REVIEW ARTICLE

AI IN MEDICINE

Jeffrey M. Drazen, M.D., *Editor*, Isaac S. Kohane, M.D., Ph.D., *Guest Editor*,
and Tze-Yun Leong, Ph.D., *Guest Editor*

Artificial Intelligence in Molecular Medicine

Bruna Gomes, M.D., and Euan A. Ashley, M.B., Ch.B., D.Phil.

From the Departments of Medicine, Genetics, and Biomedical Data Science, Stanford University, Stanford, CA (B.G., E.A.A.); and the Department of Cardiology, Pneumology, and Angiology, Heidelberg University Hospital, Heidelberg, Germany (B.G.). Dr. Ashley can be contacted at euan@stanford.edu or at Stanford University, Falk Bldg., 870 Quarry Rd., Stanford, CA 94304.

N Engl J Med 2023;388:2456-65.
DOI: 10.1056/NEJMra2204787

New methods such as genomic sequencing and mass spectrometry have prompted dramatic increases in the amount of molecular data available to scientists and health care professionals seeking more refined diagnoses and increased therapeutic precision.[1] Although the largest advances have been made in genetic sequencing of DNA and RNA, medical applications of high-dimensional measurement of proteins and metabolites are increasing.

Analytic tools have been improved in parallel to match the volume, velocity, and variety of these molecular "big data." The emergence of machine learning has proved especially valuable. In these approaches, computer systems use large amounts of data to build predictive statistical models that are iteratively improved by incorporating new data. Deep learning, a powerful subset of machine learning that includes the use of deep neural networks, has had high-profile applications in image object recognition,[2] voice recognition, autonomous driving, and virtual assistance. These approaches are now being applied in medicine to yield clinically directive medical information. In this review article, we briefly describe the methods used to generate high-dimensional molecular data and then focus on the key role that machine learning plays in the clinical application of such data.

MOLECULAR DATA GENERATION AT SCALE

A major change in our ability to measure molecules at scale has fueled the current era of individualized medicine (Fig. 1). For decades, genetic sequencing based on the technique of Sanger focused on sections of DNA or RNA that were up to a few hundred bases in length. In the early 2000s, approaches such as sequencing by synthesis (Illumina) gained traction, allowing hundreds and eventually billions of short DNA templates to be synthesized and read simultaneously. More recent methods (from Pacific Biosciences and Oxford Nanopore), which have focused on continuous sequencing of long nucleic acid molecules, have additional benefits. Whereas the Human Genome Project took 10 years to sequence one incomplete monoploid genome at a cost of several billion dollars, in 2022, a more complete human genome[3] could be sequenced in 5 hours for just a few hundred dollars.[4] This rapid acceleration in the availability of genomic data has created demand for fast processing and accurate interpretation of these data.

The process of genomic sequencing results in a computer text file in which each line represents an individually "read" molecule of DNA or RNA. For genomic sequencing, the aim is typically to generate sufficient overlapping data to cover each part of the genome 40 times. Some types of technology capture a subset of the genome and cover it many more times. This output text file is 100 to 200 gigabytes in size (similar to the hard-disk capacity of an entry-level laptop today). The reads, ranging from a few hundred to several million bases in length, are mapped to the reference genome generated by the Human Genome Project by means of the Burrows–Wheeler transform, a method derived from data-compression information theory.[5] Machine-learning or algorithmic approaches are then used to determine where the genome being analyzed differs from the reference sequence. This results in a variant call file — a text file typically 3 million to 4 million lines in length and a few megabytes in size. To prioritize the variants in the file according to, for example, their likelihood of causing a rare disease in a patient, filtering or machine-learning approaches are used.[6] For RNA sequencing, after mapping, most applications focus on quantification of gene or isoform expression rather than on sequence identity, converting read counts per gene or isoform to a standardized quantitative measure.

In contrast, mass spectrometry is the workhorse of proteomics (the study of all proteins in a cell) and metabolomics (the study of chemical processes involving metabolites in cell metabolism). It generates ions by bombarding organic or inorganic compounds with electrons and separating the resulting positively charged fragments by their mass-to-charge ratio. The first stage of mass spectrometry often involves a separation phase such as liquid chromatography, followed by a spectrometry phase. The output is in the form of spectral plots of ion signal as a function of mass-to-charge ratio. These output plots usually represent superimposed signals from at least hundreds of chemical entities that need to be decomposed into individual signals, mapped by reference to large databases of spectra from known chemical entities, and then further processed. This processing might include, for example, reassembling peptide fragments into full-length proteins.

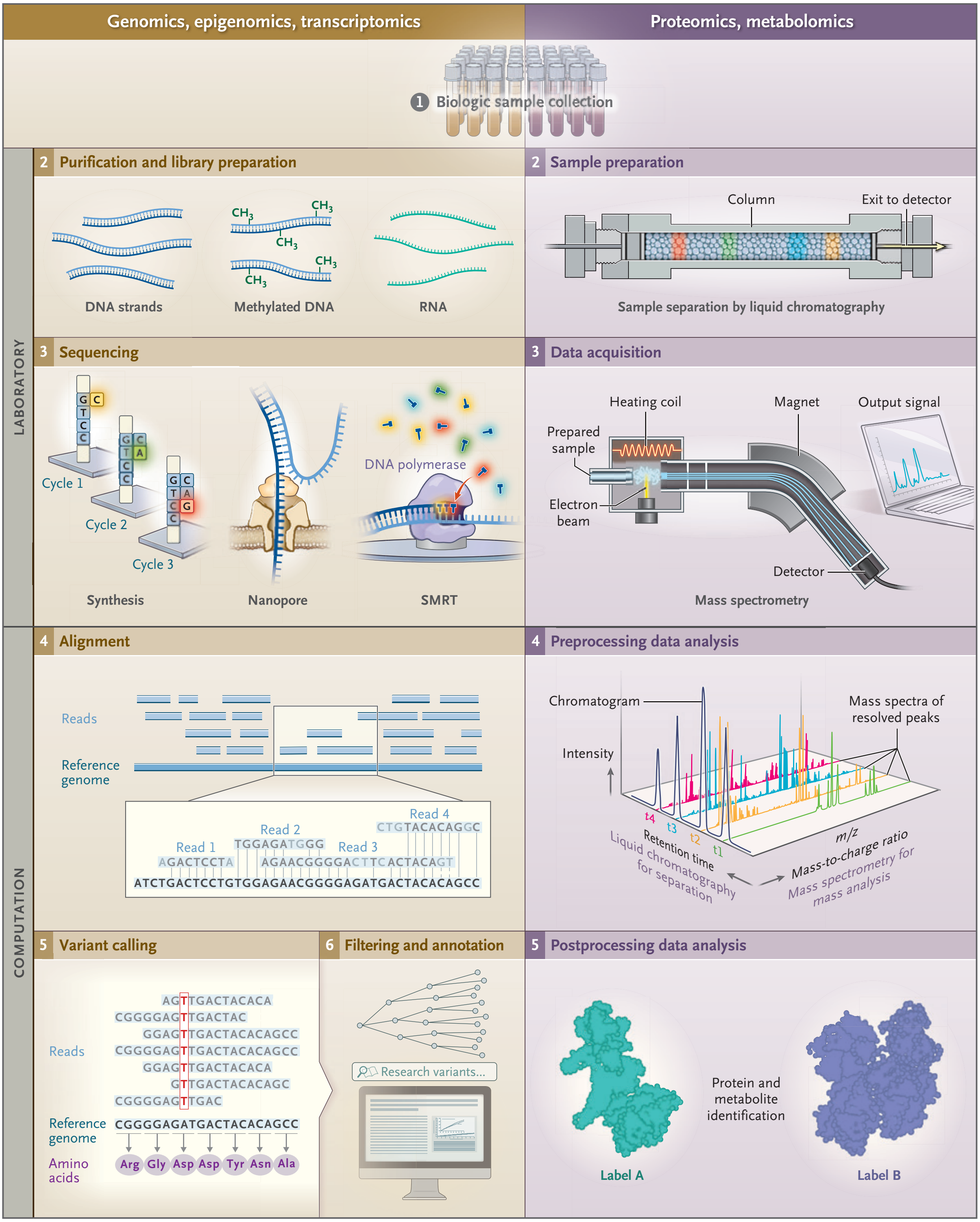
MACHINE-LEARNING APPLICATIONS IN GENOMICS

The most important advances in the application of machine learning to genomics (the study of the set of genes [the genome] in a cell) have occurred in variant calling: the process of determining where the analyte sequence (e.g., a sample from a patient) varies from the reference sequence. As individual reads are mapped to the corresponding position in the reference genome, they can be visualized as a "pile up" in which bases distinct from the reference are highlighted (Fig. 1). This visual representation facilitates rapid manual review in complex areas of the genome, an insight that led to the development of a deep-learning approach to variant identification that draws on advances in computer vision and image recognition.[7] Other approaches to variant calling use machine learning in narrower applications, such as for the technical calibration of error profiles for specific variants or regions of the genome.[8]

Deep neural networks[9] are complex, nonlinear functions fit to large data sets. Multiple layers of alternating "neuron" weights and nonlinearities transform the data into abstract and lower-dimensional representations that are useful for classification. Layers are connected through an activation function, which acts as a gatekeeper to the further propagation of individual outputs. In image tasks, pooling functions are used to downscale inputs over specific areas. Neuron weights are then refined through a process known as back-propagation, and a final classification is usually in the form of a confidence estimate for each of several output options. Convolutional neural networks are a specific form of deep neural networks, often used for image recognition, that are characterized by the process of sliding filters over the image input (Figs. 2 and 3).

With the power of neural networks and the ability to read much longer molecules of DNA, a new era in haplotyping (the mapping of DNA strands to the parental chromosome of origin) could emerge. The haplotyping approach improves the quality of variant calling by better representing the originating DNA molecules and can inform clinical management — for example, in the case of compound heterozygosity, when identification of the parent of origin of two variants at the same locus can affect patient care. Recently, unprecedented accuracy has been achieved with the use of an approach to variant calling that combines haplotyping with models optimized for sequential data, followed by the convolutional neural network approach described above (https://github.com/google/deepvariant/releases/tag/v1.5.0).[17]

The improvement in variant calling resulting from these advances has been facilitated by the National Institute of Standards and Technology through its Genome in a Bottle Consortium and by the Food and Drug Administration (FDA) through the precisionFDA initiative. Together, these groups run open "Truth Challenge" competitions with standardized samples. Results reveal continuing improvements in the accuracy of variant calling genomewide and specifically in challenging areas of the genome, such as regions encoding the major histocompatibility complex on chromosome 6. F1 accuracy scores, which



take both false positive and false negative results into account, range between 0, for the poorest outcome, and 1.0, for the best outcome. Scores higher than 0.998 have now been reported for the three most commonly applied forms of sequencing technology (with both short- and long-

Figure 1 (facing page). Data Processing for Molecular Profiling.

From tissue-sample collection to accurate clinical diagnosis, complex laboratory and computational pipelines are required to generate and analyze data with the use of new measurement techniques. Initial workflow steps commonly include sample collection and library preparation. DNA and RNA sequencing are most commonly completed through synthesis sequencing (Illumina), nanopore sequencing (Oxford Nanopore), or single-molecule real-time sequencing (SMRT, Pacific Biosciences). Each method produces output in the form of raw data that vary according to the nucleotide of focus. Computational analysis converts those raw signals into bases (A, T/U, G, C) that are then output as a text file of short or long DNA or RNA molecules. Alignment of these individual "reads" to a genome of reference is then performed, and variants are called or gene expression is quantified. For mass spectrometry applications, output is in the form of ion spectra that are mapped to known chemical entities. Further important downstream analysis includes three-dimensional structure and function predictions.

read sequencing) across all benchmark regions, including single-nucleotide and small insertion–deletion variants.[18]

Machine learning (Fig. 2) has also proved to be extremely useful in the prioritization of variants for rare disease.[19] For example, one approach used logistic regression–based machine learning within a large literature-derived data set to match phenotypes to candidate genes in order to help identify potentially causal genes for mendelian disease.[20] Another approach applies maximum likelihood estimation (an iterative method for estimating the parameters of a model) and Bayesian networks (probabilistic graphical models) to achieve the same end.[21]

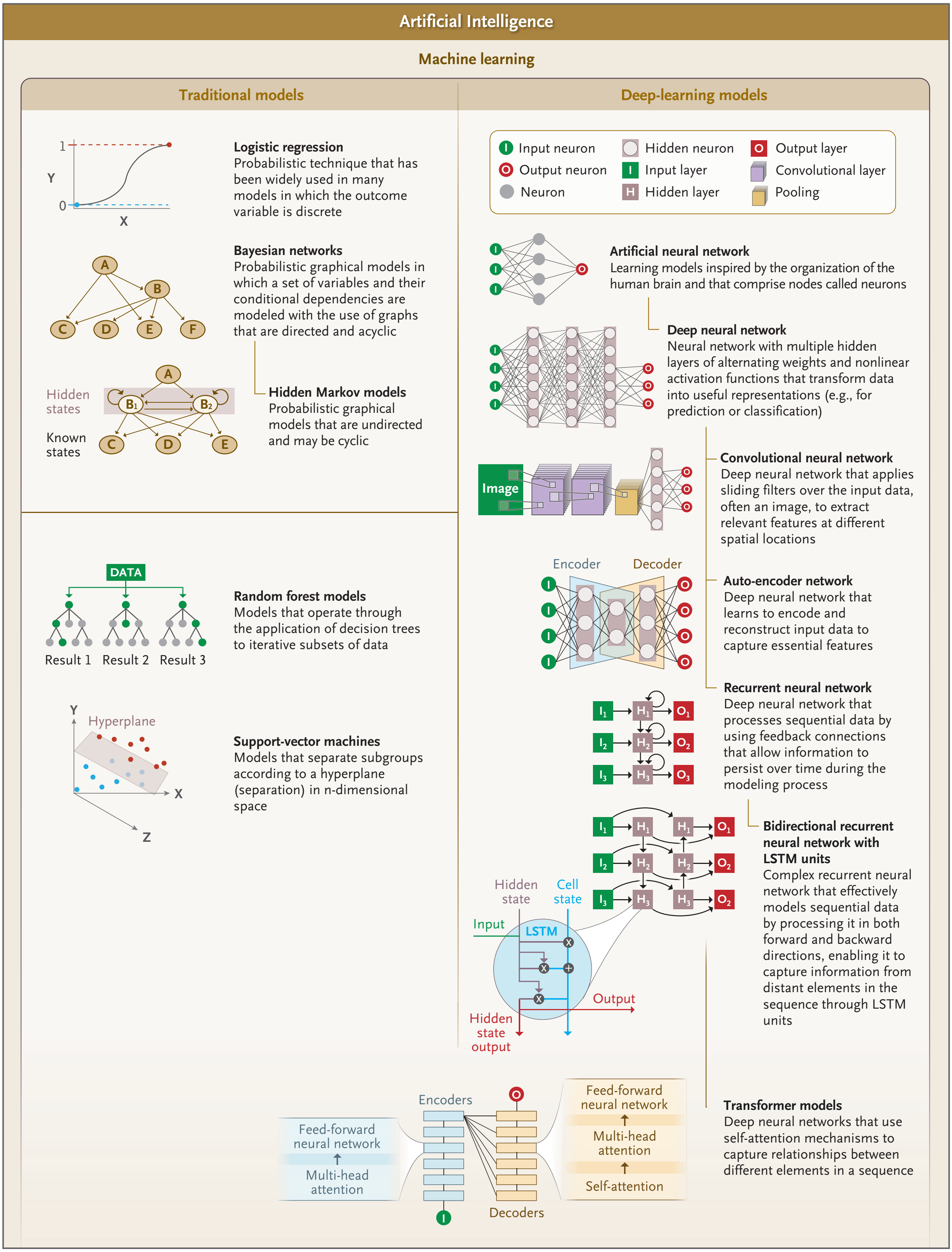
Application of all these approaches has been particularly successful in identifying rare inherited diseases, with multiple studies showing solve rates of 30 to 50% for undiagnosed genetic disease.[22,23] Using a variety of methods, the Undiagnosed Diseases Network reported a solve rate of 35% for patients in a study, one third of whom had already undergone some form of genomic sequencing.[22] In addition, in a recent study involving 13,449 probands from the United Kingdom and Ireland, the diagnosis rate was 41%.[24] Also, our team found that clinical application of nanopore long-read sequencing not only improved accuracy as compared with prior approaches[25] but also had the potential to make clinical diagnoses in a rare-disease context in less than 8 hours.[4]

In one case involving a 6-month-old infant with a seizure disorder, blood was drawn, and high-molecular-weight DNA was derived from white cells. As single DNA strands traversed hundreds of thousands of nanopore proteins, raw data representing the change in real-time current were fed to a recurrent neural network, which identified the underlying genetic sequence. These sequences were mapped to the reference sequence with the use of the Burrows–Wheeler transform. A recurrent neural network combined with a hidden Markov model then separated the parent of origin of individual DNA molecules and identified areas where the infant's genome differed from the reference sequence. Finally, a convolutional neural network processed the pile-up images, resulting in the output of 4,503,667 variant calls. A filtration scheme prioritized 29 small variants and 20 structural variants for manual review, and within 8 hours after the infant's blood had been drawn, a heterozygous truncating variant in *PCDH19* was identified as causal, establishing the diagnosis and allowing the bedside team to direct patient care according to the molecular cause of the seizures.

MACHINE LEARNING FOR TRANSCRIPTOMICS

Reading of the transcriptome (the sum of all the RNA transcripts in an organism) is being used as an additional tool to identify causal genes in rare diseases.[26] Initial efforts revealed that identifying expression outliers by comparing the expression profile of every gene with a reference range could point to otherwise unsuspected causal genes.[27] For an additive benefit, this approach was later combined with Bayesian models that predict regulatory effects for rare variants.[28] In a large cohort of patients with undiagnosed rare diseases, blood transcriptome sequencing identified causal variants in 8% of the patients.[29] Subsequently, a hierarchical Bayesian model incorporating gene expression, allele-specific expression, and alternative splicing data was used to identify genetically driven transcriptome abnormalities.[30]

Despite this progress, predicting splice junctions remains a challenging problem. One deep-learning model using a 32-layer deep neural network showed promise in improving the diagnosis of rare diseases.[31] Use of an autoencoder, a neural network that efficiently learns how to encode input data to a compressed representa-



Artificial Intelligence

Machine learning

Traditional models

Logistic regression
Probabilistic technique that has been widely used in many models in which the outcome variable is discrete

Bayesian networks
Probabilistic graphical models in which a set of variables and their conditional dependencies are modeled with the use of graphs that are directed and acyclic

Hidden Markov models
Probabilistic graphical models that are undirected and may be cyclic

Random forest models
Models that operate through the application of decision trees to iterative subsets of data

Support-vector machines
Models that separate subgroups according to a hyperplane (separation) in n-dimensional space

Deep-learning models

Artificial neural network
Learning models inspired by the organization of the human brain and that comprise nodes called neurons

Deep neural network
Neural network with multiple hidden layers of alternating weights and nonlinear activation functions that transform data into useful representations (e.g., for prediction or classification)

Convolutional neural network
Deep neural network that applies sliding filters over the input data, often an image, to extract relevant features at different spatial locations

Auto-encoder network
Deep neural network that learns to encode and reconstruct input data to capture essential features

Recurrent neural network
Deep neural network that processes sequential data by using feedback connections that allow information to persist over time during the modeling process

Bidirectional recurrent neural network with LSTM units
Complex recurrent neural network that effectively models sequential data by processing it in both forward and backward directions, enabling it to capture information from distant elements in the sequence through LSTM units

Transformer models
Deep neural networks that use self-attention mechanisms to capture relationships between different elements in a sequence

Figure 2 (facing page). Machine Learning for Biomedical Applications.

Commonly used machine-learning approaches are shown. In random forest models, decision trees are applied to iterative subsets of the data. Support-vector machines separate subgroups according to a hyperplane (separation) in n-dimensional space. Neural networks, learning models inspired by the organization of the human brain, comprise nodes called neurons. Deep learning has been particularly effective in image recognition, especially through convolutional neural networks. For sequential data, recurrent neural networks (including long short-term memory [LSTM] units) allow the network to retain memory of previous input data. Since the networks perform less well with longer input, scientists have increasingly adopted transformer[10,11] and large language models[12] that ingest sequential data as a whole, prioritizing specific areas for "attention" and incorporating weights from input data found both earlier and later in the data sequence. In a transformer model, the self-attention component assigns weights to the elements in a sequence in order to define their importance within the context. Multi-head attention builds on this idea by incorporating multiple sets of attention to capture diverse patterns and aspects of the sequence.

tion before decoding it back to a representation of the original input, has been shown to improve aberrant splicing prediction from RNA sequencing data (Fig. 2).[13]

These approaches were used in the case of a 12-year-old girl with developmental regression, tremors, and seizures. Short-read genomic sequencing identified 96 candidate gene variants, none of which appeared to be responsible for the patient's condition. The addition of a splice-outlier algorithm based on RNA sequencing of the patient's blood identified a splice-gain variant in *KCTD7*, which was not in the original list, establishing the diagnosis of progressive myoclonus epilepsy.

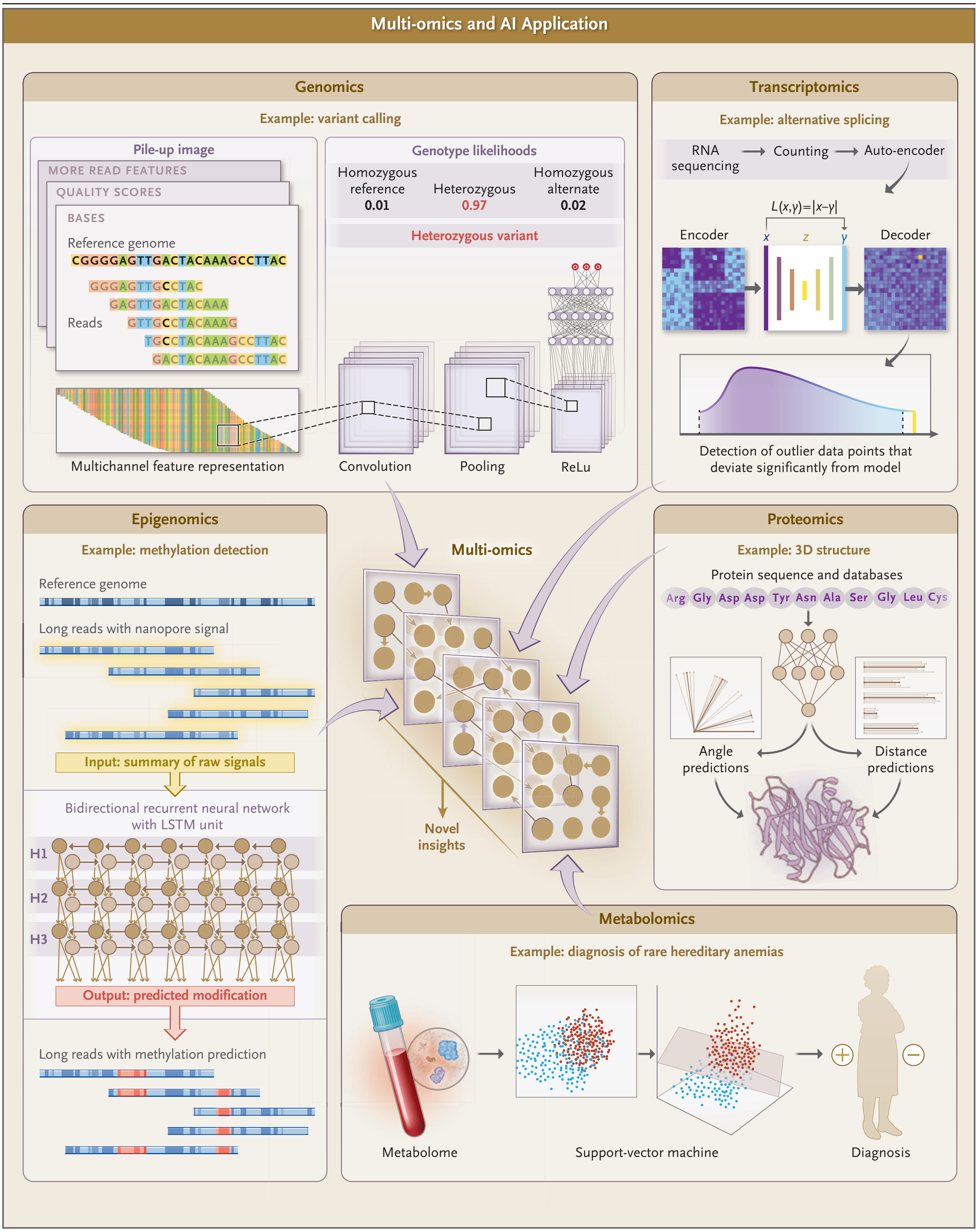
EPIGENOMIC APPLICATIONS

Epigenomics is defined as a complete set of modifications that influence gene expression. Although epigenetic mechanisms are known to play a role in certain rare and common disease presentations, characterization of chemical modifications of DNA at scale is only beginning to have an impact on clinical medicine. Long-read sequencing methods present an exciting opportunity, since they produce signals when the nucleotide passes through a protein nanopore[32] or as a base is incorporated by DNA polymerase.[33] These signals can be interpreted by machine-learning methods to call not just the nucleotide at that site but any of a series of chemical modifications of that nucleotide. These approaches do not require the bisulfite conversion of the previous standard, which has been shown to cause DNA fragmentation.[34] Because of a critical role in tissue-specific transcription, most attention has focused on the addition of a methyl group to the C5 position of cytosine residues in sequential CG dinucleotide sequences called CpG sites. Approaches involving the use of a range of neural networks, including convolutional neural networks,[35] bidirectional recurrent neural networks (Fig. 3),[14] and a combination of the two types,[36] have achieved a C statistic of more than 0.95 for the detection of methylation, outperforming previous benchmark models.[36]

MACHINE LEARNING FOR PROTEOMICS

Deep learning has facilitated substantial progress in almost all parts of the proteomics workflow.[37] With training on patterns of spectral plots from known chemical entities, deep-learning approaches have improved the prediction of spectra of candidate peptides,[38-40] which is a pivotal step for tandem mass spectrometry–based proteomics. With the use of a bidirectional, long short-term memory, deep-learning approach, in which a neural network propagates information sequentially forward and backward across an input signal, one algorithm[38] predicts peptides with a Pearson correlation coefficient greater than 0.9, surpassing the previous machine learning–based standard of 0.85.[41] Peptide retention time, which is the point in time when a peptide is eluted from the liquid chromatography column, has also been predicted accurately with the use of convolutional neural network–based tools.[42] Apart from mass spectrometry, de novo peptide sequencing[43] and protein identification have been the focus of deep-learning applications that use both convolutional neural networks and long short-term memory approaches. One tool outperformed benchmark peptide-sequencing tools, with a C statistic that was 33% higher than a previous standard,[44] and another transformer-based tool showed sequence coverage of 97.7 to 99.5%.[10,11] Moreover, large language models have recently been applied to protein-function prediction, with the aim of accelerating drug discovery.[12]

Post-translational modification of proteins in



processes such as phosphorylation is critical for protein function, regulation, and degradation,[45] but quantification remains an unsolved challenge.

Deep-learning prediction of post-translational modification sites from a protein sequence alone has been successful, with examples including

Figure 3 (facing page). Applications of Machine Learning to Omics Data.

Variant calling can be regarded as an image-classification problem (https://github.com/google/deepvariant/releases/tag/v1.5.0), whereas alternative splicing can be predicted through an auto-encoder network.[13] In the example of variant calling, the sequence data, quality scores, and other read features are encoded into a multichannel feature representation. This feature representation is then fed into a convolutional neural network to calculate genotype likelihoods for three genotype states: homozygous reference, heterozygous, or homozygous alternate. In the example shown, a heterozygous variant is identified as the most probable genotype. Prediction of methylation has benefited from bidirectional recurrent neural networks.[14] Deep-learning applications are increasing the accuracy of predictions of three-dimensional (3D) protein structures.[15] Support-vector machines in untargeted metabolomics have shown promise in the diagnosis of rare hereditary anemias.[16] Applications of learning models to combined multi-omic inputs represent the next frontier in the pursuit of precision medicine. AI denotes artificial intelligence, and ReLu rectified linear unit.

acetylation[46] and ubiquitination.[47] Predicting protein function from a peptide sequence has also recently been improved with a combination of machine-learning approaches — namely, hidden Markov models and an ensemble of convolutional neural networks.[48] The combined approaches contributed functional predictions for 360 previously unannotated human reference proteome proteins, expanding coverage of the standard protein family database by more than 9%.

In a high-profile application of deep learning to proteomics, neural network–based AlphaFold[49] (Fig. 3) won the 13th and 14th Critical Assessment of Protein Structure Prediction competitions (specifically, AlphaFold1 won the CASP13 competition and AlphaFold2 won the CASP14 competition). These were biennial, blinded competitions to benchmark progress in protein structure prediction. In the 13th competition, AlphaFold1 created high-accuracy structures for 24 of 43 free modeling domains, greatly surpassing both previous approaches and the next best method, which achieved similar accuracy for only 14 of 43 domains.[50] In the CASP14 competition, AlphaFold2 built on this progress, outperforming many competing models.[15]

The prediction of biomarkers has been a principal clinical focus for proteomics in recent years. Research has been directed at both single-marker and multimarker discovery. In one study, in which protein quantification was achieved with the use of a panel of aptamers (oligonucleotides that bind to proteins), a series of machine-learning models, including logistic regression–based models and random forests (Figs. 2 and 3), were trained to predict 11 different indicators of health that are commonly used for preventive medicine (e.g., the 5-year risk of a primary cardiovascular event) in a panel of approximately 17,000 persons with no major illnesses, from five independent cohort studies.[51] Quantification of 94 proteins predicted liver fat with a C statistic of 0.83 in a validation cohort,[51] suggesting potential near-term application for noninvasive detection of nonalcoholic fatty liver disease. A machine learning–assisted proteomics approach has also identified circulating biomarkers for alcoholic liver disease, Alzheimer's disease, and Parkinson's disease.[52]

APPLICATIONS FOR METABOLOMICS

Metabolomics focuses on the dynamics of the entire set of small molecules of an organism.[53] As compared with proteomics, which is focused on the protein complement, metabolomics includes measurements of fatty acids, lipids, organic acids, amino acids, steroids, and carbohydrates. One of the central clinical applications of metabolomics is the diagnosis of inborn errors of metabolism. Classically, the quantification of specific classes of metabolites such as purines and amino acids is undertaken with the use of individual assays, with the main limitation being a priori assumptions regarding the potentially affected pathways. Mass spectrometry–based metabolomics, in contrast, can be combined with genomic sequencing as an untargeted strategy to address the low diagnostic rate among patients presenting with typical signs of inborn errors of metabolism but with negative results of standard screening. Untargeted metabolomics led to an increase in the diagnostic rate by a factor of 6 in one cross-sectional analysis[54] and has been shown to be a useful strategy in targeting deficiencies in the nonoxidative pentose phosphate pathway.[55] In a recent study, exome sequencing combined with metabolomics improved variant classification.[56] For example, a metabolic fingerprint approach established a diagnosis of pyruvate kinase deficiency with the use of a support vector machine, which identifies subgroups by finding a hyperplane in n-dimensional space

(Figs. 2 and 3).[16] In another example, variants in genes for metalloproteins provided the training data for a multichannel convolutional neural network, which showed that mutations in the iron-binding site of metalloproteins were more closely associated with metabolic diseases than were mutations at other locations.[57]

MULTI-OMIC APPLICATIONS

As high-dimensional data from multiple types of technology are more readily available, computational approaches to combining data become more important. One of the earliest examples of a multi-omic study (i.e., an approach integrating multiple "omes" data types such as the genome or proteome) was a longitudinal analysis involving a single person that combined genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles.[58] Others have since used a multi-omic approach to build a correlation network reflecting health and disease states and by so doing have proposed novel biomarkers for cardiometabolic disease.[59] Other integrative approaches making use of deep learning have also been reported. These approaches either fuse the data early, concatenating omics data and then performing a single analysis, or fuse the data later, creating a joint model that combines output from several single omic analyses.[60] Some multi-omic approaches have proved successful in the clinical arena, such as in the identification of leucine zipper transcription factor–like 1 (*LZTFL1*) as a candidate effector gene at a coronavirus disease 2019 (Covid-19) risk locus, with the use of previously published machine-learning models, including a neural network.[61] By suggesting that increased expression of *LZTFL1* might be associated with a worse outcome, this insight reveals novel candidate targets for the prevention and treatment of Covid-19. Novel biomarkers of the response to immunotherapy have also been revealed through analysis of genomic, transcriptomic, and immunomic response data in cancer with the use of a support-vector machine.[62]

CONCLUSIONS

Over the past decade, technological advances have greatly enhanced our ability to measure fundamental biologic processes at scale. The resulting volume of data has been met with machine-learning methods that are increasingly tuned for the analysis of multidimensional biologic data sets. The outcome is a progressively detailed understanding of the molecular trajectory of disease that is now finding application in clinical medicine, with the greatest progress having been made in the diagnosis, and in some cases treatment, of rare genetic diseases. Challenges remain, including data quality, data consistency, and clinician awareness. However, as single-omic discovery gives way to multi-omic application, standardization of pipelines, expansion of benchmark metrics, and acceleration in the speed and accuracy of data processing will ensure that the potential for a far-reaching impact on precision health care is realized.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

REFERENCES

1. Ashley EA, Butte AJ, Wheeler MT, et al. Clinical assessment incorporating a personal genome. Lancet 2010;375:1525-35.
2. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580:252-6.
3. Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. Science 2022;376:44-53.
4. Gorzynski JE, Goenka SD, Shafin K, et al. Ultrarapid nanopore genome sequencing in a critical care setting. N Engl J Med 2022;386:700-2.
5. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature 2011;475:493-6.
6. Lightbody G, Haberland V, Browne F, et al. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Brief Bioinform 2019;20:1795-811.
7. Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 2018;36:983-7.
8. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491-8.
9. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press, 2016.
10. Beslic D, Tscheuschner G, Renard BY, Weller MG, Muth T. Comprehensive evaluation of peptide de novo sequencing tools for monoclonal antibody assembly. Brief Bioinform 2023;24:bbac542.
11. Yilmaz M, Fondrie W, Bittremieux W, Oh S, Noble WS. *De novo* mass spectrometry peptide sequencing with a transformer model. Proceedings of the 39th International Conference on Machine Learning, 2022 (https://proceedings.mlr.press/v162/yilmaz22a/yilmaz22a.pdf).
12. Stern A. NVIDIA expands large language models to biology. Santa Clara, CA: NVIDIA, September 20, 2022 (https://blogs.nvidia.com/blog/2022/09/20/bionemo-large-language-models-drug-discovery/).
13. Mertes C, Scheller IF, Yépez VA, et al. Detection of aberrant splicing events in RNA-seq data using FRASER. Nat Commun 2021;12:529.
14. Liu Q, Fang L, Yu G, Wang D, Xiao C-L, Wang K. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nat Commun 2019;10:2449.
15. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure predic-

tion with AlphaFold. Nature 2021;596:583-9.

16. Van Dooijeweert B, Broeks MH, Verhoeven-Duif NM, et al. Untargeted metabolic profiling in dried blood spots identifies disease fingerprint for pyruvate kinase deficiency. Haematologica 2021;106:2720-5.

17. Shafin K, Pesout T, Chang P-C, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods 2021;18:1322-32.

18. Olson ND, Wagner J, McDaniel J, et al. PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. Cell Genom 2022;2:100129.

19. Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. Front Genet 2020;11:350.

20. Birgmeier J, Haeussler M, Deisseroth CA, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. Sci Transl Med 2020;12(544):eaau9113.

21. De La Vega FM, Chowdhury S, Moore B, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome Med 2021;13:153.

22. Splinter K, Adams DR, Bacino CA, et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. N Engl J Med 2018;379:2131-9.

23. Lee H, Deignan JL, Dorrani N, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA 2014;312:1880-7.

24. Wright CF, Campbell P, Eberhardt RY, et al. Genomic diagnosis of rare pediatric disease in the United Kingdom and Ireland. N Engl J Med 2023;388:1559-71.

25. Dewey FE, Grove ME, Pan C, et al. Clinical interpretation and implications of whole-genome sequencing. JAMA 2014;311:1035-45.

26. Park CY, Zhou J, Wong AK, et al. Genome-wide landscape of RNA-binding protein target site dysregulation reveals a major impact on psychiatric disorder risk. Nat Genet 2021;53:166-73.

27. Li X, Battle A, Karczewski KJ, et al. Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. Am J Hum Genet 2014;95:245-56.

28. Li X, Kim Y, Tsang EK, et al. The impact of rare variation on gene expression across tissues. Nature 2017;550:239-43.

29. Frésard L, Smail C, Ferraro NM, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med 2019;25:911-9.

30. Ferraro NM, Strober BJ, Einson J, et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. Science 2020;369(6509):eaaz5900.

31. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting splicing from primary sequence with deep learning. Cell 2019;176(6):535-548.e24.

32. Bowden R, Davies RW, Heger A, et al. Sequencing of human genomes with nanopore technology. Nat Commun 2019;10:1869.

33. PacBio (https://www.pacb.com/).

34. Feng S, Zhong Z, Wang M, Jacobsen SE. Efficient and accurate determination of genome-wide DNA methylation patterns in Arabidopsis thaliana with enzymatic methyl sequencing. Epigenetics Chromatin 2020;13:42.

35. Tse OYO, Jiang P, Cheng SH, et al. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. Proc Natl Acad Sci U S A 2021;118(5):e2019768118.

36. Ni P, Huang N, Zhang Z, et al. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Bioinformatics 2019;35:4586-95.

37. Wen B, Zeng W-F, Liao Y, et al. Deep learning in proteomics. Proteomics 2020;20(21-22):e1900335.

38. Zhou X-X, Zeng W-F, Chi H, et al. pDeep: predicting MS/MS spectra of peptides with deep learning. Anal Chem 2017;89:12690-7.

39. Zeng W-F, Zhou X-X, Zhou W-J, Chi H, Zhan J, He S-M. MS/MS spectrum prediction for modified peptides using pDeep2 trained by transfer learning. Anal Chem 2019;91:9724-31.

40. Liu K, Li S, Wang L, Ye Y, Tang H. Full-spectrum prediction of peptides tandem mass spectra using deep neural network. Anal Chem 2020;92:4275-83.

41. Li S, Arnold RJ, Tang H, Radivojac P. On the accuracy and limits of peptide fragmentation spectrum prediction. Anal Chem 2011;83:790-6.

42. Bouwmeester R, Gabriels R, Hulstaert N, Martens L, Degroeve S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. Nat Methods 2021;18:1363-9.

43. Sinitcyn P, Richards AL, Weatheritt RJ, et al. Global detection of human variants and isoforms by deep proteome sequencing. Nat Biotechnol 2023.

44. Tran NH, Zhang X, Xin L, Shan B, Li M. De novo peptide sequencing by deep learning. Proc Natl Acad Sci U S A 2017;114:8247-52.

45. Wang Y-C, Peterson SE, Loring JF. Protein post-translational modifications and regulation of pluripotency in human stem cells. Cell Res 2014;24:143-60.

46. Zhao X, Li J, Wang R, He F, Yue L, Yin M. General and species-specific lysine acetylation site prediction using a bi-modal deep architecture. IEEE 2018;6:63560-9.

47. Fu H, Yang Y, Wang X, Wang H, Xu Y. DeepUbi: a deep learning framework for prediction of ubiquitination sites in proteins. BMC Bioinformatics 2019;20:86.

48. Bileschi ML, Belanger D, Bryant DH, et al. Using deep learning to annotate the protein universe. Nat Biotechnol 2022;40:932-7.

49. Mancino DJ. Breakthrough to nursing timeline. Imprint 2010;57:35-41.

50. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020;577:706-10.

51. Williams SA, Kivimaki M, Langenberg C, et al. Plasma protein patterns as comprehensive indicators of health. Nat Med 2019;25:1851-7.

52. Mann M, Kumar C, Zeng W-F, Strauss MT. Artificial intelligence for proteomics and biomarker discovery. Cell Syst 2021;12:759-70.

53. Lindon JC, Nicholson JK, Holmes E. The handbook of metabonomics and metabolomics. New York: Elsevier, 2011.

54. Liu N, Xiao J, Gijavanekar C, et al. Comparison of untargeted metabolomic profiling vs traditional metabolic screening to identify inborn errors of metabolism. JAMA Netw Open 2021;4(7):e2114155.

55. Shayota BJ, Donti TR, Xiao J, et al. Untargeted metabolomics as an unbiased approach to the diagnosis of inborn errors of metabolism of the non-oxidative branch of the pentose phosphate pathway. Mol Genet Metab 2020;131:147-54.

56. Alaimo JT, Glinton KE, Liu N, et al. Integrated analysis of metabolomic profiling and exome data supplements sequence variant interpretation, classification, and diagnosis. Genet Med 2020;22:1560-6.

57. Koohi-Moghadam M, Wang H, Wang Y, et al. Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach. Nature Machine Intelligence 2019;1:561-7.

58. Chen R, Mias GI, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 2012;148:1293-307.

59. Price ND, Magis AT, Earls JC, et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat Biotechnol 2017;35:747-56.

60. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv 2021;49:107739.

61. Downes DJ, Cross AR, Hua P, et al. Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. Nat Genet 2021;53:1606-15.

62. Chen S, Lai H, Zhao J, et al. The viral expression and immune status in human cancers and insights into novel biomarkers of immunotherapy. BMC Cancer 2021;21:1183.