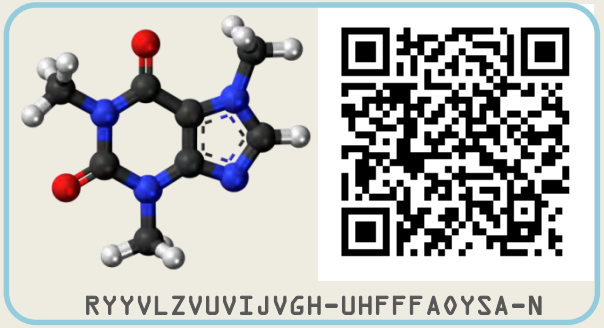


ChemChain @ InChI

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

Dr. Markus Bussen

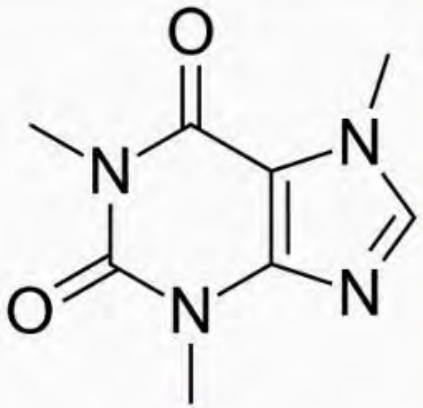
Executive summary:

open algorithmic identifiers
+ blockchain

= open authoritative identifiers

The present chemical identifier ecosystem:

Caffeine



Identifiers	
IUPAC name	
1,3,7-Trimethylpurine-2,6-dione	
CAS Number	58-08-2 ✓
PubChem CID	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 (EPA)	DTXSID0020232
ECHA InfoCard	100.000.329

Chemical and physical data	
Formula	$C_8H_{10}N_4O_2$
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]

SMILES [hide]

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

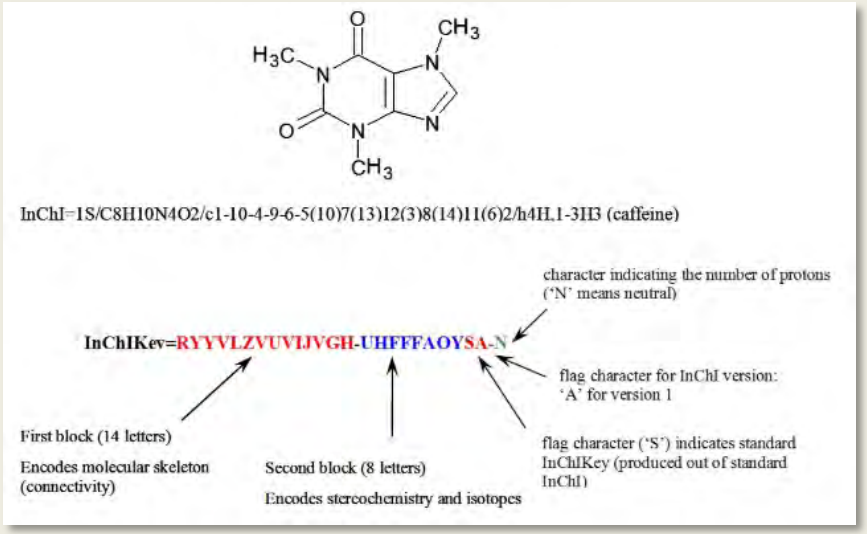
InChI [hide]

InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Key:RYYVLZVUVIJVGH-UHFFFAOYSA-N ✓

Algorithmic identifiers: InChI

InChI string and key: example caffeine



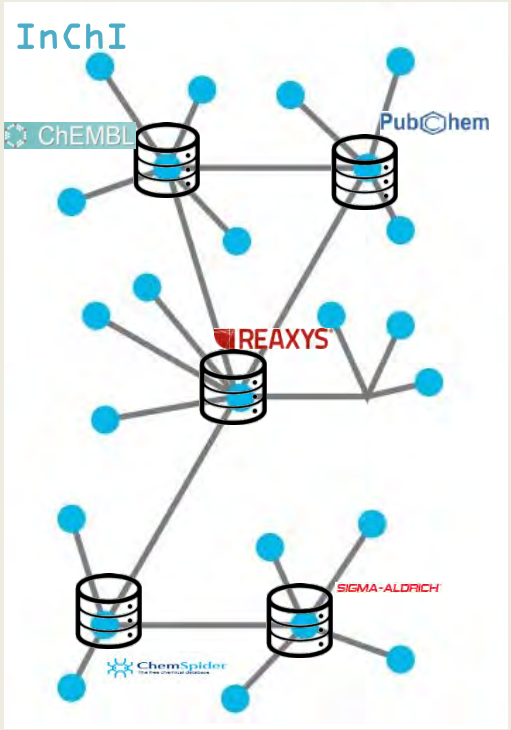
InChI is an open textual identifier for chemical substances, designed to provide a standard way to encode molecular information and to facilitate the search for such information in decentralized

Considerations regarding safety and authority of the InChI system:

- InChI versions
- Hash collusions
- 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 data

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

Database	MOL–InChI	MOL–SMILES	MOL–IUPAC
DrugBank	98.2	98.5	90.0
ChEBI	96.5	96.5	75.3
HMDB	89.3	37.2	55.7
PubChem	97.7	97.8	87.2
NPC	-	93.4	-

Akhondi et al. Journal of Cheminformatics 2012, 4:35
http://www.jcheminf.com/content/4/1/35

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

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

Saber A Akhondi[1*], Jan A Kors[1] and Sorel Muresan[2]

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 modelling and compound novelty checking. MOL files, SMILES notations, 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 standardisation. 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 standardisation.

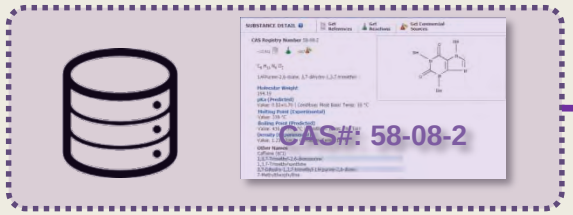
Results: The consistency between systematic chemical identifiers and their corresponding MOL representation varies greatly between data sources (37.2%-98.5%). We observed the lowest overall consistency for MOL-IUPAC names. Disregarding stereochemistry increases the consistency (84.8% to 99.9%). A wide variation in consistency also exists between MOL representations of compounds linked via cross-references (25.8% to 93.7%). Removing stereochemistry improved the consistency (47.6% to 95.6%).

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 cross-querying several databases. Regenerating systematic identifiers starting from their MOL representation and applying well-defined and documented chemistry standardisation rules to all compounds prior to creating them can dramatically increase internal consistency.

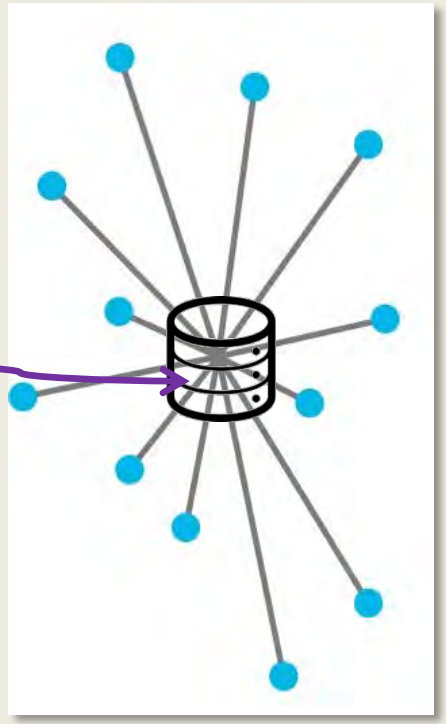
Keywords: Molecular structure, Chemical databases, Systematic chemical identifiers, Quality control, InChI, SMILES, IUPAC

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

Centralized reference identifier system: network perspective



centralized



Note:

Reference identifiers are hosted on centralized systems.

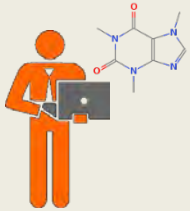
The owner of such important DB has enormous power over the users because of the dependency.

Trust models of chemical identifiers

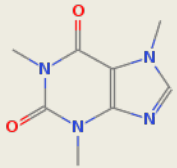
Reference identifiers

Algorithmic identifie

sender



receiver

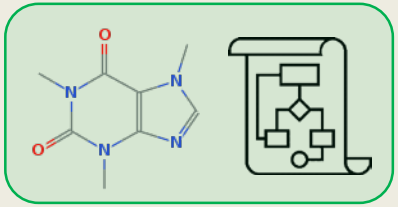


central registry

Caffeine Registry number

3 parties involved

3rd party trusted partner



Caffeine Algorithmic identifier

2 parties involved

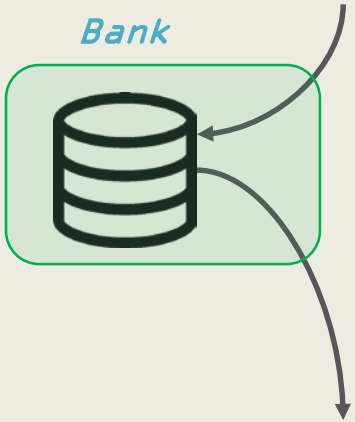
Sender needs to be trusted

Transfer of value

sender

receiver

3rd party Bank

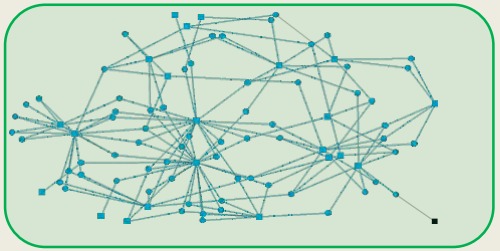


Money/value

3 parties involved

Bank trusted partner

Blockchain based

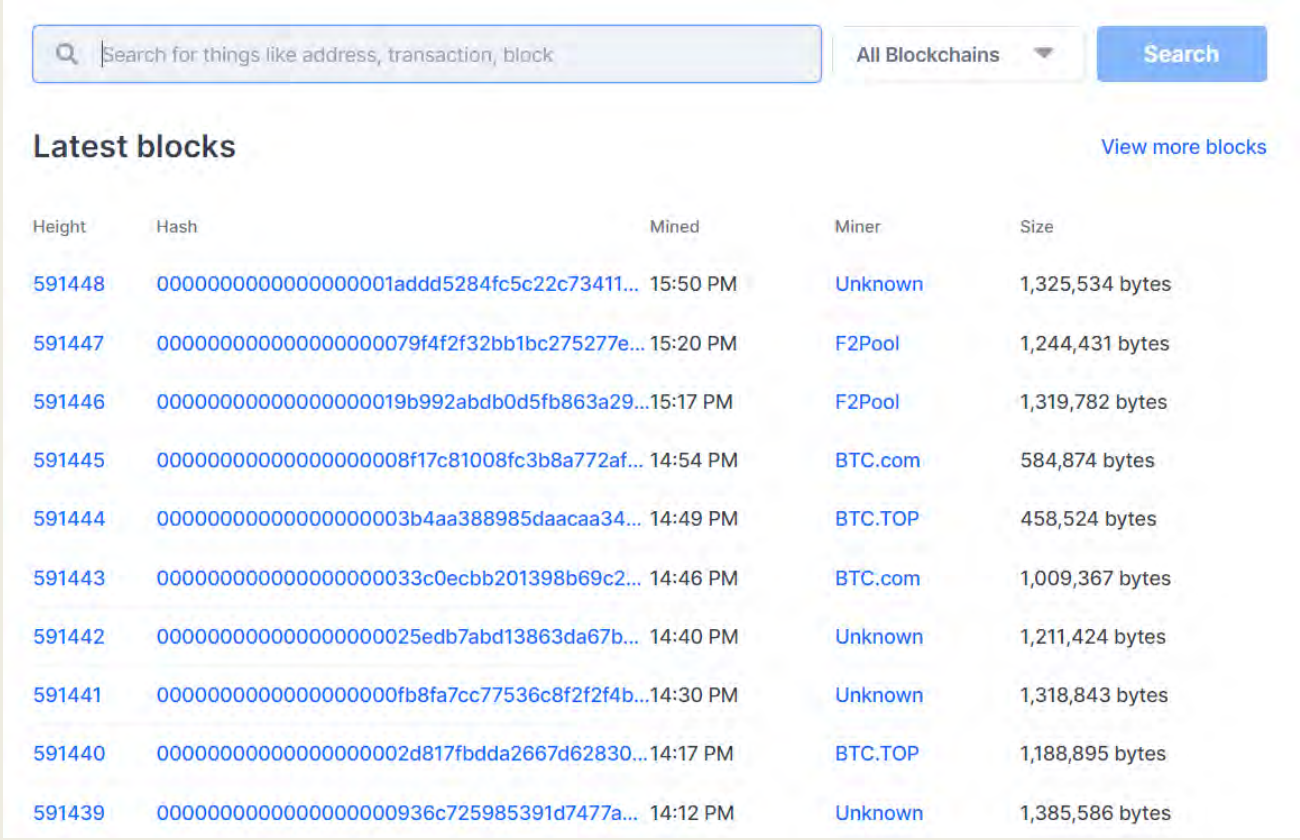


Bitcoin

2 parties + blockchain (pseudo 3rd)

Blockchain is trusted

Inspiration for ChemChain - Bitcoin explorer



Hypophysis:

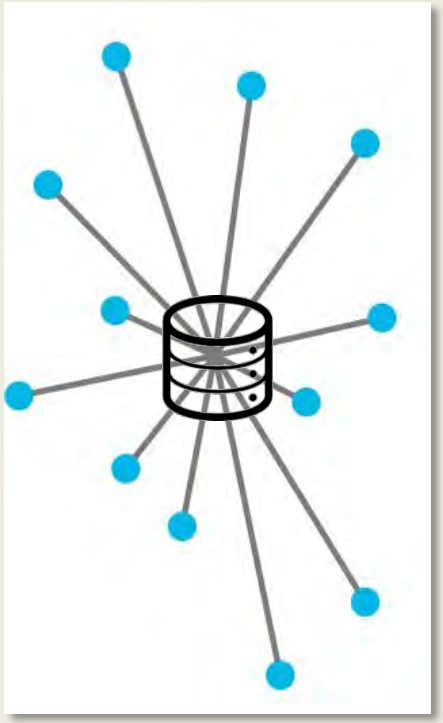
A future-proof chemical identifiers system needs to be:

authoritative, open and scalable

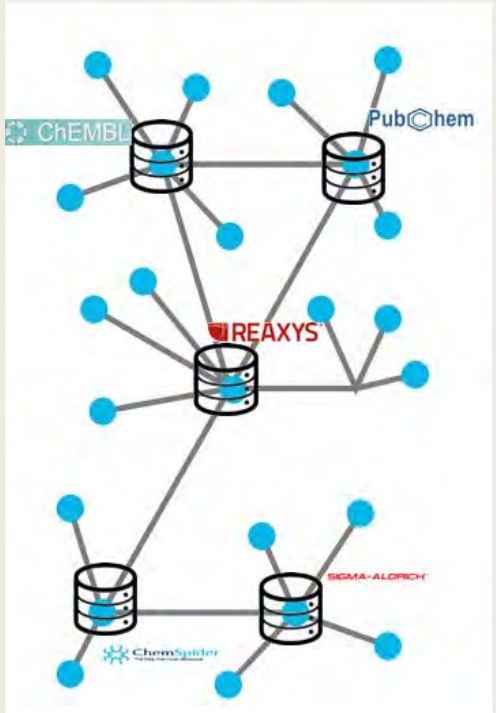
	Reference identifiers	Algorithmic identifiers	ChemChain	
- Open	NO	YES	YES	open code
- Scalable	NO	YES	YES	
- Safe	YES	NO	YES	cryptography
- Authoritative	YES	NO	YES	
- Decentralized	NO	YES	YES	distributed ledger P2P
- Cost effective	NO	YES	YES	

Chemical identifier evolution: network point of view

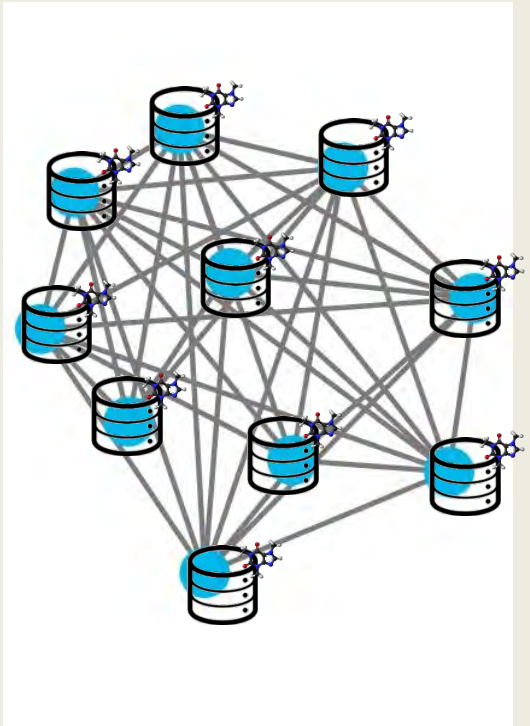
centralized



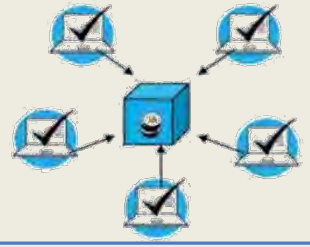
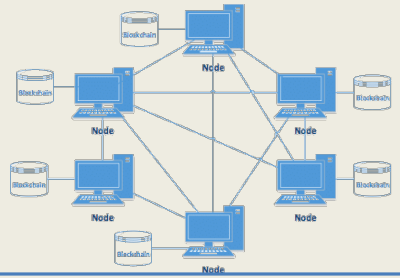
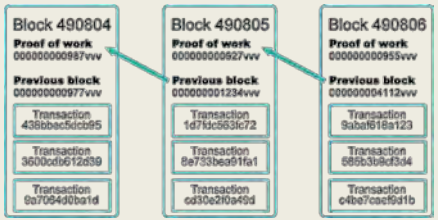
decentralized



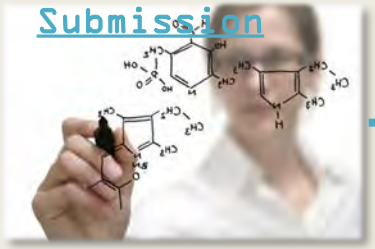
distributed

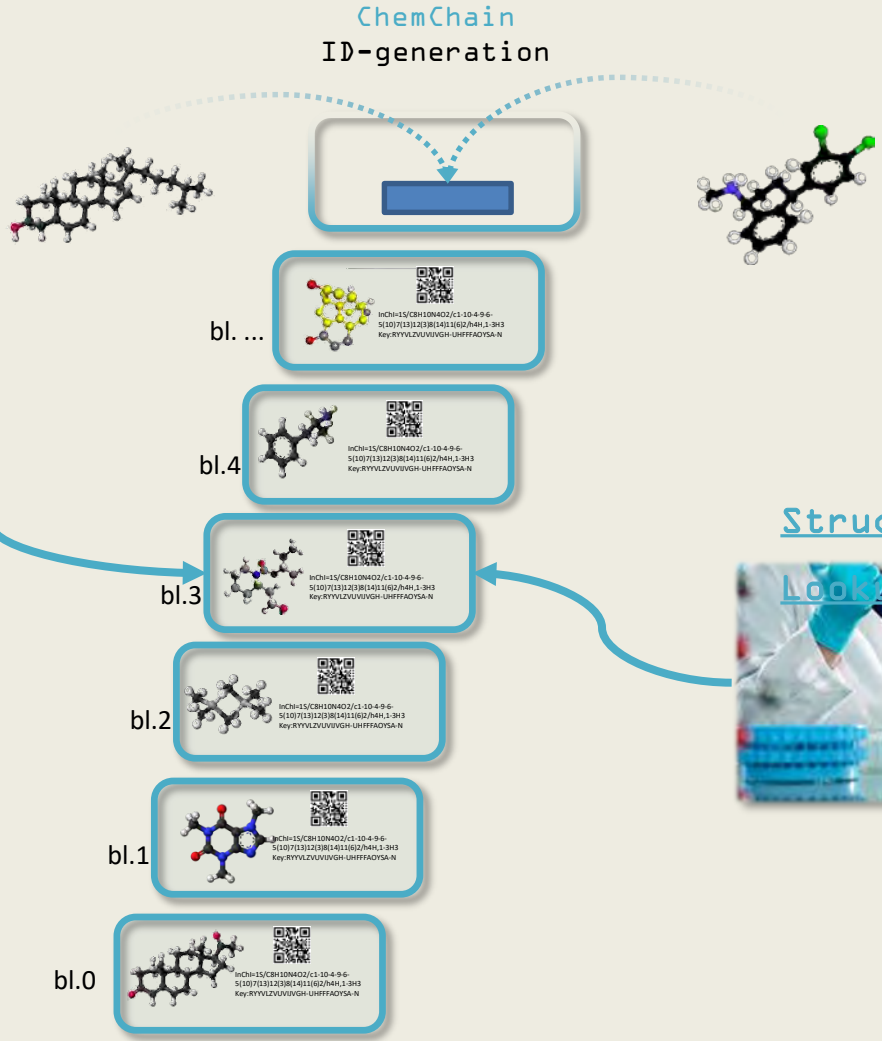


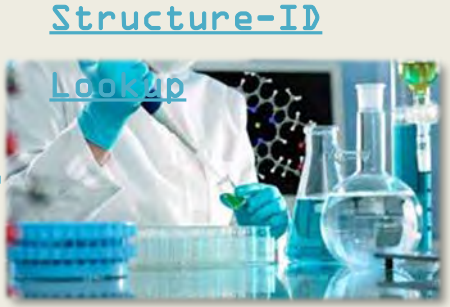
Three major blockchain components

		Bitcoin	ChemChain
Consensus Mechanism (PoW, PoS)		Consensus mechanism prevents the double-spending problem. CM keeps the network clean and honest.	Consensus mechanism controls the uniqueness and relevance of structures in ChemChain.
Peer-to-peer network		Distributed ledger technology enables every node to have the exact same copy of transactions.	Distributed ledger technology every chemist (organization) to keep a copy of structures.
Crypto/hashed blocks		Hashed blocks keep all containing transactions immutable.	Hashed blocks make it impossible to change the registered structures.

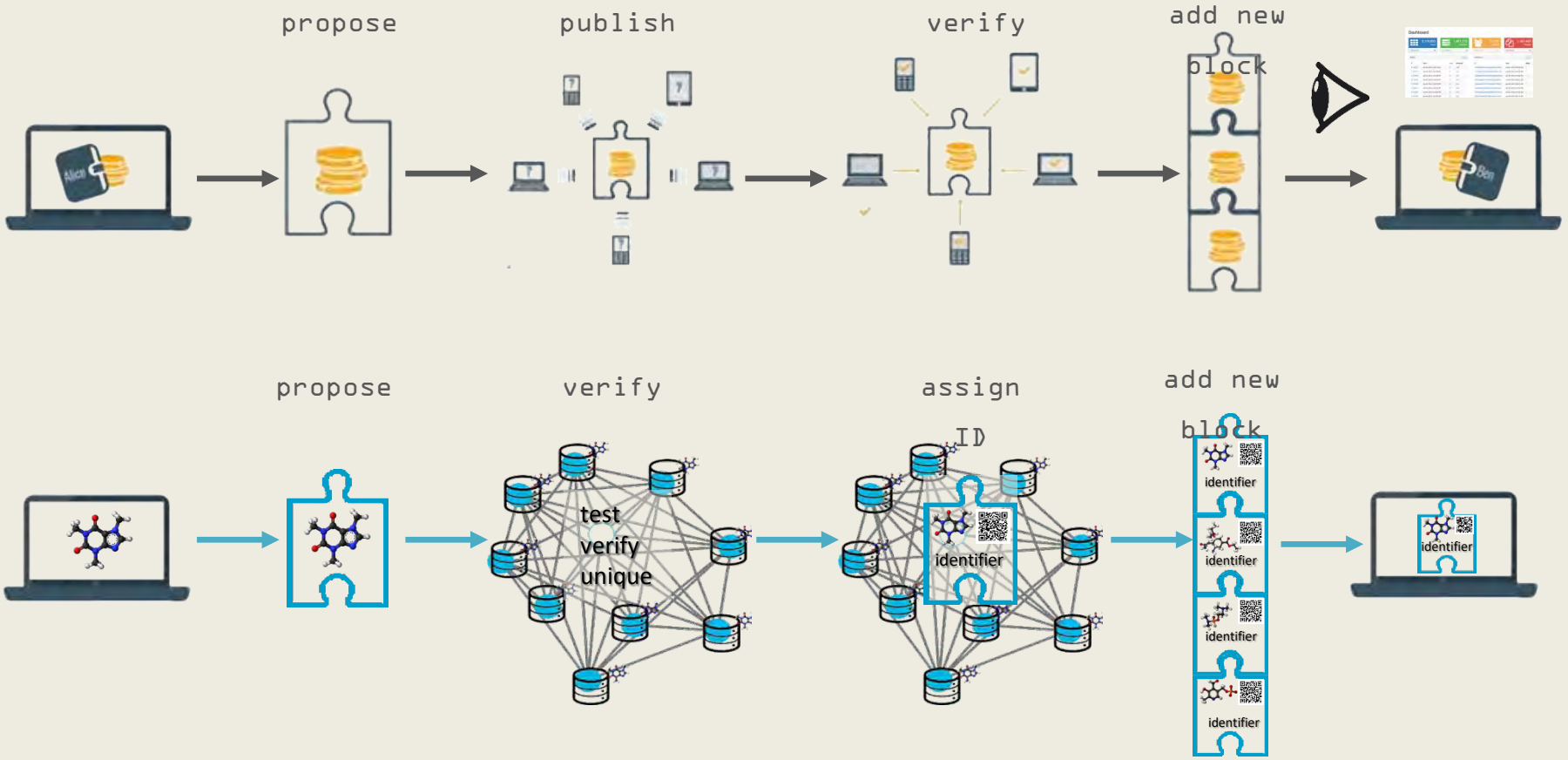
ChemChain - chemical ID Registration process





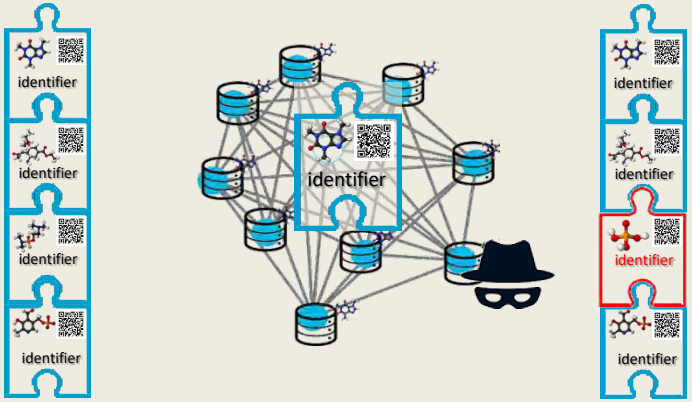


Blockchain principles of ChemChain

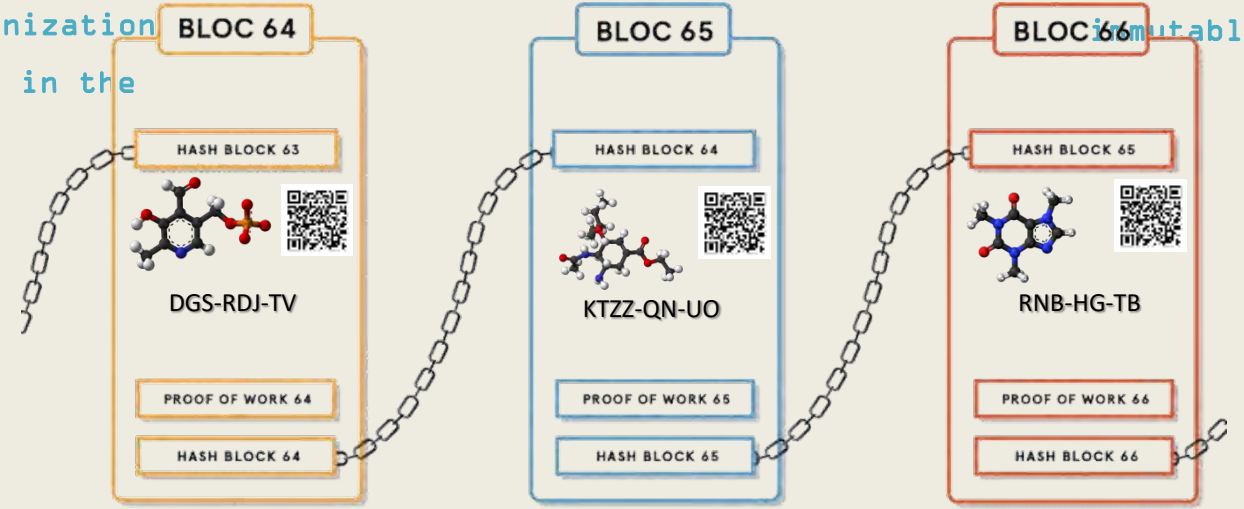


ChemChain is protected by cryptography

In a blockchain network all nodes have 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

identifier

consecutive numbers: 1,2,3,… 9999

InChI/SMILEs - style: BTEGFTGE-YTZ-H

~~CAS#: 58-08-2~~

RNB-HG-TB
3424525

personal

identifiers

Affiliation

Location

IP

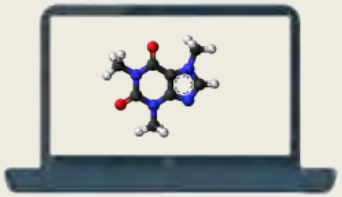
machine readable

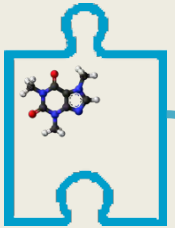
Big

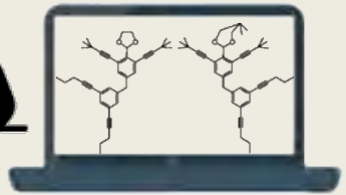
Unknown:

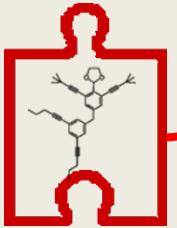
Names and

A chemistry-consensus mechanism controls ChemChain









Chemistry Consensus

IS chemically possible?

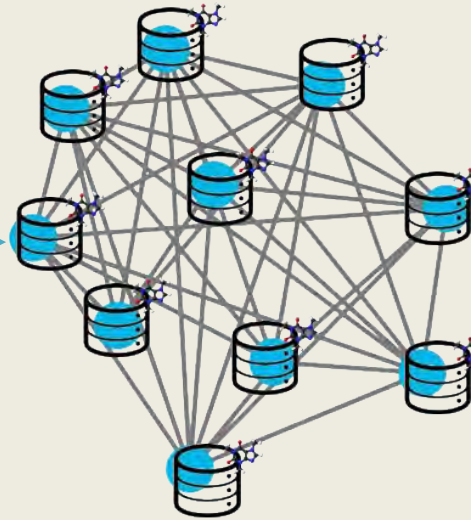
YES/NO

IS substance realistic?

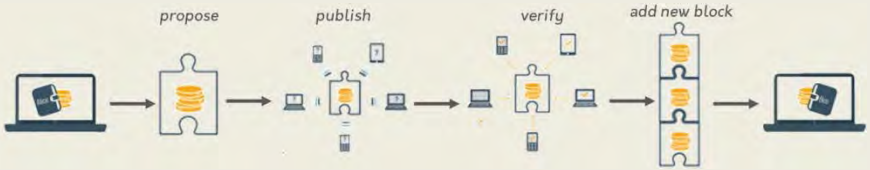
YES/NO

IS the substances used?

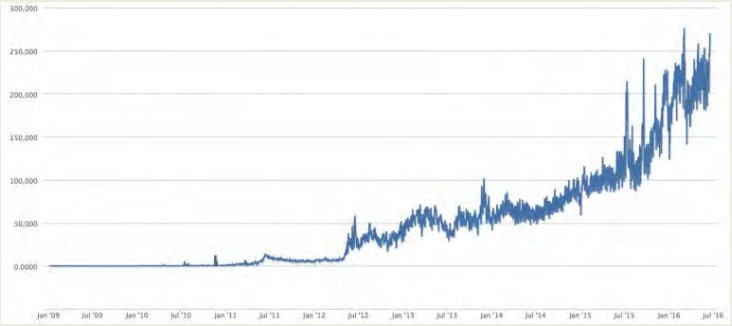
YES/NO

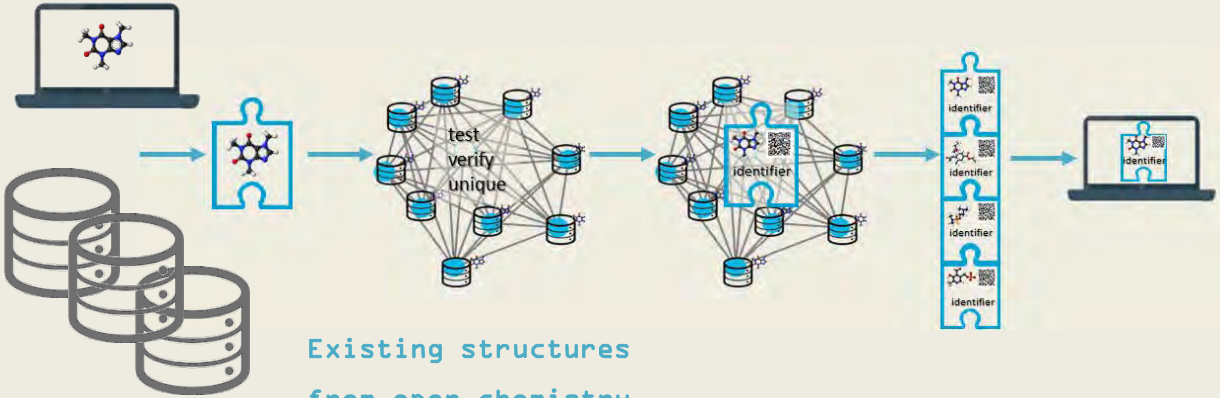


ChemChain could add existing structures quickly



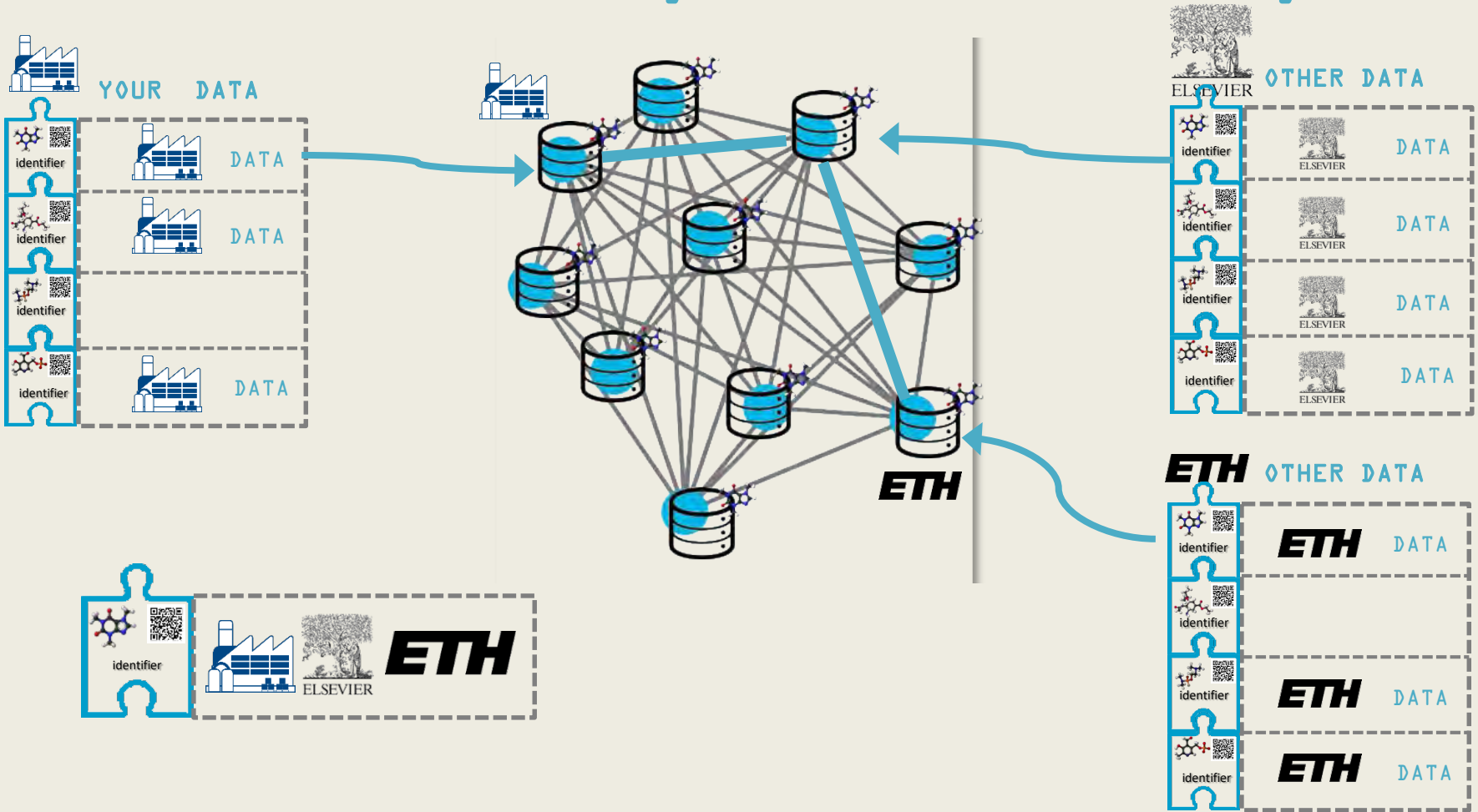
Bitcoin transactions/day





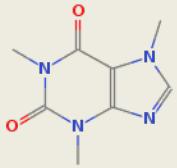
Existing structures from open chemistry sources

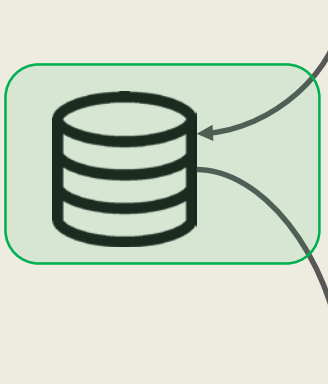
ChemChain semantically connects chemistry



ChemChain acts as a trusted authority

Reference identifiers



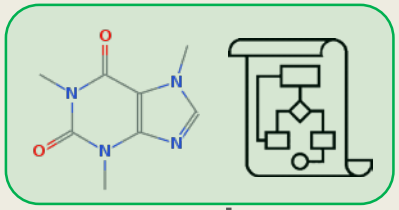


Caffeine Registry number

3 parties involved

3rd party trusted partner

Algorithmic identifiers

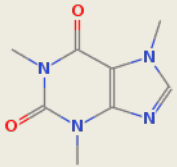


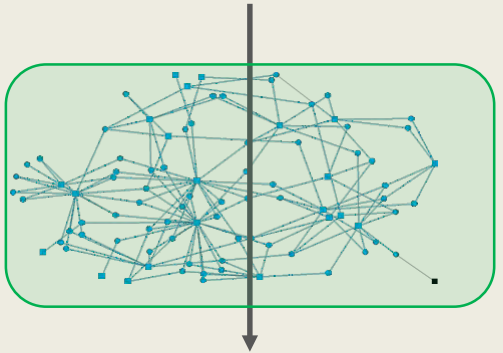
Caffeine algorithmic identifier

2 parties involved

Sender needs to be trusted

Blockchain identifie





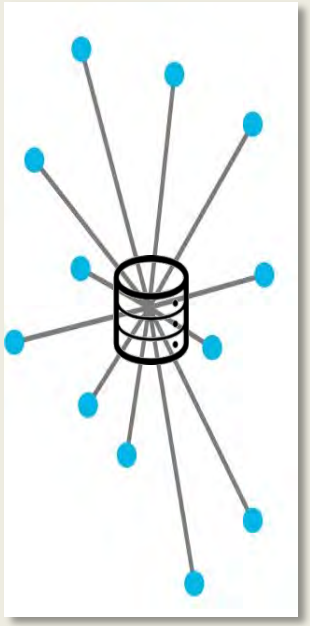
Caffeine blockchain identifier

2 parties and one BC network

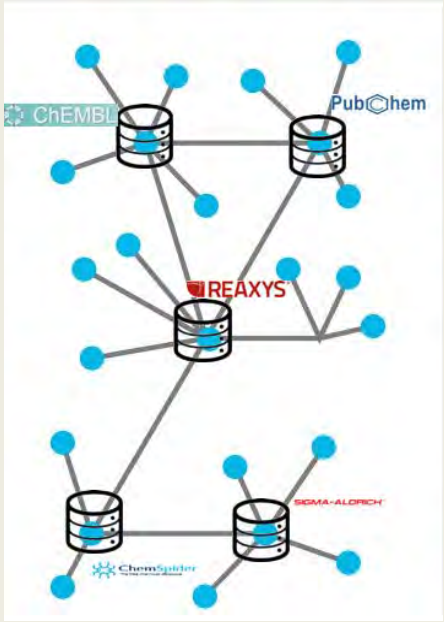
BC network is trusted

ChemChain is the better system, not the better identifier

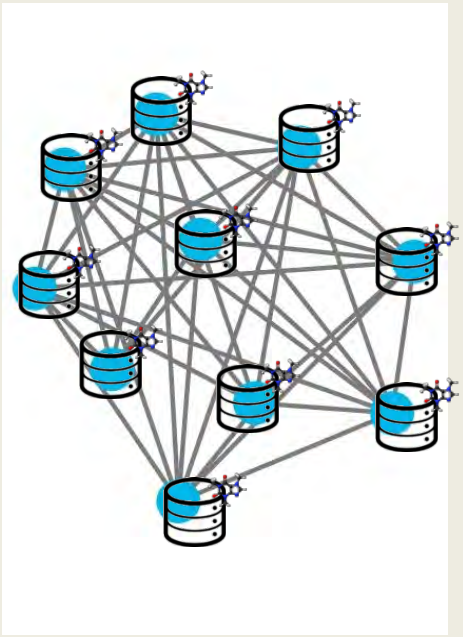
centralized



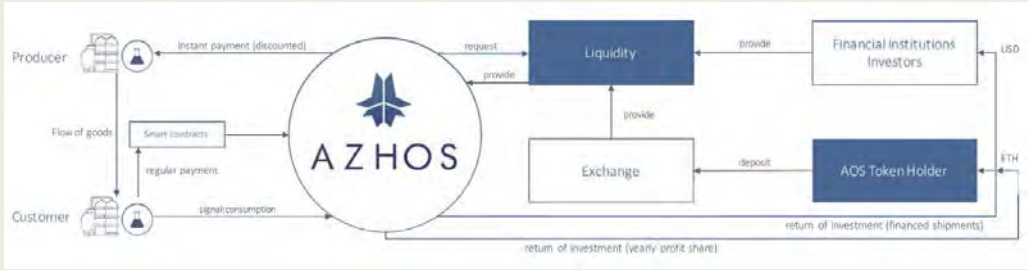
decentralized



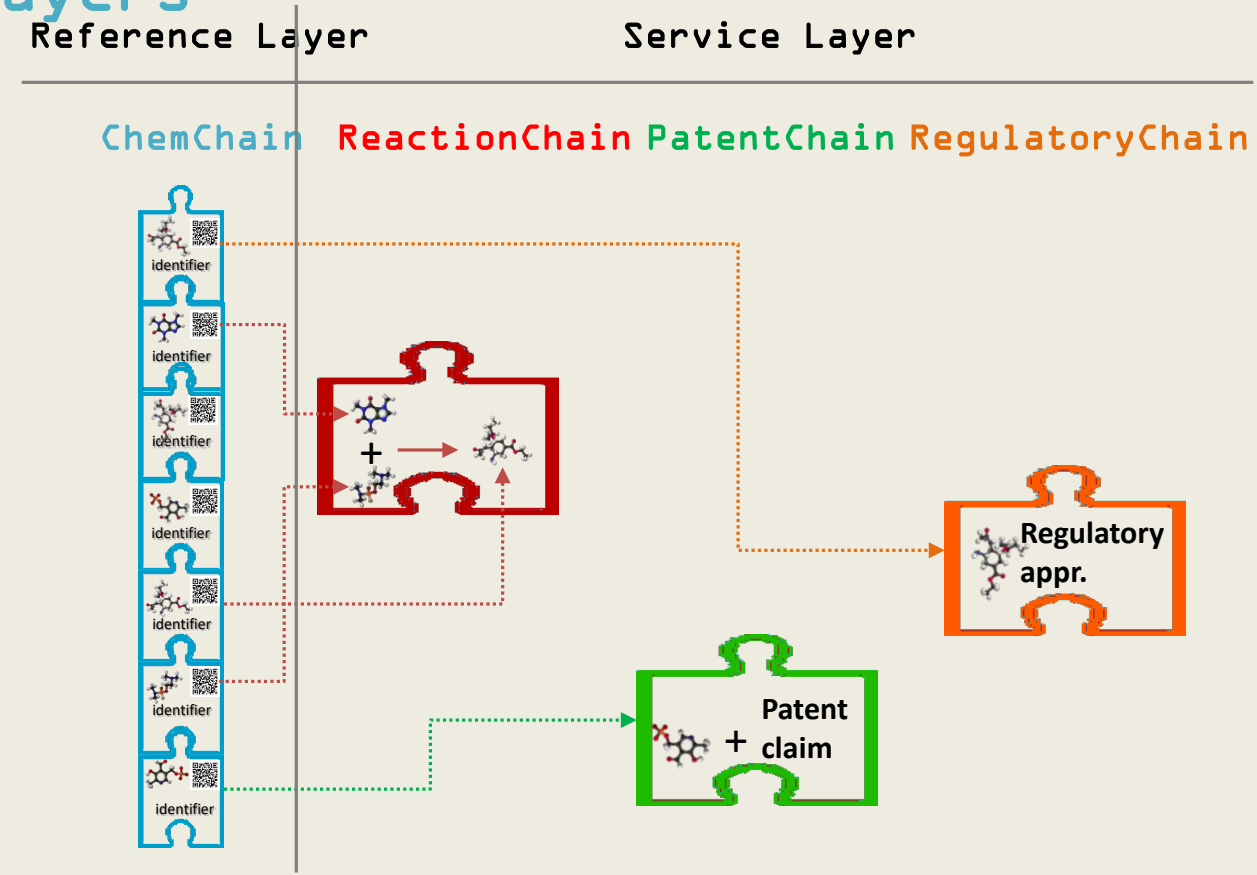
distributed



Examples for chemical/pharma blockchain projects



ChemChain as reference layer for other service layers



ChemChain contact:

Dr. Markus Bussen

markus.bussen@protonmail.ch

m.bussen@elsevier.com