Section 1: Introduction to Artificial Intelligence and Chemistry

CHAPTER 1

Introduction

NATHAN BROWN

BenevolentAI, 4-8 Maple Street, London W1T 5HD, UK
Email: nathan.brown@benevolent.ai

1.1 Introduction

Drug discovery applies to a vast range of technologies in the interest of ushering new chemical entities of disease relevance into the clinic to meet, as yet, unmet patient needs. While many of the technological methods use experiments in so-called ''wet'' laboratories, the development and application of computational methods, often called *in silico* as opposed to *in vitro* or *in vivo*, have been in wide usage for many decades. Recently, however, a renaissance of Artificial Intelligence, and specifically Machine Learning, approaches have led the vanguard of novel applications to speed drug discovery not only in efficiency gains but also in the development of improved medications for patients.

This book covers a number of different and new approaches to applying artificial intelligence and machine learning methods in drug discovery, sometimes with entirely new algorithms, but often-times building on years of research and applied anew aided by concomitant algorithmic improvements, but also significant improvements in software and hardware that allows hitherto inaccessible quantities of computational power.

The first chapters of this book consider chemical data and learning from unstructured data sources, such as those found in the literature and in patents. One of the most challenging aspects of using Artificial Intelligence in drug discovery is teasing out insights that have been discovered but often published in a way that is not amenable to further analysis – essentially

Drug Discovery Series No. 75
Artificial Intelligence in Drug Discovery
Edited by Nathan Brown

Published by the Royal Society of Chemistry, www.rsc.org

obfuscating this data through publication in formats that cannot be easily read by machines. The first of these chapters looks at the area of chemical topic modelling, which is an unsupervised learning method to extract meaning from natural language, using a methodology called natural language processing. The data extracted permits the clustering of documents based on the co-occurrences of words in these publications thereby permitting an easier approach to grouping documents based on subject potentially obviating the need for manual curation.

Later chapters of the book cover predictive modelling, specifically making use of ligand and protein structure data, respectively. However, it can often be challenging to entirely deconvolute these two data types, since aspects of each are at least implicit, if not explicit, in the other. Ligand-based predictive modelling has a long and distinguished history in drug discovery, going right back to the pioneering work of Corwin Hansch and Toshio Fujita of Quantitative Structure-Activity Relationships (QSARs) back in the 1950s and 1960s. These QSAR equations were derived manually, originally to be used to predict certain molecular properties using mathematical models. In more recent decades, the processes have moved towards using large scale data sources and libraries of molecular descriptors to automate the generation of predictive models using more modern machine learning algorithms. However, it should be recognised that the field of Machine Learning arose contemporaneously with QSARs in the late 1950s.

One of the most important aspects of predictive modelling is not using the predicted values of the endpoint independently, but rather incorporating that with consideration of whether that prediction may be reliable based on the data that are available, or even when the data available is simply insufficient to make any predictions and perhaps the best recommendation is to generate new data to inform our data space. Krstajic offers some important advice in this area with his chapter on the importance of defining the ''don't know'' answer. This advice is two-fold in importance. Firstly, it encourages diligence in application of our predictive models to drug discovery. Secondly, however, it offers an honesty in modelling to ensure that scientists from medicinal chemistry and computational methods understand more clearly where the data is limited and offer new insights and priorities for chemical synthesis and testing to bolster those datasets.

As protein structure data becomes more prevalent due to protein crystallography becoming a routine assay type in drug discovery, new methods arise that incorporate these more sizable datasets to offer more protein binding site contextual information rather than that merely implied by ligand structure data alone. Methods included in the structure-based predictive modelling chapters are predicting protein-ligand molecular recognition, approaches to using convolutional neural networks in virtual screening, and applying machine learning methods in enhancing the impact of molecular dynamics simulations.

One of the most significant challenges in synthetic organic chemistry, let alone medicinal chemistry and drug discovery, is the design and planning

of new chemical syntheses. Given a molecular target, what series of reactions and indeed conditions can be optimised to minimise materials, effort and time to produce the desired results in appropriate yield for its intended purpose in the laboratory? The challenge of synthesis planning has been something of a holy grail in the area of applying artificial intelligence methods to chemistry and drug discovery over many decades, going back to the work of E.J. Corey and others with their retrosynthetic planning working backwards from the desired product to decide which steps should make up parts of the synthesis. However, in recent years, more modern deep learning artificial intelligence, in addition to symbolic artificial intelligence methods have come to the fore. These new methods take advantage of the vast repositories of reaction data held in public databases, proprietary data held by publishers, and internal data sources at chemical companies to rapidly synthesise options of synthetic routes that have been demonstrated to be competitive with human experts. While a significant amount of research remains outstanding to develop these methods into something that is capable of working competitively with human experts, this area of research is one of the most researched areas of computational sciences applied to chemistry in recent years.

The last chapters of this book bring together many of the topics already covered earlier in the book but brings to bear this combination of methods to the more holistic approach of molecular design. The field of drug discovery is itself a multi-objective or multi-parametric challenge. Approved drugs must satisfy requirements of safety and efficacy at the intended dose, but this is itself a convolution of many different requirements and vast arrays of assays that must be considered in the eventual recommendation of not only a drug for human benefit, but also the nomination of a clinical candidate. While methods for molecular design incorporate many different approaches, the heart of the challenge is molecular graph generation: being able to recommend what to make and how to make it to satisfy the various constraints identified as important for the disease area being considered. A number of artificial intelligence and deep learning approaches have been investigated in recent years, including recurrent neural networks, junction tree variational autoencoders and other deep generative models, riding the resurgent wave of artificial intelligence methods, it is important to reflect on the challenges involved in molecular design, independent of the method used to suggest said designs. Medicinal chemistry and drug discovery have a long history of designing new molecules both manually and using automated or semi-automated ways to suggest novel designs. A significant approach is that of matched molecular pair analysis, which has historically been considered as a relatively manual approach investigating molecular replacements around a chemical series of interest. However, in more recent years, computational methods have been developed to abstract the wealth of information from data generated in various organisations to automate a data-driven process to matched molecular pairs giving rise to a vast database of molecular transforms that have concomitant levels of statistical significance in terms of modulating certain properties of interest.

As research progresses in the area of molecular design, many of the emerging methods, together with older and more established approaches, will likely combine into various hybrid systems that permit reliable and reproducible molecular design at scale. Some of the most significant challenges include those in optimising not only potency but also ADMET properties, in addition to the design of molecular structures that can indeed be made effectively and efficiently in the laboratory to test the virtual hypotheses being asked. As we draw closer the theoretical worlds of design and the practical efforts of synthesis and testing of those designed compounds, so will the modern laboratory change and adapt to facilitate more close-coupled drug discovery and development and thereby enhance efficiency and optimisation processes, which is touched on in the last main chapter of this book with a practical vision of how we can apply these methods to democratise the discovery of new chemicals.