DRUG DESIGN USING MACHINE LEARNING

Edited By
Inamuddin
Tariq Altalhi
Jorddy N. Cruz
Moamen Salah El-Deen Refat

Scrivener Publishing

WILEY

Table of Contents

List of Tables

List of Illustrations

Scrivener Publishing
100 Cummings Center, Suite 541J
Beverly, MA 01915-6106
Publishers at Scrivener
Martin Scrivener (martin@scrivenerpublishing.com)
Phillip Carmical (pcarmical@scrivenerpublishing.com)

Drug Design Using Machine Learning

Edited by

Inamuddin

Tariq Altalhi

Jorddy N. Cruz

and

Moamen Salah El-Deen Refat

WILEY

This edition first published 2022 by John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA and Scrivener Publishing LLC, 100 Cummings Center, Suite 541J, Beverly, MA 01915, USA

For more information about Scrivener publications please visit www.scrivenerpublishing.com.

Wiley Global Headquarters

111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials, or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read.

Library of Congress Cataloging-in-Publication Data

ISBN 9781394166282

Cover image: Pixabay.Com

Cover design by Russell Richardson

Set in size of 11pt and Minion Pro by Manila Typesetting Company, Makati, Philippines

Preface

Traditionally, the design of new drugs has been a long process that requires an investment of billions of dollars. In the last decades, molecular modeling techniques have been used in the pharmaceutical industry and large laboratories to accelerate this process in order to reduce the amount of money that needs to be invested. More recently, machine learning approaches to drug discovery have aroused great interest in the scientific community. This has occurred, among other reasons, due to advances in high-performance computing and the availability of an abundance of biological and chemical information on thousands of compounds. Through machine learning and artificial intelligence approaches, this information can be filtered quickly and at a low cost. From this, algorithms can be trained to be used in the different stages of drug design.

The objective of this book is to bring together several chapters that function as an overview of the use of machine learning and artificial intelligence applied to drug development. The initial chapters discuss drug-target interactions through machine learning for improving drug delivery, healthcare, and medical systems. Further chapters also provide topics on drug repurposing through machine learning, drug designing, and ultimately discuss drug combinations prescribed for patients with multiple or complex ailments. This book should be useful for information technology professionals, pharmaceutical industry workers, engineers, university students and faculty members, medical practitioners, researchers and laboratory workers who have a keen interest in the area of machine learning and artificial intelligence approaches applied to drug advancements. A chapter-by-chapter

summary of the work reported in the 12 chapters of this book follows.

– Chapter 1 describes the use of molecular recognition for drug development through various mathematical models. Molecular docking of the main elements is discussed in detail to consider which elements are important for obtaining a reliable prediction of protein-ligand complexes. The role of machine learning in molecular recognition models is also analyzed.

– Chapter 2 reviews classical and machine learning-based approaches to study target-drug interactions in the field of drug discovery, from target identification to the optimization of the lead compound. In addition, a special section discusses peptide-based drugs.

– Chapter 3 gives a brief overview of the various machine learning techniques that underpin artificial intelligence, poly-pharmacology, and drug repurposing to improve healthcare services. The way in which machine learning is used throughout the drug development process to help increase its efficacy and robustness, resulting in a significant reduction of the time and cost of bringing new drugs to market, is discussed in detail.

– Chapter 4 elaborates on the advancements in artificial intelligence technologies along with their applications, enumerates the challenges faced by these technologies that retard their full-scale implementation, and also provides an overview of their social, legal, and economic aspects. All these are discussed for various applications such as drug delivery, healthcare, and medical systems.

– Chapter 5 details the various machine learning approaches for drug repurposing. The network-based

approach, text mining-based approach, and semantics-based approach are discussed in fair detail. Furthermore, case studies of drugs repurposed through machine learning programs are also discussed.

– Chapter 6 summarizes the recent developments in the machine learning-mediated drug discovery process within industrial and academic contexts. More precisely, machine learning algorithms used for drug discovery, bioactivity prediction using machine learning, and application of machine learning in chemo-informatics are reviewed. Furthermore, an in-depth analysis of the challenges and suggestions are provided.

– Chapter 7 details the basic principles underlying supercritical fluid technology and its main techniques. Furthermore, the more representative biodegradable polymers impregnated with supercritical fluid technology are described. Finally, the state of the art of supercritical fluid technology for improving drug absorption in biopolymer and the delivery processes are thoroughly reviewed.

– Chapter 8 details the different *in vivo, in vitro* and *in silico* techniques applied to study the protein-protein interactions through the active sites. The role of the active sites of protein is also discussed. Moreover, databases and algorithms are mentioned along with their uses and advantages.

– Chapter 9 describes various machine-learning methods involved in protein redesign and engineering along with strategies ranging from designing the model to predicting hot spots. The chapter also focuses on additional support vector machines, nearest neighbor, decision trees, neural networks, Bayesian networks, ensemble learning, and deep learning.

– Chapter 10 describes computational methods used for the selection of bioactive compounds. In this context, various approaches based on transcriptomics and artificial intelligence are discussed. Additionally, methods and applications of de novo synthesis are also addressed along with its future endeavors in drug designing.

– Chapter 11 discusses the application of computer-aided techniques for the prediction of drug effectiveness and toxicity, including artificial intelligence, artificial neural networks, and machine learning. A hierarchical method for drug design is followed as drug discovery, drug design through new techniques and application, machine learning methods, deep learning, applications, and problems.

– Chapter 12 details the use of artificial intelligence in assessing the side effects to drugs. Practicing artificial intelligence necessitates the skills and awareness for data-intensive analysis, knowledge-based management, and definite challenges. A smarter future can be envisaged using artificial intelligence-guided new scientific accomplishments in the field of pharmacovigilance.

The Editors
Inamuddin
Tariq Altalhi
Jorddy N. Cruz
Moamen Salah El-Deen Refat
July 2022

1
Molecular Recognition and Machine Learning to Predict Protein-Ligand Interactions

A. Reyes Chaparro[2], J.A. Moreno-Melendres[1], A.L. Ramos-Jacques[3] and A.R. Hernandez-Martinez[2]*

[1]*PCeIM-UNAM, Centro de Física Aplicada y Tecnología Avanzada (CFATA), Universidad Nacional Autónoma de México (UNAM), Querétaro, México*

[2]*Centro de Física Aplicada y Tecnología Avanzada (CFATA), Universidad Nacional Autónoma de México (UNAM), Querétaro, México*

[3]*Edit Academy, México, Querétaro, México*

Abstract

Molecular recognition is part of several chemical-biological processes, and is the interaction between macromolecules (such as proteins and ligands) through noncovalent bonds. This phenomenon has been extensively studied for developing new drugs. Molecular modeling is an affordable method (compared with laboratory experiments) for predicting which macromolecules may interact and, through molecular docking, which will form a stable complex. Molecular docking has two main components: (1) search algorithm and (2) scoring function. The search algorithm studies the conformational space of the ligand at the binding site. The scoring function is a mathematical model that evaluates the interaction energy of each complex, and it could be empirical by using databases of ligand-protein complexes. Results of the search algorithm are satisfactory compared with experimental data, but the scoring function still must improve its performance. Due to the complexity of analysis and management of databases, accurate predictions are difficult to obtain. Machine learning can contribute to achieve better results for predicting macromolecular interactions. Computational predictions of the interaction between macromolecules complexes enhance the development of applied technology in medicine.

Keywords: Molecular docking, search algorithm, scoring function, mathematical models

1.1 Introduction

Life is based on four main groups of macromolecules: carbohydrates, lipids, proteins, and nucleic acids. Each group of molecules separately cannot create an organism, life and all its functions arise when there is a dynamic and continuous interaction between these macromolecules and small organic and inorganic molecules. Molecular recognition refers to the process in which biological macromolecules interact with each other or with small molecules through noncovalent interactions to form a complex [1]. Molecular recognition is not a physiological process, it is rather a component of many processes, such as cell signaling, genetic regulation, metabolism, immunity, and all those processes in which there is an interaction between macromolecules (or between a macromolecule and a ligand) [2, 3]. For example, in the expression of genes, there must be a molecular recognition between the inducing proteins and

the DNA chain to generate the mRNA. Subsequently, there is molecular recognition between the mRNA and the ribosomes, and thus in each of the steps until generating a functional protein, which in turn has molecular recognition according to its function [4]. Another example is the enzymatic reactions that are carried out for the energy metabolism of the carbohydrate and lipid pathways, in addition to the xenobiotic metabolism, which is carried out simultaneously in cells, all the time [5]. The chapter discusses the ligand-protein molecular recognition using data and algorithms for improving mechanisms predictions.

1.1.1 Molecular Recognition

The molecular recognition model has evolved since 1894 when Emil Fischer proposed the “lock key” model in which the ligand and the protein were kept rigid and had a highly specific interaction [6]. Later, in 1958, Koshland [7] proposed that molecules promote a conformational change in proteins and it was called the “induced fit” model. The induced fit model considers that the protein does not always have the same conformation and that different ligands could lead to different induced fit; with these new concepts, phenomena, such as noncompetitive and allosteric inhibition, could be explained [1]. More recently, the idea of conformational selection was raised, in which proteins are found naturally in different conformations and ligands have a different affinity for each conformation [8]. Current interaction models consider that the induced fit and conformational selection phenomena occur in a complementary manner (Figure 1.1) [9].

The phenomenon of molecular recognition has been studied extensively in the development of new drugs. When a new protein involved in a physiological process or disease is described, it is considered as a potential target molecule. New drugs are tested and designed for having an interaction with that target molecule to obtain a potential therapeutic effect [10]. The most widely used experimental tool for finding new molecules with potential biological activity is the high-throughput screening (HTS), which is an automated process to screen large amounts of ligands with a single protein that allows identifying molecules with activity against the target protein [11]. Although HTS is a highly optimized process, the number of existing molecules currently amounts to millions of compounds, and it is not cost-effective enough to be used as an initial tool for drug discovery screening [12]. For this reason, computational methods are of great

interest in the screening of millions of compounds at a reasonable cost. Structure-based virtual screening is one of the most widely used to screen databases of millions of compounds, based on their affinity for the target protein [13]. The only requirement is to have the three-dimensional structure of the target protein to be able to carry out the interaction tests, these are called molecular coupling.

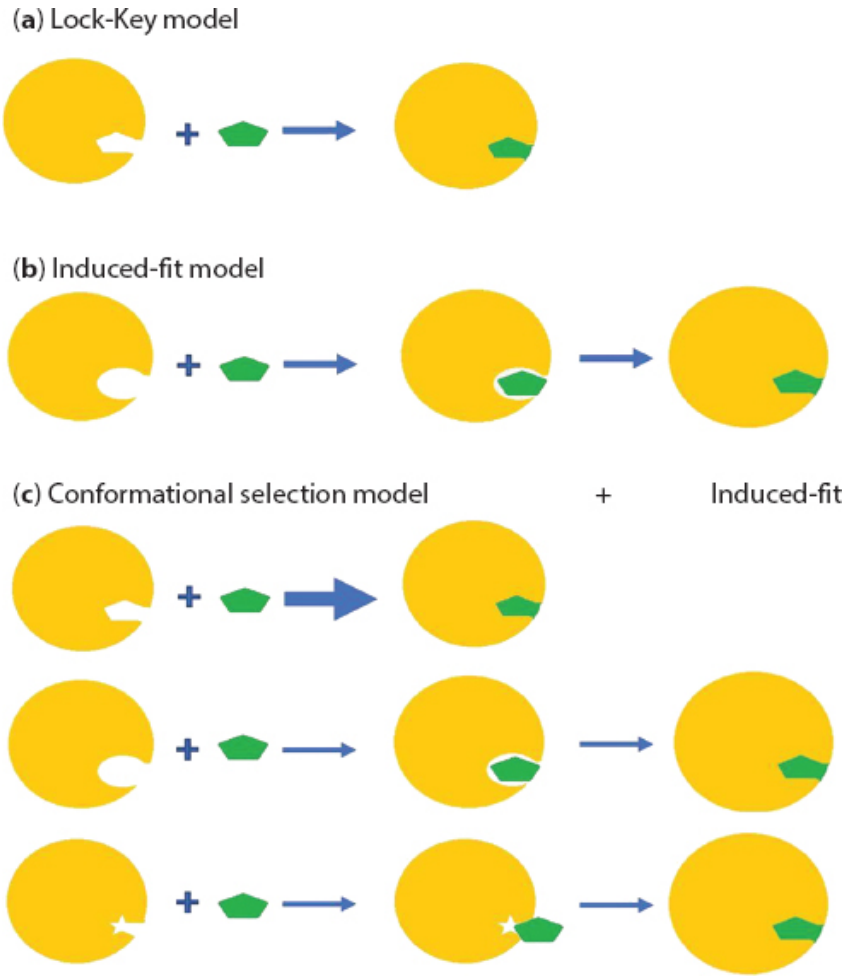

Figure 1.1 Molecular recognition models.

1.2 Molecular Docking

The use of computers for drug development is known as computer-aided drug design (CADD); many different techniques are supported by the pharmaceutical industry and universities to accelerate the development of new drugs. Some of the tools used are quantitative structure-activity relationship (QSAR) models, pharmacophore modeling, lead optimization, molecular dynamics, and molecular coupling [14]. In addition, different machine learning applications,

which will be discussed in section 1.3, have complemented the same techniques but with a machine learning algorithm.

Molecular coupling is a computational test that allows studying the interaction at the molecular level between a macromolecule, which is normally a protein and a ligand. The molecular modeling of proteins began in the 1980s, and this allowed the beginning of simulation processes, such as molecular ligand-protein recognition [15]. The results of the coupling made it possible to select a large number of substances according to their energy of interaction with the target protein. Currently, molecular docking is widely used in the search for new drugs, as a consequence of the increase in computational power and the availability of large databases of ligands and proteins [16]. Molecular docking protocols have two main components: a search algorithm that shows the conformational space of the ligand at the binding site and a scoring function that quantitatively evaluates the interaction energy of each of the conformations [17]. Finally, after sampling the conformational space and evaluating the binding energy of each pose, the conformations that present the best affinity energy are obtained. Using these results of molecular docking tests, the best molecules can be proposed to be chemically synthesized by carrying out new experimental tests (Figure 1.2) [18].

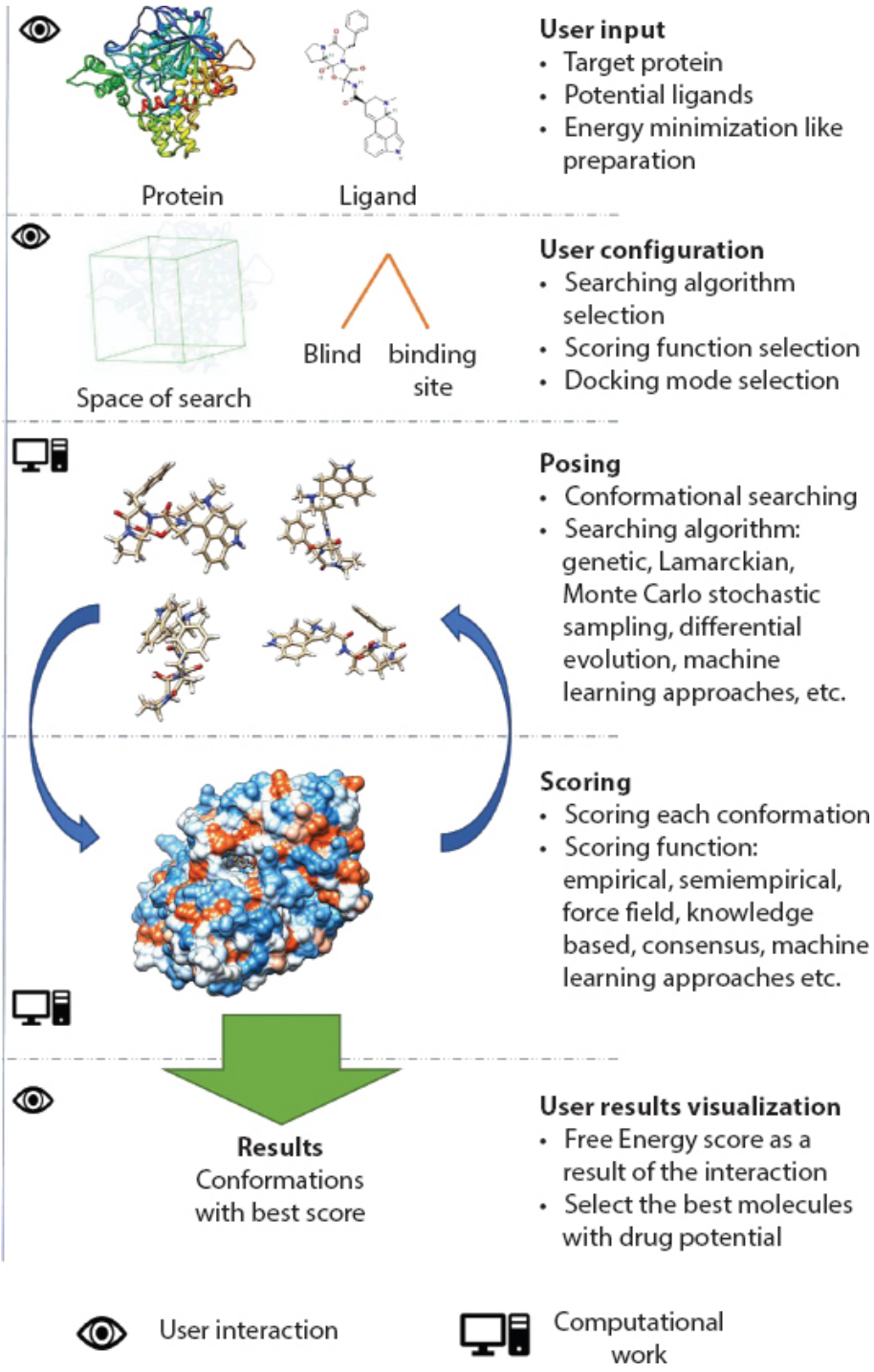

Figure 1.2 General protocol of docking work.

1.2.1 Conformational Search Algorithm

The binding pose refers to the conformation that the ligand has, with respect to the protein; in molecular docking, a large number of binding poses are evaluated trying to cover the entire conformational