LEVERAGING REPRESENTATION LEARNING FOR DRUG DISCOVERY

YOGESH KALAKOTI

DEPARTMENT OF BIOCHEMICAL ENGINEERING AND BIOTECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2023

LEVERAGING REPRESENTATION LEARNING FOR DRUG DISCOVERY

by

YOGESH KALAKOTI

Department of Biochemical Engineering and Biotechnology

submitted

in partial fulfilment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2023

CERTIFICATE

This is to certify that the thesis entitled **'Leveraging representation learning for drug discovery'** being submitted by **Mr. Yogesh Kalakoti** to the Indian Institute of Technology Delhi for the award of the degree of **'Doctor of Philosophy'**, is a record of the bonafide research work carried out by him, which has been prepared under my supervision in conformity with the rules and regulations of the Indian Institute of Technology. The research reports and the results presented in this thesis have not been submitted for any degree or diploma in any other university or institute.

Prof. D. Sundar
Institute Chair Professor
Department of Biochemical Engineering and Biotechnology
Indian Institute of Technology (IIT) Delhi

Acknowledgements

Reflecting on 2018, when I started my PhD journey, I was aware that it would be one of the most challenging and rewarding experiences of my life. This endeavour would not have been possible without individuals and institutions who have helped me along the way and made the past four and a half years.

First and foremost, I would like to extend my sincere gratitude to my supervisor, Prof. D Sundar, for his guidance, support, and encouragement throughout my research. His insights, feedback, and expertise have been invaluable in shaping my work and broadening my understanding of the subject.

I am grateful to Prof. Ravikrishnan Elangovan, Prof. Preeti Srivastava and Prof. Vivek Perumal, the members of my Students' Research Committee from, for providing support and expertise in shaping this thesis. Completing this endeavor would not have been possible without my seniors, who have greatly influenced my professional and personal growth during my time as a PhD student. I would like to sincerely acknowledge Dr. Vidhi Malik, Dr. Shayoni Dutta, Shashank Yadav and Dr. Jaspreet Kaur Dhanjal for their support, guidance, and contributions as my seniors and project partners. Their inputs, constructive criticism, and motivation have played a vital role in shaping my research work. Further, I would like to acknowledge Vipul Kumar, Dr. Dhvani Vora, Aditi Keshav and Kamlesh Kumar for their constant support and motivation as friends and lab mates.

I would also like to express my deepest gratitude to my parents for their unwavering support and encouragement throughout my academic journey. Without their love, guidance, I would not have been able to reach this point. I could strive for betterment with the support of their faith and affection. I would also like to thank my friends and fellow PhD batchmates who made my time at IIT Delhi a memorable experience.

Yogesh Kalakoti

Abstract

Deciphering the secrets of cellular machinery and disease progression is critical for developing solutions in prognosis and *in silico* drug discovery. The ability to extract quantifiable information about cellular processes allows the profiling that is necessary to devise strategies for mitigating aberrations induced in a diseased state. Advances in high-throughput technologies have allowed characterisation of this information in the form of large-scale biological datasets at multiple levels of subcellular machinery. Traditional methods used to examine this data do not have sufficient sensitivity to identify correlations between molecular patterns and distinguishing characteristics of a disease. Moreover, such profiling methods do not allow for the stratification of high-risk malignancies at a molecular level. This is one of the primary drivers of research in personalized medicine. Over the years, machine learning (ML) algorithms have found their utility in analysing biological datasets due to their ability to model complex distributions and identify data patterns. However, there is a gap in methods to effectively represent biological datasets in a way that maximises the amount of information presented to the ML agent. This thesis aims at developing frameworks capable of extracting relevant information from omics, sequence and structural datasets efficiently for downstream ML-based solutions in prognosis modelling and drug discovery. Workflows, analysis and results presented in this thesis demonstrate the utility of representation learning in developing solutions for survival estimation, motif discovery, drug-target interaction and protein fingerprinting in an effective manner.

सार

रोग निदान और इन-सिलिको दवा खोज में समाधान विकसित करने के लिए सेलुलर मशीनरी और रोग की प्रगति के रहस्यों को समझना महत्वपूर्ण है। सेलुलर प्रक्रियाओं के बारे में मात्रात्मक जानकारी निकालने की क्षमता प्रोफाइलिंग की अनुमति देती है जो रोगग्रस्त अवस्था में प्रेरित विपथन को कम करने के लिए रणनीति तैयार करने के लिए आवश्यक है। उच्च-थ्रूपुट प्रौद्योगिकियों में अग्रिमों ने उप-कोशिकीय मशीनरी के कई स्तरों पर बड़े पैमाने पर जैविक डेटासेट के रूप में इस जानकारी के लक्षण वर्णन की अनुमति दी है। इस डेटा की जांच करने के लिए उपयोग किए जाने वाले पारंपरिक तरीकों में आणविक पैटर्न और रोग की विशिष्ट विशेषताओं के बीच सहसंबंधों की पहचान करने के लिए पर्याप्त संवेदनशीलता नहीं है। इसके अलावा, इस तरह की प्रोफाइलिंग विधियां आणविक स्तर पर उच्च जोखिम वाले विकृतियों के स्तरीकरण की अनुमति नहीं देती हैं। यह वैयक्तिकृत चिकित्सा में अनुसंधान के प्राथमिक चालकों में से एक है। वर्षों से, मशीन लर्निंग एल्गोरिदम ने जैविक डेटासेट का विश्लेषण करने में अपनी उपयोगिता पाई है क्योंकि जटिल वितरणों को मॉडल करने और डेटा पैटर्न की पहचान करने की उनकी क्षमता है। हालांकि, जैविक डेटासेट को प्रभावी ढंग से प्रस्तुत करने के तरीकों में एक अंतर है जो एमएल एजेंट को प्रस्तुत जानकारी की मात्रा को अधिकतम करता है। इस थीसिस का उद्देश्य प्रोग्नोसिस मॉडलिंग और ड्रग डिस्कवरी में डाउनस्ट्रीम एमएल-आधारित समाधानों के लिए प्रभावी रूप से ओमिक्स, अनुक्रम और संरचनात्मक डेटासेट से प्रासंगिक जानकारी निकालने में सक्षम ढांचे को विकसित करना है। इस थीसिस में प्रस्तुत कार्यप्रवाह, विश्लेषण और परिणाम जीवित रहने के आकलन, मूल भाव की खोज, ड्रग-टारगेट इंटरैक्शन और प्रोटीन फ़िंगरप्रिंटिंग के प्रभावी तरीके से समाधान विकसित करने में प्रतिनिधित्व सीखने की उपयोगिता को प्रदर्शित करते हैं।

Table of contents

List of figures

Page

List of tables

List of equations

List of abbreviations

BRCA	BReast CAncer gene 1
CADD	Computer Aided Drug Design
CNN	Convolutional neural networks
CNV	Copy number variation
DNN	Deep Neural Networks
DTI	Drug Target Interaction
EMR	Electronic medical records
ENCODE	The Encyclopaedia of DNA Elements
GEVA	Gene-set Variation analysis
GNN	Graph neural networks
GPU	Graphical Processing Unit
ICGC	International Cancer Genome Consortium
LSTM	Long Short-Term Memory
LUAD	Lung Adenocarcinoma
MKL	Multiple Kernel learning
ML	Machine Learning
NCA	Neighbourhood component analysis
NLP	Natural Language Processing
PCA	Principal component analysis
PLM	Protein Language Models
RNN	Recurrent neural network
SVM	Support Vector Machines
TCGA	The Cancer Genome Atlas
APBS	Adaptive Poisson-Boltzmann Solver
MSA	Multiple sequence alignment

Acronyms of the frameworks/workflows developed in this thesis

dNNDR	Deep neural network-based drug recommendation system
SurvCNN	Survival prediction with Convolutional Neural Networks
TFactorNN	Transcription factor binding site with interpretable deep learning
TransDTI	Transformer based drug-target interaction prediction system