Weizmann Institute of Science Drives Drug Discovery with Ultra-Fast Molecular Similarity Search

Drug discovery is difficult. The average cost to develop a new drug is roughly $2.6 billion, and 90% of new drugs fail to win approval. For those few that do manage to win approval, it still takes at least 10 years to get them to market. With those sobering numbers, it is no surprise that top pharmaceutical companies are working with research centers such as The Nancy and Stephen Grand Israel National Center for Personalized Medicine (G-INCPM) at the Weizmann Institute of Science to find new ways to lower the cost and time needed to discover new drugs.

G-INCPM is an advanced research facility that provides state-of-the-art genomics, protein profiling, drug discovery, and bioinformatics research platforms and know-how. One way they are lowering drug development time and cost is through virtual screening of drug candidates, where libraries of small molecules are searched to find molecules that are most likely to be biologically active and worthy of further evaluation. This reduces the number of experiments that need to be done in a lab—significantly cutting the time and costs of drug discovery.

The Weizmann Institute, however, was struggling with the time it took them to search their small-molecule libraries. Their virtual screening generally returns many hits (molecules of interest for further study), and "if for every single hit, a similarity search takes several minutes, the task becomes exhausting," said Dr. Efrat Ben-Zeev, Computational Chemist and Cheminformatics Project Leader, Weizmann Institute of Science.

Restrictive Similarity Search Options

In addition to slow molecule similarity search, scalability was an issue for Ben-Zeev. "Previously, scaling to larger databases of molecules was difficult because of all the indexing required to build a large database," said Ben-Zeev. "It would require a database expert to index the database, and adding compounds to it became a difficult maintenance issue."

Lack of flexibility in her similarity search solution was another challenge for Ben-Zeev. As part of her virtual screening process, Ben-Zeev prefers to set the molecule similarity threshold to 0.4 or below. This allows her to build a diverse small-molecule library to serve as the foundation for her virtual screening. Unfortunately, Ben-Zeev's previous options either limited the threshold to 0.7 and above or were too slow to be of practical use when the threshold was below 0.7. This restricted her ability to work with a diverse set of molecules, thus greatly impacting her ability to discover novel compounds. Ben-Zeev realized that in order to take her drug discovery efforts to the next level, she needed to find a fast, scalable, and flexible similarity search solution where she could control the type of fingerprint, fingerprint bit size, and the threshold.

The Path to Scalability and Flexibility

GSI Technology's Associative Processing Unit (APU) was the solution.

The APU is a custom, compute-in-memory chip that combines high speed SRAM and programmable bit-logic interleaved throughout the memory. The APU computes functions directly on the data using parallel processing. The APU can be used for ultra-fast molecular structural similarity/substructure search to advance powerful existing structural similarity search algorithms in drug discovery.

GSI integrated their Python API into the Weizmann Institute's existing cheminformatics software platform (BIOVIA Pipeline Pilot) and replaced the search component in it with a GSI search component that runs on GSI's first APU chip—Gemini. "It integrated well with our software and is very easy to use," said Ben-Zeev. "We were up and running with it the first day."

"[The APU] integrated well with our software and is very easy to use. We were up and running with it the first day."

- Dr. Efrat Ben-Zeev

Gemini provides the flexibility to work with many different types of fingerprints (molecule representations used in search), such as MDL, ECFP, and FCFP. Additionally, it can work with longer fingerprints (e.g., 8192-bit fingerprints, which are better able to discriminate because they are more descriptive). Most other solutions are limited to fingerprints of 512 bits or fewer, and in many cases only offer one type of fingerprint.

Unlike Ben-Zeev's previous options, Gemini does not place restrictions on the similarity threshold, and it allows for a flexible search threshold to be set. With the APU, Ben-Zeev is able to set thresholds below 0.4, with no impact to performance. This allows for a diverse set of molecules to be returned from the initial similarity search, which greatly improves her chances of discovering novel compounds.

“By dramatically reducing the time required to search our small-molecule database, GSI’s Gemini empowers us to advance our research processes and ultimately improve human health.”

- Dr. Efrat Ben-Zeev

Reducing Molecular Search Time from Minutes to Milliseconds

With a diverse set of molecules now in hand, Ben-Zeev then performs a similarity search on a subset of those molecules (the ones that are determined to be hits through other biological assays). This similarity structure search uses a threshold of around 0.8 because, at this point, she wants to find very similar molecules to the hits and to expand the hit space. The assumption here is that structurally similar molecules exhibit similar biological activities. This is one of the steps in building a SAR (Structure-Activity-Relationship) table.

The team then explores the chemical space around this group of expanded hits in order to optimize activity and make the molecules more drug-like. Because there are generally many compounds in this expanded group, having a fast similarity search solution is critical. Previously, Ben-Zeev struggled with a similarity search solution that took several minutes for one similarity search. With Gemini, Ben-Zeev receives search results in a few hundred milliseconds, which significantly improves her ability to explore the chemical space in depth.

Fast search also allows the Weizmann Institute to efficiently scale to and explore larger databases. “Because the APU is so fast, it eliminates the need to index the database. This simplifies database management, and it makes adding compounds to the database easy,” explained Ben-Zeev. “Now we can explore very large virtual libraries, such as Enamine REAL without having to hire a database expert.” Working with large databases, such as Enamine REAL, which currently comprises over 700 million synthetically feasible molecules, allows Ben-Zeev to greatly improve her chances of finding novel compounds and achieve her ultimate goal of discovering the next important drug.

“By dramatically reducing the time required to search our small-molecule database, GSI’s Gemini empowers us to advance our research processes and ultimately improve human health,” said Ben-Zeev.

Looking ahead, Ben-Zeev is excited about the prospects of using Gemini to do batch searches, interactive searches, and for 3D similarity structure search.