Moving beyond Chemical Cartridges
by Federating our Chemical Searches

Azadi Golbamaki

Data Scientist

Syngenta Crop Protection, Chemistry

Switzerland, Stein

azadi.Golbamaki@syngenta.com

On behalf of

Marco Stenta

Computational Chemist

Syngenta Crop Protection, Chemistry

Switzerland, Stein

marco.stenta@syngenta.com

13th October 2022

Chemical Cartridge

the chemical structure is not a native database data type

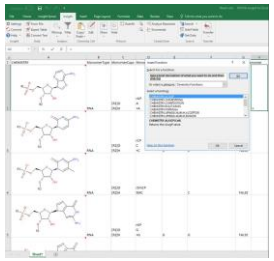
chemical cartridges enable CRUD operations with chemical structures

Chemical registration

- Create
- Read
- Update
- Delete

Chemical searches

- Identity
- similarity
- substructure
- superstrucutre



RDKit

pandas

BIOVIA

ORACLE

RDKit

Chemaxon

syngenta

Specialized software beyon chemical cartridges

BioSolveIT

infiniSee

FTreeS

Ultra-Fast searches in combinatorial space

WuXi AppTec

GalaXi **1.7 billion**

eMolecules eXplore **2.8 trillion**

Support Ukraine CHEM SPACE **201 million**

Freedom Space

Enamine REAL SPACE 31 billion

OTAVA chemicals 12 billion

CHEMriya

ARTHOR

Substrucutre, similarity [chemicals, reaction]

Reaxys® 256 million Substances
60 million reactions

Ultra-Fast searches in enumerated space

PISTACHIO **14.79** million reaction details (**4.66M** unique)
11.29 million recognized reactions (NameRxn),

SMALLWORLD

Similarty [chemicals]

Enamine REAL DB 5.5 billion substances

syngenta

That is great, but ...

Do NOT waste time testing these software,

***just buy them!** They are great!*

(cit. Marco)

A single corporate chemical cartridge is an *efficient* solution, albeit *ineffective*

- One vendor, one procurement, one support, one license, ...
- Insufficient speed, poor life-cycle (obsolescence)

Maintaining a series of specialized tools is an *effective* solution, albeit *inefficient*

- Unlimited power: Full flexibility, state-of-the-art always
- Many vendor, many procurement, many supports, many license, ...
- Difficult to integrate output: many interfaces, sources, etc.; knowledge
- Difficult to keep everything up to date (software release, databases, etc)

Users are enthusiastic…but they still do not use them.

I am puzzled.

(cit. Marco)

Our solution: forging the API of Power.

1) Cloud-based platform that hosts (or connect to) the software programs

2) Automatic chemoinformatic pipeline that consumes the data sources and transforms them into the input required by each software (who cares about data duplication?)

3) Expressive REST API that exposes all the commands of each software, allowing the platform to performs searches and retrieve the results (asynchronously)

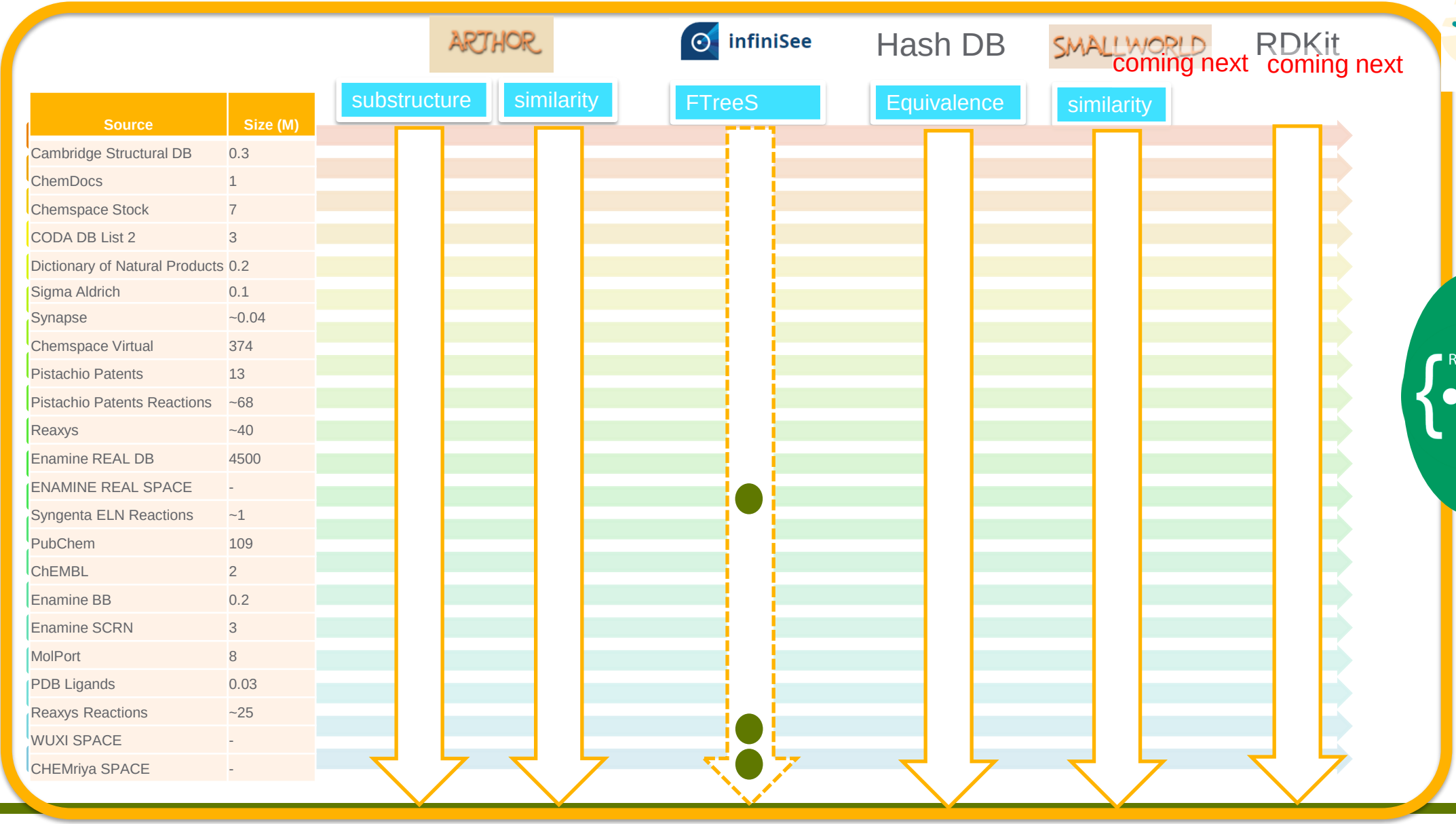
4) Python SDK: to replace or complement cartridge searches in many applications across the company

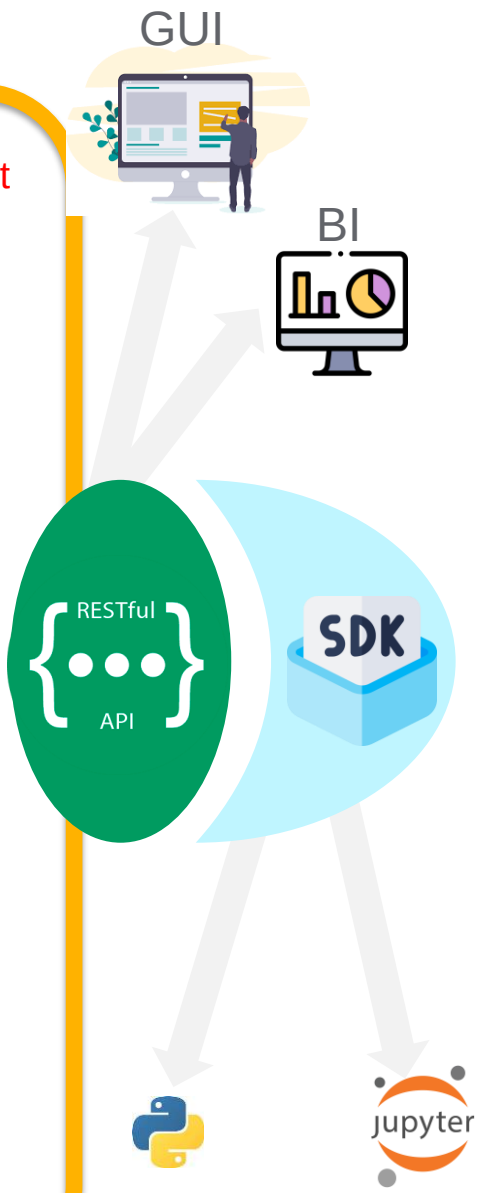
5) Save money with two cloud approaches:

Expensive and Fast	HOT space (lambda) for small-ish	(<20M) datasets and frequently used
Cheap and Slow(er)	COLD space (queue) for large	(>20M) datasets and infrequently used

FedChemS: Federated Chemical Search

Source	Size (M)
Cambridge Structural DB	0.3
ChemDocs	1
Chemspace Stock	7
CODA DB List 2	3
Dictionary of Natural Products	0.2
Sigma Aldrich	0.1
Synapse	~0.04
Chemspace Virtual	374
Pistachio Patents	13
Pistachio Patents Reactions	~68
Reaxys	~40
Enamine REAL DB	4500
ENAMINE REAL SPACE	-
Syngenta ELN Reactions	~1
PubChem	109
ChEMBL	2
Enamine BB	0.2
Enamine SCRN	3
MolPort	8
PDB Ligands	0.03
Reaxys Reactions	~25
WUXI SPACE	-
CHEMriya SPACE	-



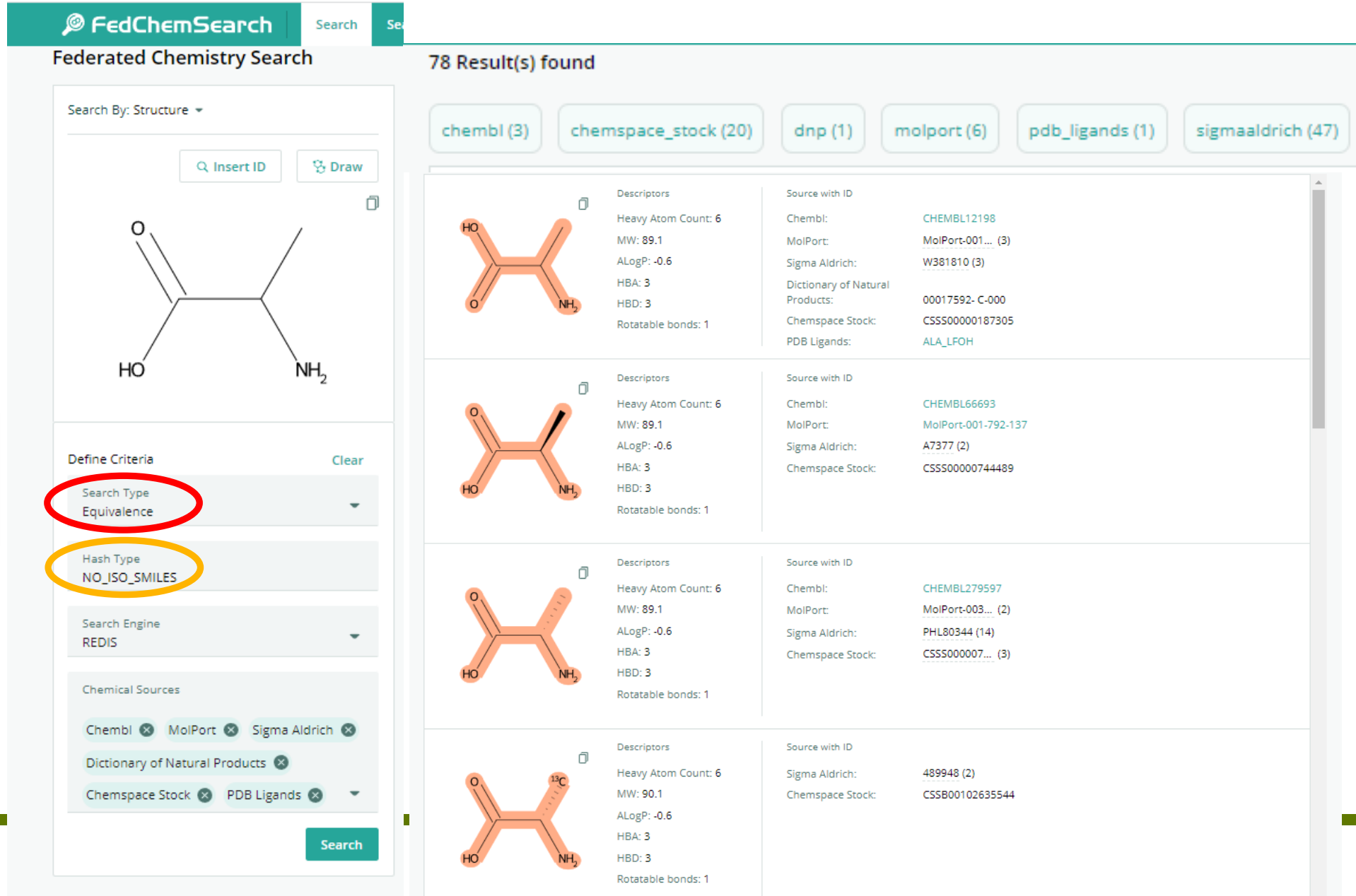


Bonus feature: from Identity to Equivalence through hashes

By using hash codes, (for molecules & reactions) we transform **strucutre searches** into **indexed string searches**

Extra-fast and useful searches (scaffold, isomers, etc.)

NextMove-RDKit hashes and custom extensions (ex pseudo-RInChI)



FedChemS Team: Chemistry x IT

Guillaume Berthon
Change Lead Digitization
of Chemical Synthesis
Syngenta

Marco Stenta
Computational Chemist
Syngenta

Vadim Kaplanov
Solution Architect
EPAM Systems

Serge P. Parel
R&D IT ELN Platform Manager
Technical Lead
Syngenta

Aleksei Samarin
Software Engineer
EPAM Systems

Thansk for your attention