ChemChain @ InChI

open chemical identifier on the blockchain - a decentralized and authoritative identifier

Dr. Markus Bussen
Executive summary:

open algorithmic identifiers + blockchain = open authoritative identifiers
The present chemical identifier ecosystem: Caffeine

<table>
<thead>
<tr>
<th>Identiﬁers</th>
<th>Chemical and physical data</th>
</tr>
</thead>
<tbody>
<tr>
<td>IUPAC name</td>
<td>Formula: $\text{C}<em>8\text{H}</em>{10}\text{N}_4\text{O}_2$</td>
</tr>
<tr>
<td>1,3,7-Trimethylpurine-2,6-dione</td>
<td>Molar mass: 194.19 g/mol g·mol$^{-1}$</td>
</tr>
<tr>
<td>CAS Number</td>
<td>58-08-2</td>
</tr>
<tr>
<td>PubChem CID</td>
<td>2519</td>
</tr>
<tr>
<td>IUPHAR/BPS</td>
<td>407</td>
</tr>
<tr>
<td>DrugBank</td>
<td>DB00201</td>
</tr>
<tr>
<td>ChemSpider</td>
<td>2424</td>
</tr>
<tr>
<td>UNII</td>
<td>3G6A5W338E</td>
</tr>
<tr>
<td>KEGG</td>
<td>D00528</td>
</tr>
<tr>
<td>ChEBI</td>
<td>CHEBI:27732</td>
</tr>
<tr>
<td>ChEMBL</td>
<td>ChEMBL113</td>
</tr>
<tr>
<td>PDB ligand</td>
<td>CFF (PDBe, RCSB PDB)</td>
</tr>
<tr>
<td>CompTox Dashboard</td>
<td>DTXSID0020232</td>
</tr>
<tr>
<td>ECHA InfoCard</td>
<td>100.000.329</td>
</tr>
<tr>
<td>SMILES</td>
<td>[hide]</td>
</tr>
<tr>
<td>InChI</td>
<td>[hide]</td>
</tr>
<tr>
<td>CAS#: 58-08-2</td>
<td>[hide]</td>
</tr>
</tbody>
</table>

**Chemistry ecosystem identifiers:**
- Reference Identifiers
- Reference Authoritative Identifiers
- Algorithmic Identifiers

- **ChemSpider:** 2424
- **PubChem:** 2519
- **IUPHAR/BPS:** 407
- **DrugBank:** DB00201
- **ChemSpider:** 2424
- **UNII:** 3G6A5W338E
- **KEGG:** D00528
- **ChEBI:** CHEBI:27732
- **ChEMBL:** ChEMBL113
- **PDB ligand:** CFF (PDBe, RCSB PDB)
- **CompTox Dashboard:** DTXSID0020232
- **ECHA InfoCard:** 100.000.329

**Molar mass:** 194.19 g/mol g·mol$^{-1}$
**3D model (JSmol):** Interactive image
**Density:** 1.23 g/cm$^3$
**Melting point:** 235 to 238 °C (455 to 460 °F) (anhydrous)$^{[8][9]}$
Algorithmic identifiers: InChI

InChI string and key: example caffeine

InChI is an open textual identifier for chemical substances, designed to provide a standard way to encode molecular information and to facilitate the search for such information in decentralized systems.

Considerations regarding safety and authority of the InChI system:
- InChI versions
- Hash collisions
- Trusting the InChI sender/provider is required.
The InChI system: from a network perspective

Decentralized DBs using InChI

Inconsistency of InChI and other systematic identifiers in different databases

Table 3 Consistency of MOLs and systematic identifiers (in % agreement) within databases

<table>
<thead>
<tr>
<th>Database</th>
<th>MOL-InChI</th>
<th>MOL-SMILES</th>
<th>MOL-IUPAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrugBank</td>
<td>982</td>
<td>98.5</td>
<td>90.0</td>
</tr>
<tr>
<td>ChEBI</td>
<td>965</td>
<td>96.5</td>
<td>75.3</td>
</tr>
<tr>
<td>HMDB</td>
<td>893</td>
<td>37.2</td>
<td>55.7</td>
</tr>
<tr>
<td>PubChem</td>
<td>977</td>
<td>97.8</td>
<td>87.2</td>
</tr>
<tr>
<td>NPC</td>
<td>-</td>
<td>93.4</td>
<td>-</td>
</tr>
</tbody>
</table>

Conclusion: algorithmic identifiers like InChI, SMILES need to be consistently applied among the different chemistry sources. However, inconsistencies exist among

Research Article: Consistency of systematic chemical identifiers within and between small-molecule databases

Abstract

Background: Correctness of structures and associated metadata within public and commercial chemical databases greatly impacts drug discovery research activities such as quantitative structure–property relationships modeling and compound novelty checking. MOL, InChI, SMILES, metadata, IUPAC names, and InChI strings are ubiquitous file formats and systematic identifiers for chemical structures. While interchangeable for many cheminformatics purposes there have been no studies on the inconsistency of these structure identifiers due to various approaches for data integration, including the use of different software and different rules for structure standardization. We have investigated the consistency of systematic identifiers of small molecules within and between some of the commonly used chemical resources, with and without structure standardization.

Results: The consistency between systematic chemical identifiers and their corresponding MOL representation varies greatly between data sources (37.2%–95.6%). We observed the lowest overall consistency for MOL-IUPAC names. Disregarding stereochemistry increases the consistency (84.8% to 95.9%). A wide variation in consistency also exists between MOL representations of compounds linked via cross-references (25.6% to 93.7%). Removing stereochemistry improved the consistency (97.1% to 95.9%).

Conclusions: We have shown that considerable inconsistency exists in structural representation and systematic chemical identifiers within and between databases. This can have a great influence especially when merging data and if systematic identifiers are used as a key index for structure integration or for querying several databases. Regenerating systematic identifiers starting from their MOL representation and applying well-defined and documented chemical standardization rules to all compounds prior to creating them can dramatically increase internal consistency.

Keywords: Molecular structure, Chemical databases, Systematic chemical identifiers, Quality control, InChI, SMILES, IUPAC
Centralized reference identifier system: network perspective

Note:
Reference identifiers are hosted on centralized systems.
The owner of such important DB has enormous power over the users because of the dependency.
Trust models of chemical identifiers

Reference identifiers

Sender

Receiver

Caffiene
Registry number

3 parties involved
3rd party trusted partner

Central registry

Algorithmic identifiers

Caffeine
Algorithmic identifier

2 parties involved
Sender needs to be trusted
Transfer of value

3rd party Bank

sender

Bank

Money/value

3 parties involved
Bank trusted partner

receiver

Blockchain based

Bitcoin
2 parties + blockchain (pseudo 3rd)
Blockchain is trusted
Inspiration for ChemChain - Bitcoin explorer

<table>
<thead>
<tr>
<th>Height</th>
<th>Hash</th>
<th>Mined</th>
<th>Miner</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>591448</td>
<td>000000000000000000000000000000001addd5284fc5c22c73411...</td>
<td>15:50 PM</td>
<td>Unknown</td>
<td>1,325,534 bytes</td>
</tr>
<tr>
<td>591447</td>
<td>00000000000000000000000000000000079f4f2f32bb1bc275277e...</td>
<td>15:20 PM</td>
<td>F2Pool</td>
<td>1,244,431 bytes</td>
</tr>
<tr>
<td>591446</td>
<td>00000000000000000000000000000000019b92aabb05fb863a29...</td>
<td>15:17 PM</td>
<td>F2Pool</td>
<td>1,319,782 bytes</td>
</tr>
<tr>
<td>591445</td>
<td>0000000000000000000000000000000008f17c81008fc3b8a772af...</td>
<td>14:54 PM</td>
<td>BTC.com</td>
<td>584,874 bytes</td>
</tr>
<tr>
<td>591444</td>
<td>0000000000000000000000000000000003b4a388985daaca34...</td>
<td>14:49 PM</td>
<td>BTC.TOP</td>
<td>458,524 bytes</td>
</tr>
<tr>
<td>591443</td>
<td>00000000000000000000000000000000033c0ecbb201398b69c2...</td>
<td>14:46 PM</td>
<td>BTC.com</td>
<td>1,009,367 bytes</td>
</tr>
<tr>
<td>591442</td>
<td>00000000000000000000000000000000025ed7abdb13863da67b...</td>
<td>14:40 PM</td>
<td>Unknown</td>
<td>1,211,424 bytes</td>
</tr>
<tr>
<td>591441</td>
<td>000000000000000000000000000000000fb8fa7cc77536c8f2f4b...</td>
<td>14:30 PM</td>
<td>Unknown</td>
<td>1,318,843 bytes</td>
</tr>
<tr>
<td>591440</td>
<td>0000000000000000000000000000000002d817fbbda2667d62830...</td>
<td>14:17 PM</td>
<td>BTC.TOP</td>
<td>1,188,895 bytes</td>
</tr>
<tr>
<td>591439</td>
<td>000000000000000000000000000000000938c725985391d7477a...</td>
<td>14:12 PM</td>
<td>Unknown</td>
<td>1,385,586 bytes</td>
</tr>
</tbody>
</table>
Hypophysis:

A future-proof chemical identifiers system needs to be: authoritative, open and scalable

<table>
<thead>
<tr>
<th></th>
<th>Reference identifiers</th>
<th>Algorithmic identifiers</th>
<th>ChemChain</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Open</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>- Scalable</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>- Safe</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>- Authoritative</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>- Decentralized</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>- Cost effective</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

- **open code**
- **cryptography**
- **distributed ledger P2P**
Chemical identifier evolution: network point of view

- centralized
- decentralized
- distributed
Three major blockchain components

**Consensus Mechanism (PoW, PoS)**

Consensus mechanism prevents the double-spending problem. CM keeps the network clean and honest.

**Peer-to-peer network**

Distributed ledger technology enables every node to have the exact same copy of transactions.

**Cryptohashed blocks**

Hashed blocks keep all containing transactions immutable.

**ChemChain**

Consensus mechanism controls the uniqueness and relevance of structures in ChemChain.

Distributed ledger technology every chemist (organization) to keep a copy of structures.

Hashed blocks make it impossible to change the registered structures.
ChemChain - chemical ID Registration process

ChemChain ID-generation

Structure

Submission

bl.1
bl.2
bl.3
bl.4
bl. ...

Structure-ID

Lookup
Blockchain principles of ChemChain
ChemChain is protected by cryptography

In a blockchain network all nodes have the same copy of the ledger/DB. Cryptography and Peer-to-peer protocols assure the consistency and synchronization of the data in the ledger/DB.

Corrupting data is practically impossible due to the economic burden of the cryptographic hash functions. Data stored in blockchains is immutable.
What information should be stored in ChemChain block?

**Structures**
- CAS#: 58-08-2
- InChI/SMILES style: BTEGFTGE-
  - RNB-HG-TB

**Identifier**
- Consecutive numbers: 1, 2, 3, ..., 9999
- InChI/SMILES - style: BTEGFTGE-
  - CAS#: 58-08-2

**Personal identifiers**
- Machine readable:

**Unknown:**
- Names and
A chemistry-consensus mechanism controls ChemChain

Chemistry Consensus

- IS chemically possible? [YES/NO]
- IS substance realistic? [YES/NO]
- IS the substances used? [YES/NO]
ChemChain could add existing structures quickly

Existing structures from open chemistry sources
ChemChain semantically connects chemistry
ChemChain acts as a trusted authority

**Caffeine Registry number**
- 3 parties involved
- 3rd party trusted partner

**Caffeine algorithmic identifier**
- 2 parties involved
- Sender needs to be trusted

**Caffeine blockchain identifier**
- 2 parties and one BC network
- BC network is trusted
ChemChain is the better system, not the better identifier.
Examples for chemical/pharma blockchain projects

**BASF and arc-net collaborate to use blockchain technology for livestock sustainability**

Florehm, Farm, New Jersey, and Belfast, Northern Ireland, April 17, 2018 – BASF CropScience and arc-net, are collaborating to use blockchain technology to capture and analyze sustainability parameters in livestock production along the value-chain, according a technology company utilizing blockchain technology to provide transparency in the agrifood industry. As part of this collaboration, BASF will use its blockchain technology tools, AppHarvest’s technology, to provide a secure, cloud-based information platform for environmental impacts along the animal production value chain.

With global pressure to reduce emissions from the agricultural and livestock sectors, customers and regulators worldwide are becoming more focused on creating sustainable systems. BASF and arc-net are positioning to support the animal production value chain meet these demands. BASF’s blockchain tool, Agilisate, and arc-net’s technology platform, AppHarvest’s technology, help to provide a secure, cloud-based information platform for environmental impacts along the animal production value chain.

**Walmart considers blockchain technology for tracing chemicals**

Potential to create ‘a new era of transparency’

21 June 2018 – Global business has a right to know, data, detailed, transparent, minority voices

US retail giant Walmart is assessing whether the digital technology ‘blockchain’ can be used to trace chemicals across some of its products and packaging.

Blockchain is a digital record keeping system that enables the creation and maintenance of a growing number of records, allowing fast backing of information. It was originally created to manage transactions through the crypto-currency Bitcoin, but has since shown potential for storing and retrieving even other forms of data.
ChemChain as reference layer for other service layers

Reference Layer                      Service Layer

ChemChain  ReactionChain  PatentChain  RegulatoryChain
ChemChain contact:

Dr. Markus Bussen

markus.bussen@protonmail.ch

m.bussen@elsevier.com