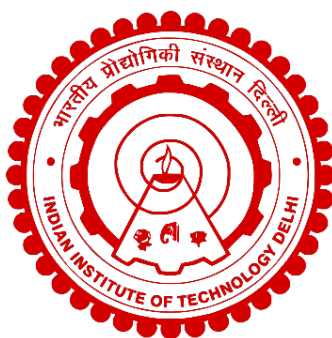


LEVERAGING REPRESENTATION LEARNING FOR DRUG DISCOVERY

YOGESH KALAKOTI



**DEPARTMENT OF BIOCHEMICAL ENGINEERING AND BIOTECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI**

JULY 2023

© INDIAN INSTITUTE OF TECHNOLOGY DELHI (IITD), NEW DELHI, 2023

LEVERAGING REPRESENTATION LEARNING FOR DRUG DISCOVERY

by

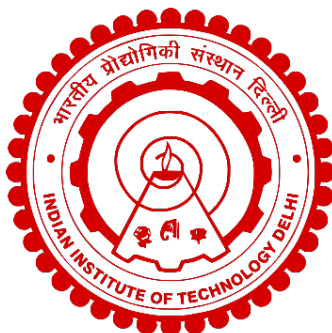
YOGESH KALAKOTI

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-श्रूपट 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

	Page
Cetrificate.....	i
Acknowledgements.....	ii
Abstract.....	iii
List of figures.....	xiii
List of tables.....	xvi
List of equations	xix
List of abbreviations	xx
CHAPTERS	
1 Introduction to biological big data and drug discovery	1
1.1 Central dogma of molecular biology	2
1.2 Biological big data and precision medicine	3
1.2.1 Precision medicine.....	3
1.2.2 High throughput technologies	3
1.2.3 Data generation and retrieval systems	5
1.3 Machine learning.....	5
1.3.1 Representation learning	7
1.3.2 Recurrent neural networks.....	8
1.3.3 Graph neural networks.....	9
1.4 Drug discovery and disease phenotype modelling.....	9
1.4.1 Efficient representation and integration of omics data: challenges and opportunities	10
1.4.2 Large language models and natural language processing for modelling drug target interactions.....	11
1.4.3 Structural datasets for developing statistical methods for modelling drug-target interactions.....	12
1.5 Challenges and opportunities in data-driven disease phenotype estimation and drug discovery	12

1.6	Definition of the problem and objectives.....	15
1.7	Thesis organisation	16
2	<i>ML-assisted omics integration for modelling cellular mechanisms and prognosis</i>	20
2.1	Introduction.....	21
2.1.1	Statistical multi-omics integration.....	21
2.1.1.1	Latent variable factorization	21
2.1.1.2	Network based methods	21
2.1.1.3	Machine learning based methods	22
2.1.2	Limitations in conventional omics integration methodologies.....	23
2.1.2.1	The curse of dimensionality	23
2.1.2.2	Data heterogeneity	23
2.1.2.3	Incomplete or missing data	24
2.1.3	Multi-omics data processing and downstream applications	24
2.2	Survival modelling in cancer and the utility of mRNA, miRNA and methylation data.....	25
2.2.1	Background.....	25
2.2.1.1	Representation learning for multi-omics data as a platform for efficient integration and downstream applications	25
2.2.2	Material and methods	26
2.2.2.1	Dataset retrieval and processing	26
2.2.2.2	Pre-processing TCGA Breast cancer patient's data	27
2.2.2.3	Data transformations and validation	28
2.2.2.4	Designing, training and optimization of a CNN-based architecture	30
2.2.2.5	Survival analysis	31
2.2.2.6	Loss function	32
2.2.2.7	Hazard probability	32
2.2.2.8	Dichotomizing Patient Groups Using Kaplan Meier Estimates	33
2.2.2.9	Performance Metrics	33
2.2.3	Results	34

2.2.3.1 Multi-omics integration improves survival prediction in BRCA patients	34
2.2.3.2 Image representations of omics data have predictive ability	34
2.2.3.3 Comparative analysis with similar methods reinforces the effectiveness of the developed framework	35
2.2.3.4 Combinatorial analysis of different combinations of omics datasets reveals significant differences in survival modelling abilities	37
2.2.4 Discussion.....	38
2.3 Modelling protein-DNA binding with ATAC-seq and ChIP-seq data	42
2.3.1 Background.....	42
2.3.2 Material and methods	43
2.3.2.1 Data retrieval and processing	43
2.3.2.2 Model architecture and training strategy	45
2.3.2.3 Model architecture	46
2.3.2.4 Attention layer	46
2.3.2.5 Benchmarking and comparison	48
2.3.2.6 Model interpretation and saliency analysis	49
2.3.2.7 K-mer enrichment analysis	49
2.3.3 Results	50
2.3.3.1 Predicting TF binding with motif occurrences	50
2.3.3.2 The developed model shows better performance than similar motif-based and ML-based methods	55
2.3.3.3 Model interpretation to identify contextual DNA features	56
2.3.3.4 K-mer enrichment to identify important sub-structures in the input sequences	58
2.3.3.5 Cell-type specific regulatory features	59
2.3.4 Discussion.....	60
2.4 Overall chapter conclusion.....	61
3 Natural language processing to model biomolecular interactions	63
3.1 Introduction.....	63
3.1.1 General premise of the drug-target interaction prediction problem	64

3.2 <i>dNNDR</i>: Deep neural network-based drug recommendation workflow.....	66
3.2.1 Background.....	66
3.2.2 Materials and methods.....	67
3.2.2.1 General framework and data processing pipeline for the drug-target interaction prediction problem	67
3.2.2.2 Chemical datasets	68
3.2.2.3 Pharmacological data	69
3.2.2.4 Model architecture	70
3.2.2.5 Input representations	70
3.2.2.6 biLSTM with attention and 1-D convolution	71
3.2.2.7 Loss and model optimization	72
3.2.2.8 Evaluation metrics	73
3.2.2.9 Baseline models for comparison	73
3.2.3 Results	74
3.2.3.1 External validation reinforces the robustness of the approach.	75
3.2.3.2 Predicting unknown drug-target interactions	76
3.2.4 Discussion.....	76
3.3 <i>TransDTI</i>: Transformer-based language models for estimating DTIs and building a drug-recommendation workflow	78
3.3.1 Background.....	78
3.3.2 Material and methods	80
3.3.2.1 Datasets and study design	80
3.3.2.2 Formulation of the problem	81
3.3.2.3 Protein and drug embeddings	82
3.3.2.4 Model architecture	82
3.3.3 Results	86
3.3.3.1 Performances of methods under evaluation and selected hyperparameters	86
3.3.3.2 Comparison of performance with other molecular descriptors and existing models	88

3.3.3.3 Models trained with random embeddings depicted the importance of choosing appropriate evaluation metrics	90
3.3.3.4 Molecules predicted by the models are at par with known inhibitors in terms of interacting potential	90
3.3.3.5 Molecular dynamics validates the effectiveness of predicted molecules	91
3.3.3.6 Lower dimensional visualization of protein embeddings	93
3.3.4 Discussion.....	93
3.4 Overall chapter conclusion.....	95
4 Biomolecular fingerprinting for downstream applications in drug-target interaction modelling	97
4.1 Introduction.....	97
4.1.1 Molecular representations for deep learning	98
4.1.2 Comparing current physics-based DTI prediction approaches.....	100
4.1.3 Limitations in current protein structure representation methods.....	101
4.1.4 Representation of bimolecular structure data	102
4.1.4.1 Topological and electrostatic properties embedded on graphs as an effective way of modelling protein surfaces	103
4.2 Introducing <i>Alphadock</i>: An end-to-end ML-based framework developed for predicting protein-ligand binding poses	104
4.2.1 Material and methods	106
4.2.1.1 Datasets and pre-processing	106
4.2.1.2 Representing protein and ligand structures as graphs	107
4.2.1.3 Graph neural networks and model architecture	108
4.2.1.4 Training	110
4.2.1.5 Benchmarking	110
4.2.1.6 Predicting binding conformers	110
4.2.2 Results	111
4.2.2.1 Statistics-based scoring function is comparable with existing physics-based scoring functions	111

4.2.2.2 Employing distance likelihoods in the scoring function aids in estimating molecular conformations with a high degree of confidence	111
4.2.2.3 Including sequence embeddings along with structural features improve performance	112
4.2.3 Discussion.....	112
4.3 Introducing <i>M-Ionic</i>: Workflow to identify and characterize metal ion binding sites	114
4.3.1 Material and methods	116
4.3.1.1 Dataset retrieval	117
4.3.1.2 Dataset pre-processing	118
4.3.2 Results	119
4.3.2.1 Performance at protein level	119
4.3.2.2 Performance at residue level	120
4.3.2.3 Impact of MSAs on model performance	121
4.3.3 Discussion.....	122
4.4 Overall chapter conclusion.....	122
5 Conclusion and future directions	123
5.1 Conclusion of the thesis	123
5.2 Future directions	124
5.2.1 Deep generative models for molecular design and lead optimisation	124
5.2.2 Estimating protein folding trajectories by protein contact-map-driven directed walks	125
6 Bibliography	127
7 Appendix	143
8 Data availability	150
9 Peer-reviewed publications	151
10 Resume of the author	152

List of figures

	<i>Page</i>
Figure 1-1: Central dogma of molecular biology and the various types of quantifiable information that can be extracted with current high throughput methods.	2
Figure 1-2: Interdependencies within multi-omics datasets and the limiting factors for conventional omics data analysis methodologies	4
Figure 1-3: Schematic overview of the hierarchies in artificial intelligence (left). Summary of multiple levels of informations that is propagated in different approaches to solving a problem (right).	6
Figure 1-4: High throughput virtual screening (HTVS) and its characterization in terms of ligand based, target based, and target-ligand based methods.	10
Figure 1-5: Basic training and validation methodology for a machine learning setup. The figure also summarises the multiple types of machine learning solutions with examples of individual methods.	14
Figure 1-6: Overall premise and challenges of phenotype estimation and drug discovery. The figure summarises the bottlenecks and opportunities at multiple levels of the prognosis modelling and <i>in silico</i> drug discovery workflows that have been addressed in the thesis.	15
Figure 1-7: Artificial intelligence and its utility in <i>in silico</i> drug discovery. The figure summarises the general pipeline that an AI-assisted drug discovery workflow follows, along with some of the technologies employed.	17
Figure 2-1: Schematic depicting the general workflow for <i>SurvCNN</i>. (A) Multi-omics data retrieval, (B) feature transformation, (C) building, training, and optimizing a deep neural network and (D) downstream analysis and inference.	27
Figure 2-2: Pipeline for the image transformation. (A) Generate lower dimensional representations of data. (B) Find a minimum bounding rectangle and re-orient. (C) Label each point in the image according to the value in omics data for pixels with multiple genes, take an average and replace the value.	28
Figure 2-3: Graphical representations of omics datasets generated via t-SNE and UMAP. Every pixel in the image can be attributed to a gene, while its intensity quantifies the corresponding expression levels.	29
Figure 2-4: Silhouette analysis on the clustered cancer genes for both the approaches (t-SNE and UMAP) helped us to optimize the parameters used for the generation of images. For t-SNE, perplexity was optimized, while number of neighbours and minimum distance were optimized for UMAP.	30
Figure 2-5: Detailed methodology with CNN architecture for survival prediction model. It includes the transformed omics data images processed using a multi-layered CNN architecture.	31
Figure 2-6: Differential analysis and survival classes. (A) Image representations for all the omics types for the two patient risk groups depicting the differential genes (marked with colours). Kaplan-Meier plots for (B) train and (C) test sets	

showing that <i>SurvCNN</i> can segregate patients into two risk groups with a high degree of confidence (log-rank p-value).....	35
Figure 2-7: Overall performance of developed workflow on different metrics of performance and projection algorithm. (A) Performance in terms of C-Index, Brier score and ICPCW (B) <i>SurvCNN</i> outperforms competing methods on nine out of twelve omics combinations tested. (C) A table describing various combinations of omics datasets used and their mapping to roman numerals.....	36
Figure 2-8: A stacked plot representing the number of patients that had an event that was predicted by the model. For the testing set, about 85% people in the low-risk set were found to be living while, about 61% of high-risk patients that were actually deceased during the given time period.....	40
Figure 2-9: Enriched terms from KEGG 2019 Human pathways, GO Biological processes and Cell type (Human gene atlas) for (A, B, C) overexpressed gene-set.	41
Figure 2-10: Data retrieval and processing pipeline that includes (A) identifying motifs for individual TFs, (B) extracting motif instances in the hg19 genome and (C) screening TF-bound sequences based on datasets from ENCODE.....	44
Figure 2-11: Model architecture for <i>TFactorNN</i>. The complete model architecture based on biLSTM and attention that was developed to predict TF-DNA binding.	47
Figure 2-12: Overall training and analysis workflow for <i>TFactorNN</i>. (A) Highlighting the research problem that given two DNA sequences containing the TF binding motif and are sterically accessible, what are the other sequential features that governs the preferential binding of TF to one sequence over the other. (B) Data processing and pre-training followed by (C) fine tuning. Ultimately, the workflow tries to (D) interpret the model predictions to identify critical sequential features.....	51
Figure 2-13: Cross model performance among all the twenty-six TFs under consideration. Heatmaps for auROC, auPR and MCC scores for (A) baseline and (B) CM models show the difference in generalizability between the two types of models. While baseline models perform well on their own data, CM models showed greater generalizability.....	54
Figure 2-14: Hierarchical clustering on cross-model Matthew's correlation coefficient (MCC) scores to ascertain if model performances correlate with the family to which the TF belongs.	55
Figure 2-15: Saliency analysis to identify important context features in the 1kb DNA sequence. Bound and unbound sequences have stark differences in region of importance throughout the 1kb sequence, which is further quantified in terms of average importance (saliency) scores. Significant differences in terms of p-value<0.001 are marked in green.....	57
Figure 2-16: Cell-type-specific enrichment of 6-mers influential for CTCF binding. Enrichments were computed using Fisher's exact tests using GC-controlled models trained separately for GM12878, HeLa-S3, or k562 cells. (A) Comparison of top-scoring 6-mers for HeLa-S3 versus k562 cells. (B) Comparison of top-scoring 6-mers for GM12878 versus k562 cells. (C) Comparison of top-scoring 6-mers for HeLa versus GM12878 cells. The	

inset table shows the auROC obtained from training each model on one cell type and using it to predict CTCF binding status in another cell type. Dashed black lines denote an adjusted P-value threshold of 0.05, based on a Bonferroni correction for the number of 6-mers tested.....59

Figure 3-1: Natural language processing (NLP) and its evolution over the years. (A) The illustration summarises the advancements in NLP and its utility in processing protein sequences along with the (B-E) basic structure of various architectures that are employed to process sequences.64

Figure 3-2: A general overview of the overall methodology. (A) Primary DTI data was collected, processed, screened, and encoded. The statistic of the screened dataset is summarized in (B). (C) An attention-biLSTM network was constructed for the protein sequences, (D) a gCNN derivative was employed for drug SMILES. (E) A fully connected feed-forward neural network taking inputs from attention-biLSTM, gCNN and (F) molecular descriptors were trained for a (G) multi-class classification problem in a five-fold cross-validated setup.66

Figure 3-3: Overall statistics of the datasets used in the study. (A, B) A general overview of multiple interactions in the datasets is depicted for drugs and targets. (C) Estimation of the sequence thresholds for drug SMILES and protein sequences. The length of SMILE strings and protein sequences does not exceed a value of about 150 and 2000 for any dataset. (D) The number of data points and an estimate of their activities for each of the three interaction-groups.....68

Figure 3-4: Comparison of performances among all the methods under observation. ROC curves for the three classes, namely, (A) active, (B) intermediate and (C) inactive, indicate the effectiveness of *dNNDR* over baseline methods in predicting the type of interaction for a given drug-target pair. (D) Validation accuracy, (E) Precision-recall curves and (F) Macro averaged F1-score also follows a similar trend and reinforces the utility of the developed methods.....74

Figure 3-5: A general overview of the overall methodology. (A) Primary DTI data was collected, processed and screened. (B) Embeddings were generated for protein and drug sequences using multiple language model-based methods. (C) Extracted embeddings were used to train a fully connected feed forward deep neural network for classification and regression task. (D) Molecular docking and dynamics workflow for validating the activity of predicted drug-target interactions.....79

Figure 3-6: Schematic representation of the seed model architecture on which all the *TransDTI* methods are based on. It uses fully connected feed-forward neural networks to process protein and molecular sequence embeddings.....83

Figure 3-7: Comparison of performances among all the methods under observation. (A) MCC and (B) R2 statistics of all the methods under consideration indicate the effectiveness of *TransDTI* over baseline methods in predicting the type of interaction for a given drug-target pair. (C) Comparison with existing methods also validates the effectiveness of the developed method. It should be noted that AlphaFold model was employed for the comparisons. (D) auROC and auPR statistics for the three classes are depicted for all the *TransDTI* models.....87

Figure 3-8: Performance of prediction models on partial random data. (A-D) Prediction losses, MAE and R2 trends for models trained on random protein and drug embeddings. (E, F) Model performance in terms of ROC and

PR curves for randomized data. (G) R2 scores for regression models trained on partial and complete random data.	90
Figure 3-9: Docking statistics for known and predicted molecules with map2k and TGFβ. Docking scores of the predicted molecules were at-par with that of the known inhibitors for map2k and TGF β	91
Figure 3-10: Interaction poses for averaged structures for all the simulated drug-target complexes for (A-D) MAP2k and (E-H) TGFβ. It can be observed that the interacting partners for the predicted compounds are similar to the ones for the known inhibitors. Also, interactions and their interacting residues are annotated.	92
Figure 3-11: TSNE mappings for proteins. It showed a significant level of clustering among protein classes even though the model was never explicitly trained to do so.	94
Figure 4-1: General workflow of <i>AlphaDock</i>. Pairwise distances between protein and ligand nodes are used as ground truth to compare against the multi-modal probability distribution predicted by the trained model.	106
Figure 4-2: Overall structure of the developed <i>AlphaDock</i> framework. (A) Protein and ligand are represented into their respective graphical forms. (B) gCNN and residual blocks to extract features from protein and ligand graphs followed by pairwise concatenation. (C) Processing of concatenated feature vectors through feed forward neural network, mixture density models and optimization algorithm to generate accurate ligand conformation.	108
Figure 4-3: Comparative analysis of the <i>AlphaDock</i> scoring function with already existing physics-based functions on the CASF 2016 benchmark dataset. The figure compiles results of comparative analysis for (A) Docking power (B) Forward screening power and (C) Reverse screening power.	111
Figure 4-4: Spearman correlation between the scores of the predicted and real conformations for all the scoring functions under consideration	113
Figure 4-5: Workflow for the residue-level ion binding prediction task using embeddings of protein sequences. Downstream validation was performed with standard benchmark datasets.	116
Figure 4-6: Log distribution of frequency of each metal-ion in the curated dataset with minimum 1000 samples	118
Figure 4-7: Performance of <i>M-ionic</i> in validation and benchmark runs. (a) Protein-level: Comparison of ROC curves for the performance of each ion type for M-Ionic trained on esm2_t33_650M_UR50D embeddings, LMetalSite and mebi-pred, (b) Residue-level: MCC scores of performances of M-Ionic (trained on esm2_t33_650M_UR50D and esm_msa1b_t12_100M_UR50S embeddings) and LMetalSite on homology reduced independent test set	120
Figure 4-8: MCC scores for the various models trained (a) esm_msa1b (t12_100M_UR50S_batch128) where esm_msa1b (t12_100M_UR50S) is trained with batch size 128 (b) esm2 (t33_650M_UR50D_batch1) is trained with esm2 (t33_650M_UR50D) model with batch size 1 (c) esm2 (t33_650M_UR50D_batch128_scrambled) trained on esm2 (t33_650M_UR50D) model with batch size 128 but with the residues within	

each batch were shuffled randomly along the length dimension such that the positions of the residues within the sequence were changed but the embedding of each residue was untouched (d) esm2 (t33_650M_UR50D_batch128_CV) is trained with esm2 (t33_650M_UR50D) model with batch size 128 as in original M-Ionic implementation (e) ProtT5 (XL_UniRef50) trained with ProtT5 (XL_UniRef50) model with batch size 128.....121

List of tables

Table 1-1: Comprehensive survey of representation learning application in biological sequences.	7
Table 1-2: Dimension reduction methods that haven been traditionally employed for multi-omics datasets.	13
Table 2-1: Overall statistics for all the omics data-types combinations for <i>SurvCNN</i>	26
Table 2-2: Pairwise t-test to validate the statistical significance of improved performance by <i>SurvCNN</i> . A positive test metric (t) denotes the superiority of <i>SurvCNN</i> . Annotations in bold mark the instances where <i>SurvCNN</i> significantly ($p < 0.005$) outperforms the competition.	37
Table 2-3: Performances comparison with different combinations of multi-omics data by pairwise paired t-test, according to C-index among five-fold cross-validation results. Note: Negative t-statistic indicates that Set1 is better than Set2. p-value < 0.05	38
Table 2-4: Performance of the survival models for multiple combinations of omics datasets and different ways to represent the data. The total performance of 48 models is depicted using three metrics: C-index, Brier score, and IPCW score.	39
Table 2-5: Summary statistics for the number of ChIP-seq experiments, number of sequences per experiment and mean sequence lengths for individual transcription factors (TFs) included in the study.	44
Table 2-6: Performance metrics for TF-binding site prediction for all the TFs under consideration.	52
Table 2-7: Effect of fine-tuning and pre-training on model performance for the baseline model.	53
Table 2-8: Performance comparison (averaged over all the TFs) with existing methods of TF binding prediction.	56
Table 2-9: Comparison of multiple model interpretation methods based on metrics such as runtime, accuracy, recall and signal-noise ratio for eight out of twenty-six TFs under consideration.	56
Table 3-1: A predictive set of compounds can be easily identifiable with AI in a relatively small amount of time and at a quarter of the cost of traditional methods.	65
Table 3-2: General summary of all the datasets used in the study.	69

Table 3-3: Key characteristics of all the models tested in the study.....	70
Table 3-4: Summary of all the models under study and the final dimensions of the data after processing	71
Table 3-5: The performance of all models under consideration are summarized. All the models five-fold cross-validated with the standard deviation mentioned alongside every result. The best performing among cross-validated models are marked in bold.	75
Table 3-6: External validation results reinforcing the effectiveness of the various <i>dNNDR</i> models.	76
Table 3-7: Novel drug target interaction predicted by <i>dNNDRb</i> . A subset of the novel predictions from drugs and proteins in the gold-standard dataset is reported here.	77
Table 3-8: Summary of all the datasets used in the study. KIBA dataset was employed for training and internal validation, while gold-standard dataset from DTI-MLCD was used for external validation.	81
Table 3-9: Summary of all the protein embedding models employed for developing the <i>TransDTI</i> methodology.	82
Table 3-10: Ten-fold cross validated performance metrics for the various models under consideration. Standard deviations for the 10-fold CV runs are in the brackets and the best performing model is marked in bold.	88
Table 3-11: External validation on gold-standard datasets shown the effectiveness of <i>TransDTI</i> against baseline and competing. All models are validated for a binary classification task.	89
Table 3-12: Residues of map2k and TGF β interacting with the known and predicted ligands in their best-docked poses.	93
Table 4-1: General terminology used and their description for easier understanding and reference.....	99
Table 4-2: Summary of the most commonly available protein structure datasets	103
Table 4-3: List of features that were embedded as node features in the graph representations of protein and drugs.	107
Table 4-4: The effect of number of gaussians in the multi density network analysis for estimating the binding conformation.	112
Table 4-5: Summary of dataset employed for training the <i>M-ionic</i> framework.	117

Table 4-6: List of current metal-ion prediction methods and the ions predicted for each method.	119
Table 4-7: Performance on distinguishing metal binding and non-binding proteins based independent test set and negative set.	120

List of equations

eq. 1-1 Supervised learning.....	6
eq. 1-2 Unsupervised learning	6
eq. 1-3 Reinforcement learning.....	6
eq. 2-1 Custom loss function for survival estimates	32
eq. 2-2 Survival probability	33
eq. 2-3 Survival rate.....	33
eq. 2-4 Sequence representation	46
eq. 2-5 Attention mechanism	47
eq. 2-6 Softmax function.....	48
eq. 2-7 F1-score	48
eq. 3-1 Sequence representation for biLSTM.....	71
eq. 3-2 Attention mechanism	72
eq. 3-3 Categorical cross entropy loss	73
eq. 3-4 F1-score	73
eq. 3-5 Exponential linear unit.....	83
eq. 3-6 Categorical cross entropy loss	83
eq. 3-7 Mean squared error	84
eq. 3-8 F1-score	84
eq. 4-1 Message passing layer of graph neural network.....	109
eq. 4-2 Readout layer of graph neural network.....	109
eq. 4-3 Dropout layer	109
eq. 4-4 ELU for mean	109
eq. 4-5 ELU for standard deviation.....	109
eq. 4-6 Total loss function for custom scoring function	110
eq. 4-7 Scoring function.....	110

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

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