U.S. FOOD & DRUG ADMINISTRATION

precisionFDA

Crowdsourced evaluation of InChI-based tautomer identification

Yulia Borodina, Ph.D.
Cheminformatician
Health Informatics,
Office of Data, Analytics, and Research (ODAR)

Disclaimer

The views and opinions presented here represent those of the speaker and should not be considered to represent advice or guidance on behalf of the Food and Drug Administration.

This initiative was made possible due to the efforts of our late friend, the analytical chemist and InChI developer Igor Pletnev

Moscow, July 2021

Outline

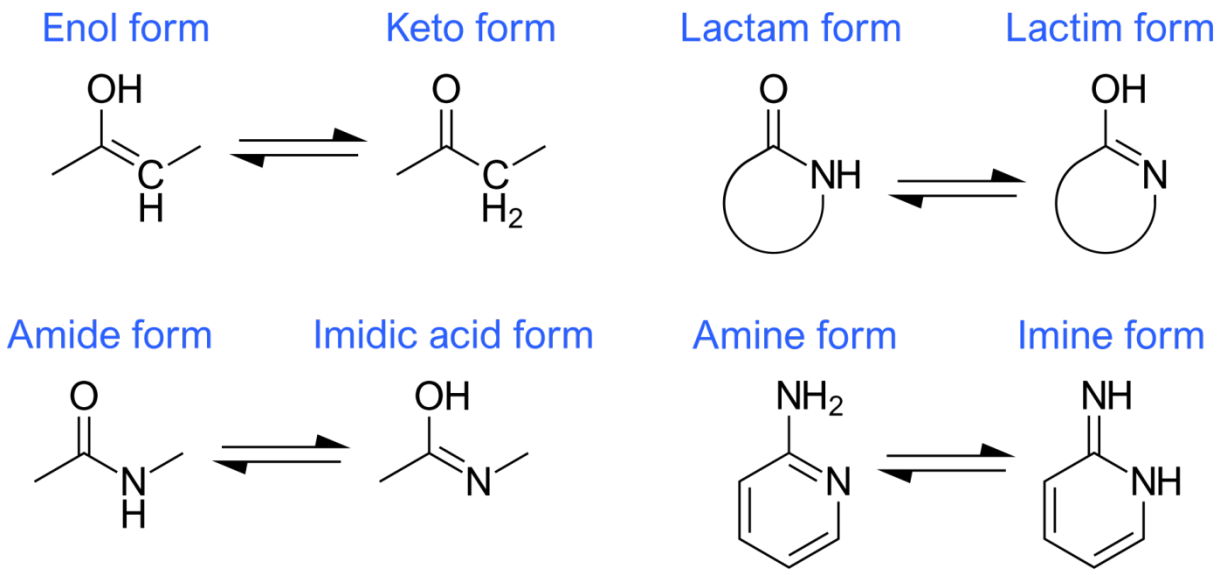
- Background
- Ideation and design
- Information for participants

Outline

- **Background**
- Ideation and design
- Information for participants

Scientific Facts

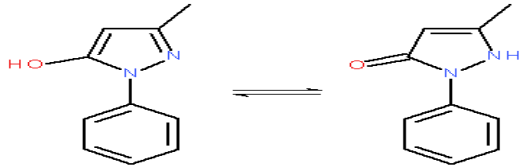
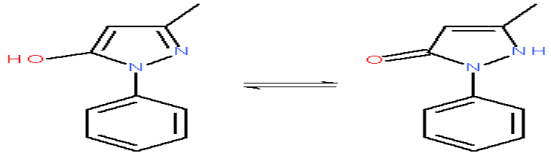
- Tautomers are structural isomers of chemical compounds that readily interconvert
- Conversion commonly results from the relocation of a hydrogen atom within the compound



https://en.wikipedia.org/wiki/Tautomer

InChI Facts

- The International Chemical Identifier (InChI) is a community **standard** for encoding molecular structures. Its hashed version is called InChIKey.
- The IUPAC Tautomer Working Group suggested a number of **modifications** to the InChI algorithm that would recognize more molecules that are tautomers of each other
- Some of the suggested modifications have been implemented by the InChI Trust in an **experimental** version of InChI
- There was no thorough testing of how well this experimental version of InChI (**Tauto InChI**) identifies tautomers along experimental results

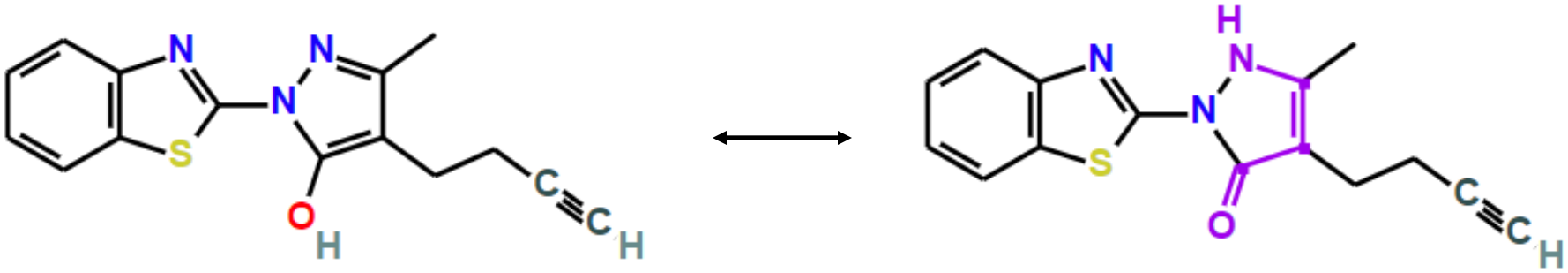
Experimental options of InChI program (RULES)	Description	Example
KET	keto-enol tautomerism recognized	
15T	1,5 H-transfer recognized	
PT_06_00	1,3 heteroatom H shift recognized	
PT_13_00	keten/ynol exchange recognized	
PT_16_00	nitroso/oxime tautomerism recognized	
PT_18_00	cyanic/iso-cyanic acids tautomerism recognized	
PT_22_00	imine via imine tautomerism recognized	
PT_39_00	nitrone/azoxy or Behrend rearrangement recognized	

Why identification of tautomers matters for FDA

- For regulation of therapeutic use and surveillance of side effects, molecules that are tautomers should be identified as the same substance
- Regulatory submissions may report one or another tautomeric isoform of a substance. This can make it difficult to find relationship between different submissions.
- Computational approaches (such as the InChI algorithm) could be used for tautomer identification

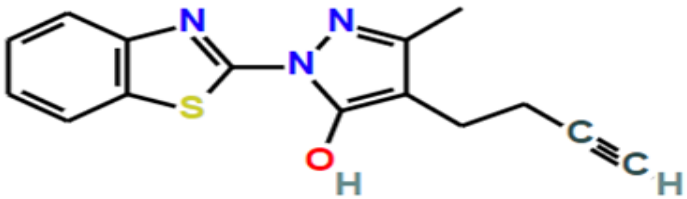
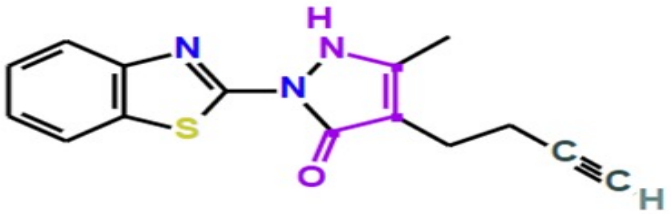
InChI can help avoid this

Tautomers sold as different products by same vendor



Identity of molecules demonstrated by NMR analysis

Guasch et al., J. Chem. Inf. Model. 2016, 56, 2149−2161

Rule set ID	Enabled options of InChI program		
1S		AZQCGJRMKGXAFT-UHFFFAOYSA-N	RVRBSNHQOVTGKP-UHFFFAOYSA-N
1	/KET	RSGMBYBMGOMIRN-UHFFFAOYNA-N	RVRBSNHQOVTGKP-UHFFFAOYSA-N
2	/15T	VWJFHHOWKWQPJY-UHFFFAOYNA-N	VWJFHHOWKWQPJY-UHFFFAOYNA-N
3	/PT_06_00	PFWUXVYFMJQDQF-UHFFFAOYNA-N	PFWUXVYFMJQDQF-UHFFFAOYNA-N
4	/PT_13_00	AZQCGJRMKGXAFT-UHFFFAOYNA-N	RVRBSNHQOVTGKP-UHFFFAOYNA-N
5	/PT_16_00	AZQCGJRMKGXAFT-UHFFFAOYNA-N	RVRBSNHQOVTGKP-UHFFFAOYNA-N
6	/PT_18_00	AZQCGJRMKGXAFT-UHFFFAOYNA-N	RVRBSNHQOVTGKP-UHFFFAOYNA-N
7	/PT_22_00	AZQCGJRMKGXAFT-UHFFFAOYNA-N	RVRBSNHQOVTGKP-UHFFFAOYNA-N
8	/PT_39_00	AZQCGJRMKGXAFT-UHFFFAOYNA-N	RVRBSNHQOVTGKP-UHFFFAOYNA-N
9	/KET /15T	DXLYIIIJVXCJHT-UHFFFAOYNA-N	DXLYIIIJVXCJHT-UHFFFAOYNA-N
10	/PT_13_00 /PT_16_00 /PT_18_00 /PT_22_00 /PT_39_00	AZQCGJRMKGXAFT-UHFFFAOYNA-N	RVRBSNHQOVTGKP-UHFFFAOYNA-N
11	/PT_06_00 /PT_13_00 /PT_16_00 /PT_18_00 /PT_22_00 /PT_39_00	PFWUXVYFMJQDQF-UHFFFAOYNA-N	PFWUXVYFMJQDQF-UHFFFAOYNA-N
12	/15T /PT_06_00 /PT_13_00 /PT_16_00 /PT_18_00 /PT_22_00 /PT_39_00	PFWUXVYFMJQDQF-UHFFFAOYNA-N	PFWUXVYFMJQDQF-UHFFFAOYNA-N
13	/KET /15T /PT_06_00 /PT_13_00 /PT_16_00 /PT_18_00 /PT_22_00 /PT_39_00	PFWUXVYFMJQDQF-UHFFFAOYNA-N	PFWUXVYFMJQDQF-UHFFFAOYNA-N

Risks of false identification need to be quantified

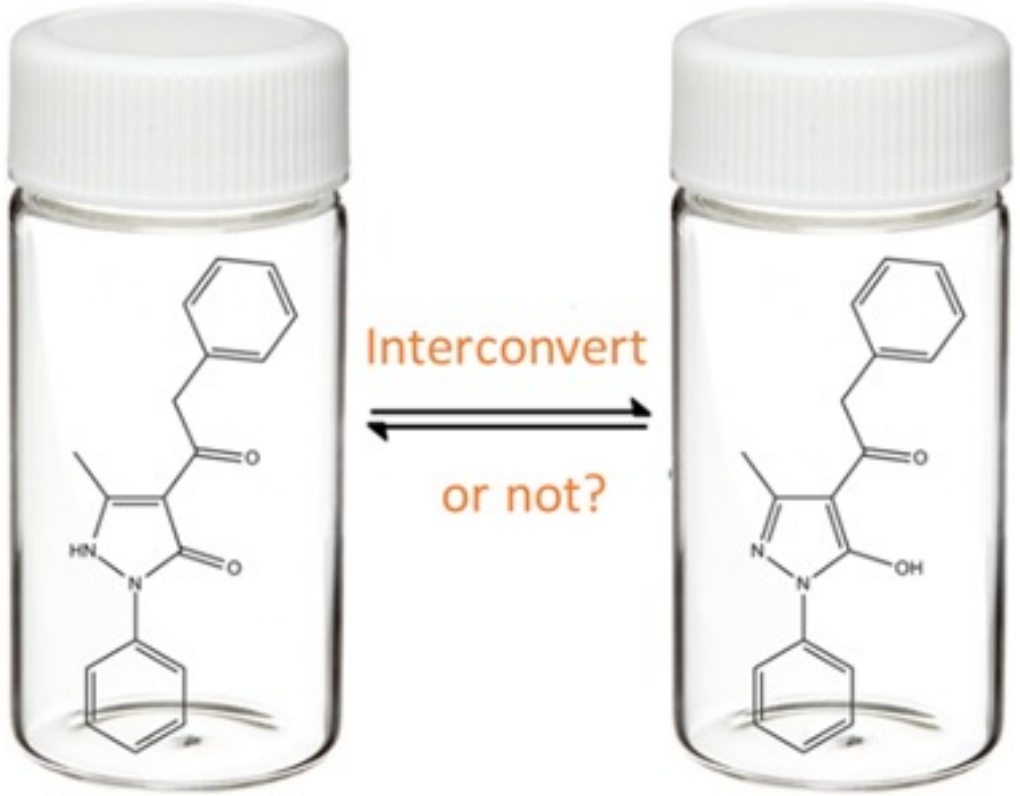
Finds false tautomers / Finds true tautomers	Rare (1-20%)	Unlikely (21-40%)	Possible (41-60%)	Likely (61-80%)	Almost certain (81-100%)
Rare (1-20%)					**Bad ruleset**
Unlikely (21-40%)					
Possible (41-60%)			**OK ruleset**		
Likely (61-80%)					
Almost certain (81-100%)	**Excellent ruleset**				

Outline

- Background
- **Ideation and design**
- Information for participants

Idea

- Invite scientific community to evaluate how well the Tauto InChI identifies tautomers
- Collect, summarize and publish the evaluation results



PCLOQTRWNKXKTO-UHFFFAOYNA-N

PCLOQTRWNKXKTO-UHFFFAOYNA-N

U.S. FOOD & DRUG ADMINISTRATION

Announcement

FDA in partnership with InChI Trust and the IUPAC Tautomer Working Group invites members of Industry, Government, and Academia to test how well the InChI algorithm agrees with experimental determination of tautomers in chemical databases

Crowdsourced evaluation of InChI-based tautomer identification

Challenge Platform: precisionFDA
https://precision.fda.gov

Challenge Launch: November 1, 2022 (tentative)
Challenge site is made available

November 1, 2022 – March 1, 2023 (tentative)
Participants will utilize the InChI tool to identify tautomers within a library of chemical structures and compare the results with known experimental or analytical results

Deadline: March 1, 2023 (tentative)
Participants will compile the results of the analysis and submit on precisionFDA

Who can participate

An individual or organization that has a dataset of small molecule structures for which there are experimental data (NMR, UV, MS, or IR spectroscopy, X-ray crystallography, etc.), computational data (energy computation), or expert knowledge, that can be used to validate tautomeric interconversion.

Outline

- Background
- Ideation and design
- **Information for participants**

U.S. FOOD & DRUG ADMINISTRATION

CLOUD BASED - HIGH PERFORMANCE - SECURE - COLLABORATIVE

What is precisionFDA?

- Secure cloud-based portal
- Developed and run by FDA
- Features a crowdsourcing model to advance data analytics and computational methods in areas that impact public health

6,000+ Users
102 Terabytes of Data
145 Unique Publicly Available Apps
31 Challenges
Cited in a Total of 74 Scientific Publications
11 Contributed/Written Scientific Publications
21 Training Videos
5 Training Workshops

ENGAGING INTERNAL AND EXTERNAL EXPERTS IN EVOLVING SCIENCE

U.S. FOOD & DRUG ADMINISTRATION

FedRAMP

FISMA COMPLIANCE

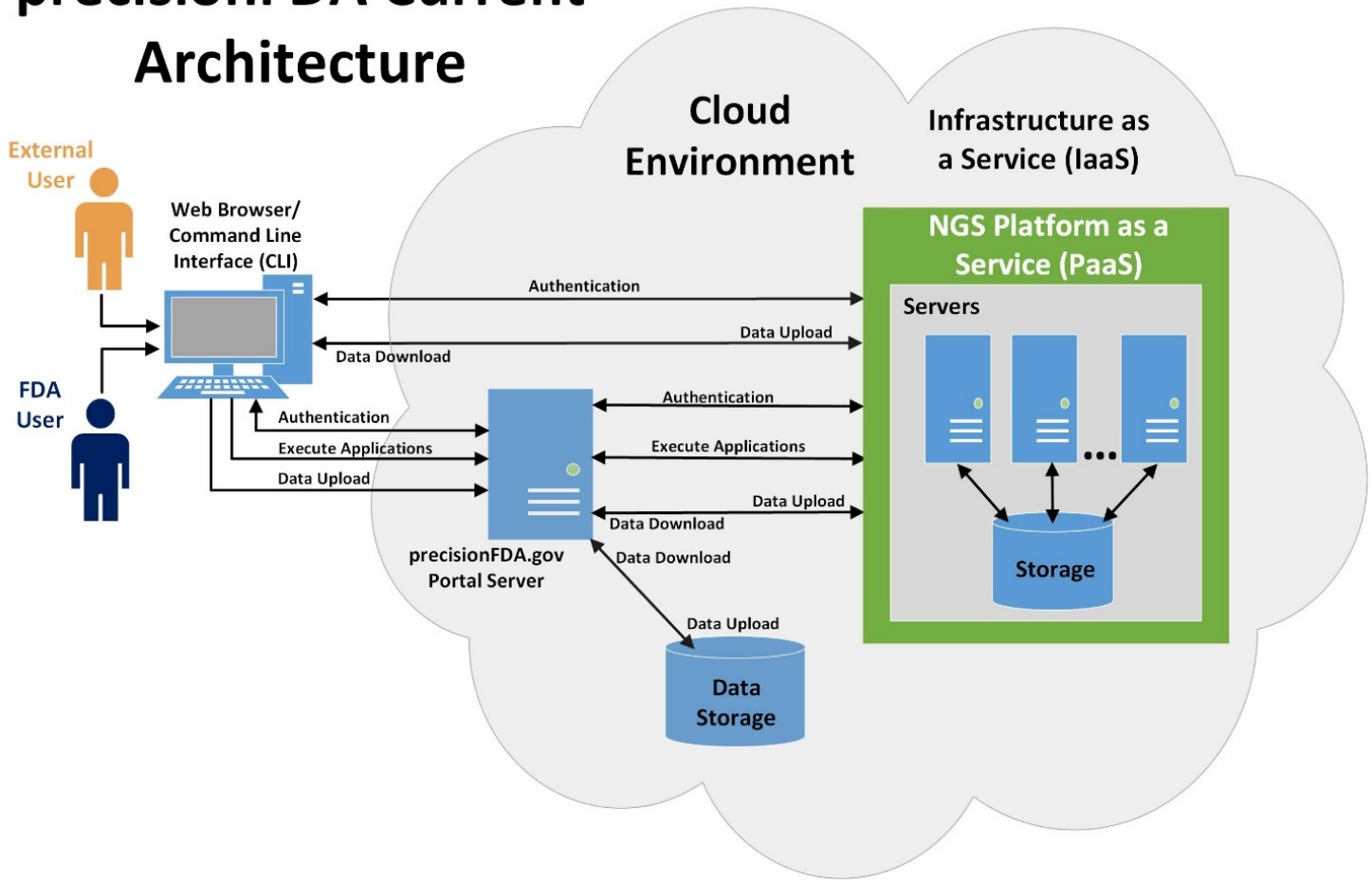
Moderate Level authorizes:

- Protected Health Information (PHI)
- Personally Identifiable Information (PII)
- Commercially sensitive information

A regulatory-grade platform:

- Encryption at rest and in-transit
- Access and collaboration controls
- Data and application chain of provenance

precisionFDA Current Architecture



Data on precisionFDA are secure and private unless shared

- All data brought to precisionFDA are private by default, i.e. cannot be viewed by anyone else
- One can use public applications with private data, while keeping that data private
- One can share data with the entire precisionFDA community or with a limited group of collaborators

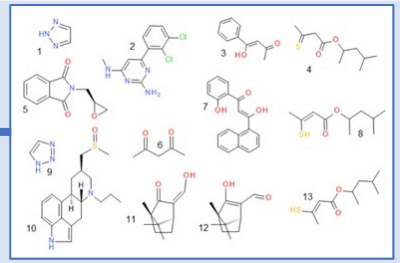
U.S. FOOD & DRUG ADMINISTRATION

precisionFDA will support

- Announcement
- Registration
- Submissions
- Framework in HPC Environment
- Computational analysis on the platform
- Computational analysis for download
- Evaluation of reports
- Publication of results

Crowdsourced Evaluation of InChI-Based Tautomer Identification

Input



Participant's Library of Chemical Structures

***Tentative* Pre-Registration: October 17, 2022**

***Tentative* Challenge Launch: November 1, 2022**

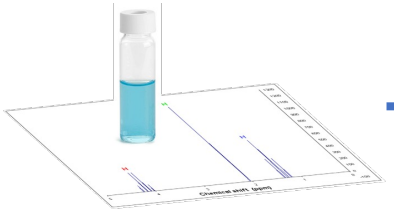
Challenge site is made available.

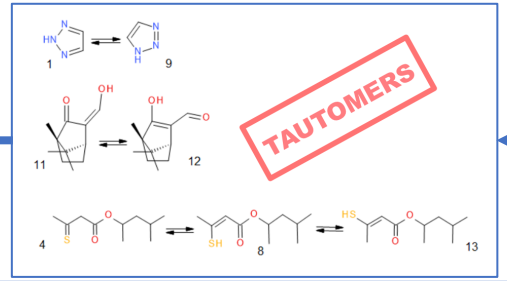
InChI-Based Computational Analysis

App on precisionFDA

OR

Docker on local systems



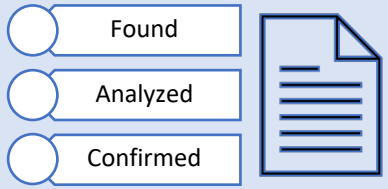


November 1, 2022 – March 1, 2023 (*Tentative*)

Participants will utilize the InChI tool to identify tautomers within a library of chemical structures and compare the results with known experimental or analytical results.

Results

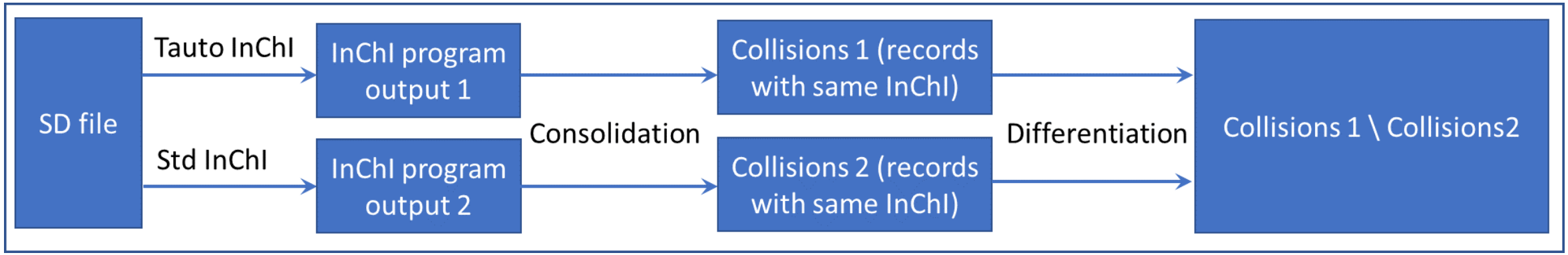
Participant Report (mandatory) and supporting material (optional)



***Tentative* Deadline: March 1, 2023**

Participants will compile the results of the analysis and submit on precisionFDA.

InChI Based Computational Analysis



Example results produced by computational analysis

InChI ruleset 6	Serial numbers of structures in SDF					
InChI=1/C20H21FN4O2/c1-12(2)17(19(22)26)23-20(27)18-15-8-3-4-9-16(15)25(24-18)11-13-6-5-7-14(21)10-13/h3-10,12H,11H2,1-2H3,(H,23,27)(H3,17,22,26)	2075	5340				→ collision 1
InChI=1/C7H12O3/c1-4-10-7(9)5(2)6(3)8/h4H2,1-3H3,(H,5,8,9)	67047	67048				→ collision 2
InChI=1/C6H10O6/c7-1-3(9)5(11)6(12)4(10)2-8/h(H5,1,4,5,8,9,12)(H5,2,3,6,7,10,11)	12790	71458	71459	71460	78180	→ collision 3

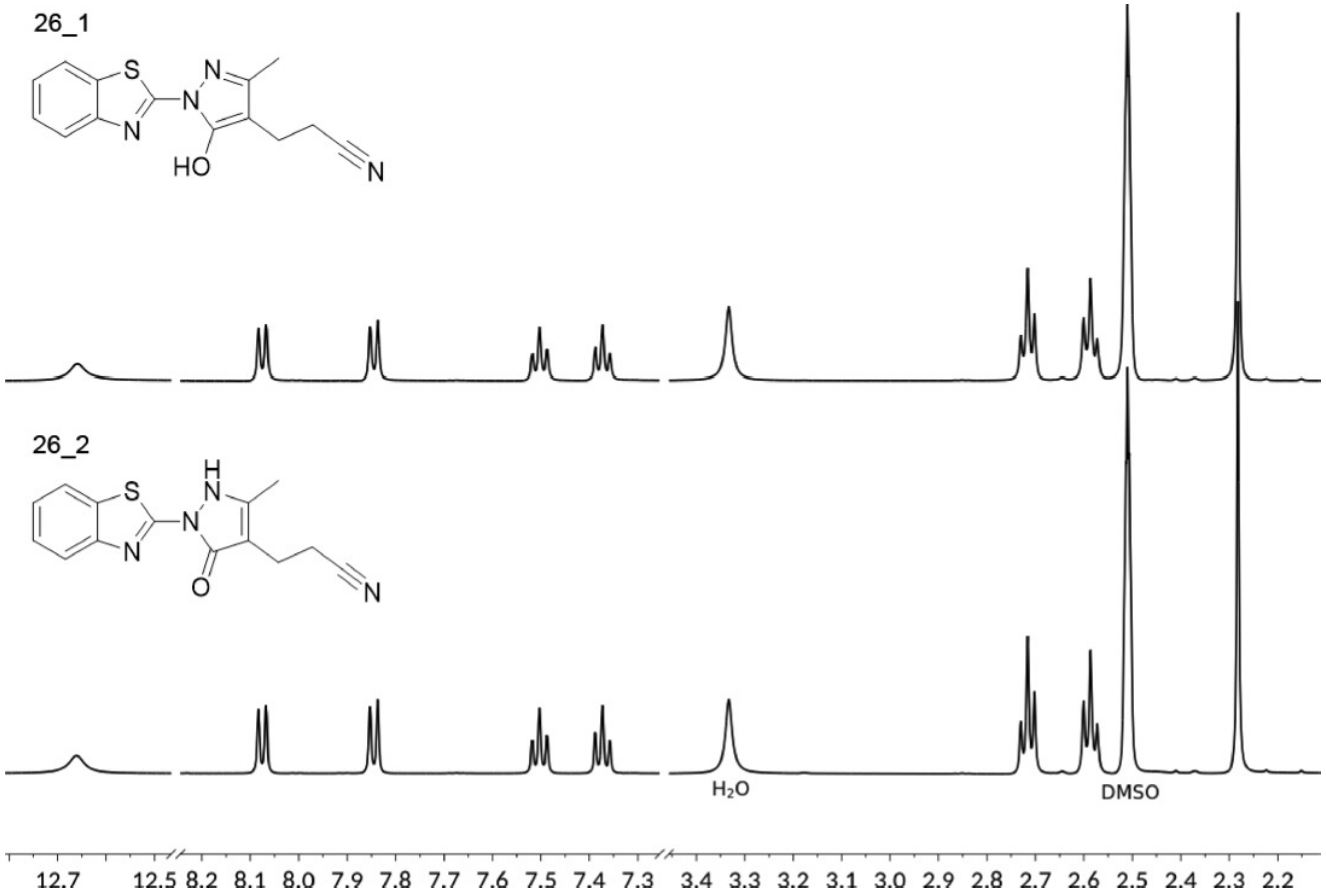
Example report produced by participant

	Ruleset 1	Ruleset 2	…	Ruleset 13
Total number of structures in SD file	180,000	180,000	…	180,000
METHOD	NMR	NMR	…	NMR
Total number of collisions	329	503	…	950
Total number of collisions analyzed using METHOD	100	200	…	…
Number of confirmed collisions	70	…	…	…
Number of disproved collisions	10	…	…	…
Number of inconclusive collisions	10	…	…	…
Number of partly confirmed collisions	10			

Example supplementary material produced by participant

InChI ruleset 6	Serial numbers of structures in SDF					NMR-based conclusion
InChI=1/C20H21FN4O2/c1-12(2)17(19(22)26)23-20(27)18-15-8-3-4-9-16(15)25(24-18)11-13-6-5-7-14(21)10-13/h3-10,12H,11H2,1-2H3,(H,23,27)(H3,17,22,26)	2075	5340				interconvert
InChI=1/C7H12O3/c1-4-10-7(9)5(2)6(3)8/h4H2,1-3H3,(H,5,8,9)	67047	67048				don’t interconvert
InChI=1/C6H10O6/c7-1-3(9)5(11)6(12)4(10)2-8/h(H5,1,4,5,8,9,12)(H5,2,3,6,7,10,11)	12790	71458	71459	71460	78180	some interconvert

Example supplementary material produced by participant



DOI: 10.1021/acs.jcim.6b00338
J. Chem. Inf. Model. **2016**, 56, 2149−2161

Benefits for participants and for InChI community

- Participants will find previously unknown tautomeric duplicates in their databases
- Participants will be invited to co-author a scientific publication
- InChI community will receive a recommendation from users that can direct further development of InChI
- InChI community will learn whether InChI can/should be used for identifying tautomers
- **A community will be formed that could help elucidate other issues of InChI such as possible additional rules, and will raise awareness of the need, and possibility, of testing cheminformatics approaches along experimental results**

Acknowledgments

Elaine Johanson (FDA)
Annette Vernon (FDA)
Letria Hall (FDA)
Sarah Prezek (Booz Allen Hamilton)
Vishal Thovarai (Booz Allen Hamilton)
Sam Westreich (DNAnexus)
Omar Serang (DNAnexus)

Igor Filippov (InChI Trust)
Steve Heller (InChI Trust)
Rudy Potenzone (InChI Trust)
Gerd Blanke (InChI Trust)
Marc Nicklaus (IUPAC WG on tautomers)
Wolf-Dietrich Ihlenfeldt (IUPAC WG on tautomers)
Tony Williams (EPA)
Charlie Lowe (EPA)
Andrey Yerin (ACD/Labs)
Yurii Moroz (Enamine)
Mitch Miller (Scientific Thinking LLC)
Lutz Weber (Ontochem)
Yuri Pevzner (New Equilibrium Biosciences)

U.S. FOOD & DRUG ADMINISTRATION

Questions?

Cheminformatics: Yulia.Borodina@fda.hhs.gov
Other: precisionfda@fda.hhs.gov

Join the community

https://precision.fda.gov