Principles of Digital Biology Individual Report

11326396

Topic: Application of AI

1 Definition of Scope and Analysis of the Problem

Genomics data plays a crucial role in understanding and managing diseases for understanding the genetic basis of diseases, precision medicine, early detection and prevention, and drug discovery.

Historically, Sanger sequencing technology only gradually researches genes (linked with the clinical presenting phenotype) individually. The development of NGS provides a much cheaper and higher-throughput alternative to sequencing DNA than traditional Sanger sequencing, as whole small genomes can be sequenced in a day [GW13].

NGS technologies developed from Single Gene Tests and targeted Gene Panels, to Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS), analyzing short fragments of nucleic acids simultaneously. However, a large number of parallel sequencing tasks cause a huge amount of genomics data that needs to be stored and processed in a short time. One of the major challenges is to efficiently transform the big genomics data into valuable knowledge (providing insights into disease mechanisms, diagnosis, and treatment), which includes the complexity of errors and the considerable execution time, making the old statistical algorithms not so efficient anymore [CCWV18].

In response to the above issues, DL is known for 1) the ability to process large-scale data (has been proven in image and video processing, natural language processing areas, etc.), 2) the ability to learn the internal structure of data, and automatically identify important features in genomics data, 3) by adjusting the model structure and parameters, it can be optimized for different tasks with high flexibility, 4) through transfer learning, the model trained on one task/dataset can be used on another task/dataset, 5) excel at identifying and modeling complex biological processes and patterns in genomics data. In summary, DL has the potential to provide solutions for this issue and has made progress on genomics data in some disease tasks.

Therefore, it is necessary to investigate the applications and effects of DL technology on genomics data (in terms of diseases), which can provide us with a reference, that is, whether exploring and applying DL in lab work can increase the efficiency and accuracy of related tasks.

2 Discussion of Strategy to Address Problem

In this section, I briefly introduce DL and its common process (help to understand its applications and limitations), then summarise DL applications in four main disease tasks, give some specific examples, and discuss the limitations of the existing methods.

2.1 What is deep learning?

Deep learning is a technology that simulates the human brain for analysis and learning by using artificial neural networks with a multi-layer structure [Sch15]. **Figure 1** shows the relationship among Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL), adapted on [JZH21]. Their relationship is emphasized as in DL applications based on genomic data, it is often not used alone, but in combination with other traditional machine learning or AI technologies, to complete genomic data analysis and disease tasks, for example, unsupervised machine learning algorithms (PCA, K-Means) are often used to pre-process the dataset (clustering) before applying DL methods.

DL models gradually abstract high-level feature representations from raw data through multi-layer processing, without manually designing features, compared with traditional ML methods, it can automatically discover effective feature representations, reducing reliance on professional knowledge.

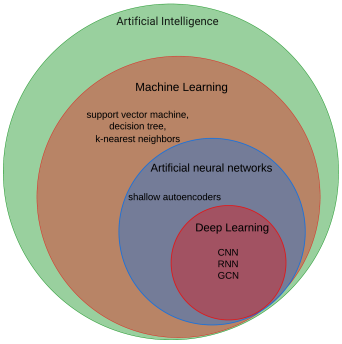

Figure 1: Relationship among AI, ML, and DL (adapted on [JZH21]).

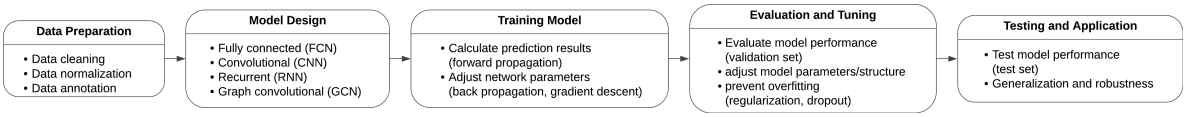

Figure 2: Common Process of deep earning models.

Figure 2 shows the common process of DL models, the content and role of every part are explained to help understand applications (successes) and limitations on genomics data in the following parts.

1) data preparation: the first thing after getting data is Exploratory Data Analysis (EDA), including summary statistics and visualization. Based on EDA, perform data cleaning, including encoding data (eg. for categorical data: one hot encoding [LSXJ18]), and deal with missing data (eg. multiple imputation and regression). Then, data normalization is performed (eg. mix-max scaling and zero mean), and data annotation is needed (for datasets without labels). Finally, as many features are uninformative in a real-world dataset, leading to more model training time and model overfitting, feature selection is necessary (eg. try multiple subsets of features or Pearson's correlation).

2) model design: Select an appropriate DL model based on data features and given tasks (usually try several models), then decide which validation method to use (train-test validation, cross-validation, and external validation) to divide the dataset into train set and test set, considering data leakage issues, and a validation set is added to help to adjust parameters in the model training process.

3) training model: Input train data into the neural network, which is composed of neurons (the unit of the network that receives inputs, perform a weighted summation of the inputs (weight and bias: used to update parameters during training), and then produce an output through a nonlinear function (activation function: eg. ReLU and Sigmoid). Generally, the Neural network has 3 layers, including the input layer (receive raw data), hidden layer (process data), and output layer (produce the final outcome). The model learns the features/structures from the former layer, and some parameters (hyperparameters) are not updated during training. The training goal is to minimize or optimize a predefined loss function.

4) evaluation and tuning: After training the model on the training set, use the validation set to evaluate the performance of the model, and adjust parameters/structures, if the training error is far greater than the validation error – overfitting, on the contrary, it is underfitting. Getting a Bias-Variance Tradeoff is important (Bias – error due to the limitations of models, Variance – error due to sensitivity and variation of data)

5) training and application: Use the test set to test the model performance, evaluation metrics include Accuracy, Sensitivity (Recall), Specificity, Precision, F1-Score (harmonic mean of the precision and recall), and ROC analysis.

2.2 DL applications(success) in four main disease tasks on genomics data

Adapted on [Kou20], **Table 1** shows the applications of widely used DL techniques on genomics data, including their inputs (omics data) and outputs (purpose/prediction).

From the inputs and outputs of **Table 1**, DL technologies have been applied to various types of genomics data, involving DNA, RNA, gene, protein, cell-line, and genome, used in different kinds of

Network type	Input	Output
RNN	miRNA-mRNA pairing expression of landmark genes DNA-seq	target prediction function of DNA Gene expression inference
LSTM	positive pre-miRNA non-miRNA histone modifications	miRNA target classify gene expression
ANN	RNA-Seq cell-line with drug response	control-cases drug response
AE	time-series gene expression cDNA microarrays gene expression mRNA and miRNA	pre-process for clustering organization of transcriptomic machinery identification/reconstruction of biological signals identification of biological patterns predict tissue-of-origin, normal or disease state and cancer type
CNN	histone modifications whole-genome sequence Single cell methylation scRNA-seq DNA-seq	classify gene expression variant caller quantitative epigenetic variation missing methylation states and detects sequence motifs transcription factor target

Table 1: 5 commonly used deep learning techniques in genetics (Adapted on [Kou20]).

tasks, which are prediction, classification, and drug response. Related to the clinical setting, I will explore 4 main disease tasks: disease prediction, disease classification, drug discovery, and precision medicine.

Applications on disease prediction: Disease prediction on genomics data is used to identify the risk of contracting a disease before birth or early in life, then diseases can be addressed early, and scarce resources can be allocated efficiently. DL methods can 1) identify patterns and variations in DNA sequences, to understand how specific genetic variations affect disease risk [APPS16]. 2) integrate information from multiple bioinformatics data sources, such as gene expression data, single nucleotide polymorphisms (SNPs), copy number variations (CNVs), and epigenetic data to provide more comprehensive disease predictions and improve prediction accuracy [MYF+18]. 3) predict the three-dimensional structure of proteins to understand how disease-related proteins change their functions due to genetic variation, thus predicting disease risk through these changes [AlQ19, SEJ+20]. DL models can be more explainable through methods such as LIME [MMT20, GBLMH21] and DeepLIFT [FFB20]. These studies underline DL's potential to supplement and, in some cases, outperform traditional methods of disease prediction, offering non-invasive, efficient, and potentially more accurate diagnostic tools.

Applications on disease classification: The main task of disease classification on genomics data involves analyzing genetic information to identify patterns, mutations, or genetic markers associated with specific diseases. DL models help to detect genetic sequence variants and classification, and the model DeepVariant [PCA+18] is a promising tool to identify variations in genomics sequence data. In specific applications, DL models 1) classify different types of cancer based on gene expression data to help identify cancer subtypes, like DeepGx [dGDL19], 2) identify variants associated with specific genetic diseases to enable rapid and accurate diagnosis, like SNPs [JNB+22], 3) based on transcriptome and gene expression data, help distinguish different categories of psychiatric disorders [LQM+22, ANT+22]. AI and DL models in disease classification have shown promising advancements in medical diagnostics and research. Studies have explored genetic disorders, cancers, and neurodegenerative conditions, demonstrating DL's potential to enhance early detection, predict disease progression, and facilitate personalized medicine.

Applications on drug discovery: Figure 3 shows machine learning applications in the drug discovery pipeline and their required data characteristics (Source: [VCC+19]). By leveraging the vast and detailed information encoded in the genome to develop new therapeutic drugs and identify novel drug targets, the main DL applications on the task include 1) identifying potential drug candidate molecules to guide drug design and screening, 2) identifying key biomarkers and potential drug targets associated with specific diseases to develop targeted treatment strategies. For the first application, the latest DL models screen for candidate drug molecules from large-scale compound libraries and

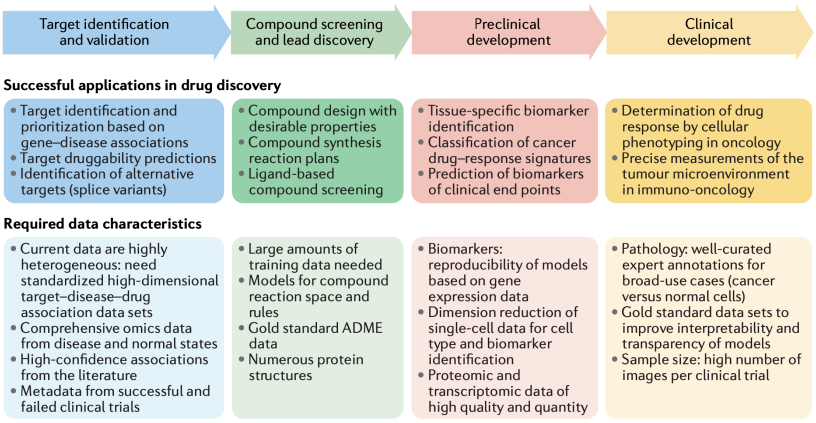

Figure 3: Machine learning applications in the drug discovery pipeline and their required data characteristics (Source: [VCC+19]).

predict Drug-Target Interaction (DTI), such as [YMH19, XHS+18]. For the second application, DL models can identify biomarkers of depression [LKL+18] and cancer [ERB+21] associated with disease progression, drug responsiveness, or efficacy from complex genomic data.

Applications on precision medicine (also known as personalized medicine): Multi-omic data are often integrated with medical history, social/behavioral determinants, and environmental knowledge to improve personalized care, the main applications are precise diagnoses, disease risk prediction, customized treatment plans design [JWW+21]. Specifically, DL methods 1) distinguish between benign and pathogenic genetic variants, helping in the accurate diagnosis of genetic conditions [DT19], which is important for diseases that have a genetic basis but present with non-specific symptoms, like PrimateAI [SGP+18]. 2) identify genetic variations and patterns associated with specific diseases to predict an individual's susceptibility to certain diseases, allowing for early intervention and preventive measures [ZTY+18]. 3) [ZBQL18] provides a list of research papers on therapeutic effects (treatment plans prediction and subgroup division), which includes susceptibility, occurrence, recurrence, survivability, and phenotype, to help design personalized treatment plans.

Survival prediction is one of the key components of precision medicine, which is valuable to doctors and patients because it can help doctors choose the appropriate treatment for their patients, especially in the fight against cancer. Survival prediction can guide therapeutic decision-making, optimize treatment plans, and improve patients' survival rates. For example, by analyzing the genetic data of a cancer patient, it's possible to predict the patient's response to certain chemotherapeutic drugs [MKS21], so an effective treatment plan can be selected.

As shown in **Table 2**, 5 specific applications of deep learning methods in survival prediction on genomics data are addressed chronologically for comparison, to gain a deeper understanding of the successes and restrictions of the current technologies.

model	task	dataset	DL network	performance
[SWCD20]	10-year survival prediction	DNA samples (GWAS) age-related macular degeneration	DNN	AUC (81.8%)
[KKK+21]	survival prediction	RNA-Seq (TCGA) (oral cancer)	DNN	AUC (97.2%)
[MKS21]	survival prediction, drug response prediction	multi-omics data (TCGA, GDSC) (breast cancer)	NCA	AUC (94%)
[WZLL22]	Long-Term survival prediction	gene expression data (TCGA) (lung cancer)	CNN	AUC (71.48%)
[SC23]	survival status classification, survival time prediction	transcriptome data (TCGA) (33 cancer types)	Auto Encoder	AUC (96%)

Table 2: 5 specific applications of deep learning methods in survival prediction on genomics data.

Compared with purely traditional statistics methods and ML algorithms (like SVM, KNN, etc.),

the DL models perform better in all tasks from **Table 2**, proving the efficiency of DL applications on survival prediction. We can see that DL methods applied to different cancer types are comparatively robust (still some fluctuations), and the prediction task is becoming more fine-grained and deeper (from basic binary survival state prediction to drug response prediction and survival time prediction), however, external validation is not used in all models, making it harder to prove the flexibility of these methods. It is worth noting that not all methods use deep learning as a model for prediction, [MKS21] uses DL to assist multi-omics integration, [KKK+21] also uses random forest and decision tree (ML models) to train the classifier, so DL is not only used to predict but also to pre-process data for statistical methods.

2.3 Limitations of DL applications in disease tasks on genomics data

Data issues: 1) The need for large, annotated datasets for model training: although genomics data is abundant, obtaining high-quality, well-annotated datasets is challenging due to the time and expense involved in data collection and annotation, particularly for rare diseases with limited available data. 2) Data heterogeneity: Genomics data comes in various forms (e.g., DNA sequences, gene expression profiles, epigenetic modifications) and from different platforms, leading to significant heterogeneity, which requires complex preprocessing and integration strategies. 3) Data privacy: Genomics data is sensitive and personal (especially in precision medicine, personal information is integrated into genomics data), which needs to meet data privacy and ethical and legal standards. 4) data bias (imbalanced labels): There are far fewer samples for a specific disease than healthy samples in genomics data, leading to biases towards the majority class in model training, so DL models perform poorly in predicting rare diseases or underrepresented conditions, limiting their clinical utility.

Model issues: 1) Computational resource requirements: Training DL models on genomics data requires significant computational resources, including powerful GPUs and large memory capacities, and lack of Software libraries specific for genomics, which limits the development of smaller institutions and research groups. 2) DL model bias (overfitting): when trained on highly complex datasets, DL models are likely to learn noise in the training data instead of generalizable patterns, limiting their practical use in clinical settings. 3) Model robustness and flexibility: Genomic data from different populations or acquired through different technologies are quite different (Like the clarity of medical images and the integrity of data features), limiting the models' applicability across diverse clinical and research settings. 4) Interpretability of DL models: In many genomic applications, researchers are more interested in the biological mechanisms revealed by the predictive model rather than the prediction accuracy itself. Unlike statistics methods and traditional ML models, DL acts as "black boxes", making it difficult to understand how they make predictions, so it's hard for clinicians and patients to trust and adopt, restricting the clinical integration of DL-based tools.

3 Strategy Recommendation

3.1 A set of recommendations to solve the limitations

Addressing data issues: 1) Enhancing data collection and annotation: Encourage collaboration between research institutions, hospitals, and biobanks to share genomics data, and utilize crowdsourcing and automated annotation tools (like ANNOVAR, VEP)to reduce the time and cost of data annotation, especially for rare diseases. 2) Managing data heterogeneity: Promote data standards and formats to facilitate the integration of multi-omics and multi-platform data, implementing advanced data integration techniques, such as multi-view learning. 3) Ensuring data privacy: Encode the privacy-related data, and adopt federated learning approaches, which train models across multiple decentralized devices or servers holding local data samples without exchanging them. 4) Reduce imbalanced labels): Utilize oversampling for minority classes and undersampling for majority classes, or give more weight to minority classes in model training.

Addressing model issues: 1) Computational resource: The major cloud-computing platforms, such as Amazon EC2, Google Cloud Engine, Microsoft Azure, and IBM Cloud can offer more flexible on-demand GPU access (still need some configuration) [ZHA+19], and [AR22] summarises a table of deep learning libraries/packages specific to Genomics application. 2) Prevent model overfitting: Apply regularization techniques, such as dropout and L1/L2 regularization, to penalize model complexity,

and use cross-validation/external-validation techniques to evaluate model performance, reducing the likelihood of overfitting. 3) Model robustness and flexibility: Consider using transfer learning, which trains a model on one task and repurposes it for a second related task, for genomics data, a pre-trained model on large-scale genomic datasets can be used to fine-tune a specific task, such as disease prediction for a particular condition, thus enhancing model flexibility and robustness [LN15]. 4) Improving interpretability of DL models: Techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) can provide insights into the contribution of each feature to the model's prediction. Visualizing and interacting with model states can explore how different genetic inputs influence predictions, and conduct ablation experiments to study the impact of each feature on the prediction/classification results, to help explain the model.

3.2 Challenges to implement these recommendations

1) Establishing effective collaboration across institutions/research groups may be hindered by competitive interests, data ownership concerns, and differing data standards in data collection. 2) Integrating diverse multi-omics data types and formats requires sophisticated computational methods and can introduce biases if not handled correctly on data heterogeneity. 3) Implementing federated learning and privacy-preserving techniques will affect data utility and model performance for data privacy. 4) DL models may have cold start problems in rare diseases, so effectively addressing imbalanced datasets without introducing bias toward minority classes will be trouble for imbalanced data. 5) Although there are some technologies to interpret DL models, they're not enough to provide obvious evidence to support the decision-making, especially in complex genomics data for interpretability. 6) Access to and the cost of computational resources can be prohibitive for some researchers and institutions for data scalability.

3.3 Conclusion

In this report, I describe why DL models can help disease tasks on genomics data, what deep learning is, and how DL models apply to genomics data, especially comparing DL applications on survival prediction, then discuss the DL limitations, and give recommendations to solve these issues and further challenges.

From the result of the investigation, I think DL models have played a role in some disease tasks based on genomics data, they are not mature enough to deliver real benefits in a clinical setting, as the robustness, scalability, and interpretability of models still have big challenges. Moreover, it can't be used to make the final decision because of the morality issues (who should be responsible for a wrong decision made by AI?).

References

[AlQ19] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell systems*, 8(4):292–301, 2019.

[ANT+22] Rosa Lundbye Allesøe, Ron Nudel, Wesley K Thompson, Yunpeng Wang, Merete Nordentoft, Anders D Børglum, David M Hougaard, Thomas Werge, Simon Rasmussen, and Michael Eriksen Benros. Deep learning–based integration of genetics with registry data for stratification of schizophrenia and depression. *Science advances*, 8(26):eabi7293, 2022.

[APPS16] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[AR22] Wardah S Alharbi and Mamoon Rashid. A review of deep learning applications in human genomics using next-generation sequencing data. *Human Genomics*, 16(1):1–20, 2022.

[CCWV18] Fabrizio Celesti, Antonio Celesti, Jiafu Wan, and Massimo Villari. Why deep learning is changing the way to approach ngs data processing: a review. *IEEE reviews in biomedical engineering*, 11:68–76, 2018.

[dGDL19] Joseph M de Guia, Madhavi Devaraj, and Carson K Leung. Deepgx: deep learning using gene expression for cancer classification. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 913–920, 2019.

[DT19] Raquel Dias and Ali Torkamani. Artificial intelligence in clinical and genomic diagnostics. *Genome medicine*, 11(1):1–12, 2019.

[ERB+21] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer*, 124(4):686–696, 2021.

[FFB20] Jelena Fiosina, Maksims Fiosins, and Stefan Bonn. Explainable deep learning for augmentation of small rna expression profiles. *Journal of Computational Biology*, 27(2):234–247, 2020.

[GBLMH21] Freddy Gabbay, Shirly Bar-Lev, Ofer Montano, and Noam Hadad. A lime-based explainable machine learning model for predicting the severity level of covid-19 diagnosed patients. *Applied Sciences*, 11(21):10417, 2021.

[GW13] Ayman Grada and Kate Weinbrecht. Next-generation sequencing: methodology and application. *Journal of Investigative Dermatology*, 133(8):1–4, 2013.

[JNB+22] Taeho Jo, Kwangsik Nho, Paula Bice, Andrew J Saykin, and Alzheimer's Disease Neuroimaging Initiative. Deep learning-based identification of genetic variants: application to alzheimer's disease classification. *Briefings in Bioinformatics*, 23(2):bbac022, 2022.

[JWW+21] Kevin B Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L Snowdon. Precision medicine, ai, and the future of personalized health care. *Clinical and translational science*, 14(1):86–93, 2021.

[JZH21] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

[KKK+21] Yeongjoo Kim, Ji Wan Kang, Junho Kang, Eun Jung Kwon, Mihyang Ha, Yoon Kyeong Kim, Hansong Lee, Je-Keun Rhee, and Yun Hak Kim. Novel deep learning-based survival prediction for oral cancer by analyzing tumor-infiltrating lymphocyte profiles through cibersort. *Oncoimmunology*, 10(1):1904573, 2021.

[Kou20] Lefteris Koumakis. Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal*, 18:1466–1473, 2020.

[LKL+18] Eugene Lin, Po-Hsiu Kuo, Yu-Li Liu, Younger W-Y Yu, Albert C Yang, and Shih-Jen Tsai. A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. *Frontiers in psychiatry*, 9:290, 2018.

[LN15] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

[LQM+22] Yichuan Liu, Hui-Qi Qu, Frank D Mentch, Jingchun Qu, Xiao Chang, Kenny Nguyen, Lifeng Tian, Joseph Glessner, Patrick MA Sleiman, and Hakon Hakonarson. Application of deep learning algorithm on whole genome sequencing data uncovers structural variants associated with multiple mental disorders in african american patients. *Molecular psychiatry*, 27(3):1469–1478, 2022.

[LSXJ18] Jia Li, Yujuan Si, Tao Xu, and Saibiao Jiang. Deep convolutional neural network based ecg classification system using information fusion and one-hot encoding techniques. *Mathematical problems in engineering*, 2018:1–10, 2018.

[MKS21] Vidhi Malik, Yogesh Kalakoti, and Durai Sundar. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *Bmc Genomics*, 22:1–11, 2021.

[MMT20] Pavan Rajkumar Magesh, Richard Delwin Myloth, and Rijo Jackson Tom. An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery. *Computers in Biology and Medicine*, 126:104041, 2020.

[MYF+18] Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature methods*, 15(4):290–298, 2018.

[PCA+18] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, et al. A universal snp and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10):983–987, 2018.

[SC23] Bo Sun and Liang Chen. Interpretable deep learning for improving cancer patient survival based on personal transcriptomes. *Scientific Reports*, 13(1):11344, 2023.

[Sch15] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[SEJ+20] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[SGP+18] Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F McRae, Yanjun Li, Jack A Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics*, 50(8):1161–1170, 2018.

[SWCD20] Tao Sun, Yue Wei, Wei Chen, and Ying Ding. Genome-wide association study-based deep learning for survival prediction. *Statistics in medicine*, 39(30):4605–4620, 2020.

[VCC+19] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

[WZLL22] Shuo Wang, Hao Zhang, Zhen Liu, and Yuanning Liu. A novel deep learning method to predict lung cancer long-term survival with biological knowledge incorporated gene expression images and clinical data. *Frontiers in Genetics*, 13:800853, 2022.

[XHS+18] Lingwei Xie, Song He, Xinyu Song, Xiaochen Bo, and Zhongnan Zhang. Deep learning-based transcriptome data classification for drug-target interaction prediction. *BMC genomics*, 19:93–102, 2018.

[YMH19] Jiaying You, Robert D McLeod, and Pingzhao Hu. Predicting drug-target interaction network using deep learning model. *Computational biology and chemistry*, 80:90–101, 2019.

[ZBQL18] Sushen Zhang, Seyed Mojtaba Hosseini Bamakan, Qiang Qu, and Sha Li. Learning for personalized medicine: a comprehensive review from a deep learning perspective. *IEEE reviews in biomedical engineering*, 12:194–208, 2018.

[ZHA+19] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nature genetics*, 51(1):12–18, 2019.

[ZTY+18] Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8):1171–1179, 2018.