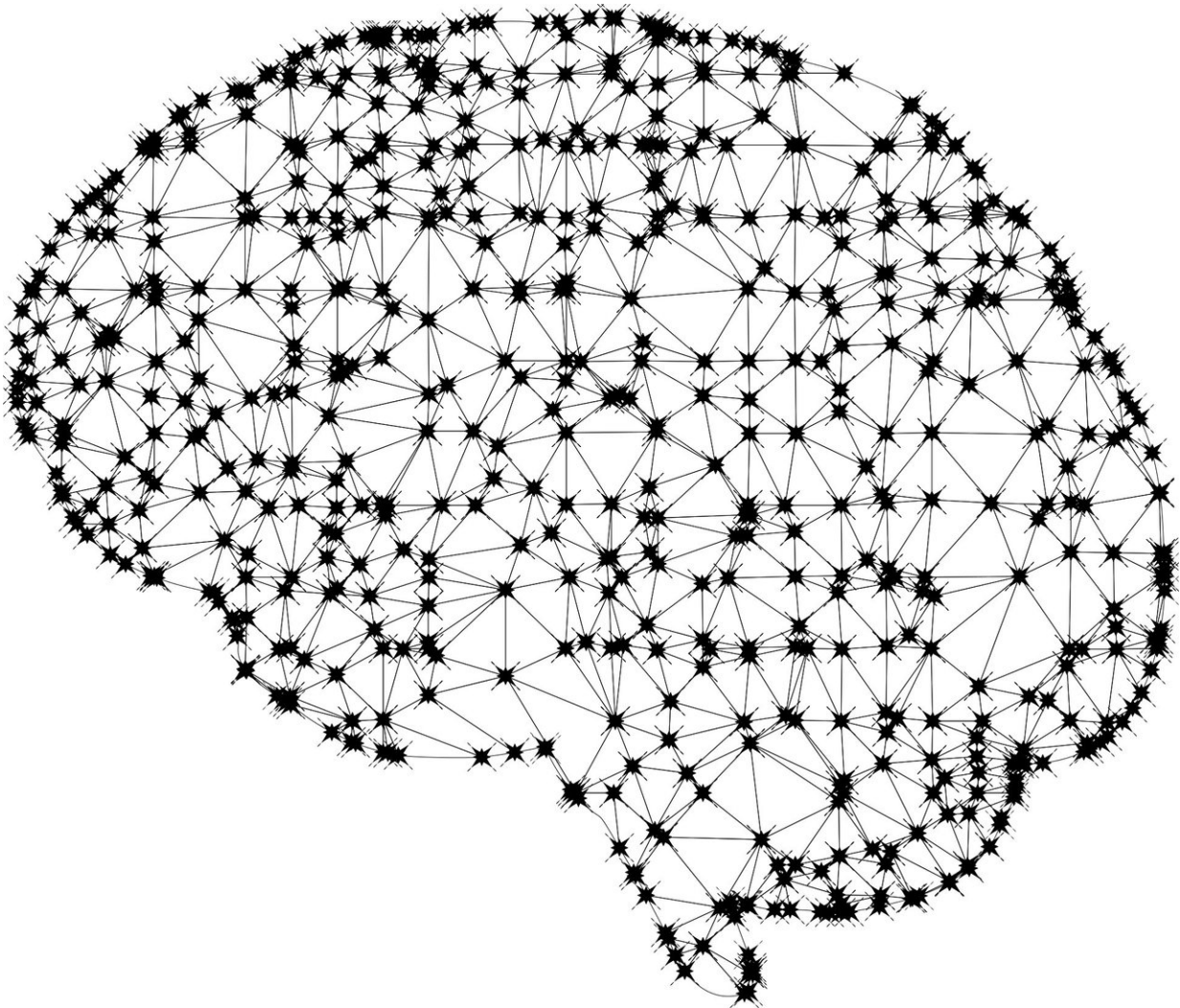# Machine learning algorithm helps in the search for new drugs

February 11 2019

Credit: CC0 Public Domain

Researchers have designed a machine learning algorithm for drug discovery which has been shown to be twice as efficient as the industry standard, which could accelerate the process of developing new treatments for disease.

The researchers, led by the University of Cambridge, used their algorithm to identify four new molecules that activate a protein which is thought to be relevant for symptoms of Alzheimer's disease and schizophrenia. The results are reported in the journal *PNAS*.

A key problem in drug discovery is predicting whether a molecule will activate a particular physiological process. It's possible to build a statistical model by searching for chemical patterns shared among molecules known to activate that process, but the data to build these models is limited because experiments are costly and it is unclear which chemical patterns are statistically significant.

"Machine learning has made significant progress in areas such as computer vision where data is abundant," said Dr. Alpha Lee from Cambridge's Cavendish Laboratory, and the study's lead author. "The next frontier is scientific applications such as drug discovery, where the amount of data is relatively limited but we do have physical insights about the problem, and the question becomes how to marry data with fundamental chemistry and physics."

The algorithm developed by Lee and his colleagues, in collaboration with biopharmaceutical company Pfizer, uses mathematics to separate pharmacologically relevant chemical patterns from irrelevant ones.

Importantly, the algorithm looks at both molecules known to be active and molecules known to be inactive, and learns to recognise which parts of the molecules are important for drug action and which parts are not. A mathematical principle known as random matrix theory gives

predictions about the statistical properties of a random and noisy dataset, which is then compared against the statistics of chemical features of active/inactive molecules to distil which chemical patterns are truly important for binding as opposed to arising simply by chance.

This methodology allows the researchers to fish out important chemical patterns not only from molecules that are active, but also from molecules that are inactive—in other words, failed experiments can now be exploited with this technique.

The researchers built a model starting with 222 active molecules, and were able to computationally screen an additional six million molecules. From this, the researchers purchased and screened the 100 most relevant molecules. From these, they identified four new molecules that activate the CHRM1 receptor, a protein that may be relevant for Alzheimer's disease and schizophrenia.

"The ability to fish out four active molecules from six million is like finding a needle in a haystack," said Lee. "A head-to-head comparison shows that our algorithm is twice as efficient as the industry standard."

Making complex organic molecules is a significant challenge in chemistry, and potential drugs abound in the space of yet-unmakeable molecules. The Cambridge researchers are currently developing algorithms that predict ways to synthesise complex organic molecules, as well as extending the machine learning methodology to materials discovery.

PNAS (2019). www.pnas.org/cgi/doi/10.1073/pnas.1810847116

Provided by University of Cambridge