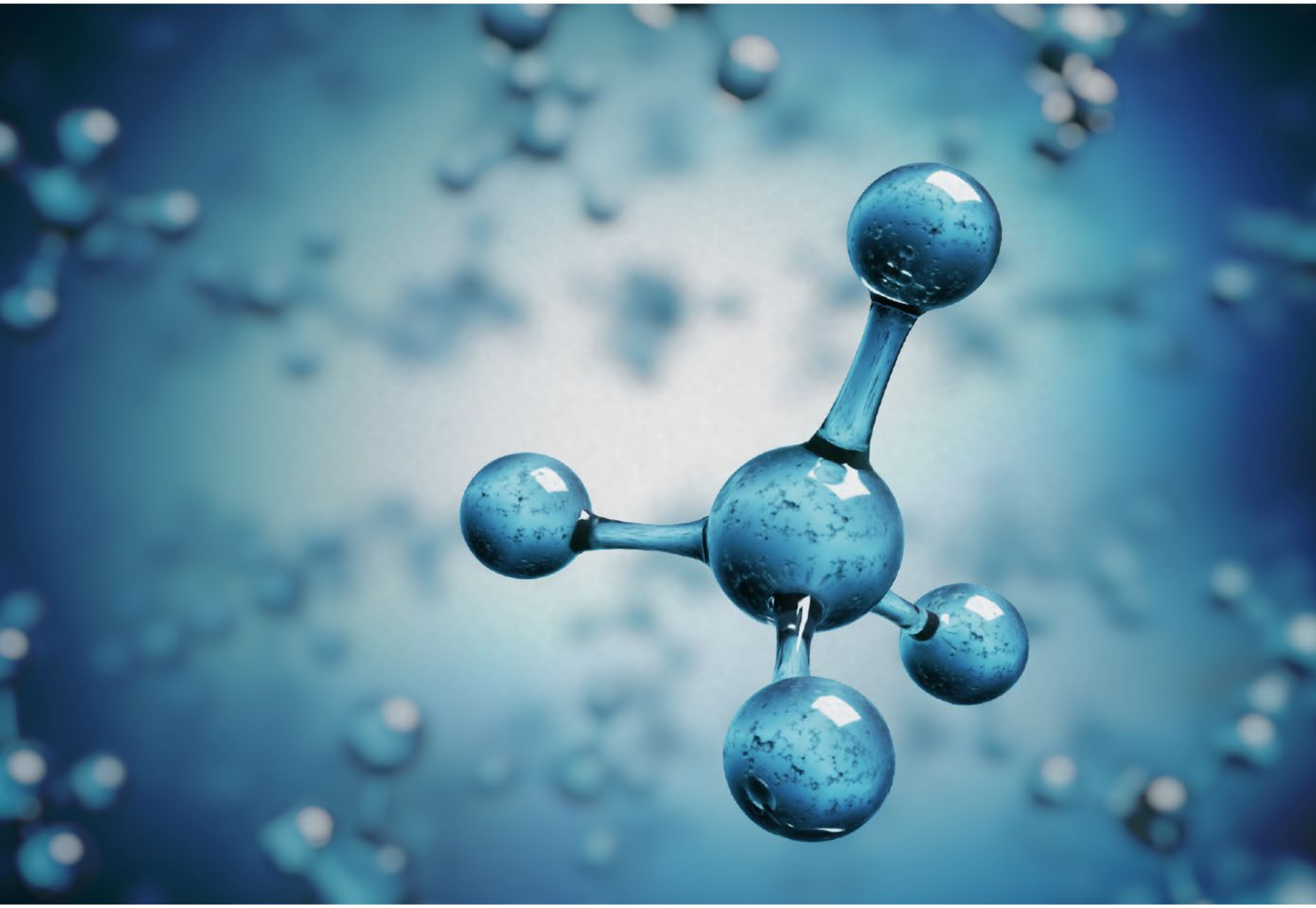
Introducing a Cheminformatics Similarity Search Solution



www.gsitechnology.com

About the APU

The Associative Processing Unit (APU) is GSI's patented processing technology and a new breed of processor. It features massive parallel data processing, compute and search, in-place, directly in the memory array. The APU's design eliminates bandwidth-costly data transfers between the memory and processor. This gives the APU a performance edge in the acceleration of similarity search applications.

The APU as a Driver of Similarity Search

Similarity search is the most general term used for a range of mechanisms which share the principle of searching (typically, very large) spaces of objects where the only available comparator is the similarity between any pair of objects. This is becoming increasingly important in an age of large information repositories where the objects contained do not possess any natural order.

Similarity search involves searching a database of vectors for similarity to a given query. The query itself must be of identical size to the database records (e.g., 512 or 1024 bits).

At GSI, we implement similarity search and top-k using parallel processing. Current processor technologies are not well suited to similarity search and top-k problems due to the high memory bandwidth demands required. For similarity search. the APU's advantage over the CPU is processing time—the APU can process a similarity search in less than 15 microseconds, with lower power consumption and latency.

Similarity Search in Cheminformatics

Cheminformatics is concerned with the development of data formats and databases containing information about different aspects of various chemical systems, and the design of tools to query those databases. Cheminformatics aims to create a user interface that assembles the different tools needed to query data as a whole. Among the main objectives of cheminformatics, are:

- Reduction of the cost and time needed to develop new drugs
- Increasing the efficacy and properties of natural substances
- Production of new protein molecules and chemical materials
- Genome sequencing

Cheminformatics uses very large databases. To date, several databases have been developed for cataloging molecule data for drug development.

These databases consist of chemical fingerprints, used to identify the features of chemical elements within a matrix, and therefore define its unique portrait in comparison to similar matrices. Molecular fingerprints have been used for a long time now in drug discovery and virtual screening. Their ease of use (requiring little to no configuration) and the speed at which substructure and similarity searches can be performed with them—paired with a virtual screening performance like other more complex methods—is the reason for their popularity.

POC: GSI Joins Forces with the Weizmann Institute and G-INCPM

Project Overview

As a proof of concept for the APU technology with similarity search applications, GSI Technology has teamed up with researchers at the *Weizmann Institute of Science*, in Rehovot, Israel, and *The Nancy & Stephen Grand Israel National Center for Personalized Medicine* (G-INCPM). The Weizmann Institute has a world class reputation in multidisciplinary scientific research, while G-INCPM is an advanced research facility that strives to promote the field of personalized medicine.

As part of ongoing research on new molecules with pharmacological properties, and with the goal of obtaining the approval of the FDA and other agencies, the team at the Weizmann Institute and G-INCPM is searching an established database of 38 million molecular fingerprints with the goal of identifying molecules that are structurally similar—or have similar properties—to their query molecule.

The Weizmann Institute & G-INCPM's Approach to Search

Researchers at the *Weizmann Institute* and *G-INCPM* have been using BIOVIA Pipeline Pilot[1] data analysis software, Oracle Database, and a CPU to fetch, process, and analyze molecular

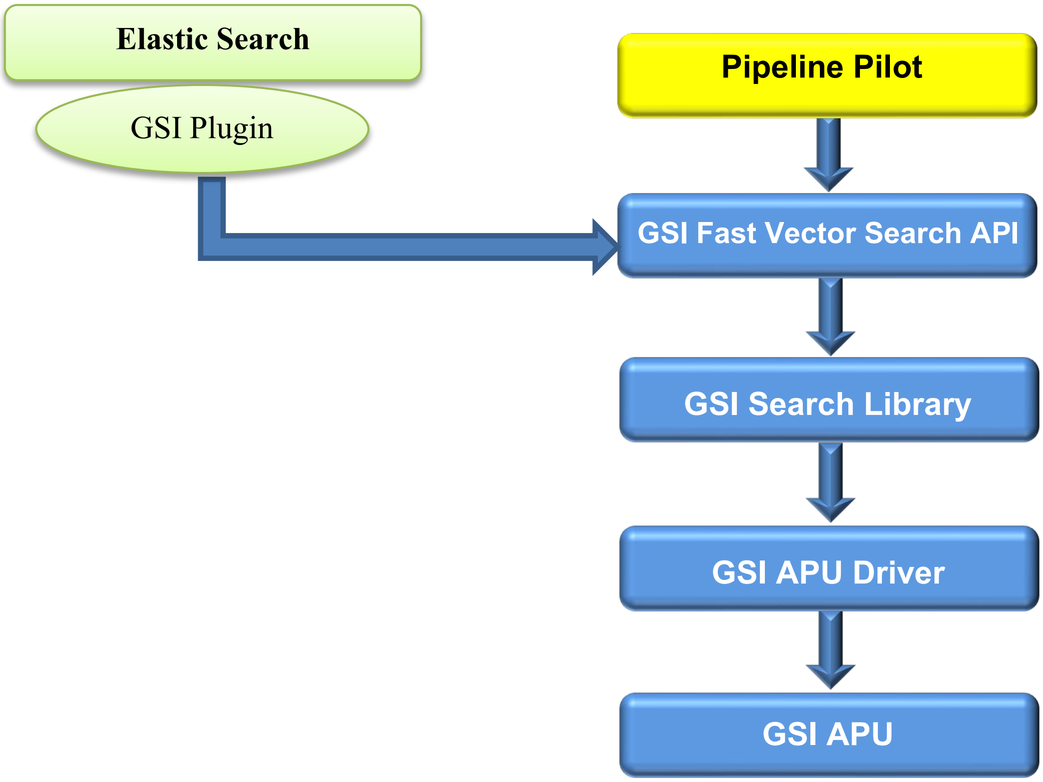
[1] BIOVIA Pipeline Pilot is a software program developed by the Accelrys® group.

structures data. However, they are finding that fetching the data with Oracle Database and performing similarity structure search with a CPU is a relatively slow and inefficient process.

A New Approach with the APU

To speed up the similarity structure search process, GSI is replacing the Oracle-based BIOVIA Tanimoto search engine with an APU-based Tanimoto search engine. GSI has replaced the BIOVIA DIRECT search component in Pipeline Pilot that calls Oracle Database and supports only MDL Public keys fingerprint type, with a new Python API that calls the APU processor and the GSI Search Library (GSL). In the GSL, GSI has implemented similarity structure search based on the Tanimoto distance measures.

The architecture diagram provides an overview of GSI’s APU similarity structure search solution. The GSI Fast Vectors search API connects Pipeline Pilot to GSL. It can equally be used by other third-party applications, such as KNIME, and Search by Elastic Search Plugin that connect to the GSI Fast Vector Search API



©2021 GSI Technology, Inc.
Rev. 1.02. 03/2021

The GSI Similarity Structure Search Process

1. The binary database (Covert to Binary by GSI Tool) is loaded ONCE into the L4[2] buffer. It will only need to be loaded again if changes are made to it. This step represents a new capability in Pipeline Pilot, developed by GSI.

2. Pipeline Pilot converts the researcher's query molecule into a folded fingerprint string type ECFP/FCFP. (Existing functionally.)

3. GSI replaces the existing Pipeline Pilot search block with a GSI Search block. This block calls the GSI Search Library (GSL) function that performs the following actions:
   a. The query is loaded into the L1 memory block.
   b. The molecule database is divided into manageable chunks of data that can fit into L1[3] memory.
   c. We use the Tanimoto algorithm to perform similarity search. It is considered an ideal method for fingerprint-based calculations of the similarity of molecular representations.

4. GSI's Python code calls the GSI Search Library's (GSL) Tanimoto distance measures, which find values and indices of k nearest entries.

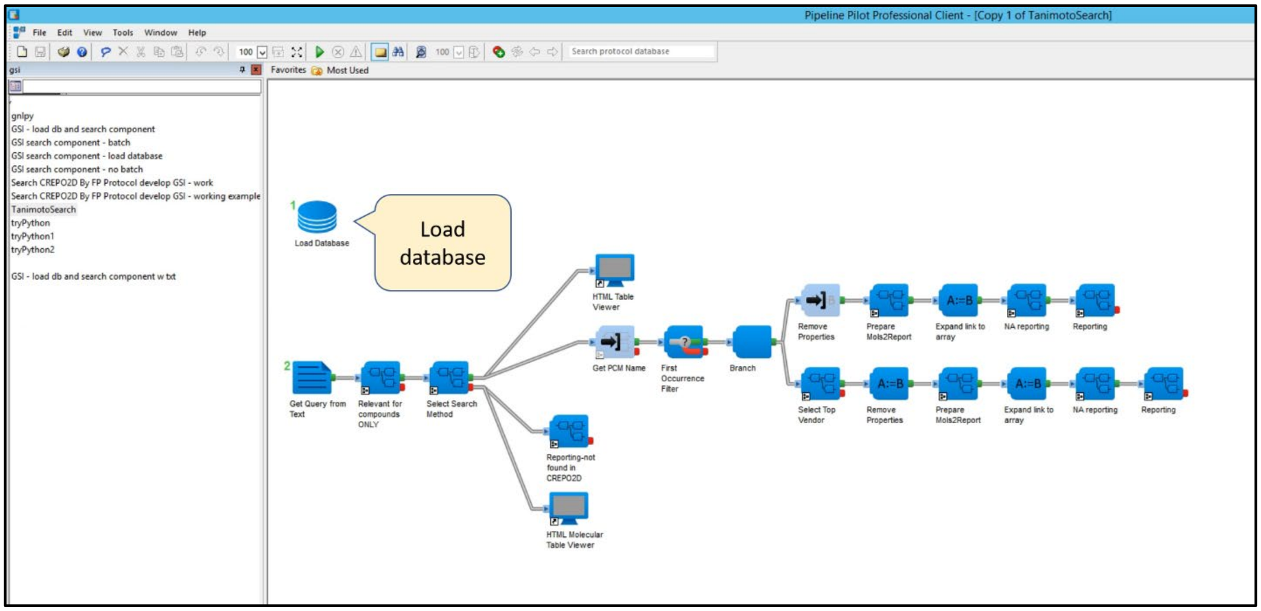
Some Screenshots from Pipeline Pilot

The following screenshots from BIOVIA's Pipeline Pilot user interface show the similarity search pipeline, including the steps where GSI's Python API code has been embedded in the pipeline, connecting GSI's APU and GSL resources.

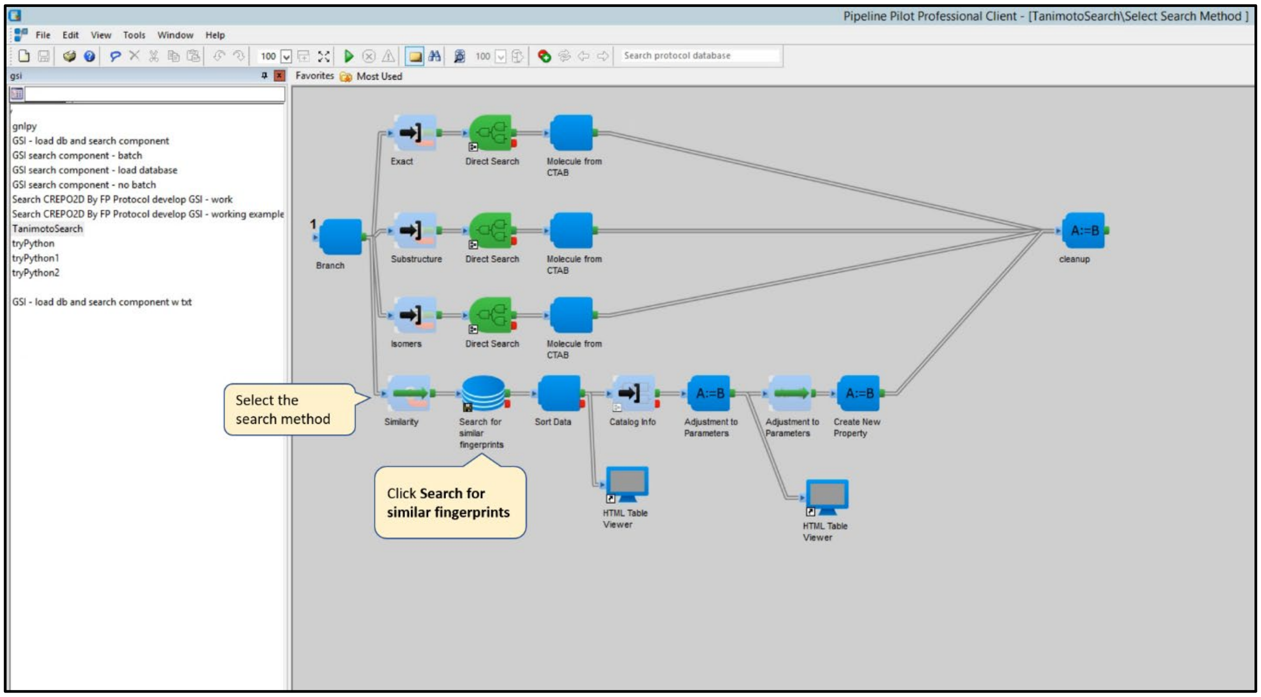
[2] L4 is the APU's DDR4 DRAM system memory, used to store the vector database of 38 million compounds.

[3] L1 is the vector memory of the APU. It serves as a closely-coupled data storage/cache to the Main Memory Block (MMB)—where data processing takes place.

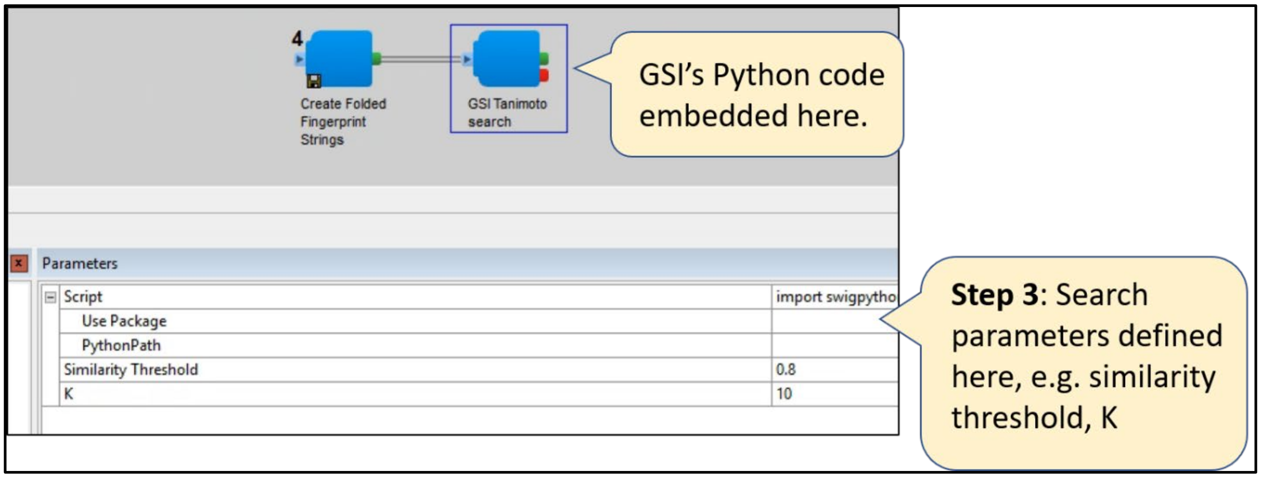
Step 1: Load database.


Pipeline Pilot Professional Client - [Copy 1 of TanimotoSearch]
File Edit View Tools Window Help
Search protocol database
gsi
Favorites Most Used
gnlpy
GSI - load db and search component
GSI search component - batch
GSI search component - load database
GSI search component - no batch
Search CREPO2D By FP Protocol develop GSI - work
Search CREPO2D By FP Protocol develop GSI - working example
TanimotoSearch
tryPython
tryPython1
tryPython2
GSI - load db and search component w txt
1
Load Database
Load database
2
Get Query from Text
Relevant for compounds ONLY
Select Search Method
HTML Table Viewer
Get PCM Name
First Occurrence Filter
Branch
Reporting-not found in CREPO2D
HTML Molecular Table Viewer
Remove Properties
Prepare Mols2Report
Expand link to array
NA reporting
Reporting
Select Top Vendor
Remove Properties
Prepare Mols2Report
Expand link to array
NA reporting
Reporting

Step 2: Select the search method (e.g., Tanimoto distance) and click Search for similar fingerprints.


Pipeline Pilot Professional Client - [TanimotoSearch\Select Search Method]
File Edit View Tools Window Help
Search protocol database
gsi
Favorites Most Used
gnlpy
GSI - load db and search component
GSI search component - batch
GSI search component - load database
GSI search component - no batch
Search CREPO2D By FP Protocol develop GSI - work
Search CREPO2D By FP Protocol develop GSI - working example
TanimotoSearch
tryPython
tryPython1
tryPython2
GSI - load db and search component w txt
1
Branch
Exact
Direct Search
Molecule from CTAB
Substructure
Direct Search
Molecule from CTAB
Isomers
Direct Search
Molecule from CTAB
cleanup
Select the search method
Similarity
Search for similar fingerprints
Sort Data
Catalog Info
Adjustment to Parameters
Adjustment to Parameters
Create New Property
HTML Table Viewer
HTML Table Viewer
Click Search for similar fingerprints

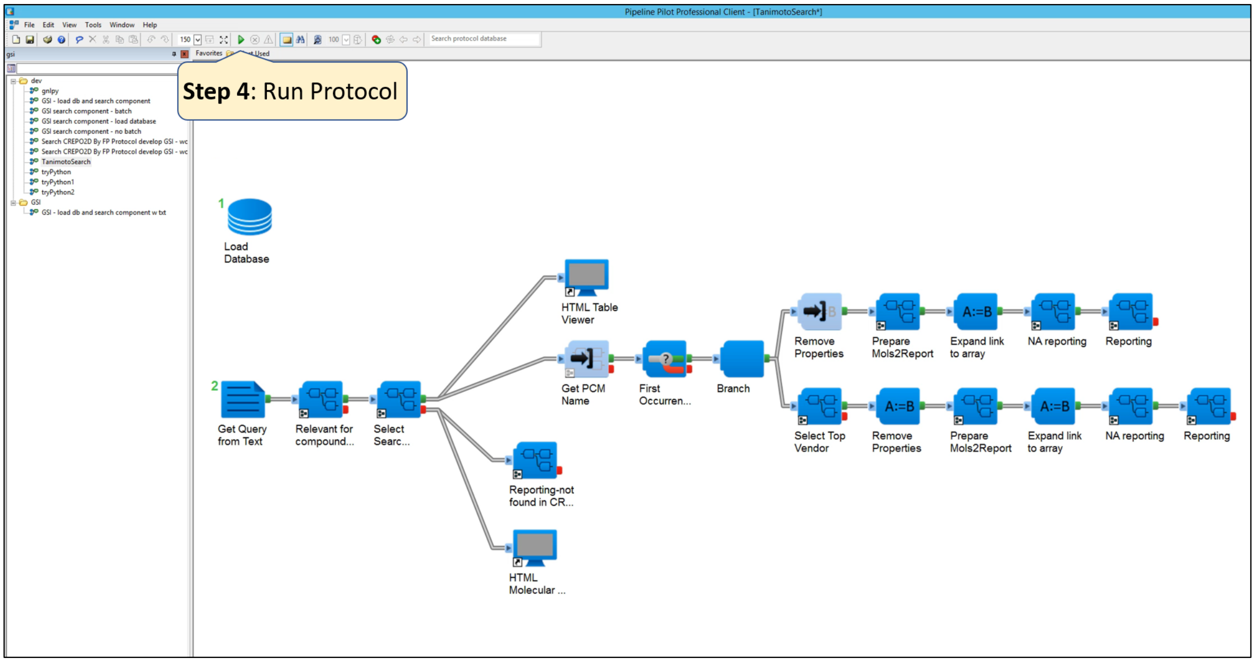
Step 3: Define the search parameters such as K, or similarity Tanimoto threshold.


4
Create Folded Fingerprint Strings
GSI Tanimoto search
GSI's Python code embedded here.
Parameters

Script	import swigpytho
Use Package	
PythonPath	
Similarity Threshold	0.8
K	10

Step 3: Search parameters defined here, e.g. similarity threshold, K

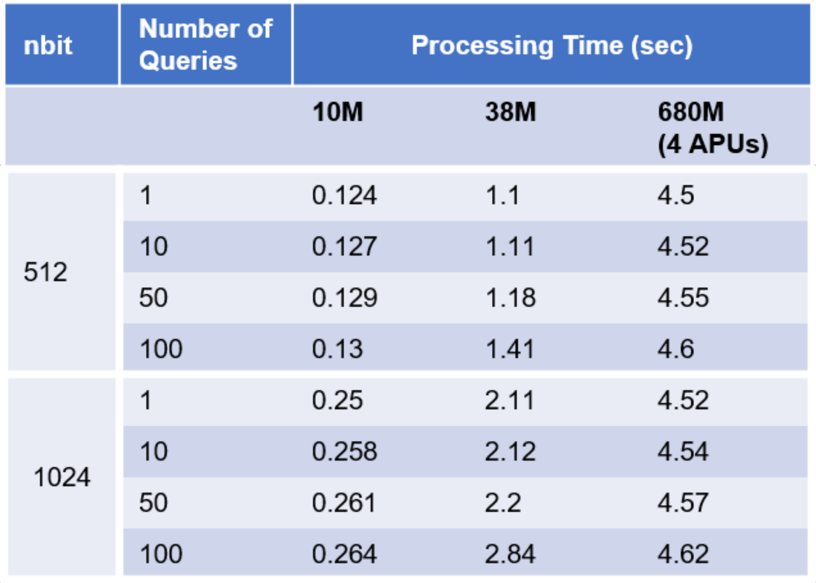
Step 4: Run the protocol.



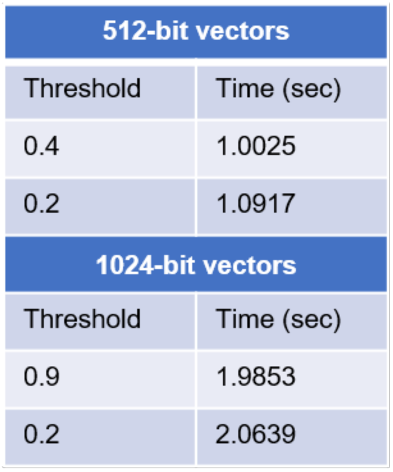
Performance and Power Consumption

The table to the left shows Gemini® APU query processing time for KNN similarity search, with 512-bit and 1024-bit vectors, 3 database sizes, and number of queries ranging from 1 to 100.

The table to the right shows selected results for query processing on a 38M database, using distance threshold similarity search.



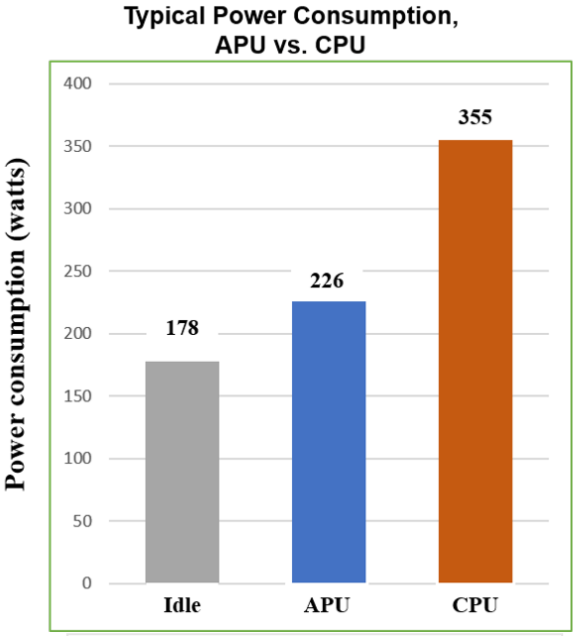
nbit	Number of Queries	Processing Time (sec)		
		10M	38M	680M (4 APUs)
512	1	0.124	1.1	4.5
	10	0.127	1.11	4.52
	50	0.129	1.18	4.55
	100	0.13	1.41	4.6
1024	1	0.25	2.11	4.52
	10	0.258	2.12	4.54
	50	0.261	2.2	4.57
	100	0.264	2.84	4.62



512-bit vectors	
Threshold	Time (sec)
0.4	1.0025
0.2	1.0917
1024-bit vectors	
Threshold	Time (sec)
0.9	1.9853
0.2	2.0639

Gemini® APU Benefits

- Support for similarity structure search on circular fingerprints at thresholds of 0.8 and under.
- Easy integration with BIOVIA – proven and fully tested.
- Gemini® can be integrated into other molecule search solutions that use the Tanimoto coefficient.
- Support for single or multiple queries on Enamine REAL database (680M compounds), all on a single APU.
- Calculation of more descriptive, larger length, folded fingerprints (bit sizes of 256, 512, 1024, and up to 8192).
- Multiple databases can be loaded into the APU's system memory. Single dataset can be chosen for a specific query.
- Batches of compounds (100s or 1000s) can be submitted simultaneously.



Gemini® reduces search power consumption by a factor or 3.5.

Future Implementations

The APU can be easily integrated with other similarity search applications, using a GSI Fast Vector search (Python or RestAPI) API.