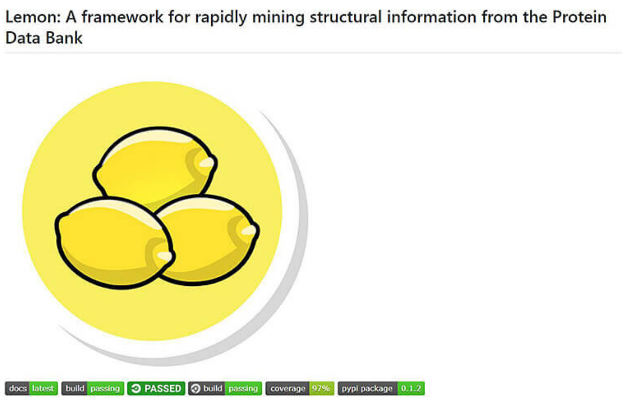
# This 'lemon' could help machine learning create better drugs

19 December 2019, by Chris Adam


Lemon: A framework for rapidly mining structural information from the Protein Data Bank

Purdue researchers have created a new system, called Lemon, for rapid mining of biomolecular interaction data to use with machine learning methods for the design of drugs. Credit: Image provided

One of the challenges in using machine learning for drug development is to create a process for the computer to extract needed information from a pool of data points. Drug scientists must pull biological data and train the software to understand how a typical human body will interact with the combinations that come together to form a medication.

Purdue University drug discovery researchers have created a new framework for mining data for training machine learning models. The framework, called Lemon, helps drug researchers better mine the Protein Data Base (PDB) – a comprehensive resource with more than 140,000 biomolecular structures and with new ones being released every week. The work is published in the Oct. 15 edition of *Bioinformatics*.

"PDB is an essential tool for the drug discovery community," said Gaurav Chopra, an assistant professor of analytical and physical chemistry in Purdue's College of Science who works with other researchers in the Purdue Institute for Drug Discovery and led the team that created Lemon. "The problem is that it can take an enormous amount of time to sort through all the accumulated data. Machine learning can help, but you still need a strong framework from which the computer can quickly analyze data to help in the creation of safe and effective drugs."

The Lemon software platform is a fast C++11 library with Python bindings that mines the PDB within minutes. Loading all traditional mmCIF files in the PDB takes about 290 minutes, but Lemon does this in about six minutes when applying a simple workflow on an 8-core machine. Lemon allows the user to write custom functions, include it as part of their software suite, and develop custom functions in a standard manner to generate unique benchmarking datasets for the entire scientific community.

"Experimental structures deposited in PDB have resulted in several advances for structural and computational biology scientific and education communities that help advance drug development and other areas," said Jonathan Fine, a Ph.D. student in chemistry who worked with Chopra to develop the platform. "We created Lemon as a one-stop-shop to quickly mine the entire data bank and pull out the useful biological information that is key for developing drugs."

Lemon got its name as it was originally designed to create benchmarking sets for drug design software and identify the lemons, biomolecular interactions that cannot be modeled well, in the PDB.

The software development work is the latest project involving health innovations from Chopra and his team. Lemon is freely available on GitHub at github.com/chopralab/lemon . Detailed documentation is available at chopralab.github.io/lemon/latest/index.html .

**More information:** Jonathan Fine et al. Lemon: a framework for rapidly mining structural information from the Protein Data Bank, *Bioinformatics* (2019). DOI: 10.1093/bioinformatics/btz178

Provided by Purdue University

APA citation: This 'lemon' could help machine learning create better drugs (2019, December 19) retrieved 21 January 2022 from https://techxplore.com/news/2019-12-lemon-machine-drugs.html