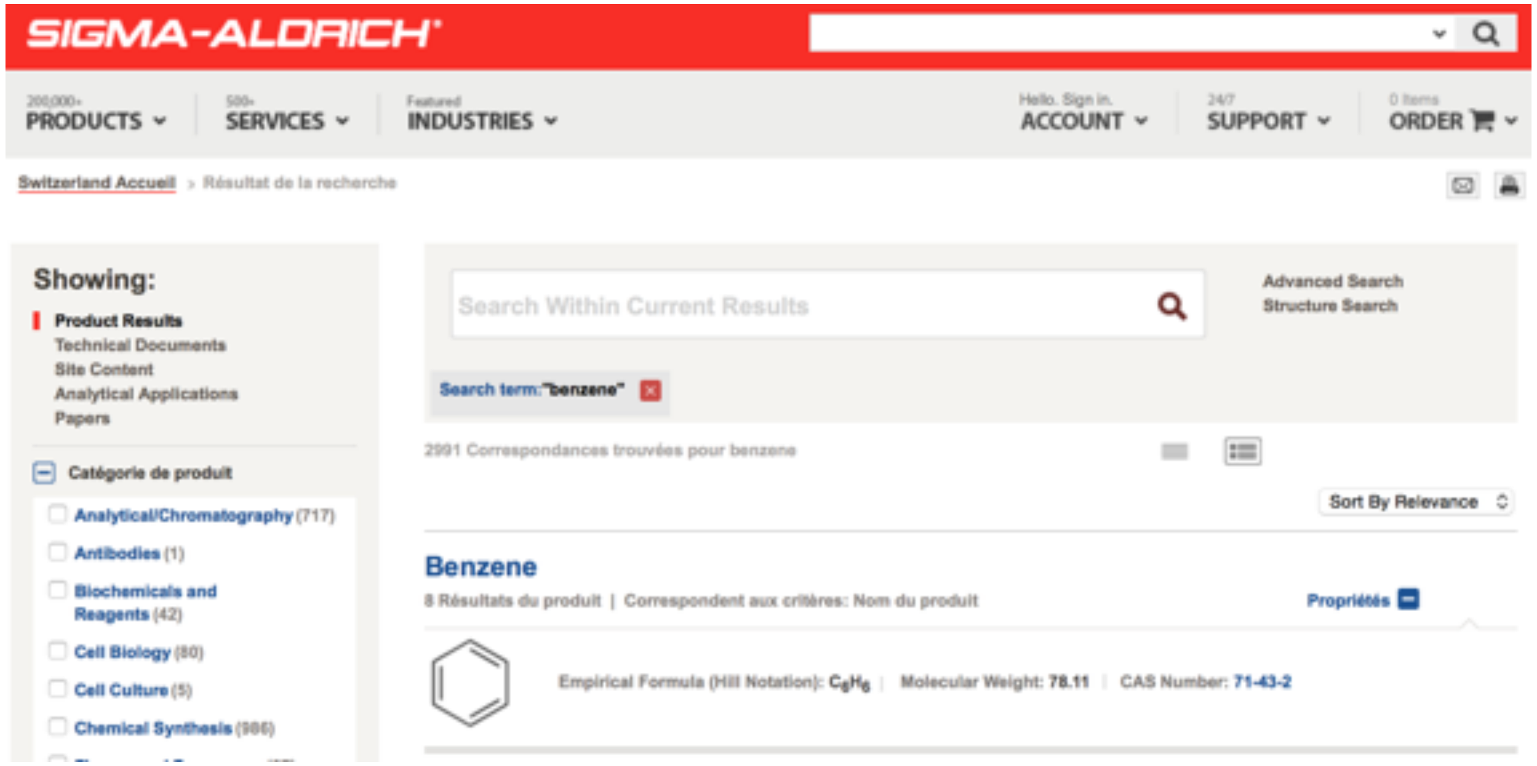
Swiss Similarity

Christophe Bovigny

Goal of the Project

- Develop a website allowing to perform ligand-based virtual screening of several libraries of small molecules
- Develop new databases based on commercial Databases
- Develop a code that allowing search and computation of 1 billion of molecule

New Databases



Showing:

- Product Results
- Technical Documents
- Site Content
- Analytical Applications
- Papers

Catégorie de produit

- □ Analytical/Chromatography (717)
- □ Antibodies (1)
- □ Biochemicals and Reagents (42)
- □ Cell Biology (80)
- □ Cell Culture (5)
- □ Chemical Synthesis (986)

Search term:"benzene" ☒

2991 Correspondances trouvées pour benzene

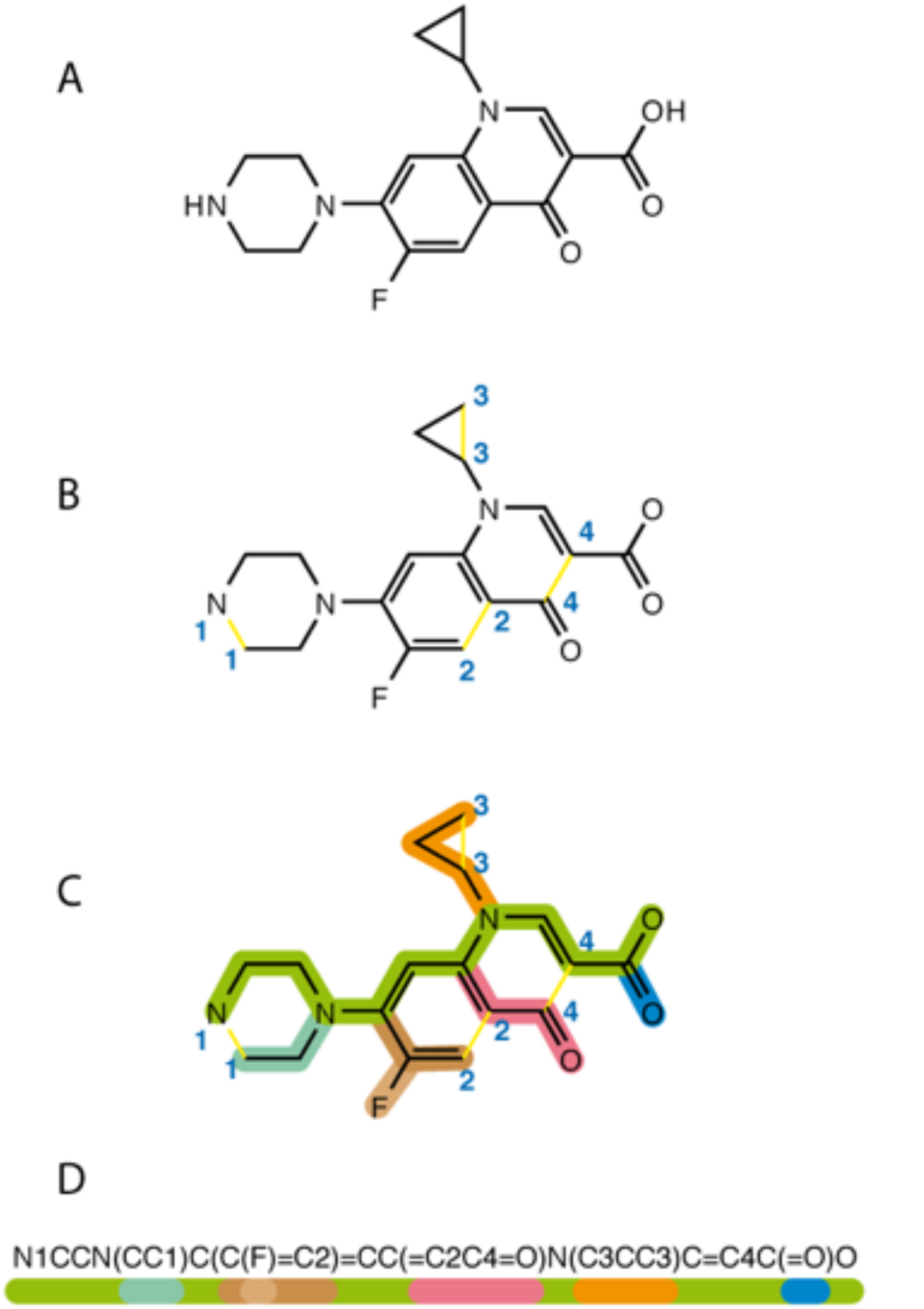
Benzene

8 Résultats du produit | Correspondent aux critères: Nom du produit

Empirical Formula (Hill Notation): C_6H_6 | Molecular Weight: 78.11 | CAS Number: 71-43-2

- Based on Sigma-Aldrich database, generate a new one by using click-chemistry reactions.
- In chemical synthesis, **click chemistry** is generating substances quickly and reliably by joining small units together.

Representation of molecules



https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

CDK API

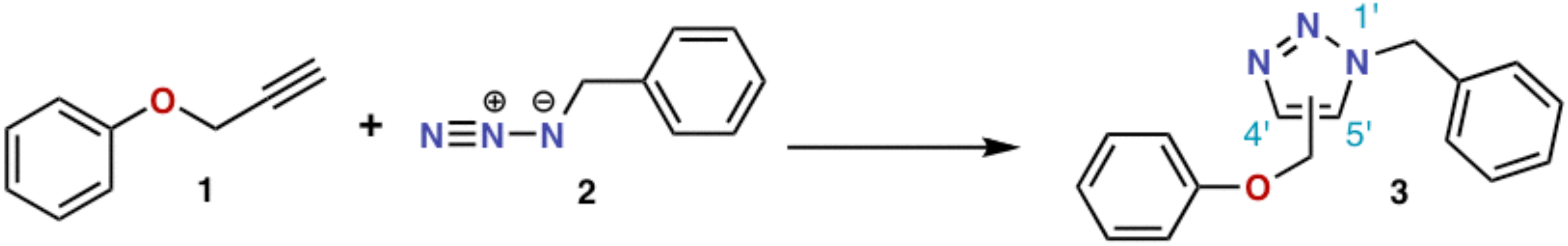
http://cdk.github.io/cdk/1.5/docs/api/

https://github.com/cdk

CDK Scala examples

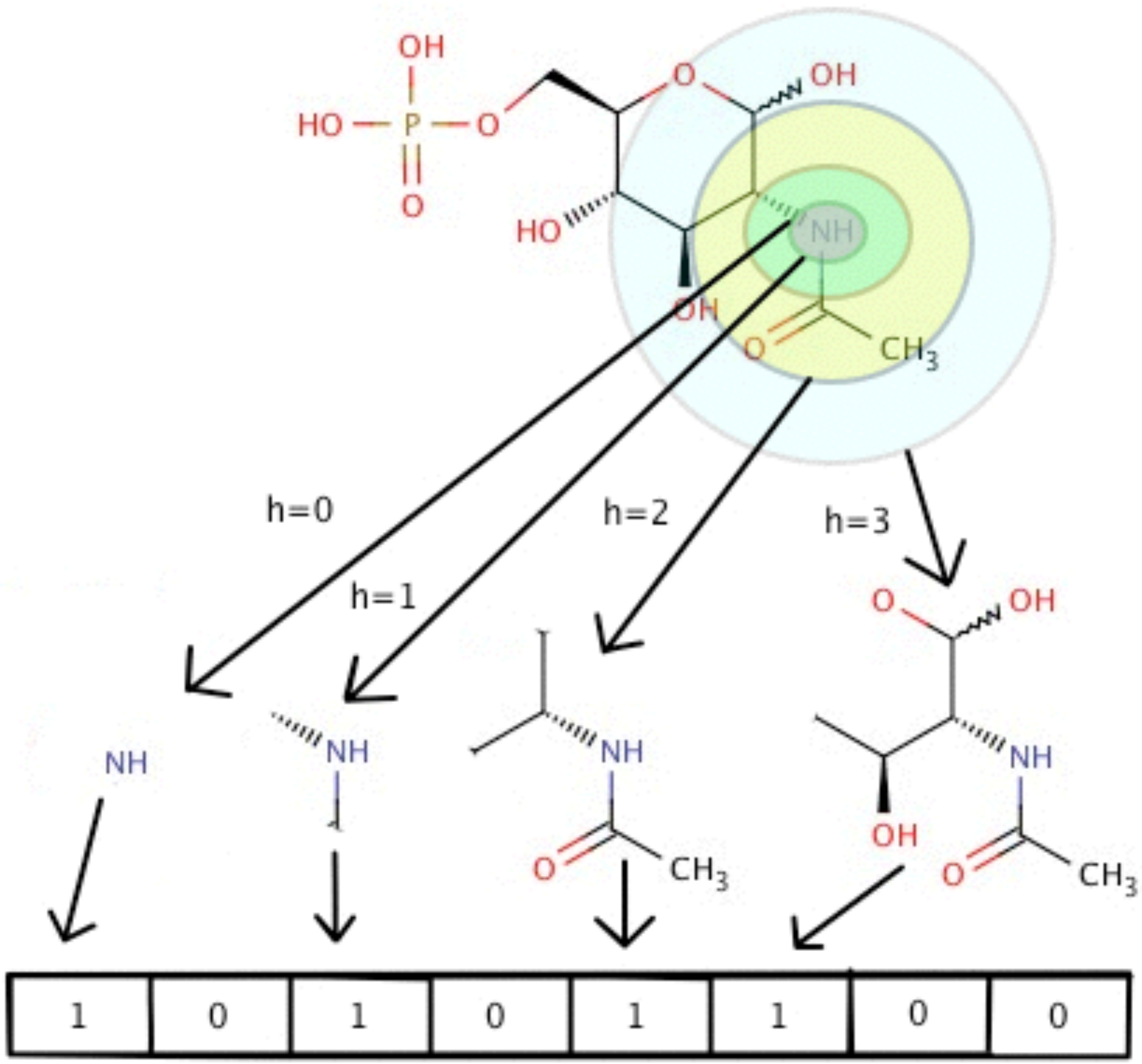
https://github.com/cdk/cdk-scala-examples

Click chemistry example



azide **2** reacts neatly with alkyne **1** to afford the triazole **3** as a mixture of 1,4-adduct and 1,5-adduct at 98 °C in 18 hours.

Fingerprint molecules



1	0	1	0	1	1	0	0

Fingerprint molecules

molecules1 (random query)	1	0	1	0	1	1	0	0
molecules2 (database)	1	0	1	0	1	1	0	0

add : 1 & 1, 0 & 0 ….

or : 1 | 1, 0 | 0 ….

z+= countBits(add)

z2 += countBits(or)

val tanimoto = (z :Float) / z2

```
def countBits(x : Int) = {
    var num = x
    var t = num & (-num)
    var bits = 0
    while ( t > 0 ) {
      bits+=1
      num = num ^ t
      t = num & (-num)
    }

    bits
  }
```

Computation Fingerprint

```
name := "CDKSpark"

version := "1.0"

scalaVersion := "2.11.7"
libraryDependencies ++= Seq(
  "org.apache.spark" % "spark-core_2.11" % "1.4.1",
  "org.apache.spark" % "spark-hive_2.11" % "1.4.1",
  "org.apache.spark" % "spark-sql_2.11" % "1.4.1",
 "org.openscience.cdk" % "cdk-core" % "1.5.11",
  "org.openscience.cdk" % "cdk-smiles" % "1.5.11",
  "org.openscience.cdk" % "cdk-silent" % "1.5.11",
  "org.openscience.cdk" % "cdk-standard" % "1.5.11",
   "org.openscience.cdk" % "cdk-fingerprint" % "1.5.11",
"com.datastax.spark" %% "spark-cassandra-connector" % "1.5.0-M1",
  "com.datastax.spark" %% "spark-cassandra-connector-java" % "1.5.0-M1"
)

resolvers += "Akka Repository" at "http://repo.akka.io/releases/"
```

build.sbt

```
// Parser molecules :
val smilesParser = new SmilesParser(
      SilentChemObjectBuilder.getInstance()
    )

// Molecules representation in SMILES

     val smiles = "N#CC(=[N]1CCC2C(C1)CCCC2)C(=O)Nc1ccccc1[N+](=O)[O-]"
   SmilesGenerator.generic().aromatic().create(smilesParser.parseSmiles(smiles))


// Fingerprint Computation for each molecules
val fpr = new CircularFingerprinter(CircularFingerprinter.CLASS_FCFP2)
val fpr1 =  fpr.getBitFingerprint(mol).asBitSet()
val fingerprinttoStore = fpr1.toLongArray.toIndexedSeq
```

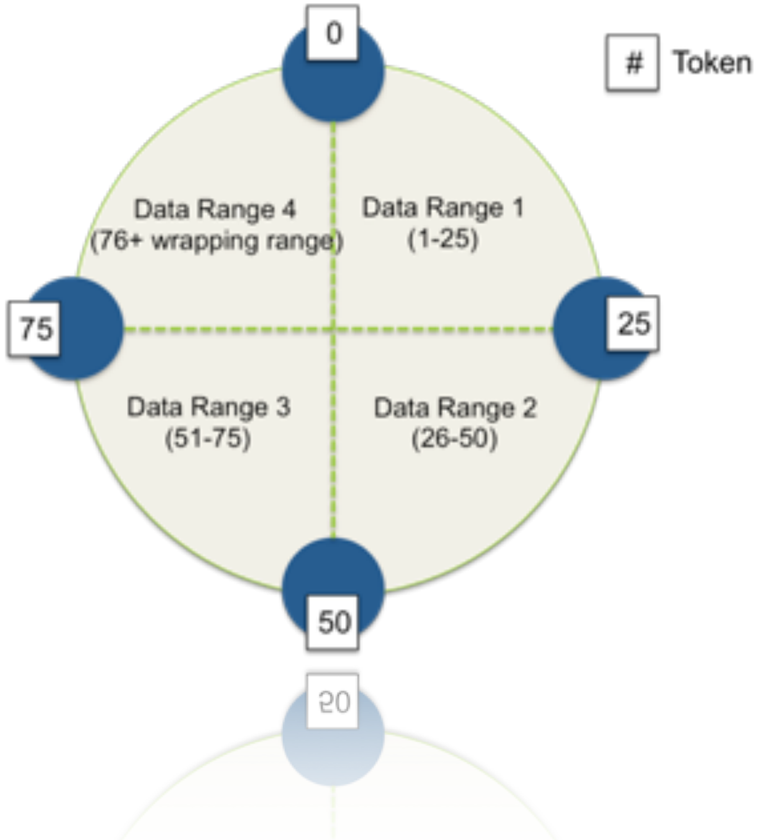
Example to compute fingerprint foreach molecule

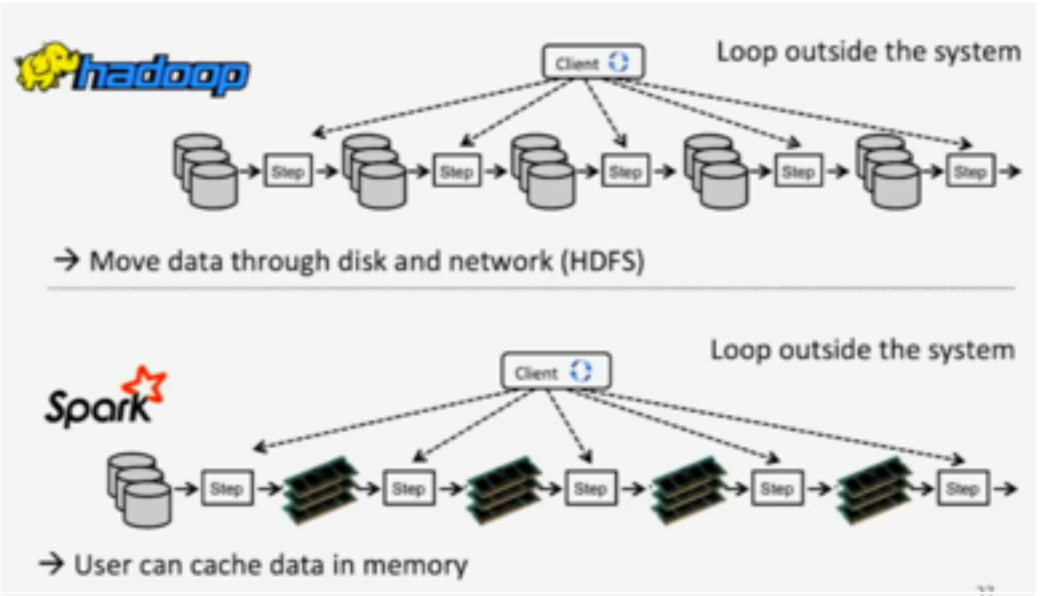
Technologies

DB(NoSQL) → Processing → Rest API

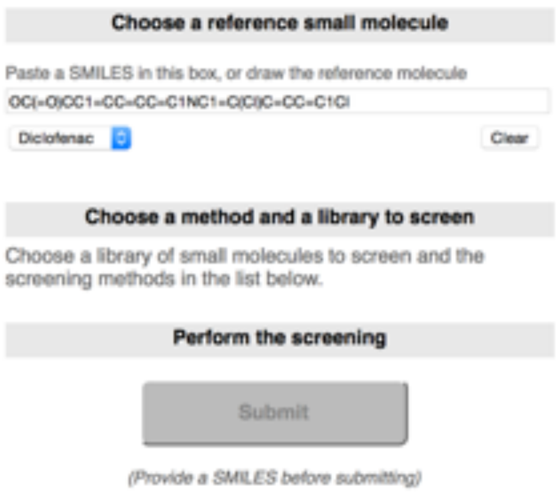
cassandra → Spark → Rest API

Scala





Swiss Similarity



Choose a reference small molecule

Paste a SMILES in this box, or draw the reference molecule

OC(=O)CC1=CC=CC=C1NC1=C(Cl)C=CC=C1Cl

Diclofenac | Clear

Choose a method and a library to screen

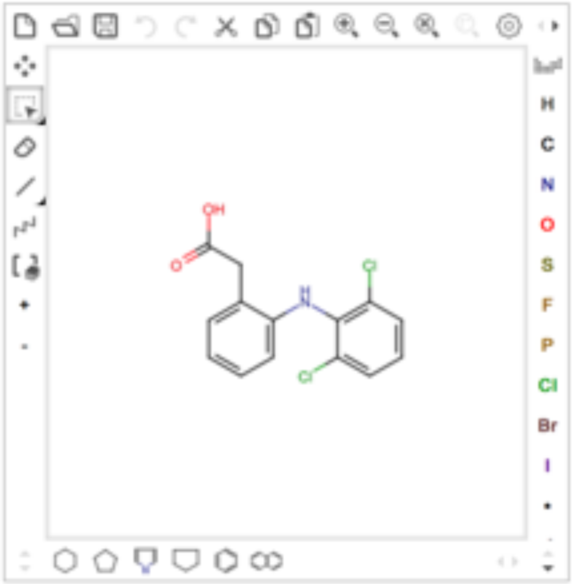
Choose a library of small molecules to screen and the screening methods in the list below.

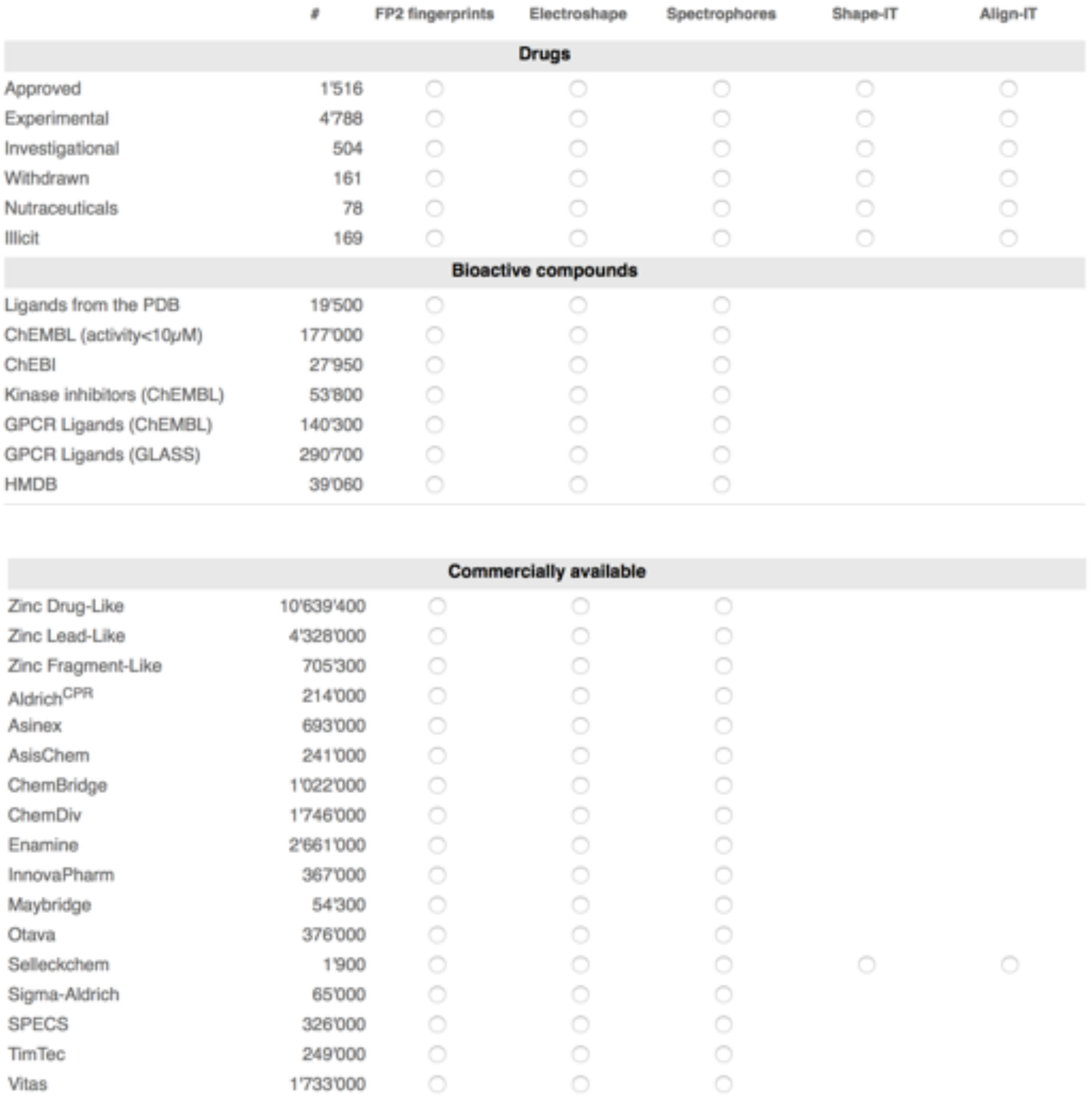
Perform the screening

Submit

(Provide a SMILES before submitting)

=





	#	FP2 fingerprints	Electroshape	Spectrophores	Shape-IT	Align-IT
Drugs						
Approved	1'516	○	○	○	○	○
Experimental	4'788	○	○	○	○	○
Investigational	504	○	○	○	○	○
Withdrawn	161	○	○	○	○	○
Nutraceuticals	78	○	○	○	○	○
Illicit	169	○	○	○	○	○
Bioactive compounds						
Ligands from the PDB	19'500	○	○	○		
ChEMBL (activity<10μM)	177'000	○	○	○		
ChEBI	27'950	○	○	○		
Kinase inhibitors (ChEMBL)	53'800	○	○	○		
GPCR Ligands (ChEMBL)	140'300	○	○	○		
GPCR Ligands (GLASS)	290'700	○	○	○		
HMDB	39'060	○	○	○		
Commercially available						
Zinc Drug-Like	10'639'400	○	○	○		
Zinc Lead-Like	4'328'000	○	○	○		
Zinc Fragment-Like	705'300	○	○	○		
Aldrich^{CPR}	214'000	○	○	○		
Asinex	693'000	○	○	○		
AsisChem	241'000	○	○	○		
ChemBridge	1'022'000	○	○	○		
ChemDiv	1'746'000	○	○	○		
Enamine	2'661'000	○	○	○		
InnovaPharm	367'000	○	○	○		
Maybridge	54'300	○	○	○		
Otava	376'000	○	○	○		
Selleckchem	1'900	○	○	○	○	○
Sigma-Aldrich	65'000	○	○	○		
SPECS	326'000	○	○	○		
TimTec	249'000	○	○	○		
Vitas	1'733'000	○	○	○		

- Develop New Databases
- Click Chemistry (Sigma)
- Use several Big Data technologies

http://www.swisssimilarity.ch