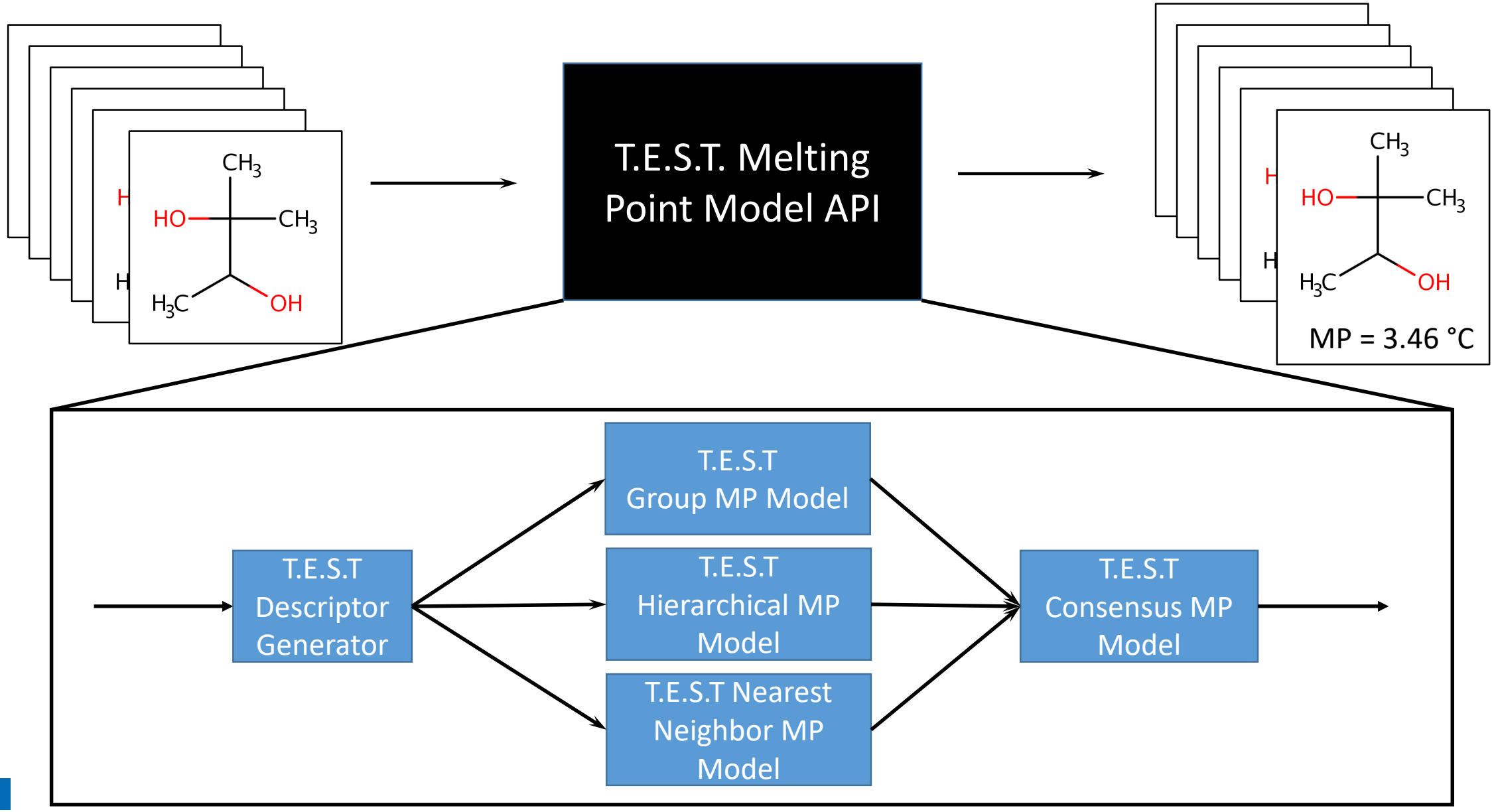
If it fails, it tells you why

Reproducible

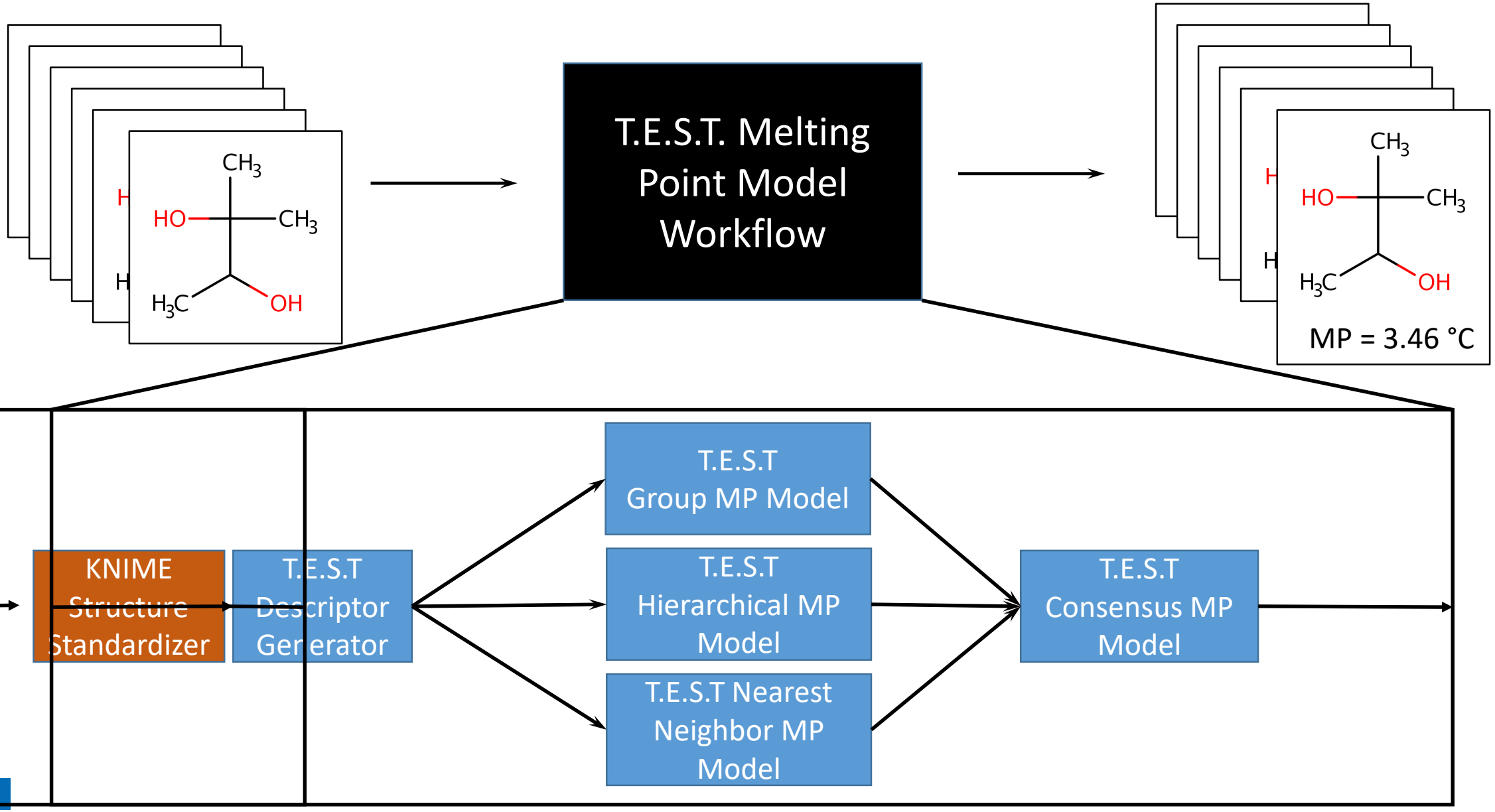
Same Chemical, Same Result

Changes to the model are clearly conveyed

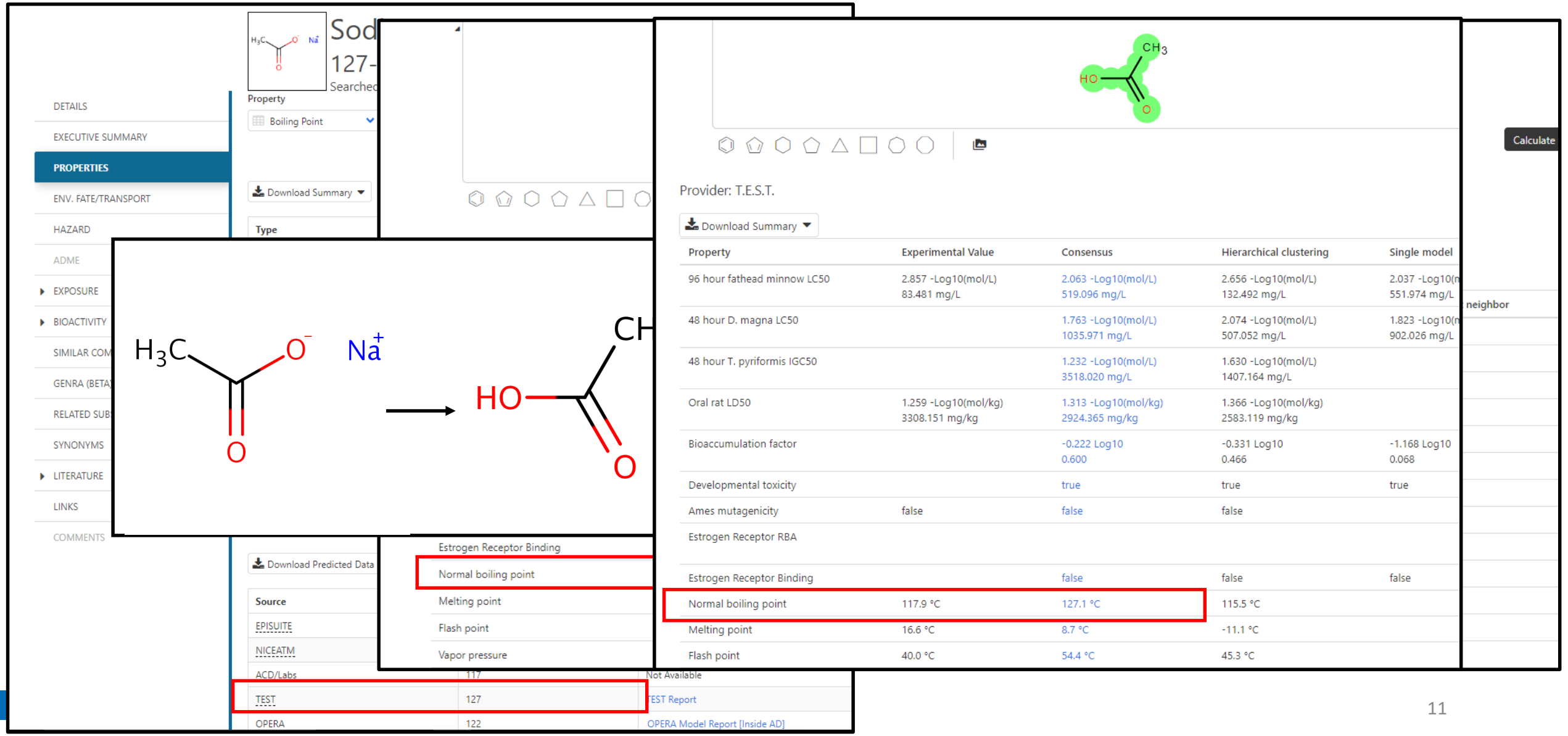
Model Prediction (T.E.S.T.)



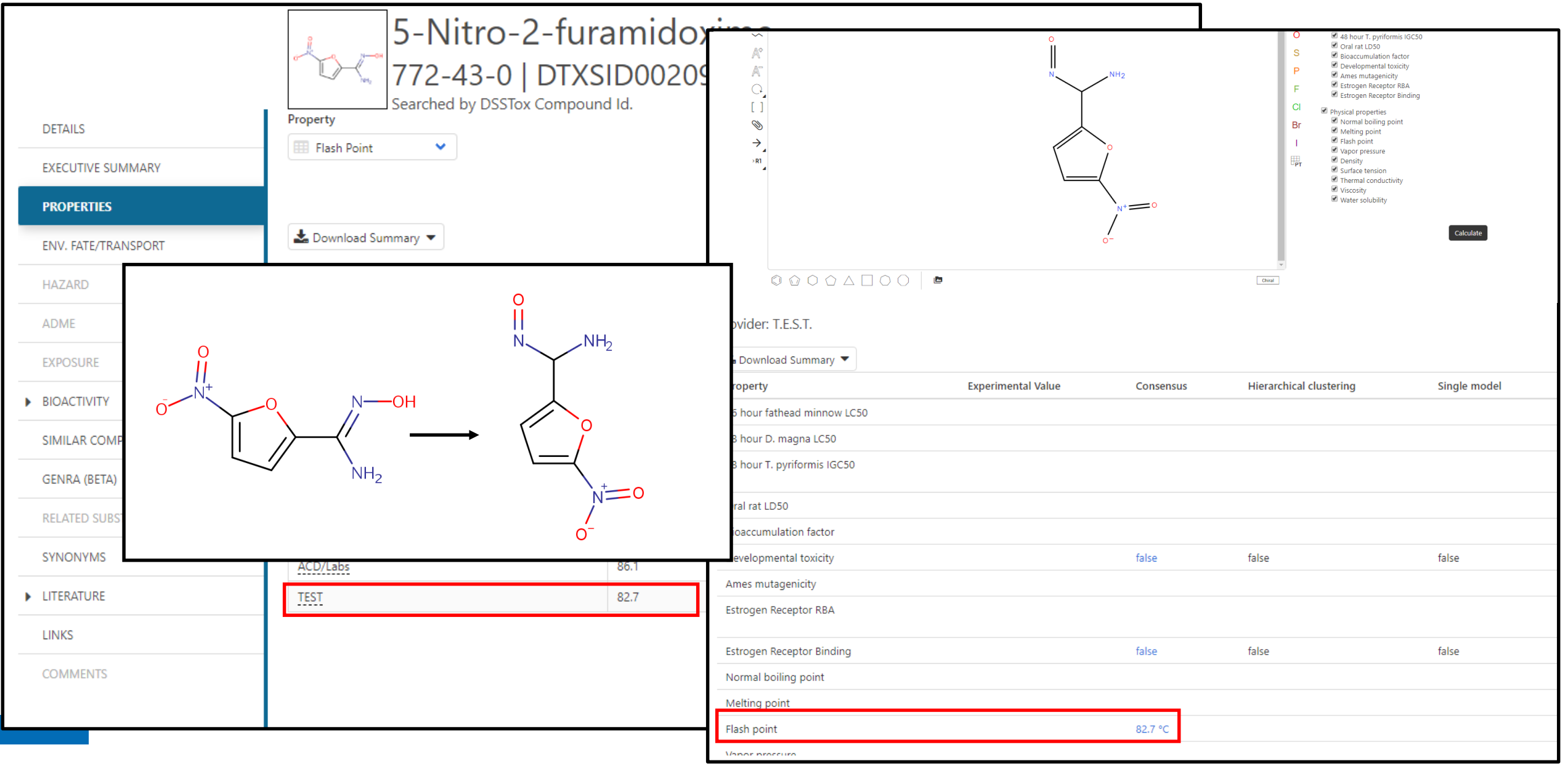
Stored Model Predictions (T.E.S.T.)



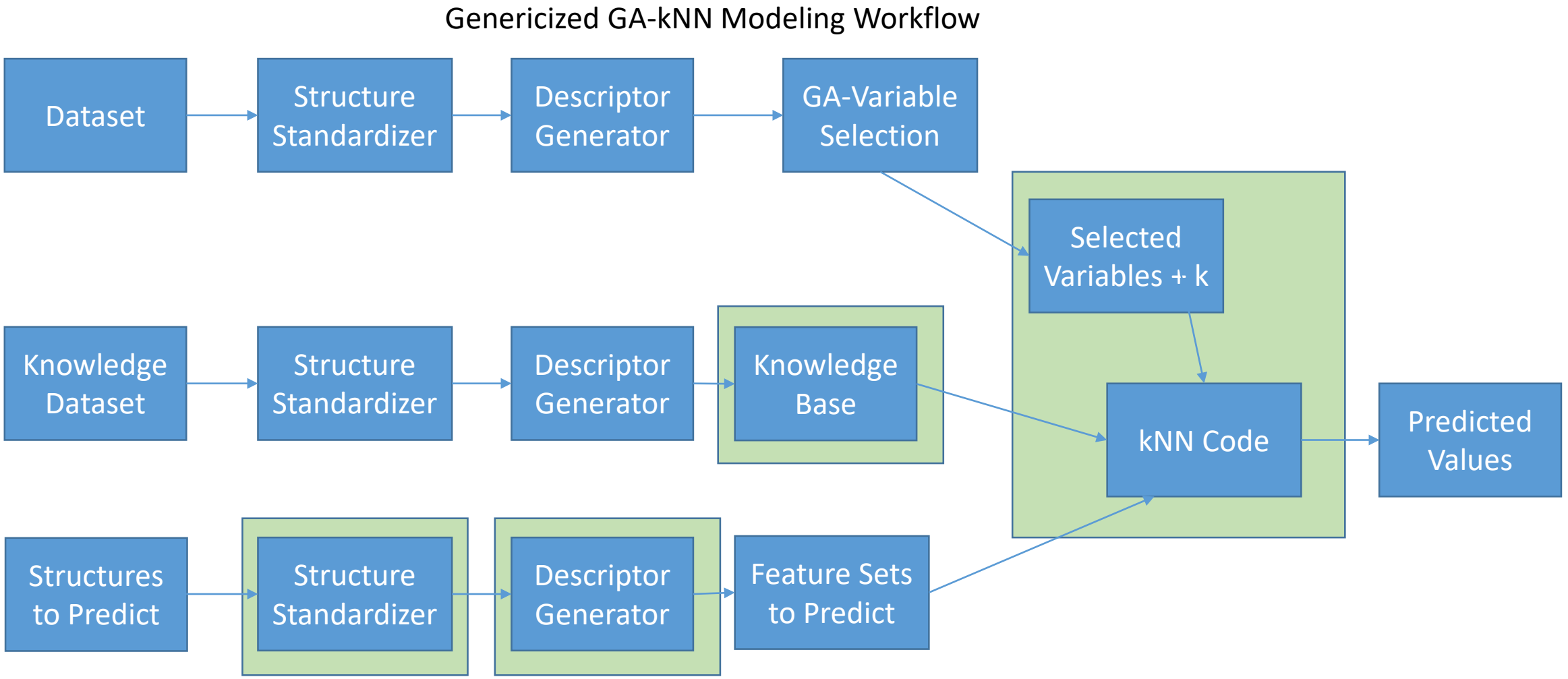
Known Issues: Prediction of Salts in T.E.S.T.



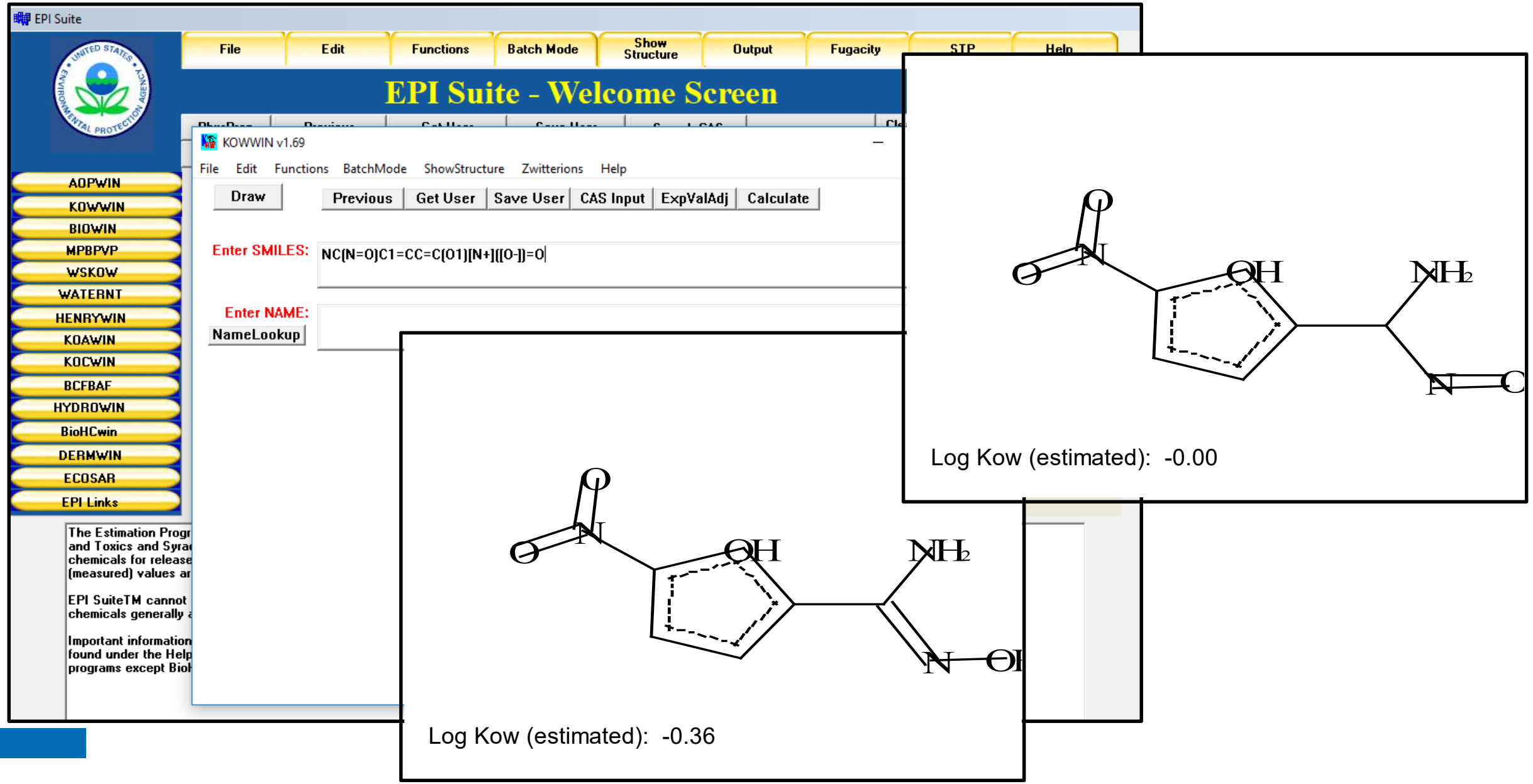
Known Issues: Prediction of Tautomers in T.E.S.T.



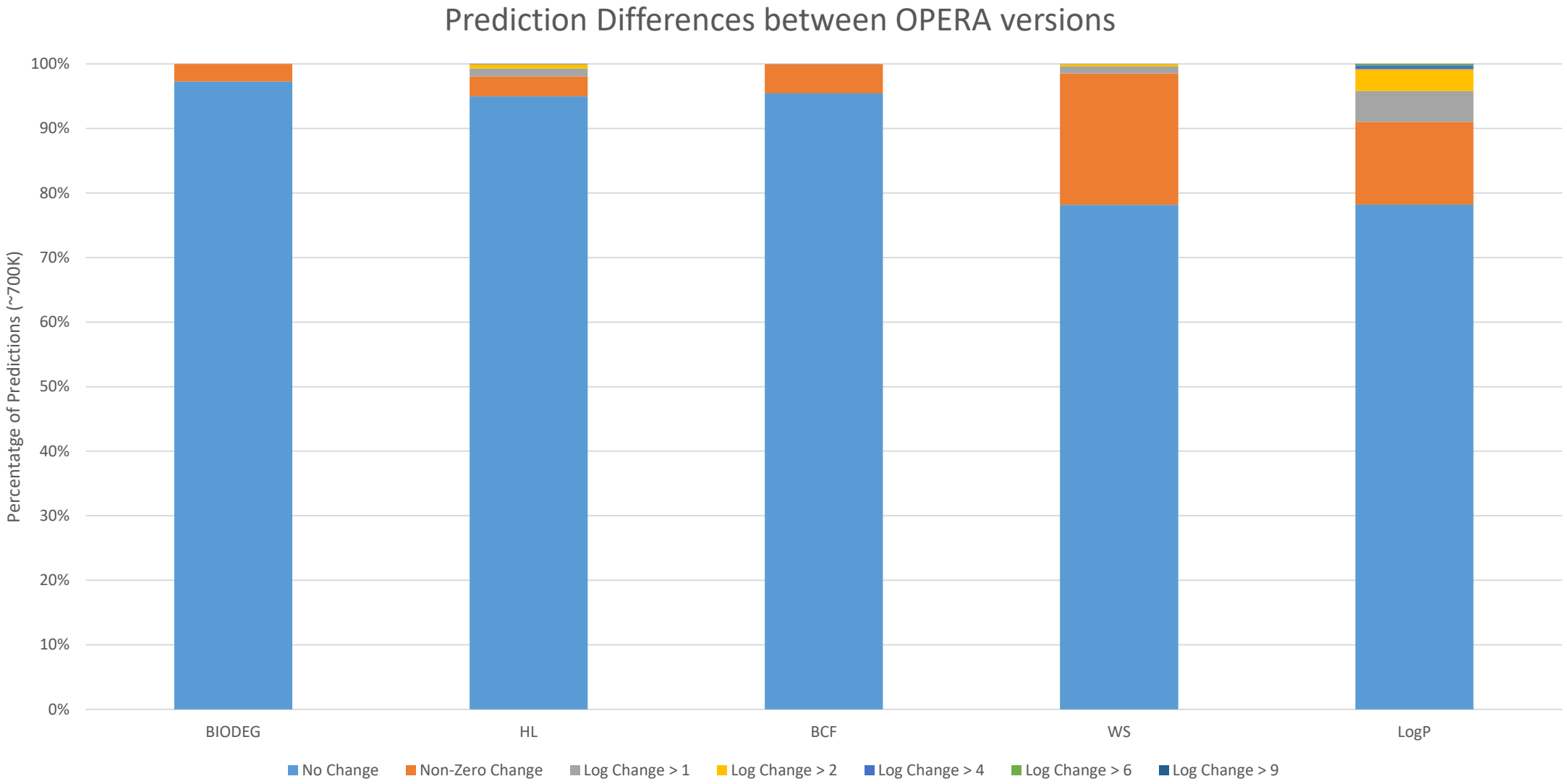
What is the Model?



Known Issues: EpiSuite Tautomers



Known Issues: OPERA 2.0 Released

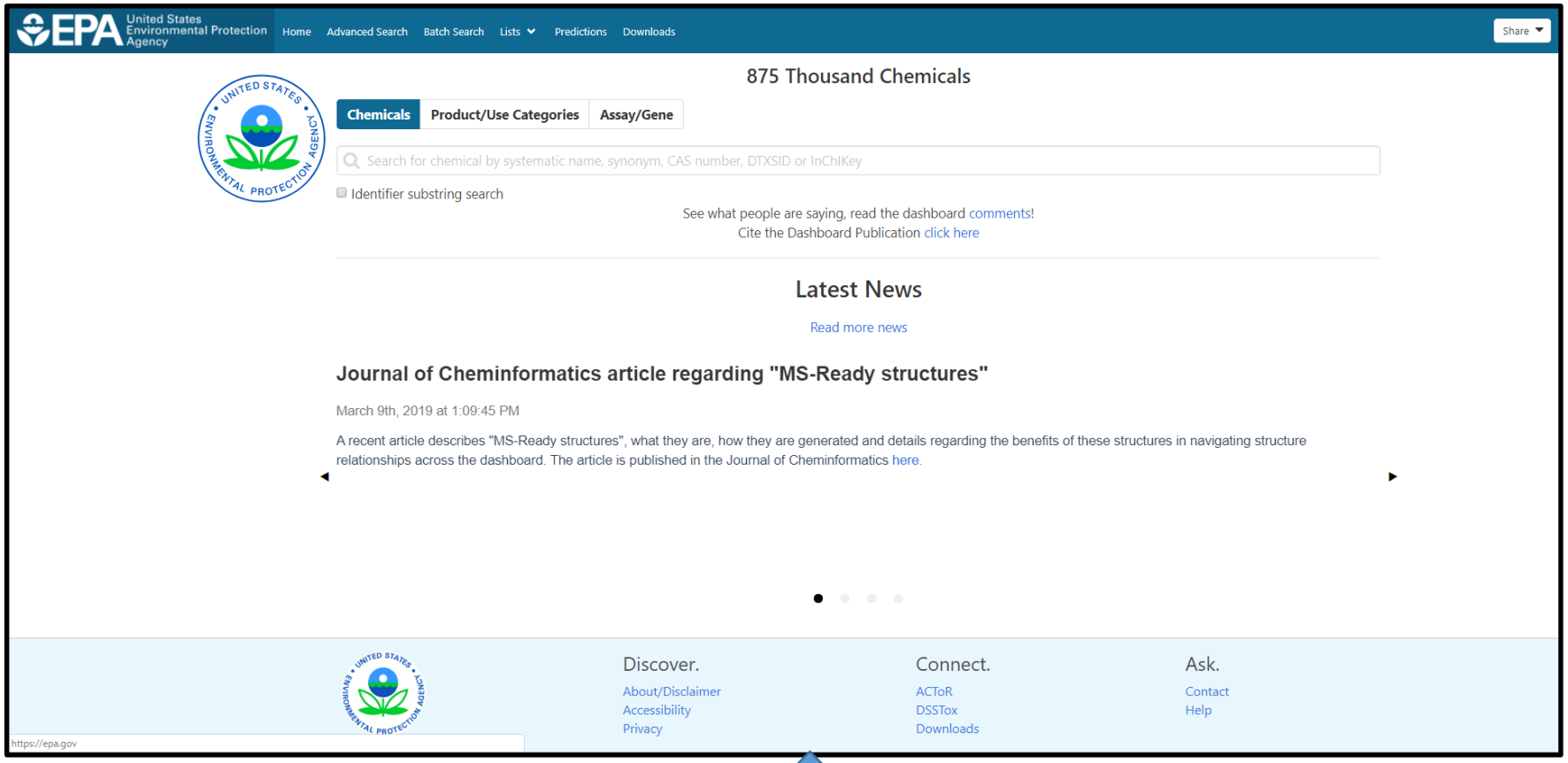


Standardization Improvements

Knowledge Base Updated


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

Current Dashboard Structure



ACToR WS

Ruby on Rails

Ruby on Rails API

ChemTrack

DSSTox

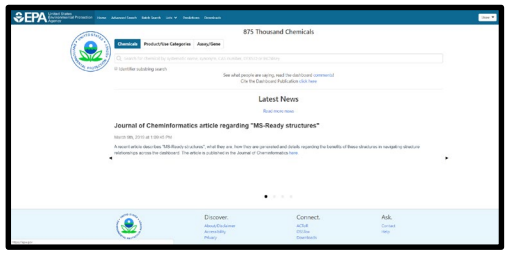
ChemProp

InVitroDB

ToxRef DB

CPDat

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)

ORD Data Lake

Data Analytics/ Learning

Generalized data model providing integration across the domains for scientific exploration

Integration Layer

ChemTrack App
ChemTrack

ChemReg Application
DSSTox
ChemProp

TCPL R Package
InVitroDB

Toxicity QC App
ToxRef DB

Facto
CPDat

Model + Pred App
Models

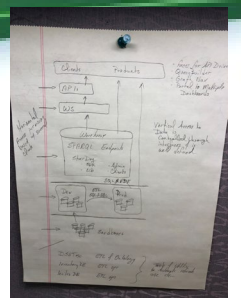
Transactional Layer

Raw data processing and data curation, quality tools

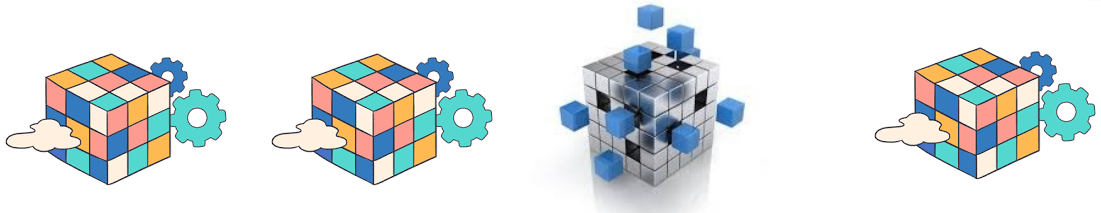
High level Information/Integration Architecture

(30K feet)

Draft



Fit-for-purpose technologies



Application/Analytics Layer

Virtual Databases (VDB's) – using Data Virtualization technology and some may be physical data especially high volume

Presentation Layer
(Fit-for-Purpose)

Chemistry BioAssays Chemical Exposure Risk* GenRA LitDb

Fit-for-purpose technologies (technology ecosystem)

Integration Layer

Data orchestration using Pentaho (PDI) / or other technologies

ChemTrack ChemReg ChemProp Unstructured/ structured Pubmed . . . .

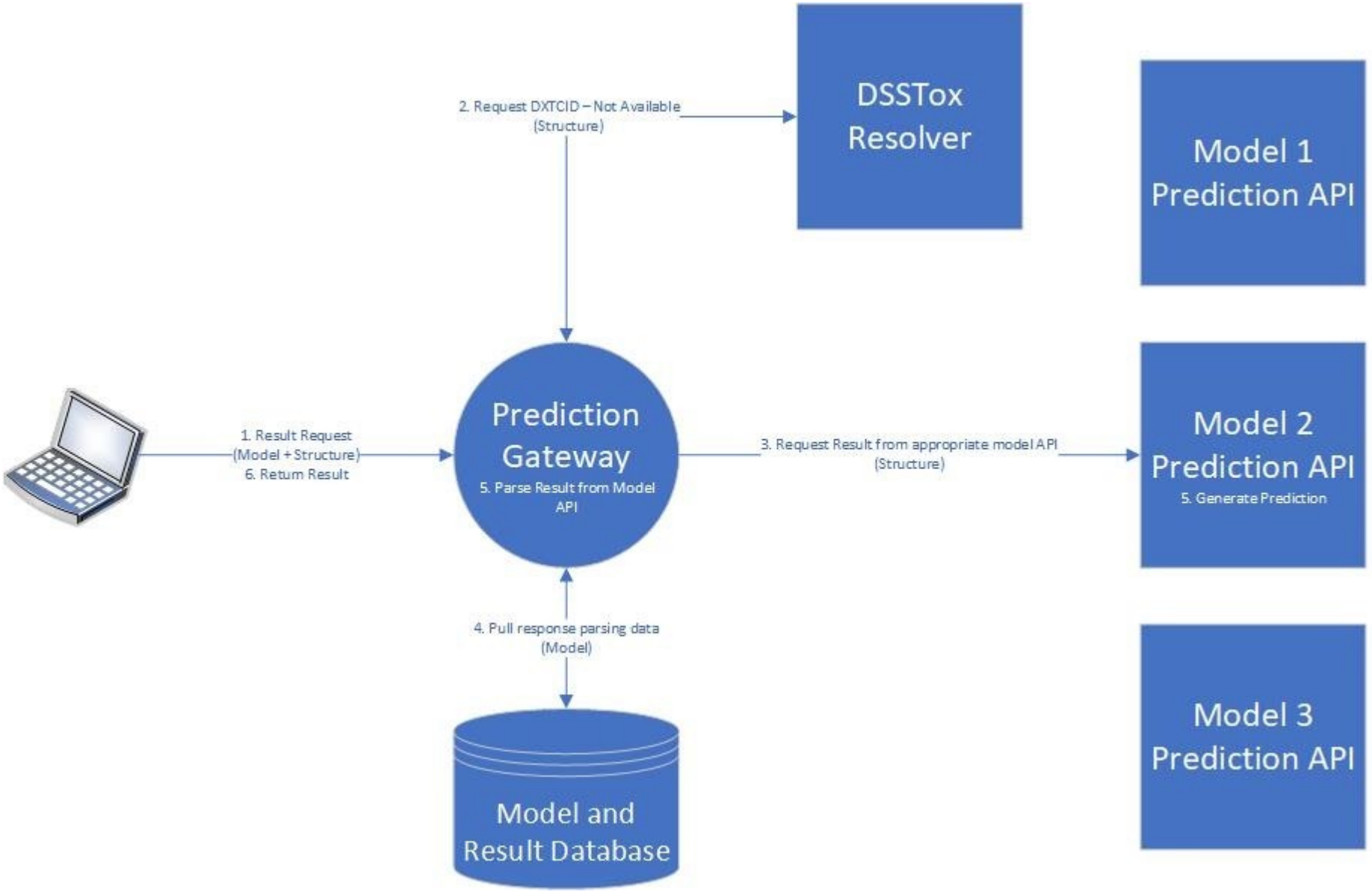
Data Collection Layer

Raw data processing and data curation, quality tools

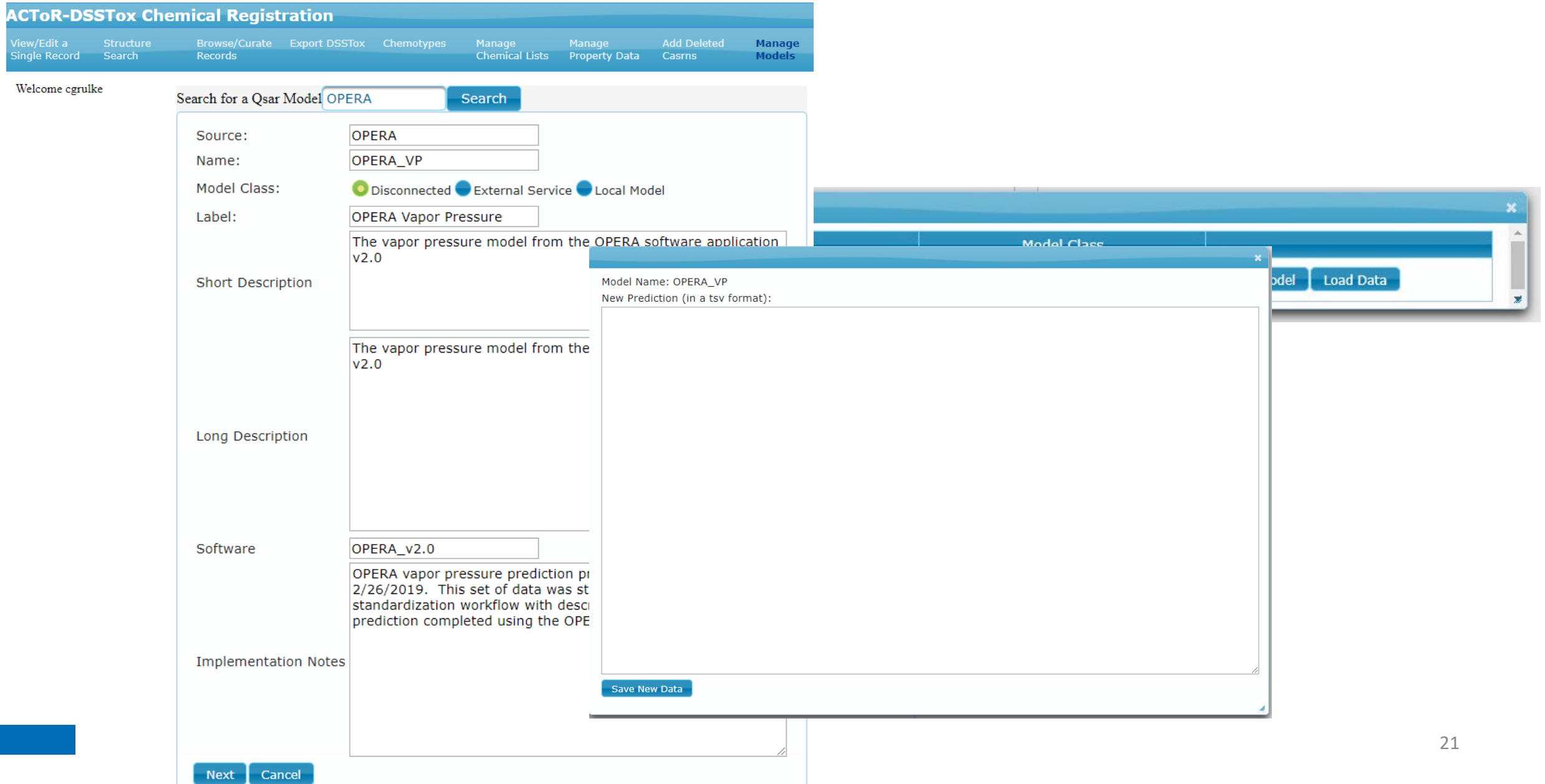
Contribution: Jeff Edwards and Amar Singh

* Decision that can be stored with consultations with key stakeholders

Model Prediction Gateway



Model registration interface



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

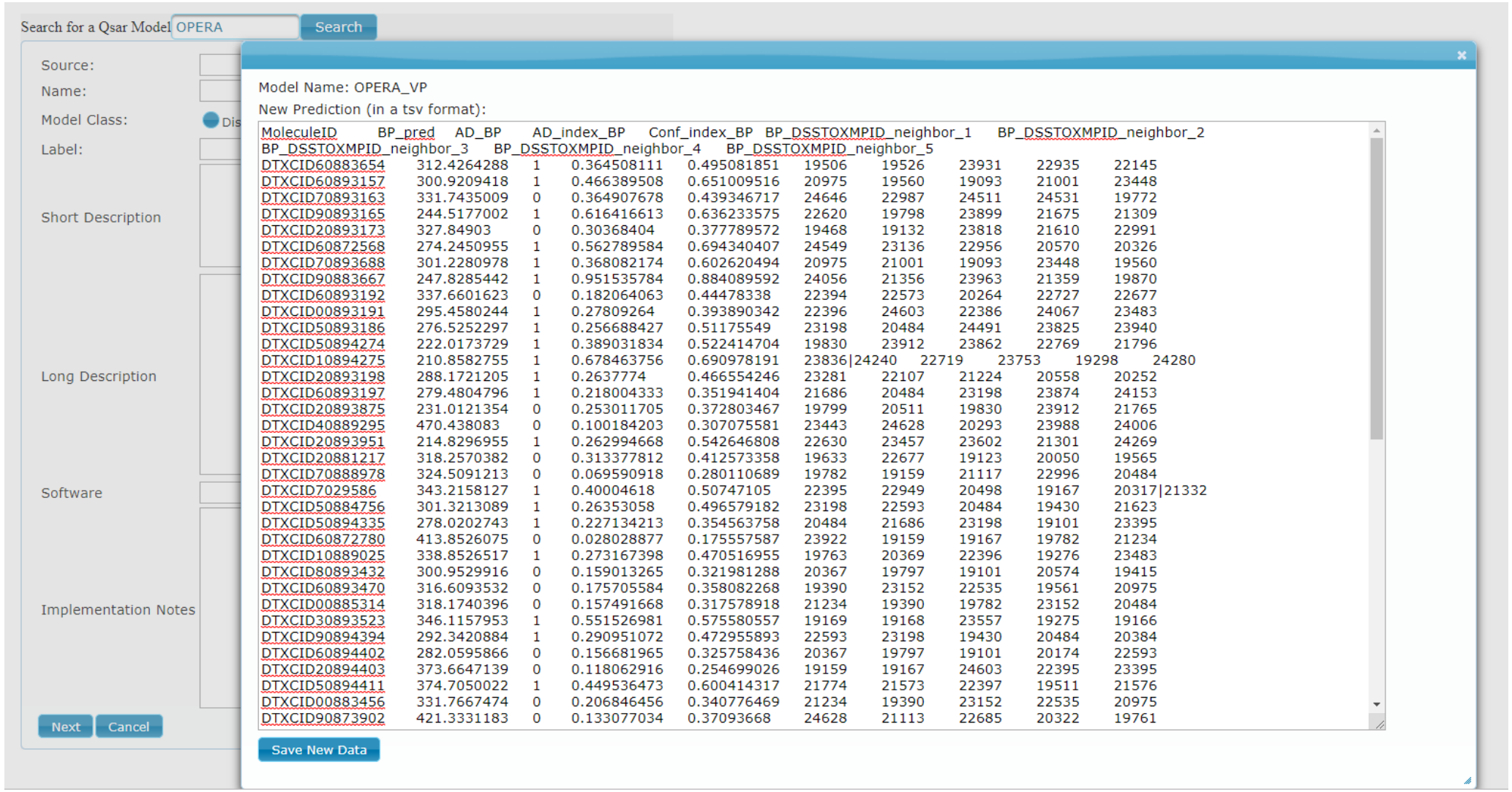
e.g., T.E.S.T Models

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



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

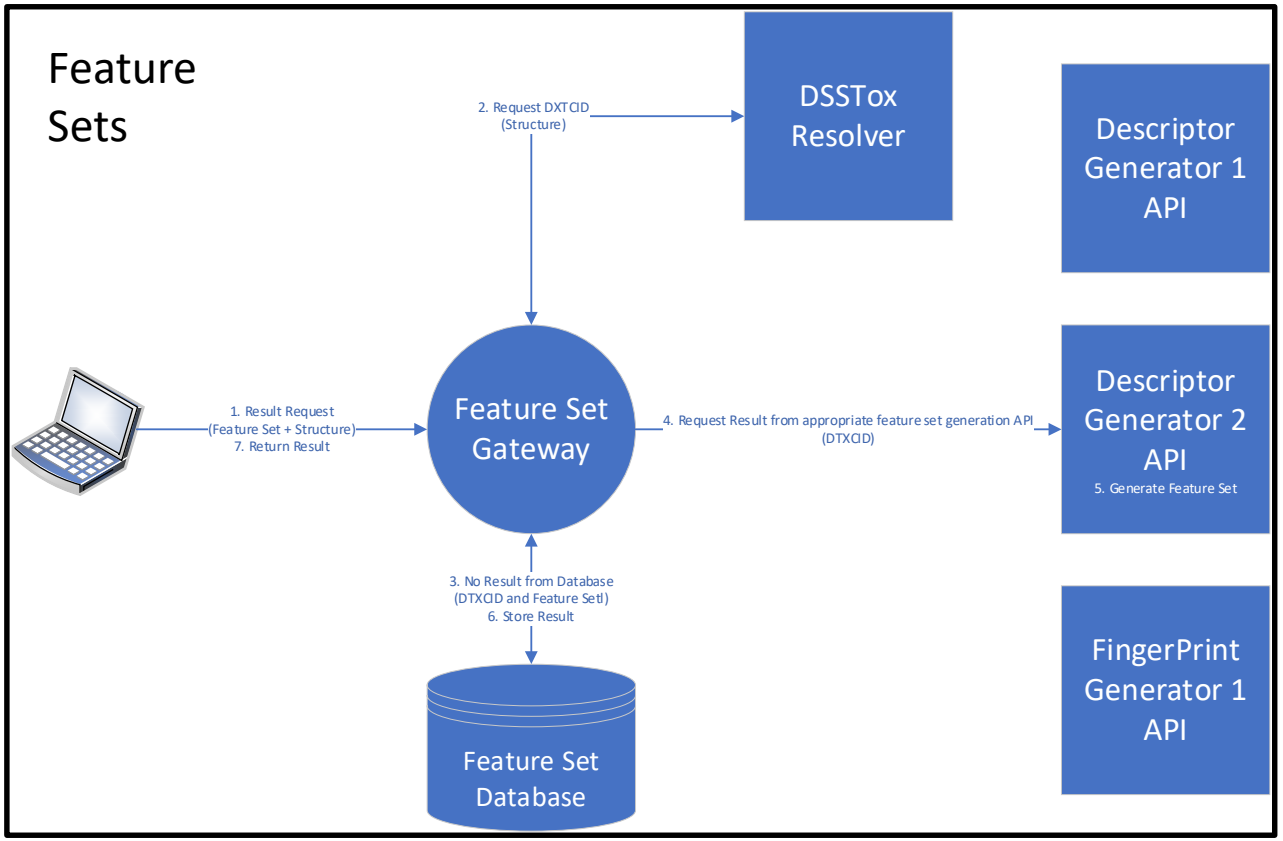
e.g., T.E.S.T Models

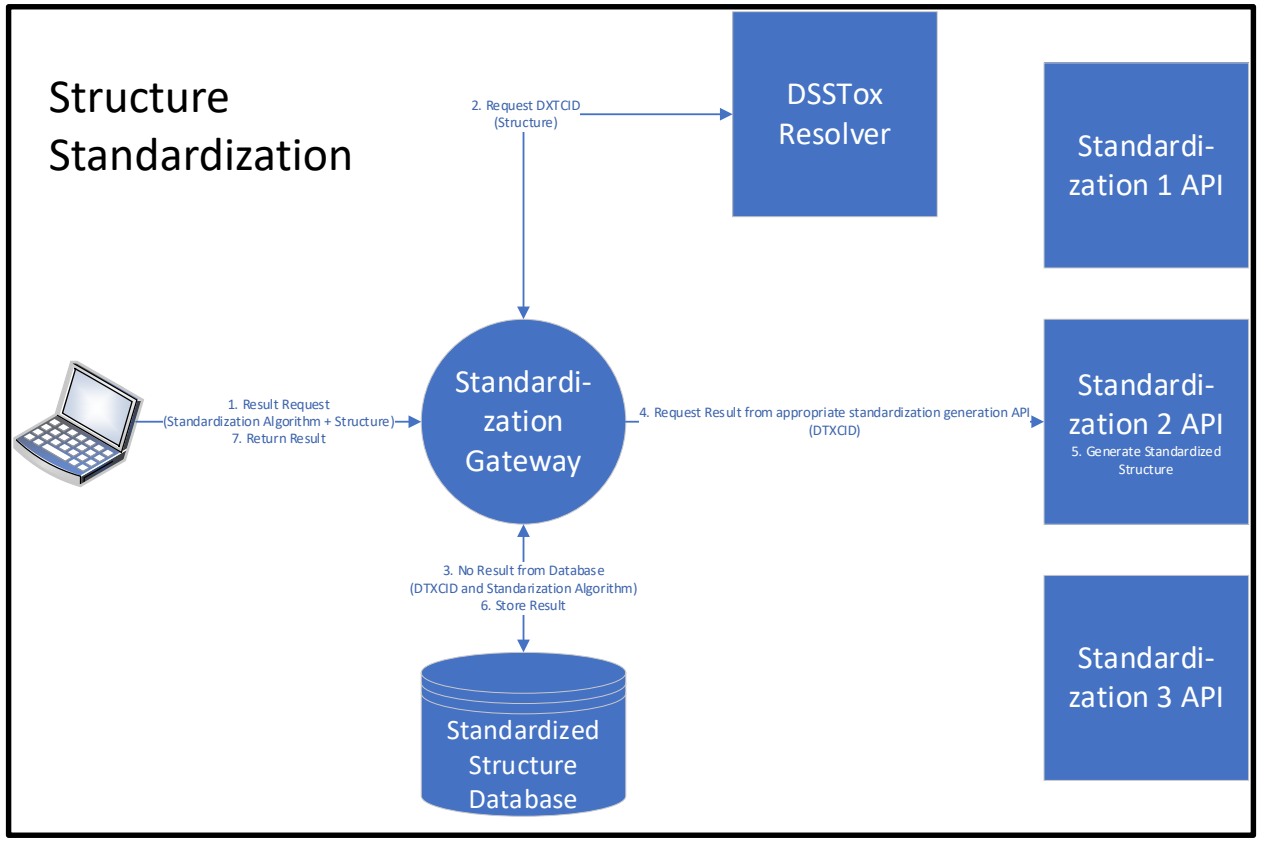
Local Model

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

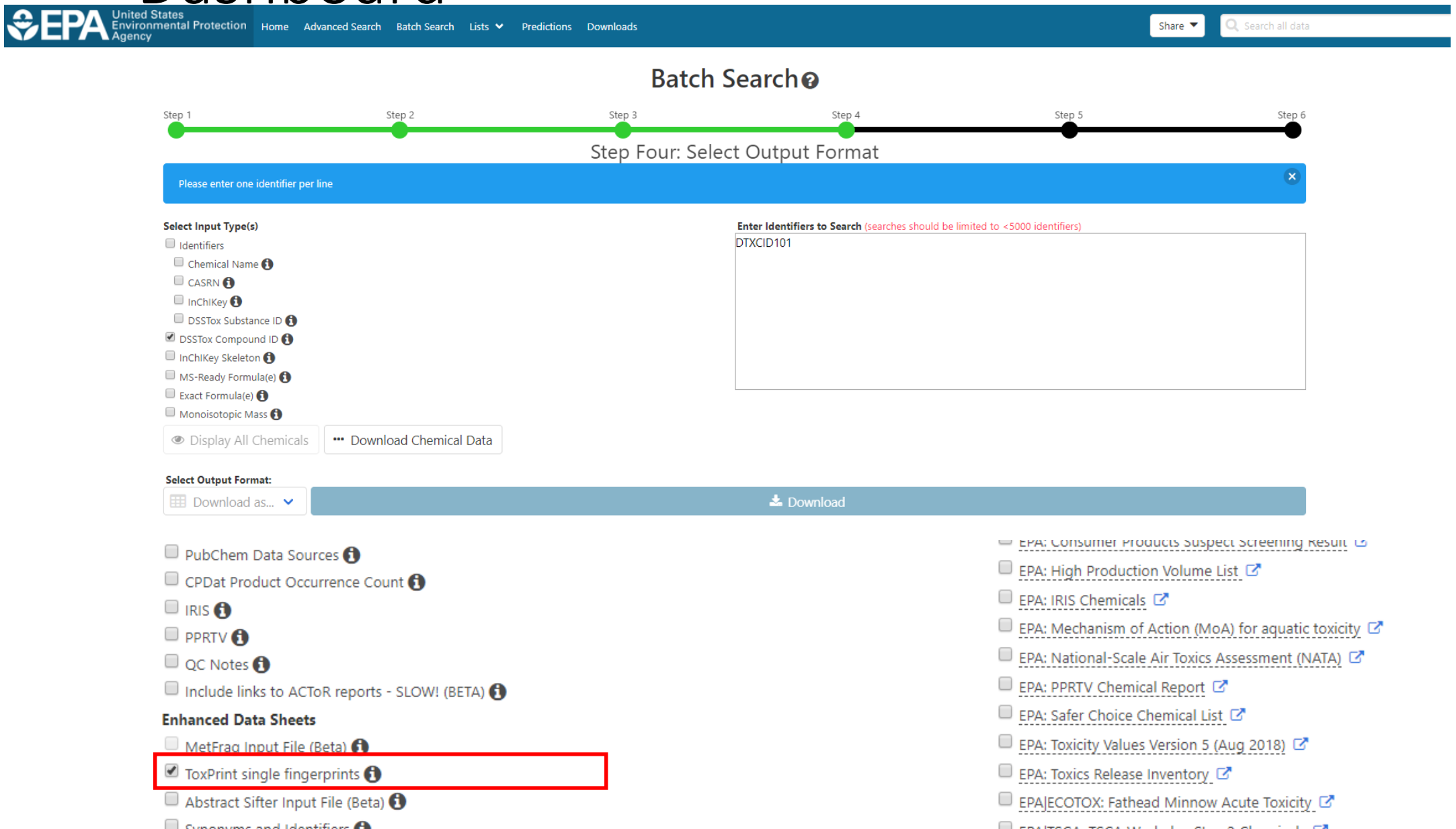
e.g., None

Supporting Service Gateways

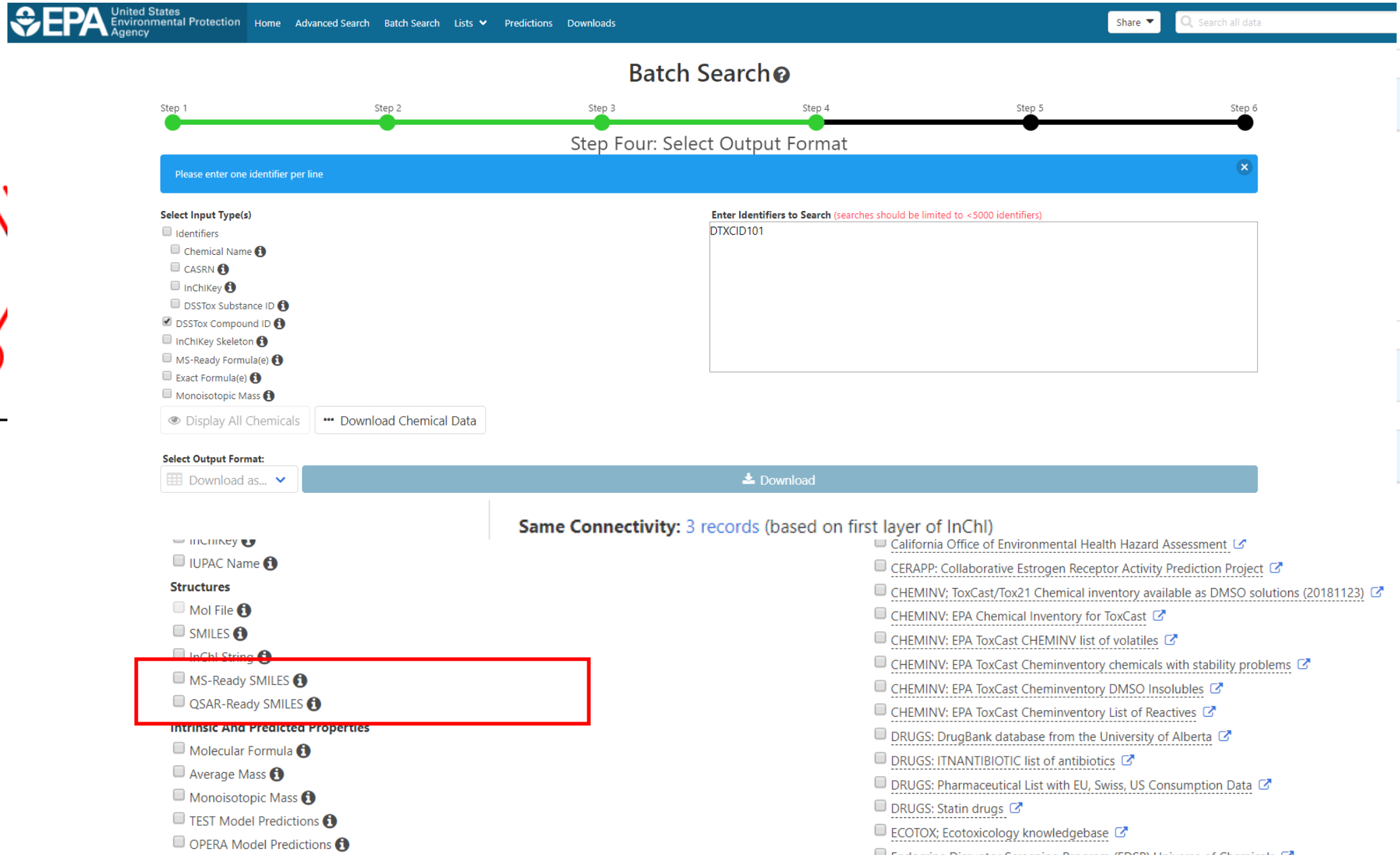




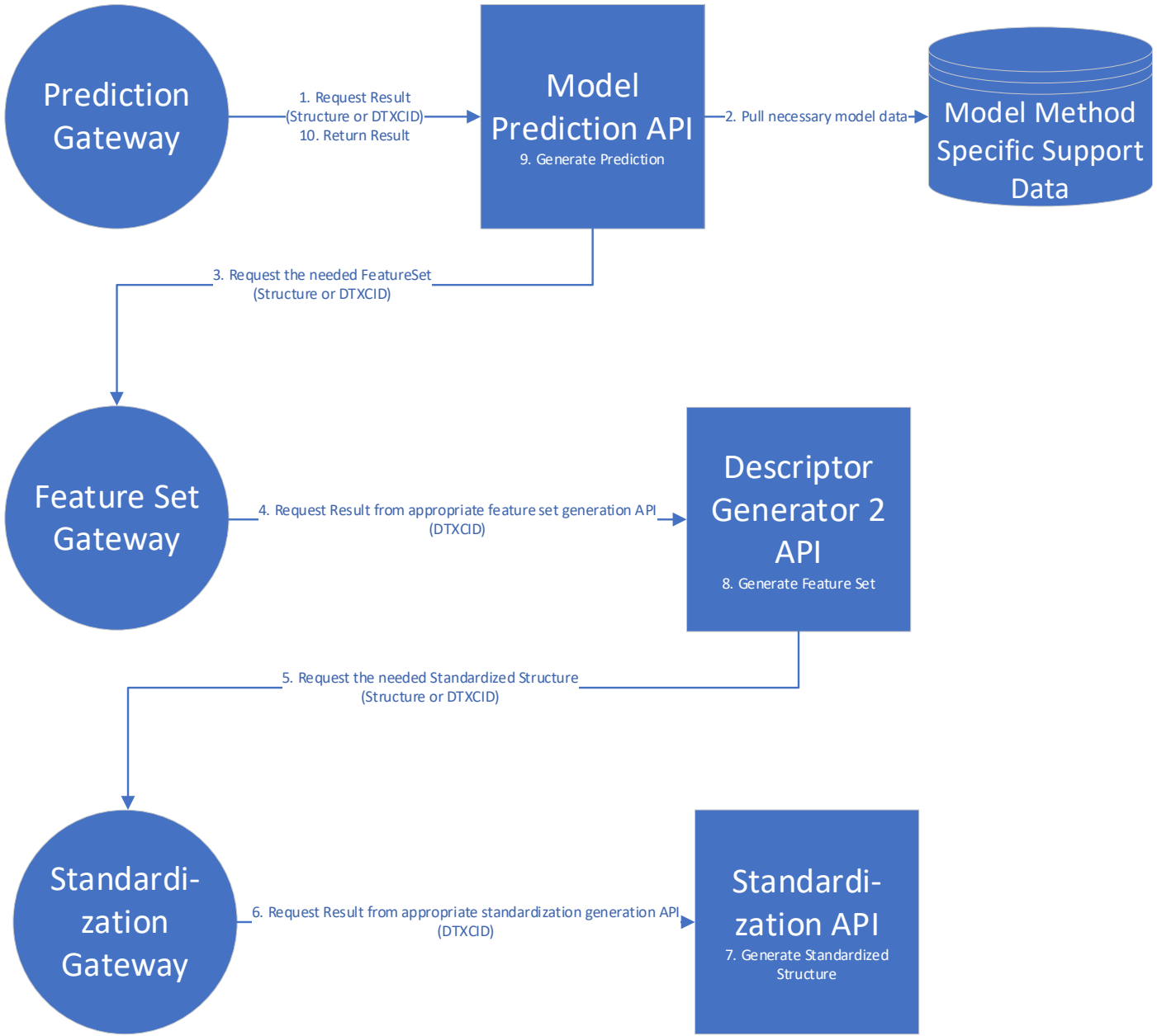
Descriptor and Fingerprint Availability in Dashboard



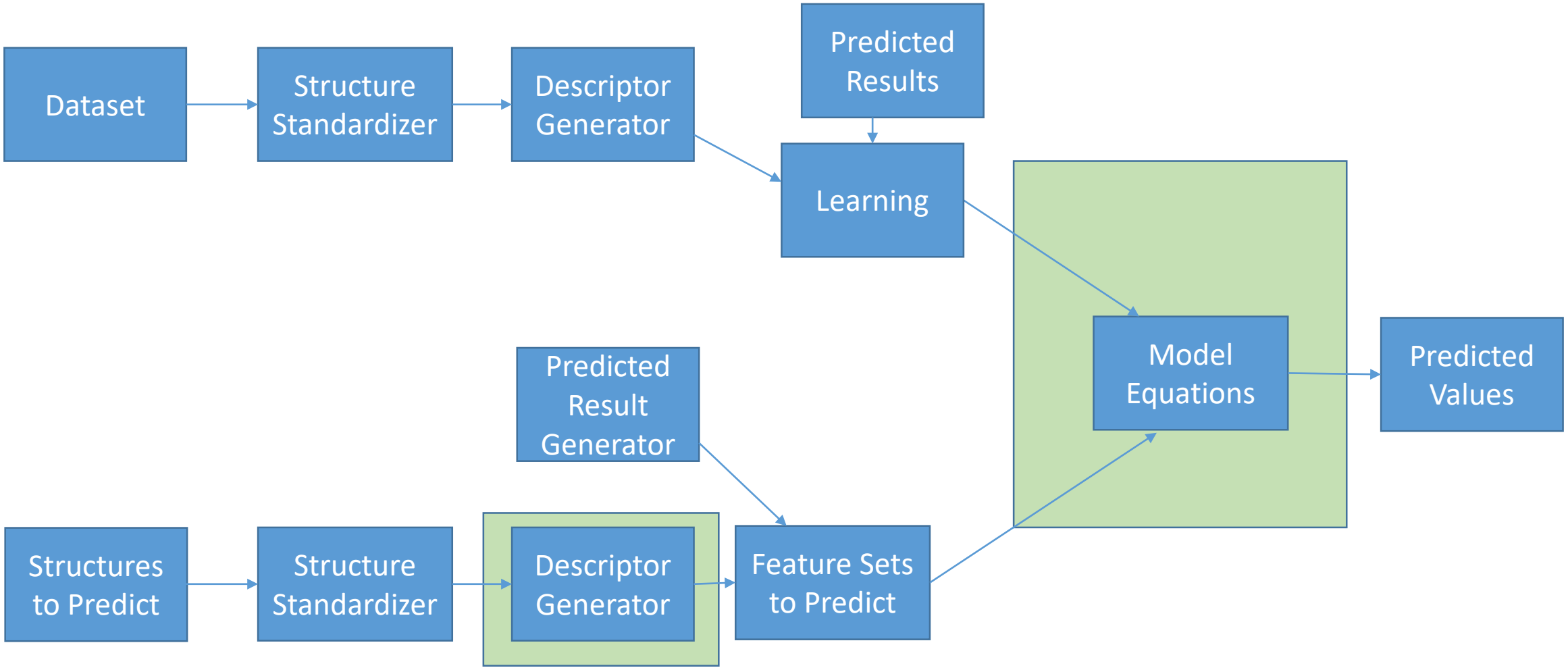
Standardization in the Dashboard



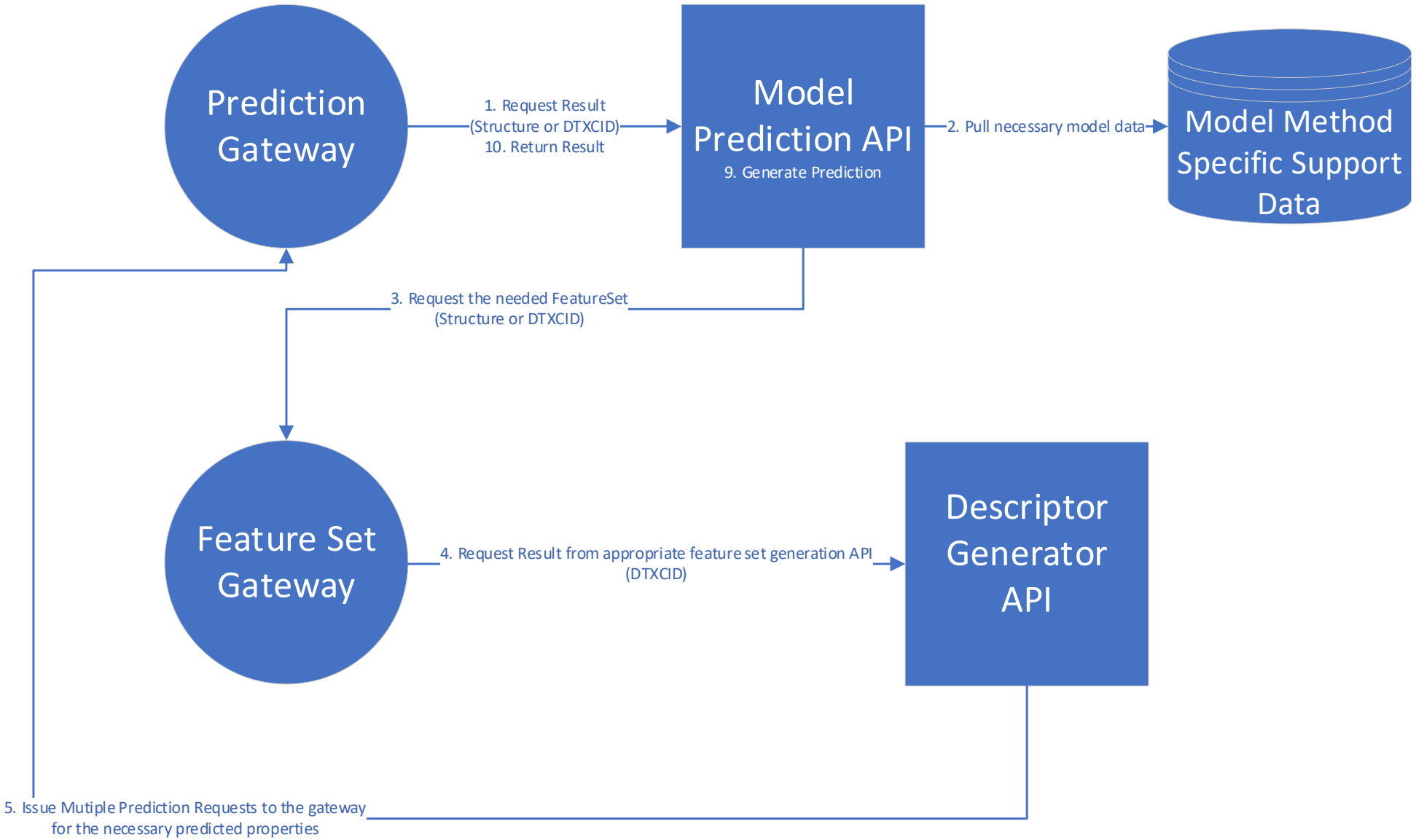
Local Model Services



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

EPA United States Environmental Protection Agency

Questions?