New machine learning method speeds up drug screening hundred-fold

The University of Eastern Finland performed a virtual search of 1.56 billion molecules to test two drug candidates. This was the world's most extensive screening of its kind.

Most drugs available today have been designed so that the target molecules are the body's own proteins. Once the structure of one member of a protein family has been determined, the structure of other proteins in the same family can be predicted through modelling. A successful drug can be developed, for example, by screening a large library to find a molecule with a three-dimensional structure enabling interaction with the target protein.

Professor **Antti Poso's** research team were looking for molecules that would react with SurA chaperone and cyclin-G-associated kinase (GAK), two candidates with medicinal effect. The project tested the HASTEN algorithm developed for the screening, and created a new machine learning model.

"These target proteins, SurA and GAK, were already known to us from existing academic research projects. The results of the massive screenings can be used in other research. We not only just validated a method but are also able to help various academic research projects," says Poso.

Chaperones contribute to protein folding and regulate protein interaction. Kinases have a role in cellular signalling, among other things.

"The SurA chaperone is related to a collaborative project with the University of Tübingen, with the aim of developing new antibiotics. Kinases, on the other hand, are a large family of proteins. Most cancer drugs are kinase inhibitors. There are some 500 types of kinase, with cyclin-G-associated kinase, or GAK, being one of them. GAK's potential lies in cancer drugs and the treatment of viral infections."

Poso's team is studying the interaction of drugs and proteins, and creating target protein models. The point at which a drug binds to a protein can usually be identified in the target protein structure, thereby making the drug work. The model can be used specifically in virtual screening. This involves searching large molecular databases for new ideas for drug development.

"Chaperone's protein structure is very different from that of kinase. So we are talking about two very different target proteins that were worth testing together."

AI predicted binding of molecules to proteins

The structural difference of two drug candidates was a key factor, because the algorithm must work in all protein families.

"Two drug candidates were used to test how the HASTEN algorithm developed by **Tuomo Kalliokoski** at Orion works in the CSC supercomputing environment. The scalability was successful."

The target protein screening was performed, for purposes of comparison, with the HASTEN algorithm and the traditional docking method. In docking, the search algorithm calculates the interactions between the protein and the drug candidate in the database. The value given by the algorithm shows how well the drug binds to the protein.

Poso's team screened 1.56 billion molecules containing the drug candidate. The molecules were screened from the REAL database of Enamine, a large Ukrainian chemical company.

"First we calculated every two-dimensional molecule drawn in the database and converted them into three-dimensional format. After that the software tried to fit each molecule inside GAK or SurA. An individual fitting can have hundreds of thousands of alternatives."

Then the researchers tested how machine learning fared compared to docking. The HASTEN algorithm was used for machine learning.

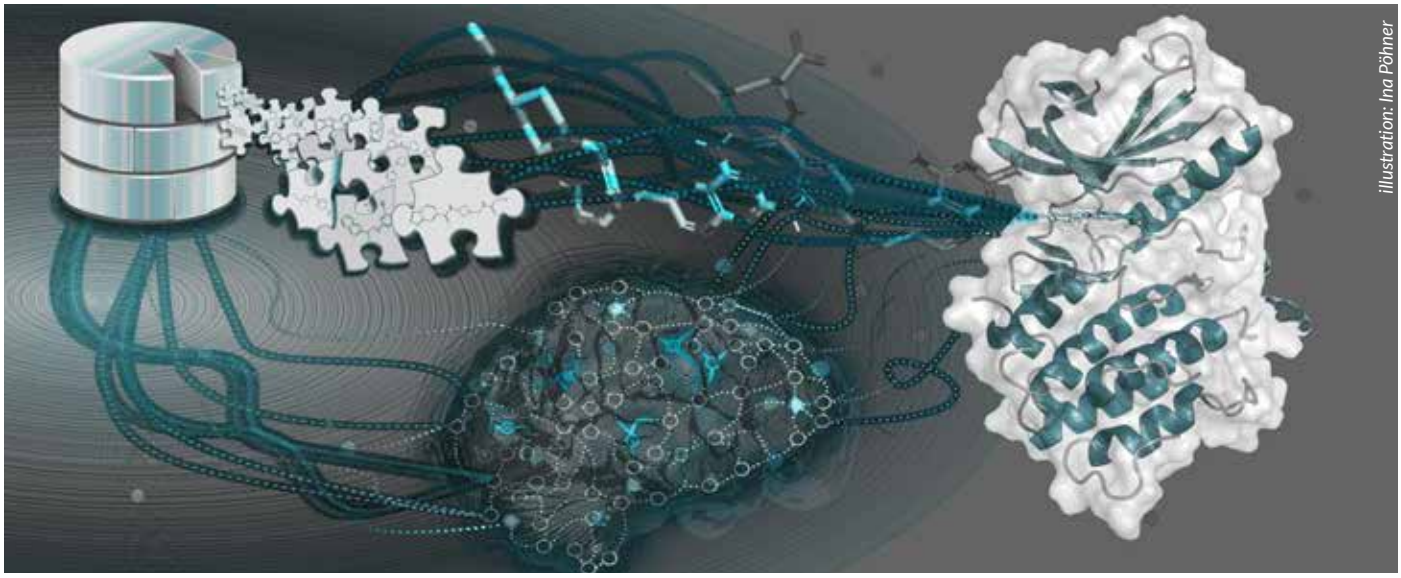

illustration: Ina Pöhner

The full version of the Enamine Real database already contains 48 billion structures. Just to convert molecular structures into three-dimensional format is time-consuming and laborious with such a huge amount of data, but with the HASTEN algorithm this stage can be "bypassed", with the material processed in a matter of days.

"We first chose a million molecules at random to see how the docking worked. We then fed the results to AI. So what the machine did was learn to predict the result on the basis of a million molecules, meaning that when a molecule has a specific shape, it docks into a specific location."

After this, all 1.56 billion molecules were fed in to the AI to predict results using the results of the initial million molecules. The ones that had the highest prediction were docked again, followed by another round of machine learning. After a few rounds the AI was able to predict docking to the accuracy of 90 per cent.

"The machine that had been trained completed the screening much more quickly than would have been possible with the traditional docking method. While the calculation of docking took a couple of months even using powerful computers, with machine learning the learning process and prediction only took a few days."

According to Poso, researchers can new routinely screen billions of molecules in the same time that previously only managed a million. And thanks to the machine learning model, billions of molecules can now be screened without a supercomputer.

"Obviously it follows that with supercomputers we can take even bigger databases and screen thousands of billions of molecules with this method."

The next thing Poso's team will be looking at is what is known as the vivid screening method.

"Instead of just predicting a single activity or docking, we can simultaneously predict a number of different properties, such as predicting a docking that can cause side effects, while maintaining solid docking to a good location."

The research made use of the supercomputing resources, data storage and tool containerisation of the Finnish ELIXIR Node, CSC – IT Center for Science.

29.8.2024 | Ari Turunen

MORE INFORMATION:

Toni Sivula, Laxman Yetukuri, Tuomo Kalliokoski, Heikki Käsnänen, Antti Poso & Ina Pöhner (2023): Machine Learning-Boosted Docking Enables the Efficient Structure-Based Virtual Screening of Giga-Scale Enumerated Chemical Libraries. J. Chem. Inf. Model. DOI: 10.1021/acs.jcim.3c01239.

HASTEN algorithm
https://github.com/TuomoKalliokoski/HASTEN

University of Eastern Finland
https://www.uef.fi/en

CSC – IT Center for Science
is a non-profit, state-owned company administered by the Ministry of Education and Culture. CSC maintains and develops the state-owned, centralised IT infrastructure.

http://www.csc.fi
https://research.csc.fi/cloud-computing

ELIXIR
builds infrastructure in support of the biological sector. It brings together the leading organisations of 21 European countries and the EMBL European Molecular Biology Laboratory to form a common infrastructure for biological information. CSC – IT Center for Science is the Finnish centre within this infrastructure.

http://www.elixir-finland.org
http://www.elixir-europe.org

ELIXIR FINLAND
Tel. +358 9 457 2821s ● e-mail: servicedesk@csc.fi
www.elixir-europe.org/about-us/who-we-are/nodes/finland

www.elixir-finland.org

ELIXIR HEAD OFFICE
EMBL-European Bioinformatics Institute
www.elixir-europe.org