International Journal of
Molecular Sciences

MDPI

Review

Applications for Deep Learning in Epilepsy Genetic Research

Robert Zeibich [1], Patrick Kwan [1,2,3,4], Terence J. O'Brien [1,2,3,4], Piero Perucca [1,2,3,5,6], Zongyuan Ge [7,8] and Alison Anderson [1,4,*]

1 Department of Neuroscience, Central Clinical School, Monash University, Melbourne, VIC 3800, Australia; robert.zeibich@monash.edu (R.Z.); patrick.kwan@monash.edu (P.K.); terence.obrien@monash.edu (T.J.O.); piero.perucca@unimelb.edu.au (P.P.)
2 Department of Neurology, Alfred Health, Melbourne, VIC 3004, Australia
3 Department of Neurology, The Royal Melbourne Hospital, The University of Melbourne, Parkville, VIC 3052, Australia
4 Department of Medicine, The Royal Melbourne Hospital, The University of Melbourne, Parkville, VIC 3052, Australia
5 Epilepsy Research Centre, Department of Medicine, Austin Health, The University of Melbourne, Melbourne, VIC 3084, Australia
6 Bladin-Berkovic Comprehensive Epilepsy Program, Department of Neurology, Austin Health, The University of Melbourne, Melbourne, VIC 3084, Australia
7 Faculty of Engineering, Monash University, Melbourne, VIC 3800, Australia; zongyuan.ge@monash.edu
8 Monash-Airdoc Research, Monash University, Melbourne, VIC 3800, Australia
* Correspondence: alison.anderson1@monash.edu

Abstract: Epilepsy is a group of brain disorders characterised by an enduring predisposition to generate unprovoked seizures. Fuelled by advances in sequencing technologies and computational approaches, more than 900 genes have now been implicated in epilepsy. The development and optimisation of tools and methods for analysing the vast quantity of genomic data is a rapidly evolving area of research. Deep learning (DL) is a subset of machine learning (ML) that brings opportunity for novel investigative strategies that can be harnessed to gain new insights into the genomic risk of people with epilepsy. DL is being harnessed to address limitations in accuracy of long-read sequencing technologies, which improve on short-read methods. Tools that predict the functional consequence of genetic variation can represent breaking ground in addressing critical knowledge gaps, while methods that integrate independent but complimentary data enhance the predictive power of genetic data. We provide an overview of these DL tools and discuss how they may be applied to the analysis of genetic data for epilepsy research.

Keywords: machine learning; deep learning; genetic epilepsy; non-protein-coding; omics data integration

Citation: Zeibich, R.; Kwan, P.; J. O'Brien, T.; Perucca, P.; Ge, Z.; Anderson, A. Applications for Deep Learning in Epilepsy Genetic Research. *Int. J. Mol. Sci.* **2023**, *24*, 14645. https://doi.org/10.3390/ijms241914645

Academic Editor: Hans van Bokhoven

Received: 23 August 2023
Revised: 11 September 2023
Accepted: 21 September 2023
Published: 27 September 2023

1. Introduction

Affecting approximately one percent of the world's population, epilepsy is one of the most common neurological disorders. Rather than being a single homogenous condition, epilepsy encompasses a group of heterogeneous brain diseases characterised by an enduring predisposition to generate epileptic seizures; hence, the term 'epilepsies' is more appropriate [1–4]. The epilepsies are classified into one of four main types according to the associated seizure types: focal epilepsies, characterised by seizures originating within brain networks limited to one hemisphere (focal seizures); generalized epilepsies, where seizures involve both hemispheres (generalised seizures); combined focal and generalized epilepsies; epilepsies of unknown type for which there is insufficient information to classify them (seizures of unknown onset) [5,6]. Within these broad groups, there are specific epilepsy syndromes that are often supported by specific aetiological findings (structural, genetic, metabolic, immune, and infectious [7].

There are substantial genetic contributions to the aetiology of many of the epilepsies, particularly in the developmental and epileptic encephalopathies (DEEs), a group of rare severe epilepsies. In this epilepsy group, approximately 50% of cases can now be attributed to either inherited or de novo variants (not inherited from parents) [8,9]. Major progress has been made in elucidating a role for genetic variance in focal epilepsies, which account for 60% of all epilepsies and which had historically been considered as largely acquired disorders [10]. Between 10 and 15% of individuals with focal epilepsy and a family history of epilepsy are attributable to germline pathogenic variants [11,12]; the 'hit rate' increases to 30–40% in those with brain malformations, which are largely accounted for by genetic variants that occur after birth (somatic mutations) [13]. The term genetic generalised epilepsies (GGEs) is used to describe patients with generalised seizure types that are presumed to have a genetic aetiology [14]. GGEs account for 15–20% of all epilepsies [14,15]. Within the GEE are two common 'absence epilepsies', childhood absence epilepsy and juvenile absence epilepsy, which together with the syndromes of juvenile myoclonic epilepsy and generalised tonic-clonic seizures alone form the idiopathic generalised epilepsies (IGEs) [14]. The main contribution to this epilepsy group appears to be from a large collection of common variants each conferring a small effect size [16].

To identify genes implicated in epilepsy (as well as in other disorders), single individuals, mother father and child trios, whole families, and large cohorts are oftentimes tested. The technologies used for testing include chromosomal microarrays (CMAs) for the detection of large structural variation in the DNA, single-nucleotide polymorphism (SNP) array data that capture common variation and sequencing or whole genome sequencing (WGS). The use of whole exome sequencing (WES) and WGS is increasingly feasible as the costs for short-read (150 base pairs) technology will continue to decrease [17]. Technologies that read longer lengths (5000–30,000 base pairs) generated by Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing and the Oxford Nanopore Technology (ONT) platform can improve on short-read sequencing for tasks such as detecting structural variation [18], but are often challenged by increased error rates [19]. The development of bioinformatic methods to resolve the error rates is important as these tools have attractive clinical utility given the ability to sequence in real time; a pipeline using the ONT PromethION found candidate disease variants within 8 h of blood draw [20] and could potentially be used to detect variants relevant to pharmacogenomics.

The processing of genomic data to yield useful clinical information can be broadly grouped into three steps: variant detection, variant annotation and interpretation, and classification or prediction tasks (see Figure 1 and Box 1).

The development of methods and tools for processing and interpreting sequencing data is an active area of research and as with many other fields, machine learning is playing a transformational role. Machine learning (ML) is a type of artificial intelligence that uses statistical algorithms to determine an output from a given input. Deep learning (DL) is a subfield of ML that was inspired by the complexity of neuronal networks in the brain and uses artificial neural network architectures to perform ML tasks [24]. DL models have multiple layers of nonlinear processing units that progressively learn representations of the data needed for detection or classification [25]. In epilepsy, DL has been successfully applied for the study of magnetic resonance imaging (MRI) [26], diffusion kurtosis imaging (DKIs) [27], and electroencephalogram (EEG) [28]. Here, we review DL methods and tools (see Box 2) with the capacity to address challenges or enhance genomic research tasks as outlined in Figure 1.

Box 1. Brief glossary of genetic terms.

Protein coding regions: DNA regions, also known as exons, that contain the instructions for producing a protein (regions within genes).

Intronic regions: DNA regions that are located within introns (non-coding regions within genes).

Intergenic regions: DNA regions that are located between genes (non-coding regions between genes).

Variants: Differences in DNA sequences compared to a reference genome.

Single-nucleotide variants (SNVs): Genetic variation that occurs when a single nucleotide differs from the reference genome.

Insertions and deletions (Indels): Genetic variations that involve the insertion or deletion of DNA sequences.

Structural variants (SVs): Genetic variations that involve large-scale changes in the structure of the genome, including deletions, insertions, duplications, insertions, inversions, or translocations.

Copy number variants (CNVs): Structural genetic variations that involve changes in the number of copies of a particular DNA region.

Rare variants: Rare genetic variations that are generally considered to be present in less than 1% of the population.

Common variants: Common genetic variations that are generally considered to be present in at least 1% or 5% of the population.

Mosaic variants: Genetic variations that are not expressed in all germline, somatic, or both cells.

Candidate disease variants: Genetic variations that are hypothesized to be associated with a particular trait or disease based on prior knowledge/evidence.

Variant interpretation: The process of assessing the significance or pathogenicity (likely benign, benign, variant of uncertain significance, likely pathogenic, or pathogenic) and determining the potential functional consequences of genetic variations.

Variant browser: Software tools or web-based interface that allows the visualization of genetic variations.

Polygenic risk scores (PRS): Calculated score based on multiple common genetic variations, which estimates the risk of developing a particular trait or disease.

Risk allele: Genetic variation within a gene that is associated with an increased risk of a specific trait or disease.

Gene panels: Targeted sets of genes used to assess rare genetic variations of large effect size based on their suspected involvement in a particular trait or disease.

Whole genome: Genome representing the complete set of genetic information (DNA sequence) of an organism.

Whole exome: Subset of a genome containing only protein-coding regions (exons) of genes.

Single-nucleotide polymorphism (SNP) array data: Data generated using SNP arrays to detect and analyse SNVs across a genome.

Chromosomal microarrays (CMAs): Technique used to detect chromosomal abnormalities or copy number variations in the genome.

Short-read sequencing data: Data generated through next-generation sequencing technologies that produce DNA or RNA fragments ranging from a few dozen to a few hundred base pairs in length.

Long-read sequencing data: Data generated through sequencing technologies that produce DNA or RNA fragments ranging from several thousand to tens of thousands of base pairs in length.

Gene expression data: Data that quantifies the abundance or activity of genes in a specific sample or tissue.

Genotype data: Data representing genetic information at specific genomic positions, typically focusing on SNVs or small insertions and deletions.

Protein–protein and gene co-expression networks: Networks that represent the interactions or co-expression patterns between proteins or genes, respectively.

Genome-wide association study (GWAS): Study designed to identify common genetic variations raising the risk for traits, diseases, or phenotypes on a genome-wide scale. Typically involves genotyping or sequencing on a large scale and comparing common genetic variations between cases and controls to identify statistically significant associations.

Box 2. Brief glossary of DL-related terms.

Supervised Machine Learning: Models are trained on labelled samples to subsequently classify the labels of unseen samples. Commonly used algorithms are Decision Trees, Linear Regression, Logistic Regression, and Support Vector Machines (SVM).

Unsupervised Machine Learning: Models that identify clusters in data from un-labelled data. Commonly used algorithms are Principal Component Analysis (PCA) and K-means clustering. Semi-supervised Machine Learning models are trained on labelled and un-labelled samples.

Deep Learning (DL): DL uses artificial neural network architectures to perform ML tasks. The most basic neural network encompasses three layers with nodes or neurons: an input, hidden, and output layer. Once the number of hidden layers is increased, the network increases in depth and is then considered to be a deep neural network.

Convolutional Neural Networks (CNN): A type of network used for spatial data, images, video, or sequences. The CNN extracts local patterns (e.g., edges or sequence patterns) by applying a filtering mechanism whereby the output from the previous layer is passed to the next layer.

Recurrent Neural Networks (RNN): Networks that can process sequential or temporal data and are, therefore, predominantly used for speech recognition, natural language processing (NLP) or processing time series data.

Transfer learning: Methods that improve the learning process by transferring previously learnt knowledge (e.g., parameters, weights, or output data) from an existing ML model or external information (source) to a new ML model.

Data augmentation: Methods that improve the performance of DL by adding newly synthetically created or slightly modified samples to increase the number of training samples.

Representation learning: Algorithms can learn representations (or features) of raw data that improve the performance of a model or be used as input to a model. Deep learning allows multiple transformations, yielding more abstract and potentially more useful representations.

Overfitting and Underfitting: Terms applied to a model that performs poorly when tested on new data. The model might perform well on test data but not on new unseen data (overfitting), or make incorrect predictions with new data (underfitting).

Dimensionality Reduction: Methods applied before modelling to reduce the number of features while preserving the relationships between features in data.

Feature Learning or Feature Engineering: Techniques that map the data into a more consolidated or lower dimension while trying to preserve the input information.

Generative Adversarial Networks (GANs): Networks that encompass two neural networks, a discriminator and a generator network, which are both trained together. The generator produces data samples, and the discriminator differentiates whether a given sample is real or generated by the generator.

Autoencoder: An unsupervised learning architecture that encompasses two neural networks, a feature extractor (encoder) and a feature validator (decoder) that work together to generate a representation of the data with reduced dimensionality.

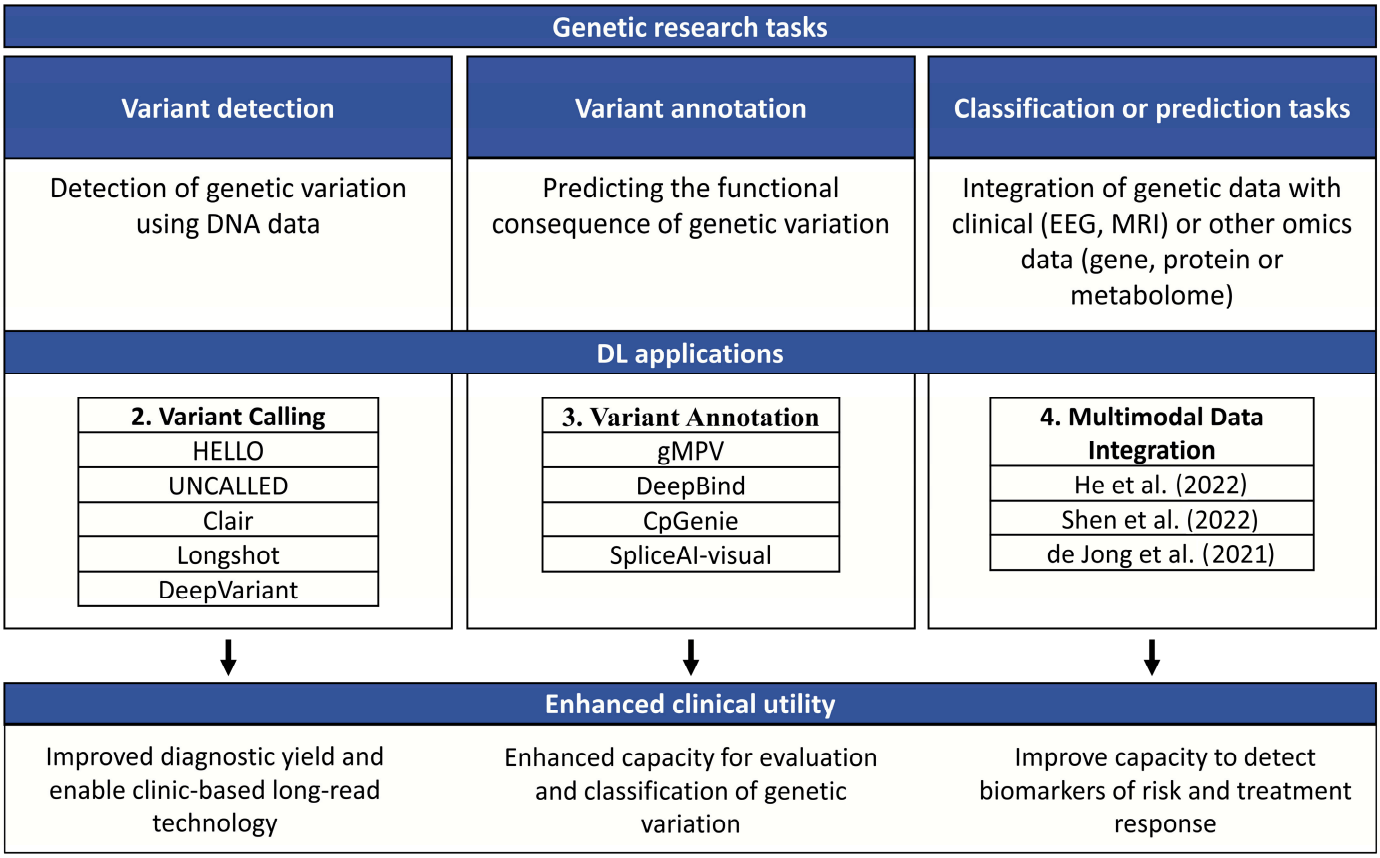

Figure 1. Routine clinical and genetic research tasks that can be enhanced by DL [21–23].

2. Variant Calling

The Genome Analysis Toolkit (GATK) is a software pipeline developed at the Broad Institute [29] for processing short-read next-generation sequencing data. It has become the industry 'gold standard' for identifying single-nucleotide variants (SNVs) and short insertions and deletions (indels) from short-read germline DNA sequence data. In recent years, several new DL tools for variant calling, which involves comparing the sequence of an individual against a reference genome to identify differences within the genome, have been introduced. Short-read callers include DeepVariant [30] and HELLO (Hybrid and stand-alone Estimation of small genomic variants) [31]. DeepVariant [30] generates an image by mapping the raw read data to a 2D data matrix and then uses the images to call variants based on predictions for candidate alternative alleles at each site. In contrast, the HELLO algorithm uses a deep neural network and customized variant inference functions to make predictions for each candidate allele by analysing the reads supporting each allele. The authors of HELLO claim that their method outperforms DeepVariant, in terms of a reduced number of insertion and deletion errors and accuracy [31]. Clair [32] is a recently introduced deep neural network caller that was specifically designed to reduce the error rate inherent in long-read sequencing data but performs less well, relative to DeepVariant, on short-read data. The authors claim that Clair outperforms their earlier DL tool called Clairvoyant, and two other long-read callers introduced in 2019 named Longshot [33] and Medaka [32]. The ONT MiniON platform uses a method known as ReadUntil which allows unwanted sequences to be depleted during the sequence run. UNCALLED is a variant caller that exploits the ReadUntil method. When tested using a panel of human cancer genes, it was found to have high precision in the detection of SNVs, indels, structural variations and DNA methylation [34].

As new tools are introduced, improved performance is typically demonstrated by comparison against earlier published methods and using standard genomes for bench-

marking made available through, for instance, the gene in a bottle consortium (https://www.nist.gov/programs-projects/genome-bottle; accessed on 20 September 2023). Evidence of how these tools perform in real-world research and clinical settings is limited, and it will take time to generate sufficient data for robust evaluation. A comparison between the GATK pipeline and DeepVariant for the detection of pathogenic variants from exome data from independent cohorts of cancer patients found that DeepVariant had higher sensitivity and specificity [35]. In epilepsy, the adoption of WGS is gaining momentum. A systematic review and meta-analysis of genetic testing options in epilepsy found WGS to have the highest diagnostic yield (48%) [9]. Costain, Cordeiro, et al. [36] found WGS gave the highest yield for childhood epilepsy. Ostrander, Butterfield, et al. [37] evaluated the effectiveness of WGS for clinical diagnosis and gene discovery in early infantile epileptic encephalopathy and concluded that, in comparison to standard approaches involving multiple genetic tests, WGS saves time and costs while also allowing for the evaluation of non-coding and CNVs. The American Society of Genetic Counsellors also recommends genetic testing, preferentially exome or genome sequencing or multi-gene panel testing, for individuals with unexplained epilepsy, irrespective of age [38]. The GREP (Genomic sequencing for Refractory EPilepsy) study [39], a randomised controlled trial, aims to assess the utility and cost-effectiveness of WGS for refractory epilepsy in children and adults will provide a valuable 'ground truth' of patient outcomes for future performance testing. The EMPOWER-1: A Multi-site Clinical Cohort Research Study to Reduce Health Inequality in the United Kingdom is using WGS to identify candidate genetic variants that may underpin observed disparities in treatment failure for 19 disease areas, including epilepsy [40].

3. Variant Annotation

Variant annotation tools provide information about the likely functional consequence of genetic variants. This information is then used to evaluate, interpret, and classify variants in accordance with the American College of Medical Genetics (ACMG) guidelines (see Figure 2) [41].

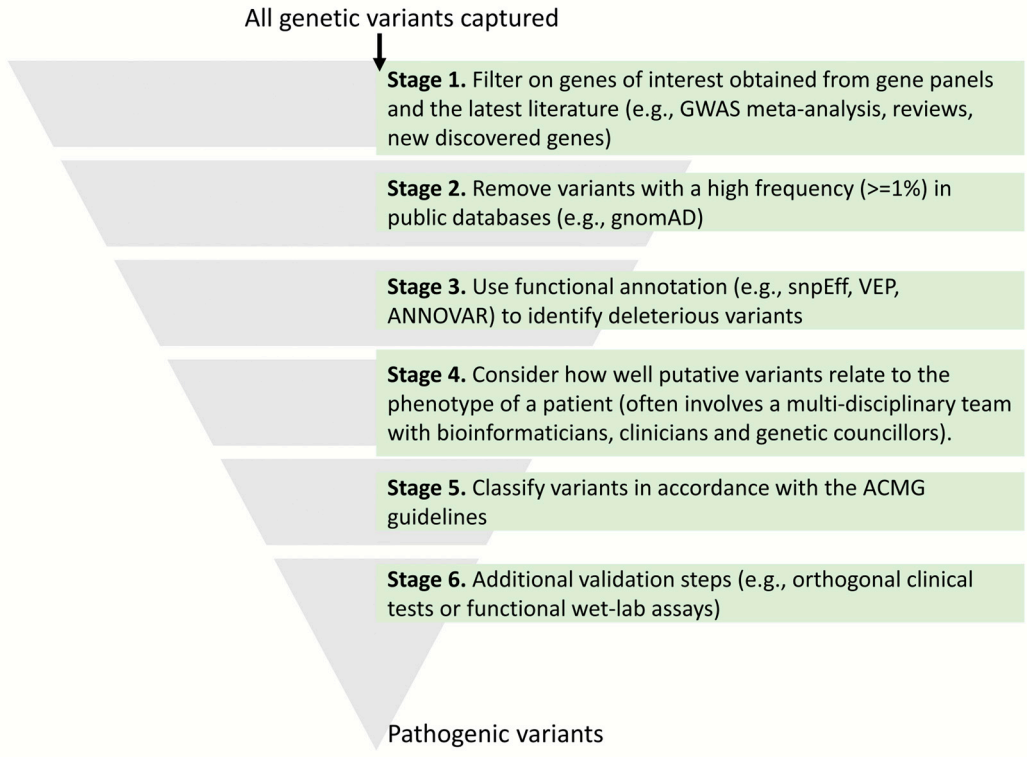

Figure 2. Overview of how genetic variants are filtered and screened.

Tools commonly used for annotation include the Ensembl Variant Effect Predictor (VEP) [42], ANNOVAR (ANNOtate VARiation) [43], and snpEff [44]. The annotation information generated includes, for example, scores generated by the PolyPhen-2 (Polymorphism Phenotyping v2) tool that indicate the possible impact of an amino acid substitution on the structure and function of a human protein [42–44]. Some tools, such as the Franklin by genoox web-based tool (https://franklin.genoox.com/clinical-db/home; accessed on 20 September 2023) introduced in 2019, and the MobiDetails online tool for DNA variant interpretation [45] introduced in 2020, automatically classify variants based on the American College of Medical Genetics (ACMG) guidelines [41]. MobiDetails is freely available for academic use (https://mobidetails.iurc.montp.inserm.fr/MD/; accessed on 20 September 2023) and is designed to work on mobile telephones or tablets by medical staff. This 'real time' interpretation of genetic variation will be necessary to keep pace with the evolution of 'clinic-friendly' real-time DNA sequencing tools.

3.1. Protein-Coding Variants

Numerous tools have been developed to predict whether amino acid substitutions result in disease. Output from these tools is typically used as 'supporting' evidence for variant interpretation and classification [46]. The ACMG recommends that this evidence be counted only if all the in-silico tools used to predict are concordant in their findings (e.g., all predict deleterious or all predict benign effect). This is problematic as tools vary in accuracy with more recent algorithms have higher predictive power over older poorer poorer-performing algorithms. Also, concordant in silico predictions have been found to be opposite to the evidence provided by other sources leading to an error in variant interpretation, that has been termed 'false concordance' [47]. Metapredictors that incorporate older algorithms as features (e.g., REVEL) may be a better alternative, but still not ideal given the inefficiencies of older algorithms and duplication of information [47].

DL and innovative modelling architectures hold promise for improving prediction. For example, the graphical missense variant pathogenicity predictor (gMPV,) is a tool for predicting the impact of missense variants. It uses a novel graph attention neuronal network that pools information on the variant with information on the local protein context (see Figure 3) [48].

3.2. Intronic and Intergenic Regions

Annotation information remains limited for variants that fall within intronic and intergenic DNA regions, yet these regions contain sequences with critically important roles such as splicing sites, binding sites for transcription factors, and CpG islands where DNA methylation occurs. Methylation within enhancer regions can have a similar function to methylation within gene promoter regions but needs yet to be fully understood [49,50]. Since DL tools are helping to address this knowledge gap, we provide below some examples of how DL tools predict the impact on these mechanisms, controlling how genes are transcribed and turned on or off.

3.2.1. Transcription Factor Binding

Transcription factors (TFs) are proteins that bind to short DNA sequences (usually 6–12 bases) and act as 'master regulators' of cell type-specific gene expression [51]. Polymorphisms in TF binding sites comprise only 8% of the genome, yet represent 31% of all trait-associated polymorphisms [52]. Genetic variants within the binding sites can modify the strength of binding (binding affinity), prevent one or more TFs from binding at all, or permit aberrant TF binding [53]. A role for variant-disrupted TF binding has been evidenced in epilepsy, but remains poorly understood. In an animal model of epilepsy, a single long seizure was found to increase the levels of the TF neuron restrictive silencer factor (NRSF). The effect on genes regulated by NRSF was dependent on the binding affinity of this repressor at its target binding sites, with mid-range binding frequencies rendering genes sensitive to moderate fluctuations with deleterious effects [54]. The BCL11A TF was

found to have the strongest association signal for the GGEs in the International League Against Epilepsy (ILAE) genome-wide association study (GAWS) [55]. Given that there are large numbers of TFs and their matched binding sites are widely dispersed throughout the genome, it is likely that this kind of genetic risk is highly heterogeneous and remains elusive. DeepBind [56] is a tool that can be used to investigate the impact of genetic variation on TF binding. This tool was trained using information from multiple sources including protein binding microarrays (PBMs), ribonucleic acid complete (RNAcomplete) data, chromatin immunoprecipitation sequencing (ChIP-seq) and powerful assays that determine the in vitro binding specificities of proteins (SELEX). This tool can be used to predict the binding affinity of 617 human-related TFs across the entire genome. The algorithm produces a score that represents the binding of a specific TF at specific DNA loci [53]. By comparing binding affinity scores generated using reference DNA and DNA-carrying variants at any specific binding site, this tool can be used to determine the impact of genomic variance on TF activity. This information can then be incorporated into annotation tools.

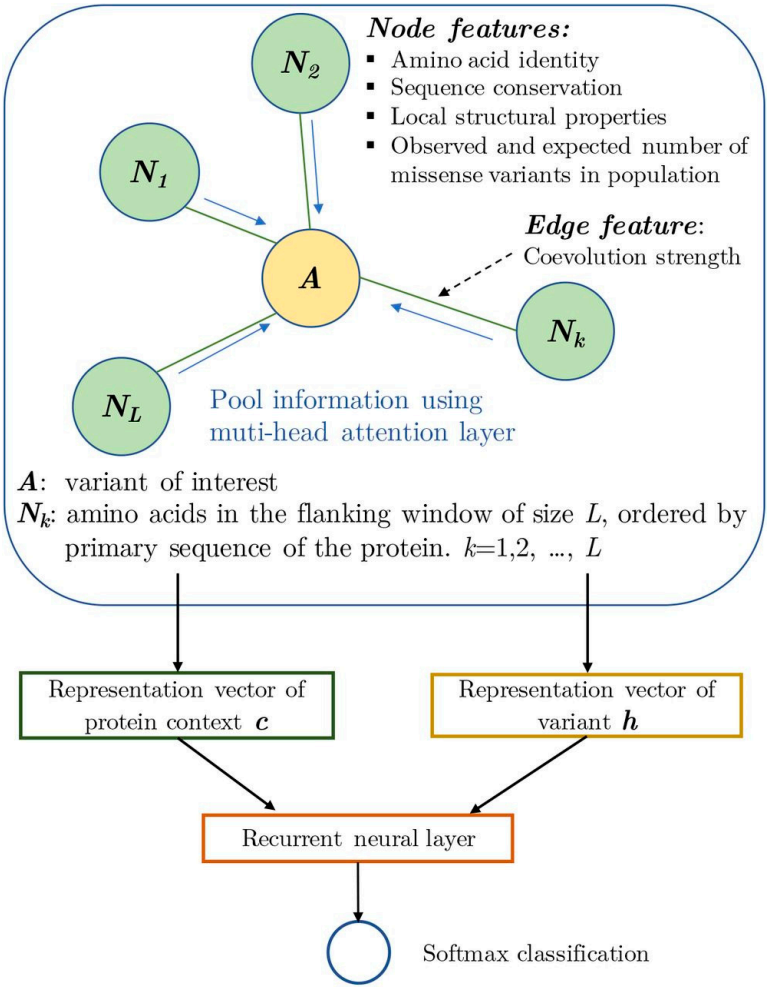

Figure 3. Overview of the graphical methodology that pools together feature information about a protein, represented as green nodes (N_1, N_2, N_L, N_k) in the network and genetic variation, represented as a yellow node (A). Edges, the lines that link nodes, can also contain information such as the strength of coevolution across species; copied and text adjusted from [48] licensed under Creative Commons BY 4.0 (CC BY NC ND 4.0). Full terms at https://creativecommons.org/licenses/by-nc-nd/4.0/.

3.2.2. DNA Methylation

DNA methylation involves the covalent addition of a methyl group at CpG sites and is the most-studied and best-understood of the epigenetic marks [57]. Variants that

associate with changes in DNA methylation have been variously named methylation quantitative trait loci or meQTLs, mQTLs, or metQTLs [57]. Aberrant DNA methylation has been reported in patients with focal epilepsy [58], including temporal lobe epilepsy (TLE) [59] and, specifically, TLE with hippocampal sclerosis [60]. Animal models of epilepsy have implicated dysregulation of DNA methylation across different rat strains [61,62] and show that status epilepticus modifies the methylation status of the glutamate receptor Grin2b [63]. CpGenie [64] is a tool that was trained on methylation sequencing data to predict the methylation status at CpG sites in the DNA. It can evaluate the impact of a variant on methylation status by comparing predicted methylation at CpG sites flanking a variant, with predictions based on reference DNA at the same locus.

3.2.3. Splicing

A large number of pathogenic variants affect how mRNA is spliced [65]. These variants can fall between splice-specific dinucleotides (GT and AG) or within other regions in the DNA where their presence introduces alternative splicing activity (cryptic splice variants) [66]. A role for splice variants in epilepsy is evidenced for genes associated with the DEEs and focal epilepsy including but not limited to STXBP1 [67], SCN1A [68], DNM1 [69], DEPDC5 [70], and GRIN2B [71]. Tools that predict variant impact have been available for a long time, but are constantly being improved (e.g., SpliceFinder introduced in 2019 [72] and SpliceViNCI in 2021 [73]). SpliceAI [66] is a DL predictor that is widely adopted and included in current annotation tools and data repositories including the Genome Aggregation Database (gnomAD) [74]. SpliceAI-visual is a recently introduced improved version that has been trained on additional data and overcomes limitations in the earlier version. It is freely available, compatible with the commonly used variant browsers (Integrative Genomics Viewer [IGV] and UCSC) and has been integrated into the MobiDetails variant interpretation tool [75]. This evolution makes it challenging for researchers to know which tool to use and highlight the importance of keeping variant curation pipelines up to date.

In summary, DL-based tools for predicting the functional impact of genetic variation are rapidly evolving and bring opportunities to break ground on annotation knowledge gaps and the 'false concordance' issue. This work is supported by the emergence of tool repositories that are made accessible through the open-source community platform, GitHub, which fosters community-based software development and helps to ensure robust and reusable code and methods. An example is the Kipoi repository [76], which provides tools that are specific to genomics. The tools are typically pre-trained for specific tasks and ready to use on new data. Bioinformaticians and software developers can also use them in a building block fashion to generate more comprehensive models or solve more complex tasks [76].

4. Multimodal Data Integration

In the DL literature, the terms 'data fusion', 'information fusion', and 'multi-view' and 'multimodal' learning are commonly used to describe models that integrate different data modalities [77,78]. The integration of genetic data, such as GWAS summary statistics, with data or information from complementary sources can lead to novel insights. Johnson et al. (2015) employed what they called a 'systems genetic approach' whereby they integrated network analysis of global gene expression in the hippocampi of TLE patients with GWAS data. Similarly, they integrated genetic risk data for psychiatric disease and behavioural traits with gene expression data in a rat model, to gain insight into neurodevelopmental outcomes following gestational exposure to the known teratogenic anti-seizure medication valproic acid [79].

de Jong, Cutcutache, et al. [23] explored how genetic and clinical data could be integrated into predicting response to brivaracetam, a new antiseizure medication. They generated several 'genomic features' from genotype data to use as input to their models including polygenic risk scores, a dichotomous variable for the presence of a specific gene variant, and burden scores for gene sets in which the genes were relevant to either epilepsy

or the drug mode of action. They evaluated one DL and four ML integration strategies and found that the latter performed best. They considered that the poor performance of the DL method was likely due to the large number of features and the small number of samples [23].

Methodologies for multimodal modelling are rapidly evolving, although the application of these methods in medical research and clinical setting is still in its nascent stages. Li, Wang, et al. [80] provide a proof-of-concept for the integration of SNVs and brain image data to delineate schizophrenic patients and healthy controls. The model architecture was designed to detect complex nonlinear relationships between SNV and the image data. In epilepsy, Shen et al. [22], used a convolutional neural network (CNN) to predict post-stroke epilepsy. The model inputs include EEG signal data and the frequency of gene mutations in genes associated with stroke.

He and Xie [81] developed a method for the prediction of anti-cancer drug sensitivity named Cross-LEvel Information Transmission (CLEIT). As the name suggests the underlying method transmits information 'learned' from modelling one data type (e.g., gene expression) to a model of a second data type (e.g., genotypes) to improve the predictive capacity of the latter. This is an example of a strategy known as transfer learning. This method demonstrates that integration of genotype data with outputs learned from gene expression data can considerably improve performance over prediction with genotype data alone.

Integrating information and knowledge can also improve performance when modelling a single data type. Marini, Limongelli, et al. [82] used a method they describe as a 'data fusion network' to distinguish patients affected by epilepsy from controls. This method merged domain knowledge, gene, pathway, and protein–protein interaction data related to genes in an epilepsy gene panel. The integration of multiple data/knowledge sources can also be used to generate features for input to a model. This typically involves representation learning. For each data or information source, the algorithms discover the best representation of the raw data, then multiple representations are combined in a predictive model. This approach was utilised to develop a Multi-Graph Representation learning-based Ensemble Learning method (MGREL) for gene–disease association prediction [83].

5. Challenges and Potential Solutions

Insufficient data: DL algorithms work best when there are large numbers of samples available for training which is often not the case. Data augmentation [84] and transfer learning [85] are technical strategies that can be adopted to overcome insufficient datasets. Data augmentation involves the generation of new or slightly modified samples to increase the sample numbers for training the algorithms, which, in turn, improves the performance of the models [84]. Data augmentation has been successfully applied to epileptic seizure prediction from EEG data [21], and to delineate the seizure-onset zone in individuals with focal epilepsy [86]. See Habashi, Azab, et al. [87] for a full review. In genomics, Wei, Li, et al. [84] used data augmentation of genetic cancer data to demonstrate how this strategy improves cancer classification.

Transfer learning improves the learning process by transferring previously learnt knowledge (e.g., parameters, weights, or output data) from one model to another, as in the CLEIT method [81]. In general, by applying transfer learning, a new ML model can be trained in less time with fewer data but with increased predictive performance relevant to an ML model trained on available data in the target domain alone [76,88]. Si, Zhang, et al. [89] explored the utility of transfer learning combined with diffusion MRI to predict juvenile myoclonic epilepsy in a small sample of participants (15 with juvenile myoclonic epilepsy and 15 healthy controls). In genomics, Liu, Meng, et al. [90] propose a deep transfer learning model for the calculation of PRSs and Tan and Shen [91] show how transfer learning can be applied for in silico confirmation of rare CNVs identified from sequencing data.

High dimensionality: In genomics, the number of variables typically far exceeds the number of samples and too many features can lead to overfitting and poor model

performance. This challenge is also commonly referred to as the curse of data dimensionality [92]. Conventional dimensionality reduction uses linear and non-linear transformation, including spectral [93] and kernel [94] methods to reduce the number of features into a lower dimension [95]. In genomics, important features can be lost when a linear transformation is applied [96]. The architecture of DL models supports non-linear dimensional reduction [95]. Representation learning is an important and increasingly applied solution. Representation learning can, for example, capture global features in networks representing protein–protein interactions and gene co-expression networks. Topological network information and a representation learning techniques have been used to identify genes associated with cerebral ischemic stroke [97], to prioritize candidate genes for complex diseases with gene networks and GWAS data [98], and to identify gene-phenotype relations from the biomedical literature [99].

Low interpretability: The condensed data representations generated by DL are challenging to interpret, leading them to be labelled as 'black-box' methods. The interpretability of a model can deteriorate as the complexity of the neural network (e.g., depth, width, integration of data modalities) is increased. Work on improving this issue is ongoing. Techniques that confer an 'interpretable' outcome have been demonstrated in a model that learned features from SNP and brain imaging data from a neurodevelopmental cohort [100]. This method identified specific SNPs and functional connectivity in the brain.

Further reading: For a broader summary of DL strategies, see Stahlschmidt, Ulfenborg and Synnergren [101], and the book "Deep Learning" by Bengio, Goodfellow and Courville [102].Stahlschmidt, Ulfenborg and Synnergren [101], and the book "Deep Learning" by Bengio, Goodfellow and Courville [102].

6. Conclusions

In this review, we have highlighted important roles for DL for variant calling, variant interpretation, and multimodal techniques for classification and prediction tasks relevant to epilepsy genetics research. The rapid developments in this field make it challenging for researchers and clinicians to keep abreast and to make the best tool and method selections. This is important as it has been shown that the choice of the variant caller or in silico tools used for variant annotation has an impact on diagnostic yield [47]. The ACMG guidelines criteria for in silico support for pathogenicity (PP3) and benign effects (BP4) are based on in silico predictor tool concordance but this becomes increasingly challenging as more predictors are tested. A recent evaluation of the impact of these criteria found that on the removal of the PP3 criterion, 14% of pathogenic and 24% of likely pathogenic variants were downgraded to likely pathogenic and VUS, respectively, while the removal of BP4 changed the classification of 64% of variants from benign to a variant of unknown significance (VUS) [103]. It is suggested that variant classification be considered a dynamic process, whereby there is benefit in the re-evaluation of previously classified variants as tools and methods evolve [104]. The rapid development of tools also creates a pressing requirement for benchmark datasets.

Repositories such as Kipoi aim to make it easier for researchers to adopt DL methods. Additional DL repositories relevant to genomic research include Selene [105], pyster [106], and Janguu [107]. Unlike commercial software, these tools require bioinformatics or coding expertise, and technical support may be limited, slowing down uptake by researchers and clinicians. Another caveat is that many of these tools have been trained on large publicly available data, in particular, ENCODE [108], but not all tissues or cell types are represented in these training data [109].

Multimodal modelling is an important next step in genomic research, with relevance to precision medicine. de Jong, Cutcutache, et al. [23] demonstrated how the inclusion of genetic information could improve predictive power and the CLEIT model by He and Xie [81] demonstrated how DL methods could better model the nuance and complexity in the relationships between features across data modalities. Publicly available repositories of

information and genomic data will be increasingly important for enabling the training and evaluation of DL applications.

Author Contributions: R.Z. and A.A. conceptualised the review. R.Z. drafted and revised the manuscript for intellectual content. P.K., T.J.O., P.P. and A.A. provided scientific direction. P.K., T.J.O., P.P. and Z.G. critically revised the manuscript for intellectual content. All authors have read and agreed to the published version of the manuscript.

Funding: P.K. was supported by a Practitioner Fellowship from the Medical Research Future Fund (MRF1136427). P.P. is supported by an Emerging Leadership 2 Investigator Grant from the NHMRC (APP2017651), The University of Melbourne, Monash University, the Weary Dunlop Medical Research Foundation, Brain Australia, and the Norman Beischer Medical Research Foundation. He has received speaker honoraria or consultancy fees to his institution from Chiesi, Eisai, LivaNova, Novartis, Sun Pharma, Supernus, The Limbic, and UCB Pharma, outside the submitted work. He is an Associate Editor for Epilepsia Open. T.J.O. is supported by a NHMRC Investigator Leadership grant (APP1176426) and his institution has received research funding and/or consulting fees from Eisai, UCB Pharma, Biogen, Supernus, LivaNova, ES Therapeutics, Kinoxis Therapeutics, and Epidurex.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Moshe, S.L.; Perucca, E.; Ryvlin, P.; Tomson, T. Epilepsy: New advances. *Lancet* **2015**, *385*, 884–898. [CrossRef] [PubMed]
2. Beghi, E. The Epidemiology of Epilepsy. *Neuroepidemiology* **2020**, *54*, 185–191. [CrossRef] [PubMed]
3. Dhiman, V. Molecular Genetics of Epilepsy: A Clinician's Perspective. *Ann. Indian Acad. Neurol.* **2017**, *20*, 96–102. [CrossRef] [PubMed]
4. Mullan, K.A.; Anderson, A.; Illing, P.T.; Kwan, P.; Purcell, A.W.; Mifsud, N.A. HLA-associated antiepileptic drug-induced cutaneous adverse reactions. *HLA* **2019**, *93*, 417–435. [CrossRef]
5. Fisher, R.S.; Cross, J.H.; French, J.A.; Higurashi, N.; Hirsch, E.; Jansen, F.E.; Lagae, L.; Moshé, S.L.; Peltola, J.; Roulet Perez, E.; et al. Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. *Epilepsia* **2017**, *58*, 522–530. [CrossRef] [PubMed]
6. Scheffer, I.E.; Berkovic, S.; Capovilla, G.; Connolly, M.B.; French, J.; Guilhoto, L.; Hirsch, E.; Jain, S.; Mathern, G.W.; Moshe, S.L.; et al. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. *Epilepsia* **2017**, *58*, 512–521. [CrossRef]
7. Wirrell, E.; Tinuper, P.; Perucca, E.; Moshé, S.L. Introduction to the epilepsy syndrome papers. *Epilepsia* **2022**, *63*, 1330–1332. [CrossRef]
8. Perucca, P.; Bahlo, M.; Berkovic, S.F. The Genetics of Epilepsy. *Annu. Rev. Genom. Hum. Genet.* **2020**, *21*, 205–230. [CrossRef]
9. Sheidley, B.R.; Malinowski, J.; Bergner, A.L.; Bier, L.; Gloss, D.S.; Mu, W.; Mulhern, M.M.; Partack, E.J.; Poduri, A. Genetic testing for the epilepsies: A systematic review. *Epilepsia* **2022**, *63*, 375–387. [CrossRef]
10. Perucca, P. Genetics of Focal Epilepsies: What Do We Know and Where Are We Heading? *Epilepsy Curr.* **2018**, *18*, 356–362. [CrossRef]
11. Perucca, P.; Scheffer, I.E.; Harvey, A.S.; James, P.A.; Lunke, S.; Thorne, N.; Gaff, C.; Regan, B.M.; Damiano, J.A.; Hildebrand, M.S.; et al. Real-world utility of whole exome sequencing with targeted gene analysis for focal epilepsy. *Epilepsy Res.* **2017**, *131*, 1–8. [CrossRef] [PubMed]
12. Epi4K-Consortium. Ultra-rare genetic variation in common epilepsies: A case-control sequencing study. *Lancet Neurol.* **2017**, *16*, 135–143. [CrossRef] [PubMed]
13. Lai, D.; Gade, M.; Yang, E.; Koh, H.Y.; Lu, J.; Walley, N.M.; Buckley, A.F.; Sands, T.T.; Akman, C.I.; Mikati, M.A.; et al. Somatic variants in diverse genes leads to a spectrum of focal cortical malformations. *Brain* **2022**, *145*, 2704–2720. [CrossRef] [PubMed]
14. Nilo, A.; Gelisse, P.; Crespel, A. Genetic/idiopathic generalized epilepsies: Not so good as that! *Rev. Neurol.* **2020**, *176*, 427–438. [CrossRef] [PubMed]
15. Jallon, P.; Latour, P. Epidemiology of idiopathic generalized epilepsies. *Epilepsia* **2005**, *46* (Suppl. S9), 10–14. [CrossRef]
16. Stevelink, R.; Campbell, C.; Chen, S.; Abou-Khalil, B.; Adesoji, O.M.; Afawi, Z.; Amadori, E.; Anderson, A.; Anderson, J.; Andrade, D.M.; et al. GWAS meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture. *Nat. Genet.* **2023**, *55*, 1471–1482. [CrossRef]
17. Pervez, M.T.; Hasnain, M.J.U.; Abbas, S.H.; Moustafa, M.F.; Aslam, N.; Shah, S.S.M. A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. *BioMed Res. Int.* **2022**, *2022*, 3457806. [CrossRef]

18. McClinton, B.; Crinnion, L.A.; McKibbin, M.; Mukherjee, R.; Poulter, J.A.; Smith, C.E.L.; Ali, M.; Watson, C.M.; Inglehearn, C.F.; Toomes, C. Targeted nanopore sequencing enables complete characterisation of structural deletions initially identified using exon-based short-read sequencing strategies. *Mol. Genet. Genom. Med.* **2023**, *11*, e2164. [CrossRef]
19. Amarasinghe, S.L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M.E.; Gouil, Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **2020**, *21*, 30. [CrossRef]
20. Goenka, S.D.; Gorzynski, J.E.; Shafin, K.; Fisk, D.G.; Pesout, T.; Jensen, T.D.; Monlong, J.; Chang, P.-C.; Baid, G.; Bernstein, J.A.; et al. Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. *Nat. Biotechnol.* **2022**, *40*, 1035–1041. [CrossRef]
21. He, D.; Xie, L. A cross-level information transmission network for hierarchical omics data integration and phenotype prediction from a new genotype. *Bioinformatics* **2022**, *38*, 204–210. [CrossRef] [PubMed]
22. Shen, D.; Deng, Y.; Lin, C.; Li, J.; Lin, X.; Zou, C. Clinical Characteristics and Gene Mutation Analysis of Poststroke Epilepsy. *Contrast Media Mol. Imaging* **2022**, *2022*, 4801037. [CrossRef] [PubMed]
23. de Jong, J.; Cutcutache, I.; Page, M.; Elmoufti, S.; Dilley, C.; Fröhlich, H.; Armstrong, M. Towards realizing the vision of precision medicine: AI based prediction of clinical drug response. *Brain* **2021**, *144*, 1738–1750. [CrossRef] [PubMed]
24. Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40–55. [CrossRef]
25. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
26. Ozdemir, M.A.; Cura, O.K.; Akan, A. Epileptic eeg classification by using time-frequency images for deep learning. *Int. J. Neural Syst.* **2021**, *31*, 2150026. [CrossRef]
27. Huang, J.; Xu, J.; Kang, L.; Zhang, T. Identifying epilepsy based on deep learning using DKI images. *Front. Hum. Neurosci.* **2020**, *14*, 590815. [CrossRef]
28. Nhu, D.; Janmohamed, M.; Antonic-Baker, A.; Perucca, P.; O'Brien, T.J.; Gilligan, A.K.; Kwan, P.; Tan, C.W.; Kuhlmann, L. Deep learning for automated epileptiform discharge detection from scalp EEG: A systematic review. *J. Neural Eng.* **2022**, *19*, 051002. [CrossRef]
29. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef]
30. Poplin, R.; Chang, P.-C.; Alexander, D.; Schwartz, S.; Colthurst, T.; Ku, A.; Newburger, D.; Dijamco, J.; Nguyen, N.; Afshar, P.T.; et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **2018**, *36*, 983–987. [CrossRef]
31. Ramachandran, A.; Lumetta, S.S.; Klee, E.W.; Chen, D. HELLO: Improved neural network architectures and methodologies for small variant calling. *BMC Bioinform.* **2021**, *22*, 404. [CrossRef] [PubMed]
32. Luo, R.; Wong, C.-L.; Wong, Y.-S.; Tang, C.-I.; Liu, C.-M.; Leung, C.-M.; Lam, T.-W. Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nat. Mach. Intell.* **2020**, *2*, 220–227. [CrossRef]
33. Edge, P.; Bansal, V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* **2019**, *10*, 4660. [CrossRef] [PubMed]
34. Kovaka, S.; Fan, Y.; Ni, B.; Timp, W.; Schatz, M.C. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* **2021**, *39*, 431–441. [CrossRef]
35. AlDubayan, S.H.; Conway, J.R.; Camp, S.Y.; Witkowski, L.; Kofman, E.; Reardon, B.; Han, S.; Moore, N.; Elmarakeby, H.; Salari, K.; et al. Detection of Pathogenic Variants with Germline Genetic Testing Using Deep Learning vs Standard Methods in Patients with Prostate Cancer and Melanoma. *JAMA* **2020**, *324*, 1957–1969. [CrossRef]
36. Costain, G.; Cordeiro, D.; Matviychuk, D.; Mercimek-Andrews, S. Clinical Application of Targeted Next-Generation Sequencing Panels and Whole Exome Sequencing in Childhood Epilepsy. *Neuroscience* **2019**, *418*, 291–310. [CrossRef]
37. Ostrander, B.E.P.; Butterfield, R.J.; Pedersen, B.S.; Farrell, A.J.; Layer, R.M.; Ward, A.; Miller, C.; DiSera, T.; Filloux, F.M.; Candee, M.S.; et al. Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy. *NPJ Genom. Med.* **2018**, *3*, 22. [CrossRef]
38. Smith, L.; Malinowski, J.; Ceulemans, S.; Peck, K.; Walton, N.; Sheidley, B.R.; Lippa, N. Genetic testing and counseling for the unexplained epilepsies: An evidence-based practice guideline of the National Society of Genetic Counselors. *J. Genet. Couns.* **2023**, *32*, 266–280. [CrossRef]
39. Kwan, P.; Todaro, M. Genomic Sequencing for Refractory EPilepsy (GREP). 2021. Available online: https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=375633&isReview (accessed on 10 August 2023).
40. Kamran, M. EMPOWER-1: A Multi-Site Clinical Cohort Research Study to Reduce Health Inequality—Full Text View—ClinicalTrials.gov. 2021. Available online: https://classic.clinicaltrials.gov/ct2/show/NCT03987633?term=WGS&cond=Epilepsy&draw=2&rank=1 (accessed on 9 August 2023).
41. Richards, S.; Aziz, N.; Bale, S.; Bick, D.; Das, S.; Gastier-Foster, J.; Grody, W.W.; Hegde, M.; Lyon, E.; Spector, E.; et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **2015**, *17*, 405–424. [CrossRef]
42. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.S.; Thormann, A.; Flicek, P.; Cunningham, F. The Ensembl Variant Effect Predictor. *Genome Biol.* **2016**, *17*, 122. [CrossRef]

43. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef] [PubMed]
44. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **2012**, *6*, 80–92. [CrossRef] [PubMed]
45. Baux, D.; Van Goethem, C.; Ardouin, O.; Guignard, T.; Bergougnoux, A.; Koenig, M.; Roux, A.-F. MobiDetails: Online DNA variants interpretation. *Eur. J. Hum. Genet.* **2021**, *29*, 356–360. [CrossRef] [PubMed]
46. Bean, L.J.H.; Hegde, M.R. Clinical implications and considerations for evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Med.* **2017**, *9*, 111. [CrossRef] [PubMed]
47. Ghosh, R.; Oak, N.; Plon, S.E. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol.* **2017**, *18*, 225. [CrossRef]
48. Zhang, H.; Xu, M.S.; Fan, X.; Chung, W.K.; Shen, Y. Predicting functional effect of missense variants using graph attention neural networks. *Nat. Mach. Intell.* **2022**, *4*, 1017–1028. [CrossRef]
49. Messerschmidt, D.M.; Knowles, B.B.; Solter, D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev.* **2014**, *28*, 812–828. [CrossRef]
50. Aran, D.; Sabato, S.; Hellman, A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* **2013**, *14*, R21. [CrossRef]
51. Crocker, J.; Noon, E.P.; Stern, D.L. The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. *Curr. Top. Dev. Biol.* **2016**, *117*, 455–469. [CrossRef]
52. Nishizaki, S.S.; Ng, N.; Dong, S.; Porter, R.S.; Morterud, C.; Williams, C.; Asman, C.; Switzenberg, J.A.; Boyle, A.P. Predicting the effects of SNPs on transcription factor binding affinity. *Bioinformatics* **2020**, *36*, 364–372. [CrossRef]
53. Tseng, C.C.; Wong, M.C.; Liao, W.T.; Chen, C.J.; Lee, S.C.; Yen, J.H.; Chang, S.J. Genetic Variants in Transcription Factor Binding Sites in Humans: Triggered by Natural Selection and Triggers of Diseases. *Int. J. Mol. Sci.* **2021**, *22*, 4184. [CrossRef] [PubMed]
54. McClelland, S.; Brennan, G.P.; Dubé, C.; Rajpara, S.; Iyer, S.; Richichi, C.; Bernard, C.; Baram, T.Z. The transcription factor NRSF contributes to epileptogenesis by selective repression of a subset of target genes. *Elife* **2014**, *3*, e01267. [CrossRef] [PubMed]
55. International League Against Epilepsy Consortium on Complex Epilepsies; Berkovic, S.F.; Cavalleri, G.L.; Koeleman, B.P. Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-specific genetic architecture. *medRxiv* **2022**. [CrossRef]
56. Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **2015**, *33*, 831–838. [CrossRef] [PubMed]
57. Villicaña, S.; Bell, J.T. Genetic impacts on DNA methylation: Research findings and future perspectives. *Genome Biol.* **2021**, *22*, 127. [CrossRef]
58. Belhedi, N.; Perroud, N.; Karege, F.; Vessaz, M.; Malafosse, A.; Salzmann, A. Increased CPA6 promoter methylation in focal epilepsy and in febrile seizures. *Epilepsy Res.* **2014**, *108*, 144–148. [CrossRef]
59. Kobow, K.; Jeske, I.; Hildebrandt, M.; Hauke, J.; Hahnen, E.; Buslei, R.; Buchfelder, M.; Weigel, D.; Stefan, H.; Kasper, B.; et al. Increased reelin promoter methylation is associated with granule cell dispersion in human temporal lobe epilepsy. *J. Neuropathol. Exp. Neurol.* **2009**, *68*, 356–364. [CrossRef]
60. Miller-Delaney, S.F.; Bryan, K.; Das, S.; McKiernan, R.C.; Bray, I.M.; Reynolds, J.P.; Gwinn, R.; Stallings, R.L.; Henshall, D.C. Differential DNA methylation profiles of coding and non-coding genes define hippocampal sclerosis in human temporal lobe epilepsy. *Brain* **2015**, *138*, 616–631. [CrossRef]
61. Dębski, K.J.; Pitkanen, A.; Puhakka, N.; Bot, A.M.; Khurana, I.; Harikrishnan, K.N.; Ziemann, M.; Kaspi, A.; El-Osta, A.; Lukasiuk, K.; et al. Etiology matters—Genomic DNA Methylation Patterns in Three Rat Models of Acquired Epilepsy. *Sci. Rep.* **2016**, *6*, 25668. [CrossRef]
62. Kiese, K.; Jablonski, J.; Hackenbracht, J.; Wrosch, J.K.; Groemer, T.W.; Kornhuber, J.; Blümcke, I.; Kobow, K. Epigenetic control of epilepsy target genes contributes to a cellular memory of epileptogenesis in cultured rat hippocampal neurons. *Acta Neuropathol. Commun.* **2017**, *5*, 79. [CrossRef]
63. Ryley Parrish, R.; Albertson, A.J.; Buckingham, S.C.; Hablitz, J.J.; Mascia, K.L.; Davis Haselden, W.; Lubin, F.D. Status epilepticus triggers early and late alterations in brain-derived neurotrophic factor and NMDA glutamate receptor Grin2b DNA methylation levels in the hippocampus. *Neuroscience* **2013**, *248*, 602–619. [CrossRef] [PubMed]
64. Zeng, H.; Gifford, D.K. Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Res.* **2017**, *45*, e99. [CrossRef] [PubMed]
65. López-Bigas, N.; Audit, B.; Ouzounis, C.; Parra, G.; Guigó, R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* **2005**, *579*, 1900–1903. [CrossRef]
66. Jaganathan, K.; Kyriazopoulou Panagiotopoulou, S.; McRae, J.F.; Darbandi, S.F.; Knowles, D.; Li, Y.I.; Kosmicki, J.A.; Arbelaez, J.; Cui, W.; Schwartz, G.B.; et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **2019**, *176*, 535–548.e524. [CrossRef]
67. Stamberger, H.; Crosiers, D.; Balagura, G.; Bonardi, C.M.; Basu, A.; Cantalupo, G.; Chiesa, V.; Christensen, J.; Bernardina, B.D.; Ellis, C.A.; et al. Natural History Study of STXBP1-Developmental and Epileptic Encephalopathy Into Adulthood. *Neurology* **2022**, *99*, e221–e233. [CrossRef] [PubMed]

68. Carvill, G.L.; Engel, K.L.; Ramamurthy, A.; Cochran, J.N.; Roovers, J.; Stamberger, H.; Lim, N.; Schneider, A.L.; Hollingsworth, G.; Holder, D.H. Aberrant inclusion of a poison exon causes dravet syndrome and related SCN1A-associated genetic epilepsies. *Am. J. Hum. Genet.* **2018**, *103*, 1022–1029. [CrossRef]
69. Parthasarathy, S.; Ruggiero, S.M.; Gelot, A.; Soardi, F.C.; Ribeiro, B.F.; Pires, D.E.; Ascher, D.B.; Schmitt, A.; Rambaud, C.; Represa, A. A recurrent de novo splice site variant involving DNM1 exon 10a causes developmental and epileptic encephalopathy through a dominant-negative mechanism. *Am. J. Hum. Genet.* **2022**, *109*, 2253–2269. [CrossRef]
70. Tsai, M.H.; Chan, C.K.; Chang, Y.C.; Yu, Y.T.; Chuang, S.T.; Fan, W.L.; Li, S.C.; Fu, T.Y.; Chang, W.N.; Liou, C.W. DEPDC5 mutations in familial and sporadic focal epilepsy. *Clin. Genet.* **2017**, *92*, 397–404. [CrossRef]
71. Lemke, J.R.; Hendrickx, R.; Geider, K.; Laube, B.; Schwake, M.; Harvey, R.J.; James, V.M.; Pepler, A.; Steiner, I.; Hörtnagel, K. GRIN2B mutations in West syndrome and intellectual disability with focal epilepsy. *Ann. Neurol.* **2014**, *75*, 147–154. [CrossRef]
72. Wang, R.; Wang, Z.; Wang, J.; Li, S. SpliceFinder: Ab initio prediction of splice sites using convolutional neural network. *BMC Bioinform.* **2019**, *20*, 652. [CrossRef]
73. Dutta, A.; Singh, K.K.; Anand, A. SpliceViNCI: Visualizing the splicing of non-canonical introns through recurrent neural networks. *bioRxiv* **2020**. [CrossRef] [PubMed]
74. Chen, S.; Francioli, L.C.; Goodrich, J.K.; Collins, R.L.; Kanai, M.; Wang, Q.; Alföldi, J.; Watts, N.A.; Vittal, C.; Gauthier, L.D.; et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv* **2022**. [CrossRef]
75. de Sainte Agathe, J.-M.; Filser, M.; Isidor, B.; Besnard, T.; Gueguen, P.; Perrin, A.; Van Goethem, C.; Verebi, C.; Masingue, M.; Rendu, J.; et al. SpliceAI-visual: A free online tool to improve SpliceAI splicing variant interpretation. *Hum. Genom.* **2023**, *17*, 7. [CrossRef] [PubMed]
76. Avsec, Ž.; Kreuzhuber, R.; Israeli, J.; Xu, N.; Cheng, J.; Shrikumar, A.; Banerjee, A.; Kim, D.S.; Beier, T.; Urban, L.; et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **2019**, *37*, 592–600. [CrossRef]
77. Nguyen, N.D.; Wang, D. Multiview learning for understanding functional multiomics. *PLoS Comput. Biol.* **2020**, *16*, e1007677. [CrossRef]
78. Castanedo, F. A Review of Data Fusion Techniques. *Sci. World J.* **2013**, *2013*, 704504. [CrossRef]
79. Feleke, R.; Jazayeri, D.; Abouzeid, M.; Powell, K.L.; Srivastava, P.K.; O'Brien, T.J.; Jones, N.C.; Johnson, M.R. Integrative genomics reveals pathogenic mediator of valproate-induced neurodevelopmental disability. *Brain* **2022**, *145*, 3832–3842. [CrossRef]
80. Li, G.; Wang, C.; Han, D.P.; Zhang, Y.P.; Peng, P.; Calhoun, V.D.; Wang, Y.P. Deep Principal Correlated Auto-Encoders with Application to Imaging and Genomics Data Integration. *IEEE Access* **2020**, *8*, 20093–20107. [CrossRef]
81. Marini, S.; Limongelli, I.; Rizzo, E.; Malovini, A.; Errichiello, E.; Vetro, A.; Da, T.; Zuffardi, O.; Bellazzi, R. A Data Fusion Approach to Enhance Association Study in Epilepsy. *PLoS ONE* **2016**, *11*, e0164940. [CrossRef]
82. Wang, Z.; Gu, Y.; Zheng, S.; Yang, L.; Li, J. MGREL: A multi-graph representation learning-based ensemble learning method for gene-disease association prediction. *Comput. Biol. Med.* **2023**, *155*, 106642. [CrossRef]
83. Wei, K.; Li, T.; Huang, F.; Chen, J.; He, Z. Cancer classification with data augmentation based on generative adversarial networks. *Front. Comput. Sci.* **2021**, *16*, 162601. [CrossRef]
84. Iman, M.; Arabnia, H.R.; Rasheed, K. A Review of Deep Transfer Learning and Recent Advancements. *Technologies* **2023**, *11*, 40. [CrossRef]
85. He, P.; Wang, L.; Cui, Y.; Wang, R.; Wu, D. Unsupervised feature learning based on autoencoder for epileptic seizures prediction. *Appl. Intell.* **2023**, *53*, 20766–20784. [CrossRef]
86. Zhao, X.; Sole-Casals, J.; Sugano, H.; Tanaka, T. Seizure onset zone classification based on imbalanced iEEG with data augmentation. *J. Neural Eng.* **2022**, *19*, 065001. [CrossRef] [PubMed]
87. Habashi, A.G.; Azab, A.M.; Eldawlatly, S.; Aly, G.M. Generative adversarial networks in EEG analysis: An overview. *J. NeuroEng. Rehabil.* **2023**, *20*, 40. [CrossRef] [PubMed]
88. Eraslan, G.; Avsec, Z.; Gagneur, J.; Theis, F.J. Deep learning: New computational modelling techniques for genomics. *Nat. Rev. Genet.* **2019**, *20*, 389–403. [CrossRef]
89. Si, X.; Zhang, X.; Zhou, Y.; Sun, Y.; Jin, W.; Yin, S.; Zhao, X.; Li, Q.; Ming, D. Automated Detection of Juvenile Myoclonic Epilepsy using CNN based Transfer Learning in Diffusion MRI—Test. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 1679–1682. [CrossRef]
90. Liu, L.; Meng, Q.; Weng, C.; Lu, Q.; Wang, T.; Wen, Y. Explainable deep transfer learning model for disease risk prediction using high-dimensional genomic data. *PLoS Comput. Biol.* **2022**, *18*, e1010328. [CrossRef]
91. Tan, R.; Shen, Y. Accurate in silico confirmation of rare copy number variant calls from exome sequencing data using transfer learning. *Nucleic Acids Res.* **2022**, *50*, e123. [CrossRef]
92. Wörheide, M.A.; Krumsiek, J.; Kastenmüller, G.; Arnold, M. Multi-omics integration in biomedical research—A metabolomics-centric review. *Anal. Chim. Acta* **2021**, *1141*, 144–162. [CrossRef]
93. Ng, A.; Jordan, M.; Weiss, Y. On spectral clustering: Analysis and an algorithm. In Proceedings of the 15th Annual Conference on Neural Information Processing Systems (NIPS 2001), Vancouver, BC, Canada, 3–8 December 2022; MIT Press: Cambridge, MA, USA, 2002.
94. Hofmann, T.; Schölkopf, B.; Smola, A.J. Kernel methods in machine learning. *Ann. Stat.* **2008**, *36*, 1171–1220. [CrossRef]

95. Karim, M.R.; Beyan, O.; Zappa, A.; Costa, I.G.; Rebholz-Schuhmann, D.; Cochez, M.; Decker, S. Deep learning-based clustering approaches for bioinformatics. *Brief Bioinform.* **2021**, *22*, 393–415. [CrossRef] [PubMed]
96. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
97. Liu, H.; Hou, L.; Xu, S.; Li, H.; Chen, X.; Gao, J.; Wang, Z.; Han, B.; Liu, X.; Wan, S. Discovering Cerebral Ischemic Stroke Associated Genes Based on Network Representation Learning. *Front. Genet.* **2021**, *12*, 728333. [CrossRef] [PubMed]
98. Wu, M.; Zeng, W.; Liu, W.; Lv, H.; Chen, T.; Jiang, R. Leveraging multiple gene networks to prioritize GWAS candidate genes via network representation learning. *Methods* **2018**, *145*, 41–50. [CrossRef]
99. Xing, W.; Qi, J.; Yuan, X.; Li, L.; Zhang, X.; Fu, Y.; Xiong, S.; Hu, L.; Peng, J. A gene–phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach. *Bioinformatics* **2018**, *34*, i386–i394. [CrossRef]
100. Hu, W.; Meng, X.; Bai, Y.; Zhang, A.; Qu, G.; Cai, B.; Zhang, G.; Wilson, T.W.; Stephen, J.M.; Calhoun, V.D.; et al. Interpretable Multimodal Fusion Networks Reveal Mechanisms of Brain Cognition. *IEEE Trans. Med. Imaging* **2021**, *40*, 1474–1483. [CrossRef]
101. Stahlschmidt, S.R.; Ulfenborg, B.; Synnergren, J. Multimodal deep learning for biomedical data fusion: A review. *Brief. Bioinform.* **2022**, *23*, bbab569. [CrossRef]
102. Bengio, Y.; Goodfellow, I.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2017; Volume 1.
103. Wilcox, E.H.; Sarmady, M.; Wulf, B.; Wright, M.W.; Rehm, H.L.; Biesecker, L.G.; Abou Tayoun, A.N. Evaluating the impact of in silico predictors on clinical variant classification. *Genet. Med.* **2022**, *24*, 924–930. [CrossRef]
104. Deignan, J.L.; Chung, W.K.; Kearney, H.M.; Monaghan, K.G.; Rehder, C.W.; Chao, E.C.; ACMG Laboratory Quality Assurance Committee. Points to consider in the reevaluation and reanalysis of genomictest results: A statement of the American College of Medical Genetics and Genomics(ACMG). *Genet. Med.* **2019**, *21*, 1267–1270. [CrossRef]
105. Chen, K.M.; Cofer, E.M.; Zhou, J.; Troyanskaya, O.G. Selene: A PyTorch-based deep learning library for sequence data. *Nat. Methods* **2019**, *16*, 315–318. [CrossRef]
106. Budach, S.; Marsico, A. Pysster: Classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. *Bioinformatics* **2018**, *34*, 3035–3037. [CrossRef] [PubMed]
107. Kopp, W.; Monti, R.; Tamburrini, A.; Ohler, U.; Akalin, A. Deep learning for genomics using Janggu. *Nat. Commun.* **2020**, *11*, 3488. [CrossRef] [PubMed]
108. Dunham, I.; Kundaje, A.; Aldred, S.F.; Collins, P.J.; Davis, C.A.; Doyle, F.; Epstein, C.B.; Frietze, S.; Harrow, J.; Kaul, R.; et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74. [CrossRef]
109. Lai, B.; Qian, S.; Zhang, H.; Zhang, S.; Kozlova, A.; Duan, J.; Xu, J.; He, X. Annotating functional effects of non-coding variants in neuropsychiatric cell types by deep transfer learning. *PLoS Comput. Biol.* **2022**, *18*, e1010011. [CrossRef] [PubMed]