NATURAL PRODUCT DRUG DISCOVERY IN THE AGE OF METABOLOMICS AND MACHINE LEARNING.

Bryan H. Norman,[1] Marvin S. Yu,[2] Kendra Hightower,[1] Gennady Voronov,[2] Abe Frandsen,[2] Brian Bargh,[2] David Healey,[2] Rose Lightheart,[2] Tobias Kind,[2] Pieter C. Dorrestein,[3] Hannah Gordon,[4] Joe Rokicki,[2] Pelle Simpson,[2] Laurel Wright,[2] Elizabeth Spencer,[2] Alexander Kislukhin,[2] Megan Schroeder,[2] Matthew Guthrie,[2] Jordan Toutounchian,[2] Thomas Butler,[2] August Allen,[2] Viswa Colluru[2]

[1]Lead Generation, [2]Platform Sciences, [4]Product, Enveda Biosciences, Boulder, CO.
[3]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA.

Natural products have been a rich source of drugs for centuries and were a mainstay of lead generation strategies in drug discovery for much of the 20th century. Nonetheless, multiple challenges led to their diminished use in most lead generation workflows. Challenges exist along 3 axes: inability to quickly prioritize lead-like structures (requiring chemical annotation), inability to confidently identify the bioactive molecule in a mixture (requiring accurate biological annotation) and inability to access enough material to enable preclinical and clinical development (requiring material access). Technological advances in separation sciences, automation, tandem mass spectrometry (MS2), metabolomics, machine learning (ML) and synthetic biology have individually, and in combination, addressed many of the challenges associated with natural product-based drug discovery. Herein, we describe our work in the systemization of complex high throughput workflows that allow us to rapidly identify lead-like bioactive molecules in complex plant mixtures without the need for isolation. To address the first two challenges noted above, we (i) developed ML-based tools to directly predict medicinally relevant chemical properties as well as structure from MS2 spectra alone, and (ii) custom-built a high-throughput plant fractionation protocol that enables bioactivity correlation analyses of complex mixtures to annotate the structure and function of plant metabolites. The former is accomplished using an ML model we developed called MS2Mol, which can predict, with high accuracy, the structure of known and novel molecules from MS2 spectral signatures. Furthermore, we can also predict physicochemical properties, such as LogP, TPSA, Fsp3 with high accuracy using a related ML-derived tool called MS2Prop. Together, these methods allow the high throughput annotation of function and structure without the need for expensive and time-consuming isolation. We have used these methods in lead generation campaigns to successfully identify novel chemical substrates, and present both the methods and results of our applications in drug discovery workflows.