Tautomeric Conflicts in Forty Small-Molecule Databases

Devendra K. Dhaked[1], Marc C. Nicklaus[1*]

[1]Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, NIH, Frederick, MD 21702, USA.

ABSTRACT

We have analyzed forty different databases ranging in size from a few thousand to nearly 100 million molecules, comprising a total of over 200 million structures, for their tautomeric conflicts. A tautomeric conflict is defined as an occurrence of two or more structures within a data set identified by the tautomeric rules applied as being tautomers of each other. We tested a total of 119 detailed tautomeric transform rules expressed as SMIRKS, out of which 79 yielded at least one conflict. The databases analyzed spanned a wide variety of types including large aggregating databases, drug collections, and experimentally based structure collections. Almost all databases analyzed showed intra-database tautomeric conflicts. The conflict rates as percentage of the database were typically in the few tenths of a percent range, which for the largest databases amounts to >100,000 cases per database.

INTRODUCTION

Tautomerism is a ubiquitous phenomenon in chemistry. Estimates of what percentage of structures in a small-molecule database are capable of some kind of tautomerism have typically been above 10%[1,2] and if analyzed by comprehensive chemoinformatics rules, – were found to be as high as 66% of the cases.[3] Our recent study using 80+ tautomeric rules indicated the presence of an average of 71% tautomerism-capable molecules across 9 databases from a collection of >400 million structures.[4] It is therefore no surprise that existing databases may show cases where molecules are listed as separate record in it that rule-based approaches define as tautomers of each other. We have termed such cases "tautomer conflicts."

We previously analyzed over 150 databases, totaling more than 103 million structures, for the tautomeric conflicts in each database as well as across all databases combined.[3] We also analyzed a medium-size database (~6 million records) for its tautomeric conflicts by both chemoinformatics and experimental analysis, which computationally identified more than 31,000 conflicts, out of which more than 100 were experimentally tested.[5] This showed that in most cases, the chemoinformatics rules had correctly predicted the analytical identity of the tautomeric tuple members.

For this study, we significantly broadened both the number of tautomeric transform rules applied,[4] as well as the total number of compounds studied.

We need to strongly emphasize that finding a tautomer conflict in a database is not identical with having identified a problem in the database. There may be many reasons why tautomers of each other may be present in different database records. A tautomer conflict as identified in this study should therefore rather be seen as an alert to perhaps take a closer look at tautomer tuples to investigate if they truly create an issue in the database.

DATA AND METHODS

Databases

We downloaded 40 databases, ranging in size from nearly 100 million to a few thousand structures (Table 1). The total number of (non-unique) structures analyzed is more than 210 million.

Table 1. Databases used in this study.

Database	**Type**	**Molecule count in database**	**Date accessed**	**Way of access <download URL; or other way of getting it>**
PubChem	A	96,502,282	Oct 2018	https://pubchem.ncbi.nlm.nih.gov/
eMolecules	B, S, A	23,217,671	Nov 2019	https://www.emolecules.com/info/plus/download-database
SureChEMBL (Patents)	P	19,334,472	Dec 2019	https://www.surechembl.org/search/
MCULE	B, O	18,445,974	Nov 2019	https://mcule.com/database/
Ambinter	S	9,837,722	Nov 2019	http://www.ambinter.com/
AMS Screening (2019 Q4)	S	8,122,497	Dec 2019	Obtained by subscription to ChemNavigator databases
MolPort	B, S	7,615,837	Nov 2019	https://www.molport.com/
Enamine	B, S, O	3,856,359	Nov 2019	https://enamine.net/
InterChim	S	2,881,727	June 2019	http://www.interchim.com/
AsisChem	S	2,107,628	Dec 2019	https://asischem.com/
ChEMBL	L	1,820,035	Dec 2019	https://www.ebi.ac.uk/chembl/
Princeton Biomolecular Research	B, S	1,780,907	Nov 2019	http://www.princetonbio.com/
Life Chemicals	B, F, O	1,445,268	Nov 2019	https://lifechemicals.com/
ChemDiv	B, O	1,444,422	Nov 2019	https://www.chemdiv.com/

Vitas-M Laboratory	B, S	1,441,235	Nov 2019	https://vitasmlab.biz/
Chemspace	B, S, F	1,381,383	Nov 2019	https://chem-space.com/
ChemBridge	S	1,243,316	June 2019	https://www.chembridge.com/
Innovapharm	B, S, F, O	1,123,797	Dec 2019	https://innovapharm.com.ua/
TimTec	B, S, O	1,086,513	Nov 2019	https://www.timtec.net/ and https://www.timtec.net/home/download-databases.html
Alinda	B, S	929,626	Dec 2019	http://www.alinda.ru/synthes_en.html
US EPA	R	848,945	Dec 2019	https://comptox.epa.gov/dashboard/downloads
UORSY	S, F, O	681,161	Dec 2019	https://uorsy.com/
Asinex	B, S, F	642,206	Nov 2019	http://www.asinex.com/
InterBioScreen	B, S, F, O	572,055	Nov 2019	https://www.ibscreen.com/
Otava	B, S, F, O	566,491	Dec 2019	https://www.otavachemicals.com/
CSD Organic	E	319,204	Dec 2018	https://www.ccdc.cam.ac.uk/
HTS Biochemie	B, S, O	290,883	Nov 2019	http://www.hts-biochemie.de/hts-en/index.php
ChemBank	E	1,530,003	Jan 2020	https://data.broadinstitute.org/chembank/assay (Full set of molecules privately shared with us[6])
Key Organic	B, S, F, O	202,241	Nov 2019	https://www.keyorganics.net/downloads-bionet-databases/
ChemBlock	B, S	131,662	June 2019	http://www.chemical-block.com/
Maybridge	B, S, F	126,414	Nov 2019	https://www.maybridge.com/portal/alias__Rainbow/lang__en/tabID__177/DesktopDefault.

HMDB	M	113,983	Nov 2019	http://www.hmdb.ca/downloads
ChEBI	A	103,104	Nov 2019	https://www.ebi.ac.uk/chebi/downloadsForward.do
FDA DailyMed	R	60,095	Feb 2020	ftp://public.nlm.nih.gov/nlmdata/.dailymed/substance_indexing_spl_files.zip
PDB ligand	E	29,877	Aug 2019	http://ligand-expo.rcsb.org/ld-download.html
Combiphos Catalyst	O	19,583	Dec 2019	https://www.combiphos.com/products
MolMall-MDPI	O	15,309	Dec 2019	http://www.molmall.net/download.html
Cayman	O	13,860	Nov 2019	https://www.caymanchem.com/
DrugBank (5.1.4)	D	10,632	Aug 2019	https://www.drugbank.ca/releases/5-1-4
DrugCentral	D	4,531	Dec 2019	http://drugcentral.org/download

Note: “A”: Aggregating database; “S”: Screening sample supplier; “L”: Literature extractions; “R”: Regulatory agency database; “E”: Experimental results database; “D”: Drug database; “M”: Metabolite. “B”: Building blocks; “F”: Fragment or scaffold library; “P”: Patent; “O”: Other.

This set spans a wide variety of types of databases: large aggregated databases (e.g., PubChem); screening sample (and other commercially available) catalogs (e.g., Ambinter); databases of regulatory agencies (e.g., US EPA); experimental results (e.g., CSD Organic); drug databases (e.g., DrugBank); human metabolite (e. g. HMDB); patent (e.g. SureChEMBL) and others.

Algorithmic Approaches and Software Used

The analyses of tautomeric conflicts were performed with the chemoinformatics toolkit CACTVS.[7] Version 3.4.8.12 of CACTVS was used. CACTVS allows the user to calculate a number of identifiers (hashcodes) that are sensitive to different chemical features such as formal charges in the input structure, presence of isotopically labeled atoms, stereochemistry, etc. One of these features is tautomerism, i.e. if tautomerism invariance is turned on, the identifier returned by CACTVS is the same for all possible tautomers that can be enumerated based on the tautomeric rule set active at the time of execution, otherwise different tautomers receive different identifiers. One such tautomer-invariant identifier is called E_TAUTO_HASH128 (the “E_” standing for: Ensemble property, "ensemble" being the CACTVS term for a compound structure) with 128 bit length (the default hashcode length is 64 bit). Other related hashcodes we computed (named with the terminology "ISOTOPE": sensitive to isotopes; "STEREO": sentive to stereochemistry; "TAUTO": invariant to tautomers) were: tautomer-invariant hashcodes E_ISOTOPE_TAUTO_HASH128, E_STEREO_TAUTO_HASH128, E_ISOTOPE_STEREO_TAUTO_HASH128; and tautomer-sensitive hashcodes E_HASH128, E_ISOTOPE_HASH128, E_STEREO_HASH128, and E_ISOTOPE_STEREO_HASH128 .

Tautomerism Rules Used

We started with 86 previously defined rules[4] encompassing a wide variety of prototropic, ring-chain, and valence tautomeric transforms. They have been thoroughly analyzed as to their prevalence in a variety of databases (including some used in this paper), defined of how many structures in a database are amenable to each rule, i.e. can have more than one tautomer based on this rule; as to their comparison with current InChI (which has some but significantly incomplete handling of tautomerism in its current version, 1.05); and as to other properties. We refer the reader to that analysis[4] for these detailed analyses. To these 86 SMIRKS, we added 33 SMIRKS from 11 types of ring⇌chain rules[8] (encoded in a total of 38 SMIRKS). Five of these ring-chain SMIRKS were already part of 86 SMIRKS of our previous study,[4] which led to a total of 119 SMIRKS used in this study. The entire list of transforms is included as Supplementary Information Table S1. We note that this set can be subdivided into “common” and “rare” rules, simply based on the prevalence observed, with common rules matching millions of structures[4] in a large database such as PubChem. Fewer than 20 rules are common in this sense. Since a tautomeric conflict requires at least two structures in a database that can interconvert according to our rules but have been labeled as different records by the database provider, i.e. is essentially a square function of the occurrence rate, it is obvious that common rules are much more likely to yield tautomeric conflicts than rare ones.

Definition and Determination of Tautomeric Conflicts

We define a tautomeric conflict as the occurrence, within a database, of two or more records labeled by the database provider as structurally different entries, whereas our tautomeric rules indicate that these structures are tautomers of each other. We determine such conflicts by searching for compounds in the database that have the same tautomer-invariant but different tautomer-sensitive hashcodes. Specifically, we used E_ISOTOPE_STEREO_HASH128 and E_ISOTOPE_STEREO_TAUTO_HASH128 to search for tautomeric conflicts. It is important to use stereo-sensitive hashcodes. Non-stereo-sensitive hashcodes ignore stereochemistry – anywhere in the molecule. Tautomerism may make some stereogenic centers non-persistent (see below) but many other stereocenters are not affected by tautomerism. Non-stereo-sensitive hashcodes would therefore equate stereoisomers which each other that truly are not tautomers of each other. At the same time, and perhaps somewhat counterintuitively, non-stereo-sensitive hashcodes typically lead to lower conflict counts because they have already projected different stereoisomers onto each other that may be found to be non-persistent based on our rules and thus may receive the same tautomer-invariant hashcodes, while they are often listed as different records in databases.

Determining Tautomeric Rules for Conflicts

We determined the single transform or sequence of transforms connecting the tuple members with each other, employing an approach bases on tautomeric network as published before[5], however here using all 119 rules, not just the 20 standard (prototropic) rules of CACTVS. Briefly, we first enumerate all possible tautomers from each tautomeric tuple; then, we generate a tautomer network among those enumerated tautomers. In such a network, one typically finds several pathways that connect one tautomer to the other by different tautomeric transforms. Finally, we search for the shortest pathway, defined by the smallest number of transformation steps within the tautomeric pair.

RESULTS

Intra-Database Conflict Counts

Table 2 shows the conflict counts for the databases analyzed. While the majority of conflicts stem from tautomer pairs, we also looked for, and found numerous cases of higher-order tuples. We checked for up to 25-tuples. The larger the database, the higher the chance for higher-order tuples. It is therefore no surprise that PubChem leads the pack with two cases of a 25-tuple. The total number of intra-database conflicts based on a merger of all 40 databases was 5,401,803. Table 2 also shows the conflict rate, defined as the number of conflicts divided by the number of molecules in the database (as shown in Table 1). We note that the number of database records involved in the conflicts is at least twice as high since the lowest-order tautomer tuple is a pair. The conflict rates typically fall in the range between 0.1% and 1%, with the extrema being 11.59% at the high end and 0.0% at the low end, and a median of 0.23%. The average conflict rate was about 0.77%. If we remove the outlier of 11.59% for ChemBank, the median and the average reduce to 0.21% and 0.48%, respectively. This means that between half and one percent of a database is typically involved in tautomer conflicts. We note that this number is close to the result of our 2010 study,[3] which had yielded an overall tautomeric conflict rate of 0.3%. We found no conflicts for two databases: Uorsy and Combiphos Catalyst.

Table 2. Tautomeric conflict counts within each database (intra-database conflicts).

			Tautomer conflict tuple size													
Database	**Total conflicts**	**Conflict Rate (%)**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
PubChem	2,491,470	2.58	1,447,010	936,500	64,785	27,934	7,098	3,772	1,544	1,355	527	413	185	109	63	53
eMolecules	199,658	0.86	179,414	18,804	1,116	219	63	16	5	17	1	2	1	-	-	-
SureChEMBL (Patents)	304,621	1.58	259,825	39,408	3,450	1,140	388	182	108	40	22	16	17	6	2	3
MCULE	29,666	0.16	25,953	3,684	21	2	4	2	-	-	-	-	-	-	-	-
Ambinter	36,559	0.37	34,830	1,692	30	5	1	-	1	-	-	-	-	-	-	-
AMS Screening 19_Q4	71,274	0.88	66,852	4,264	142	10	3	1	-	-	1	1	-	-	-	-
MolPort	138,846	1.82	131,095	6,887	754	92	14	2	-	-	1	1	-	-	-	-
Enamine	4,252	0.11	2,313	1,935	4	-	-	-	-	-	-	-	-	-	-	-
InterChim	9,440	0.33	9,259	173	8	-	-	-	-	-	-	-	-	-	-	-
AsisChem	480	0.02	480	-	-	-	-	-	-	-	-	-	-	-	-	-
ChEMBL	10,525	0.58	9,346	1,057	94	21	6	1	-	-	-	-	-	-	-	-
Princeton Biomolecular Research	33,318	1.87	32,632	668	17	1	-	-	-	-	-	-	-	-	-	-
Life Chemicals	870	0.06	868	2	-	-	-	-	-	-	-	-	-	-	-	-
ChemDiv	614	0.04	602	12	-	-	-	-	-	-	-	-	-	-	-	-
Vitas-M Laboratory	21,639	1.50	21,084	536	17	2	-	-	-	-	-	-	-	-	-	-

Chemspace	2,926	0.21	2,403	503	16	2	2	-	-	-	-	-	-	-	-	-
ChemBridge	388	0.03	386	2	-	-	-	-	-	-	-	-	-	-	-	-
Innovapharm	3,716	0.33	3,611	101	4	-	-	-	-	-	-	-	-	-	-	-
TimTec	1,092	0.10	1,074	18	-	-	-	-	-	-	-	-	-	-	-	-
Alinda	488	0.05	484	4	-	-	-	-	-	-	-	-	-	-	-	-
US EPA	2,777	0.33	2,478	251	38	6	-	2	-	2	-	-	-	-	-	-
UORSY	125	0.02	125	-	-	-	-	-	-	-	-	-	-	-	-	-
Asinex	194	0.03	174	20	-	-	-	-	-	-	-	-	-	-	-	-
InterBioScreen	275	0.05	272	3	-	-	-	-	-	-	-	-	-	-	-	-
Otava	336	0.06	315	17	4	-	-	-	-	-	-	-	-	-	-	-
CSD Organic	452	0.14	443	8	-	1	-	-	-	-	-	-	-	-	-	-
HTS Biochemie	324	0.11	324	-	-	-	-	-	-	-	-	-	-	-	-	-
ChemBank	33,527	11.59	33,051	429	39	8	-	-	-	-	-	-	-	-	-	-
Key Organic	309	0.15	248	60	-	1	-	-	-	-	-	-	-	-	-	-
ChemBlock	139	0.11	139	-	-	-	-	-	-	-	-	-	-	-	-	-
Maybridge	54	0.04	53	1	-	-	-	-	-	-	-	-	-	-	-	-
HMDB	284	0.25	246	33	4	-	1	-	-	-	-	-	-	-	-	-

ChEBI	1,385	1.34	1,103	215	33	13	8	6	-	2	-	1	1	1	2	-
FDA DailyMed	556	0.93	513	24	9	4	3	1	-	1	1	-	-	-	-	-
PDB ligand	167	0.56	148	13	4	-	1	1	-	-	-	-	-	-	-	-
Combiphos Catalyst	4	0.02	4	-	-	-	-	-	-	-	-	-	-	-	-	-
MolMall-MDPI	62	0.40	59	2	1	-	-	-	-	-	-	-	-	-	-	-
Cayman	61	0.44	49	10	2	-	-	-	-	-	-	-	-	-	-	-
DrugBank (5.1.4)	43	0.40	38	3	1	1	-	-	-	-	-	-	-	-	-	-
DrugCentral	9	0.20	9	-	-	-	-	-	-	-	-	-	-	-	-	-

We tested tuple sizes up to 25 but show tuple sizes only up to 15 to limit the size of the table. The entire table with all tuple sizes is provided as Supporting Information Table S2.

Inter-Database Conflicts

We determined inter-database conflicts by checking, for each tautomeric molecule of a database, for the presence of alternative tautomer(s) in all other 39 databases. We note that these alternative tautomers may be fully or partially the same across these 39 databases but they had to be different from the query tautomer of the tested database. Table 3 shows the inter-database conflicts, again broken down by tuple sizes similar to Table 2. It is not surprising that the number of inter-databases conflicts is generally higher than that of the intra-databases conflicts (e.g. for PubChem >5 million for the former vs. about 1.7 million for the latter) given that structure normalization is most likely different between databases. We also calculated the number of conflicts across all 40 databases combined. Among these 201 million non-unique structures, we found 5,401,803 conflicts.

Table 3. Tautomeric conflict counts for each database relative to all other databases (inter-database conflicts).

	Inter-database tautomer conflict tuple sizes														
Database	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
PubChem	5,135,589	257,072	18,108	4,573	1,749	1,141	543	283	398	91	106	61	116	13	21
eMolecules	843,642	580,741	13,679	19,775	1,766	1,523	628	1,109	150	120	77	68	26	18	8
SureChEMBL (Patents)	698,216	134,812	12,056	5,545	1,440	1,030	534	398	231	102	120	75	82	24	18
MCULE	2,566,895	151,734	85,015	5,680	2,463	739	314	126	90	34	18	4	13	3	-
Ambinter	1,057,664	546,596	16,194	17,823	2,511	1,964	680	1,290	218	240	101	59	33	42	18
AMS Screening (2019 Q4)	1,391,234	538,174	13,708	18,622	1,719	1,548	687	1,057	191	158	69	58	29	28	14
MolPort	851,328	469,692	10,831	17,277	1,566	1,382	532	1,039	106	93	60	54	19	12	4
Enamine	214,666	25,102	14,707	1,252	522	74	65	17	14	4	4	-	4	-	-
InterChim	214,882	156,473	4,189	12,018	619	891	443	1,328	60	64	61	57	12	9	1
AsisChem	255,351	37,918	9,908	2,205	350	90	18	9	6	2	2	1	-	-	-
ChEMBL	286,958	18,199	14,197	2,310	679	357	204	139	73	38	20	7	16	3	4
Princeton Biomolecular Research	548,344	30,672	43,184	3,457	1,317	441	262	104	48	25	10	3	4	-	-
Life Chemicals	44,273	2,493	2,880	311	82	26	48	19	4	3	1	-	1	-	-
ChemDiv	121,076	9,033	13,053	1,360	366	158	78	31	15	5	3	-	-	-	-
Vitas-M Laboratory	685,190	35,014	42,302	3,172	1,291	408	196	95	41	22	10	3	3	-	-
Chemspace	48,927	4,246	2,047	281	74	29	20	7	8	2	3	-	-	-	-
ChemBridge	103,393	83,338	1,200	2,873	128	156	55	320	24	21	14	7	4	2	-
Innovapharm	311,573	42,806	19,657	1,427	232	95	145	32	15	5	-	-	2	-	-
TimTec	379,678	20,845	25,251	2,341	691	221	134	63	30	10	4	1	3	-	-
Alinda	193,984	7,646	11,492	780	252	71	47	16	8	2	2	-	-	-	-
US EPA	135,800	9,458	3,017	619	194	146	44	29	21	14	17	11	9	5	1

UORSY	79,804	4,300	4,252	459	86	19	61	11	3	3	1	-	2	-	-
Asinex	140,134	7,192	14,932	844	208	54	47	22	9	7	2	-	-	-	-
InterBioScreen	298,326	17,003	17,766	1,460	386	149	122	65	20	11	2	1	-	1	-
Otava	109,420	9,777	10,017	934	74	77	72	19	6	4	-	1	1	-	-
CSD Organic	5,671	1,105	123	56	13	5	5	2	-	-	-	1	1	1	1
HTS Biochemie	2,999	1,604	7	6	2	-	-	-	-	1	-	-	-	-	-
ChemBank	233,927	20,239	14,644	1,896	517	241	127	74	39	17	9	1	7	4	3
Key Organic	32,194	8,184	387	297	42	43	18	16	6	2	1	2	1	-	1
ChemBlock	23,469	2,043	3,905	426	150	59	22	9	5	-	1	-	-	-	-
Maybridge	22,287	1,883	1,413	90	29	10	1	1	-	1	1	-	-	-	-
HMDB	8,353	822	272	74	54	25	14	5	9	8	3	1	1	4	1
ChEBI	31,223	3,104	781	291	165	80	28	25	23	17	14	9	2	2	1
FDA DailyMed	16,968	1,754	743	215	119	76	32	23	10	6	16	2	2	1	-
PDB ligand	4,229	708	145	72	40	23	18	6	11	3	-	3	-	-	-
Combiphos Catalyst	1,900	141	33	4	-	1	-	-	-	-	-	-	-	-	-
MolMall-MDPI	5,100	374	262	76	17	6	3	2	-	1	1	-	-	-	-
Cayman	2,687	789	100	61	15	12	8	3	2	4	1	3	3	1	-
DrugBank (5.1.4)	7,566	657	281	90	29	18	8	9	8	2	4	2	-	-	1
DrugCentral	2,325	253	50	21	7	5	-	2	1	1	2	1	-	1	-

Transforms Associated with Conflicts

Table 4 shows the counts of transforms we were able to associate with intra-database conflicts identified based on E_ISOTOPE_HASH128 and E_ISOTOPE_TAUTO_HASH128 . We had to use non-stereo sensitive hashcodes for this analysis because the explicit handling of stereogenic elements in the STEREO hashcodes prevents the straightforward generation of tautomeric pathways (see above), i.e. one tautomer may not converge to another tautomer. A total of 73 out of the 119 tested rules yielded at least one case of tautomeric conflict, be it a single- or multiple-rule conflict. We note that the vast majority of conflicts are based on one of the common rules as mentioned above, possibly in combination with one or a few other rules, these being in the majority common rules themselves. We do however find some conflicts based on rare rules, including more than 1,000 cases involving new rules going beyond the 20 standard CACTVS rules: 435 single-rule conflicts, 1,277 multiple-rule conflicts..

Table 4. Counts of associated rules in single-rule conflicts as well as rule combinations associated with conflicts

Rule number	**Single-rule conflicts**[a]	**Combined or/and alternative rule conflicts**[b]
PT_02_00	1,337	31,868
PT_03_00	316	68,591
PT_04_00	-	15,889
PT_05_00	6	122,429
PT_06_00	190,015	271,174
PT_07_00	40,094	43,028
PT_08_00	4	25,528
PT_09_00	6,124	59,634
PT_10_00	98	4,807
PT_11_00	57	7,670
PT_11_01	6	346
PT_11_02	5	369
PT_11_03	0	81
PT_11_04	0	35
PT_12_00	-	13,199
PT_13_00	443	51
PT_14_00	-	387
PT_15_00	1	387
PT_16_00	989	10,297
PT_17_00	-	87
PT_18_00	113	49
PT_19_00	64	20
PT_20_00	26	7
PT_21_00	1,389	61
PT_22_00	60	383

PT_23_00	254	42
PT_24_00	-	4
PT_27_00	58	40
PT_28_00	-	79
PT_29_00	-	58
PT_29_01	3	248
PT_32_00	-	42
PT_33_00	-	45
PT_35_00	-	2
PT_36_00	-	2
PT_37_00	6	7
PT_39_00	2	3
PT_41_00	-	4
PT_42_00	17	206
PT_44_00	-	6
PT_45_00	35	102
PT_47_00	-	4
RC_01_00	-	4
RC_02_00	-	1
RC_03_00	-	26
RC_03_01	-	19
RC_03_02	-	4
RC_04_00	-	9
RC_04_01	-	10
RC_04_02	-	13
RC_04_03	-	7
RC_05_00	-	3
RC_05_01	-	4
RC_05_02	-	3
RC_05_03	-	3
RC_06_00	-	10
RC_06_01	-	9
RC_06_02	-	3
RC_07_00	-	4
RC_07_01	-	4
RC_07_02	-	6
RC_07_03	-	55
RC_09_00	-	44
RC_09_01	-	23
RC_10_00	-	22

RC_10_01	-	22
RC_10_02	-	13
RC_11_00	-	2
RC_11_01	-	2
RC_11_02	-	2
RC_20_00	-	1
VT_02_00	-	3
VT_06_00	-	1

[a] Examining only single-rule transformations without any alternative rule possibility. [b] Examining transformations involving alternative rules and/or multistep transformations.

We note that the rule combinations are mostly involving prototropic transforms, rarely ring-chain transforms and valence transforms.

Percentage of conflicts involving non-persistent stereocenters

We have previously shown[3] that tautomerism can change the stereochemistry of a compound: The changing location of a double bond may add or eliminate the presence of an E/Z stereo bond. Likewise, migration of a double bond to an sp^3 hybridized atom that was chiral removes this chirality, which a further tautomeric isomerization step can re-establish with the opposite chirality, effectively creating racemization of this stereo center. Similarly, ring-chain equilibrium at an sp^2 center can lead to generation of enantiomers, which was recently observed in two investigational molecules, where two ring tautomers of opposite chirality were in equilibrium with the chain form for each compound.[9]

The analysis of this effect in the context of this study was less straightforward than one may think. The issue is that tautomeric structures of the same molecule typically have different atom numbers, which makes atom-atom mapping challenging to determine whether a specific stereogenic center was affected by tautomerism or not. We instead compared the tautomeric overlaps based on stereo-sensitive hashcode identifiers with those based on stereo-invariant identifiers (called a "stereo tauto conflict" below). This entails the risk that structures with stereocenters that had different R/S or E/Z geometry, but did not participate in the tautomeric interconversion, and had no tautomerism-affected stereo centers, are erroneously counted. In detail, we extracted the subset of stereo tauto conflicts out of the overall conflicts in the following way: If the E_ISOTOPE_STEREO_HASH128 values were different from each other, and different from E_ISOTOPE_HASH128, and that one was the same as the E_ISOTOPE_STEREO_TAUTO_HASH128 value, then we designated this a case of stereo tauto conflict (being aware that there is a small possibility of mis-designation as outlined above)

Table 5. Percentage of conflicts involving non-persistent stereocenters

Database	**Stereo conflict rate(%)**
PubChem	67.47
eMolecules	27.01
SureChEMBL (Patents)	38.66
MCULE	22.04
Ambinter	8.54
AMS Screening (2019 Q4)	43.56
MolPort	26.22
Enamine	50.80
InterChim	59.26
AsisChem	44.17
ChEMBL	78.30
Princeton Biomolecular Research	12.01
Life Chemicals	0.92
ChemDiv	9.77
Vitas-M Laboratory	15.73
Chemspace	39.95
ChemBridge	92.01
Innovapharm	49.46
TimTec	22.34
Alinda	55.12
US EPA	21.89
UORSY	0.00
Asinex	85.57
InterBioScreen	21.09
Otava	12.50
CSD Organic	89.16
HTS Biochemie	0.31
ChemBank	7.52
Key Organic	59.55
ChemBlock	16.55
Maybridge	12.96
HMDB	31.34

ChEBI	78.34
FDA DailyMed	97.12
PDB ligand	83.23
Combiphos Catalyst	0.00
MolMall-MDPI	69.35
Cayman	91.80
DrugBank (5.1.4)	62.79
DrugCentral	100.00
40 DBs merged	43.14

Table 5 shows a large spread of the percentages of conflicts involving non-persistent stereocenters, ranging from 0 to 100 %. This suggests that different databases treat stereochemistry differently, as well as that they differ in the types of compounds they contain that are affected by this effect. The average of the percentages of conflicts involving non-persistent stereocenters was 42.6%, and 43.1% when all databases were merged into one. We therefore see that this "destruction" of stereo centers by tautomerism is a common effect. The table with the absolute numbers of conflicts for each database is provided as Supporting Information Table S3.

DISCUSSION

The fact that the overall intra-database tautomeric conflict rate has changed little of the course of 10 years[3] indicates that not much appears to have changed in the recent past with how tautomerism is treated by database managers and providers.

We need to re-emphasize that identifying a tautomer conflict in a database does not per se indicate a problem with that database. There may be numerous reasons why several tautomers of the same molecule may be present in the same database. This includes the possibility that different tautomers were truly present in different samples, i.e. that the applied chemoinformatics rule was too “aggressive”; or that – especially in experimental databases – different conditions of measurement[10] or sample preparation (e.g. in X-ray crystallography) yielded different tautomeric states.

Aggregating databases such as PubChem, combining data sets from hundreds of different sources[11] are particularly prone to tautomeric conflicts. Strong normalization of input structures can reduce this effect but is fraught with some risks, including original submitters complaining, “What did you do to our structures?” We have previously analyzed a large database of commercially available samples[5] and will therefore not further discuss this class of databases (labeled as “S” in Table 1).

As examples of our analyses, we present a handful of tautomeric conflicts we found for databases in the classes of regulatory agency databases (“R”), experimental results databases (“E”), and drug databases (“D”).

US EPA

The Distributed Structure-Searchable Toxicity (DSSTox) database of the US Environmental Protection Agency’s (EPA) is a chemical database for predictive toxicology. Each chemical is mapped with bioassay and physicochemical property and toxicity data. The database preparation was described by the authors as being based on uniquely mapped identifiers (i.e., CAS RN, name and structure) and with rejection of entries that have any two identical identifiers.[12] We found more than 2,777 tautomeric conflicts (see Supplementary Information Table S2) among the around 848,000 entries of this database. A few examples are shown in Figure 1.

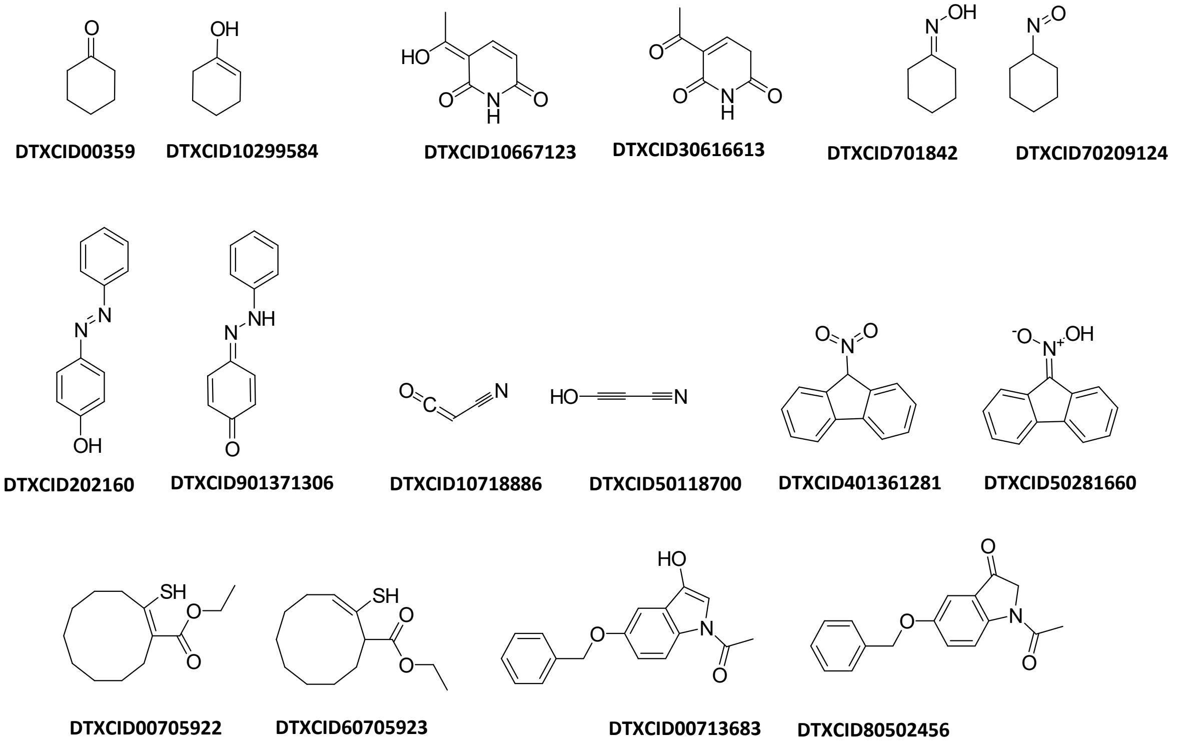


Figure 1. Tautomeric conflict examples from EPA DSSTox.

FDA DailyMed

FDA has been posting Substance Index SPL files for small-molecule based substances and biological substances (therapeutic proteins and biological organisms). From SPL files downloaded from the DailyMed website[13], around 61,000 small molecules were collected from those SPL files that contained only a single small molecule. In these files each molecule has its own Unique Ingredient Identifier (UNII).[14] Note that these structures are not limited to approved drugs but contain a wide variety of substances of interest to the FDA. We found a total of 556 tautomeric conflicts in this structure set. Figure 2 shows a few examples of tautomeric conflicts of structures having different UNIIs. We note, however, that we did not find any tautomeric conflicts amongactive ingredients in approved drugs (except for possible interchange of stereocenters upon tautomeric interconversion, see below). The conflicts we found involved other substances, which may indicate that registration of drug ingredients may undergo more comprehensive scrutiny than other substances of interest to FDA.

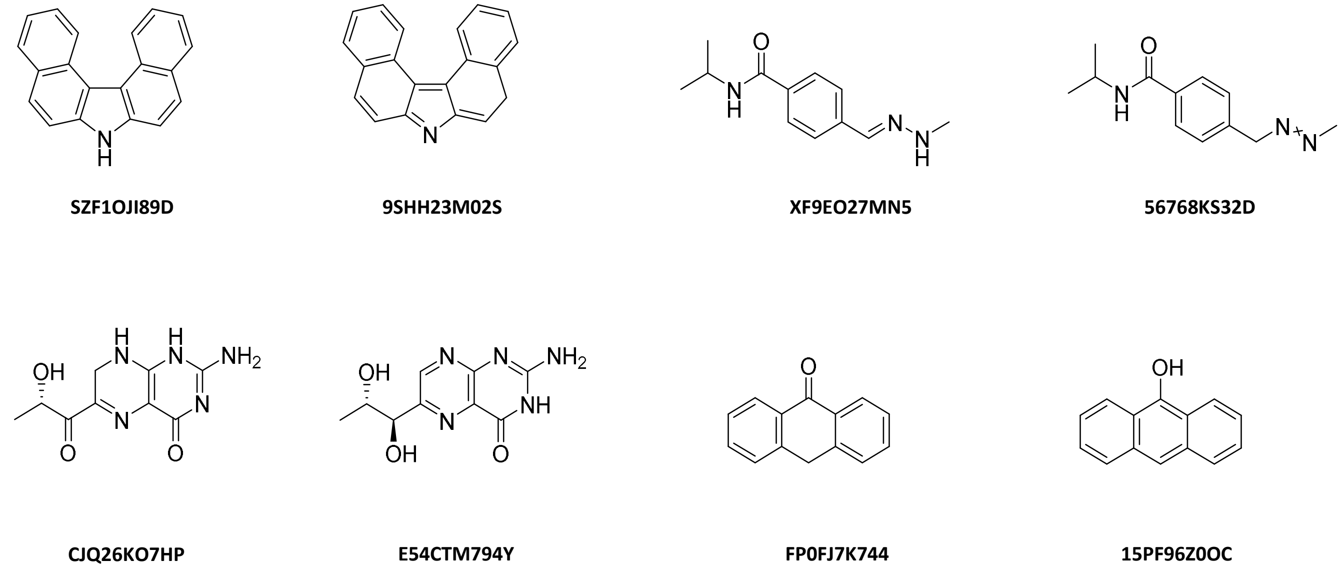


Figure 2. Tautomeric conflict examples from FDA Substance Indexing SPL files downloaded from DailyMed, with UNII identifiers shown.

CSD Organic

For the tautomeric analysis of the CSD, the collection of small molecule structures solved by X-ray crystallography and neutron diffraction methods, we limited ourselves to only its organic molecules. From around 319,000 organic molecules, we found 452 cases where molecules exist in one or more alternative forms. In other databases, we term such pairs “tautomeric conflicts.” However, in the case of the CSD they may be better called “alternative forms” in the same database because these are experimentally solved structures in which for the most part the hydrogens are resolved. Figure 3 shows some of the examples involving 1,3, 1,5, and 1,7 H-shifts, respectively.

In the WUQCOA/WEMRUD pair, the H-shift occurs between an N atom and an otherwise uncommonly seen Te atom. Other pairs involve hydrogen migration through aromatic phenyl and pyridine ring system (POFWOX/NEDMUF01 and AHPAZP10/CALWUA01).

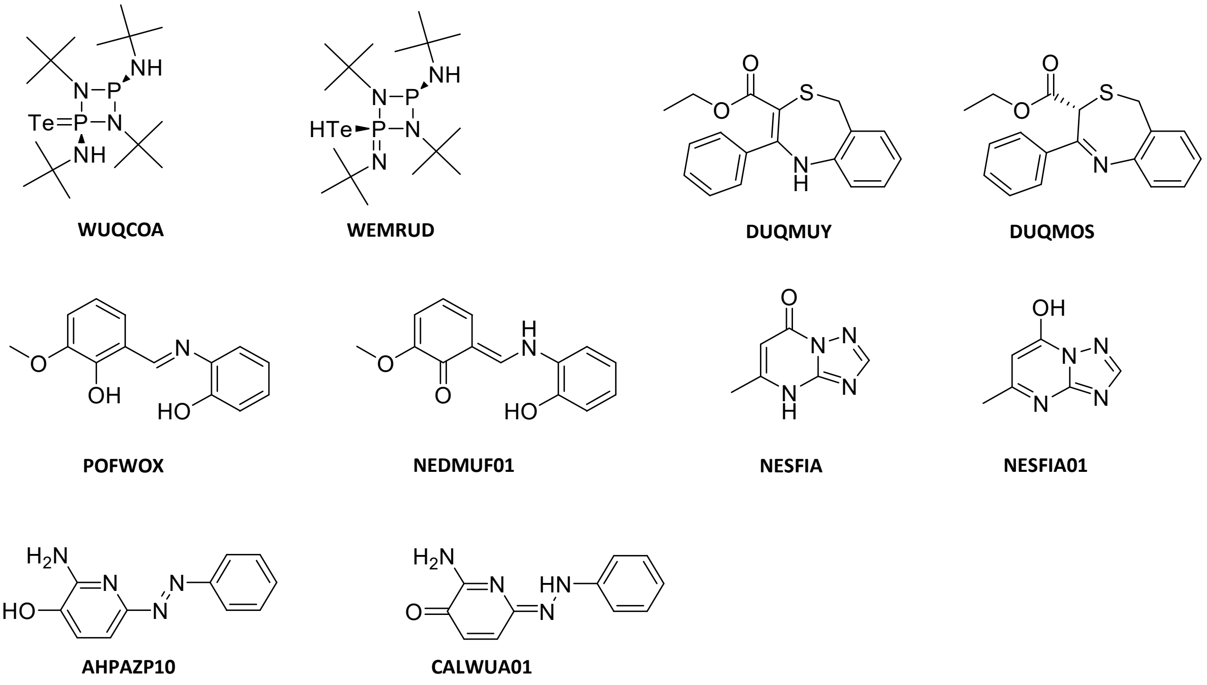


Figure 3. Alternative pair examples from CSD.

PDB Ligand Expo

The PDB Ligand Expo database showed 167 tautomeric conflicts out of about 29,000 small molecule components reported in PDB entries bound to various macromolecules (protein, DNA). These different molecules are called “Unique Ligands” by the PDB. We could not find mentioning of a the possibility of presence of tautomeric molecules within this ligand set.

For example, from a set of four tautomers, three tautomers (50M, 53M and 53L) are reported to be present in 5CDO (http://www.rcsb.org/structure/5CDO) and 1 tautomer (54Q) exists in 5CDM (http://www.rcsb.org/structure/5CDM) (Figure 4). One point to mention here is that all of these four tautomers have same standard InChI, thus InChI-based resolution of this case of tautomerism would not have been possible.

It is usually difficult to determine the tautomeric state of a bound molecule. Pyridoxamine-5-phosphate-hydroxyisoxazole is an experimental molecule that is reported in three different tautomeric forms: 7TS, LCS and PMH in one (5U3F), four (6QP1, 4OMA, 1D7U and 4D9E) and two PDB entries (1XQK and 1XQL), respectively.

For some molecules, one tautomer is reported to have high preference compared to another one. For example, the keto form of uric acid is reported in 22 PDB entries whereas its hydroxy form occurs in only two PDB entries.

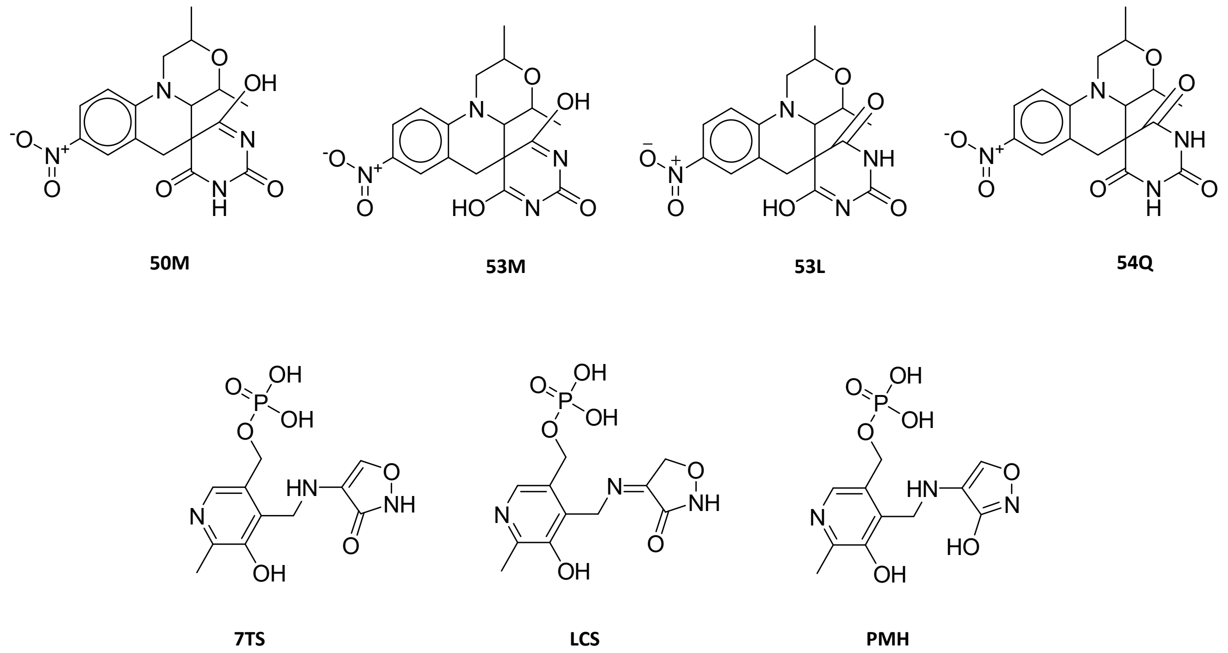
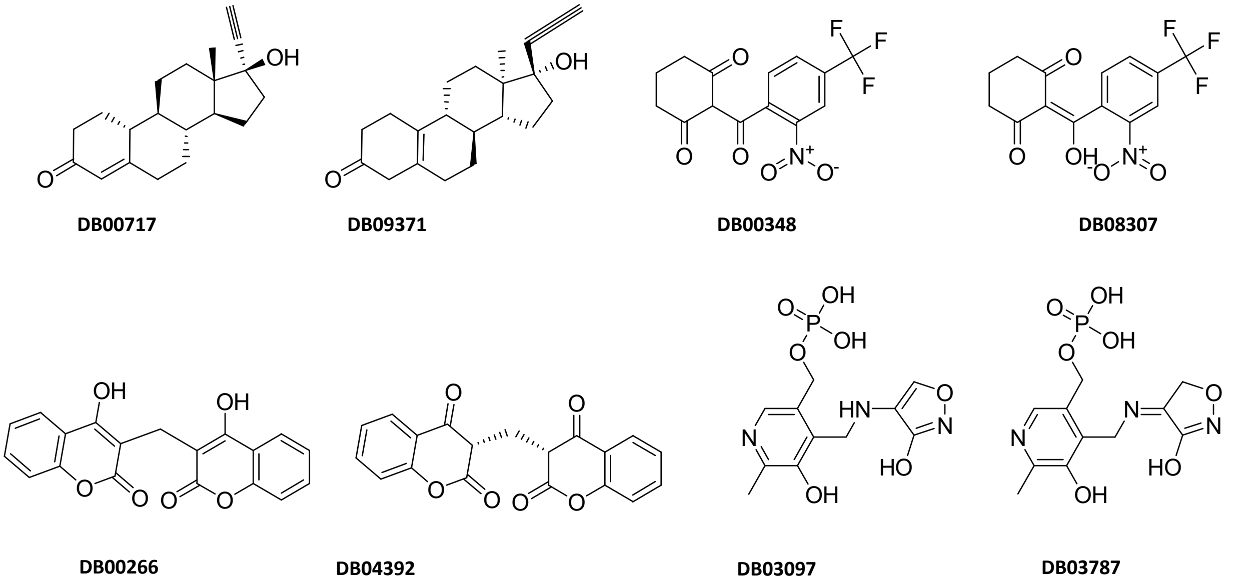


Figure 4. Tautomeric conflict examples among PDB ligands.

DrugBank

This database contains information on drugs and their targets. The main entries are from approved small molecule drugs, approved biologics (proteins, peptides, vaccines, and allergenics), nutraceuticals and experimental (discovery-phase) drugs. The major portion of DrugBank's information is related to chemical data and rest to the drugs' targets.[15]

DB00717 (Norethisterone) and DB09371 (Norethynodrel) are approved progesterone-based drugs. They are tautomers of each other based on our rules. (They also have different UNIIs: T18F433X4S and 88181ACA0M, respectively.) In other cases, one experimental drug shows conflict with other approved or/and experimental drug. DB00266 (Dicoumarol) and DB04392 (only the IUPAC name given in DrugBank) are an approved and an experimental drug, respectively, which form a tautomeric conflict. Similarly, an approved drug (DB00348, Nitisinone) is in tautomeric conflict with an experimental drug (DB08307). We note that some of the conflicts are based on stereocenters that are not tautomerism-persistent according to our rules.[3]



DB00717 DB09371 DB00348 DB08307

DB00266 DB04392 DB03097 DB03787

Figure 5. Tautomeric conflict examples among DrugBank molecules.

CONCLUSION

Tautomeric conflicts are still found in virtually any database we look at. More comprehensive treatment of tautomerism therefore remains a desirable advance in chemistry, especially in database management and compound registration systems.

SUPPORTING INFORMATION

Table S1 contains all 119 transforms used in this paper, shown in SMIRKS format. It is an extension of a rule set we published previously.[4] Table S2 shows the tautomeric conflict counts within each database for all tuple sizes, up to 25. Table S3 shows the conflicts involving non-persistent stereocenters ("stereo conflicts") for all tuple sizes, up to 25.

AUTHOR INFORMATION

Corresponding Author

* Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, NIH, Frederick, MD 21702, USA. E-mail: mn1@mail.nih.gov.

Current Affiliation

Devendra K Dhaked: Department of Pharmacoinformatics, National Institute of Pharmaceutical Education and Research (NIPER), Kolkata – 700054 India

ORCID

Devendra K. Dhaked: 0000-0001-6349-7649

Marc C. Nicklaus: 0000-0002-4775-7030

Competing interests

The authors declare that they have no competing interests.

Funding Sources

The authors received funding from the NCI, NIH, Intramural Research Program.

REFERENCES

(1) Cruz-Cabeza, A. J.; Groom, C. R. Identification, Classification and Relative Stability of Tautomers in the Cambridge Structural Database. *CrystEngComm* **2011**, *13* (1), 93–98. https://doi.org/10.1039/C0CE00123F.

(2) Milletti, F.; Storchi, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases. *J. Chem. Inf. Model.* **2009**, *49* (1), 68–75. https://doi.org/10.1021/ci800340j.

(3) Sitzmann, M.; Ihlenfeldt, W.-D.; Nicklaus, M. C. Tautomerism in Large Databases. *J. Comput. Aided Mol. Des.* **2010**, *24* (6–7), 521–551. https://doi.org/10.1007/s10822-010-9346-4.

(4) Dhaked, D. K.; Ihlenfeldt, W.-D.; Patel, H.; Delannée, V.; Nicklaus, M. C. Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2. *J. Chem. Inf. Model.* **2020**, *60* (3), 1253–1275. https://doi.org/10.1021/acs.jcim.9b01080.

(5) Guasch, L.; Yapamudiyansel, W.; Peach, M. L.; Kelley, J. A.; Barchi, J. J.; Nicklaus, M. C. Experimental and Chemoinformatics Study of Tautomerism in a Database of Commercially Available Screening Samples. *J. Chem. Inf. Model.* **2016**, *56* (11), 2149–2161. https://doi.org/10.1021/acs.jcim.6b00338.

(6) Clemon, P. Private Communication. **2020**.

(7) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Model.* **1994**, *34* (1), 109–116. https://doi.org/10.1021/ci00017a013.

(8) Guasch, L.; Sitzmann, M.; Nicklaus, M. C. Enumeration of Ring–Chain Tautomers Based on SMIRKS Rules. *J. Chem. Inf. Model.* **2014**, *54* (9), 2423–2432. https://doi.org/10.1021/ci500363p.

(9) Ottosson, J. E.; Gränfors, M.; van Pelt, S.; Langborg Weinmann, A.; Nilsson Lill, S. O.; Hulthe, G.; Grönberg, G. Characterization and Demonstration of Drug Compound Ring-Chain Tautomer Formation and Its Impacts on Quality Control. *J. Pharm. Biomed. Anal.* **2021**, *198*, 114020. https://doi.org/10.1016/j.jpba.2021.114020.

(10) Dhaked, D. K.; Guasch, L.; Nicklaus, M. C. Tautomer Database: A Comprehensive Resource for Tautomerism Analyses. *J. Chem. Inf. Model.* **2020**, *60* (3), 1090–1100. https://doi.org/10.1021/acs.jcim.9b01156.

(11) PubChem Data Sources https://pubchem.ncbi.nlm.nih.gov/sources/ (accessed 2020 -04 -09).

(12) Grulke, C. M.; Williams, A. J.; Thillanadarajah, I.; Richard, A. M. EPA's DSSTox Database: History of Development of a Curated Chemistry Resource Supporting Computational Toxicology Research. *Comput. Toxicol.* **2019**, *12*, 100096. https://doi.org/10.1016/j.comtox.2019.100096.

(13) FDA DailyMed ftp://public.nlm.nih.gov/nlmdata/.dailymed/substance_indexing_spl_files.zip.

(14) FDA's Global Substance Registration System https://www.fda.gov/industry/fda-resources-data-standards/fdas-global-substance-registration-system (accessed 2020 -05 -18).

(15) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037.