Principles of Artificial Neural Networks and Machine Learning for Bioinformatics Applications

2023-08-26

Principles of Artificial Neural Networks and Machine Learning for Bioinformatics Applications

Konstantinos Krampis*[1], Eric Ross[2], Olorunseun O. Ogunwobi[1], Grace Ma,[3] Raja Mazumder,[4] Claudia Wultsch[1]

[1]Belfer Research Facility, Biological Sciences, Hunter College, City University of New York, NY, USA [2]Fox Chase Cancer Center, Philadephia, PA, USA [3]Center for Asian Health, Lewis Katz School of Medicine, Temple University, Philadelphia, PA, USA [4]Biochemistry and Molecular Biology, George Washington University, Washington D.C., USA

*Corresponding Author, *kk104@hunter.cuny.edu*

ABSTRACT

With the exponential growth of machine learning and development of Artificial Neural Network (ANNs) in recent years, there is great opportunity to leverage this approach and accelarate biological discoveries through applications in the analysis of high-throughput data. Various types of datasets, including protein or gene interaction networks, molecular structures, and cellular signalling pathways, have already been utilized for machine learning by training ANNs for inference and pattern classification. However, unlike regular data structures commonly used in the fields of computer science and engineering, bioinformatics datasets present challenges that require unique algorithmic approaches. The recent development of geometric and deep learning approaches within the machine learning field holds great promise for accelerating the analysis of complex bioinformatics datasets. Here, we demonstrate the principles of ANNs and their significance for bioinformatics machine learning by presenting the underlying mathematical and statistical foundations from group theory, symmetry, and linear algebra. Furthermore, the structure and functions of ANN algorithms, which constitute the core principles of artificial intelligence, are explained in relation to the bioinformatics data domain. In summary, this manuscript provides guidance for researchers to understand the principles necessary for practicing machine learning and artificial intelligence, with special considerations for bioinformatics applications.

*Keywords:**machine learning, artificial intelligence, bioinformatics, cancer biology, neural networks, symmetry, group theory, algorithms* biology, neural networks, symmetry, group theory, algorithms_

SIMPLE SUMMARY

Here, we provide an overview of the foundational formalisms of Artificial Neural Networks (ANNs), which serve as the basis for Artificial Intelligence within the broader field of of Machine Learning. The review is from the perspective of bioinformatics data, and multiple examples showcasing the applications of these formalisms to experimental scenarios are presented herein. The mathematical formalisms are explained

in detail, offering biologists who are not Machine Learning experts the opportunity to understand the algorithmic basis of Artificial Intelligence as it relates to bioinformatics applications.

INTRODUCTION

In summary, Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) are interconnected concepts with distinct differences: AI is centered around developing machines capable of performing tasks that require human intelligence, ML empowers computers to learn from data and make predictions without explicit programming, and DL employs deep neural networks to discern patterns from complex datasets. AI encompasses both ML and DL, which function as subsets of AI. ML algorithms learn patterns from data to facilitate accurate predictions or decisions and can be categorized into supervised, unsupervised, and reinforcement learning. DL algorithms, drawing inspiration from the human brain, utilize deep neural networks to learn and extract patterns from large-scale datasets. DL has shown success in domains such as image and speech recognition, natural language processing (NLP), and autonomous driving.

In the last decade, technologies such as genomic sequencing have led to an exponential increase [1] in the data describing the molecular elements, structure, and function of biological systems. Additionally, data digitization and generation across varied fields such as physics, software development, and social media [2], has yielded complex datasets of scales previously unavailable to scientists. AI also provides many opportunities for healthcare, ranging from clinical decision-support systems to deep-learning based health information management systems. This abundance of data has played a pivotal role in the rapid progress of machine learning, deep learning, and artificial intelligence. As a result, we now have algorithms that can be trained to extract insights from data with a level of sophistication that closely resembles human intuition.

While researchers have developed hundreds of successful algorithms, there are currently a few overarching principles to systematically organize machine learning algorithms. In a seminal `proto-book` by Bronstein et al. [3], various systematization principles for different Artificial Neural Network (ANN) architectures and deep learning algorithms were presented. These principles are founded on the concepts of symmetry and mathematical group theory. Symmetry and invariance are central concepts in physics, mathematics, and biological systems. Since the early 20^{th} century, it has been established that fundamental principles of nature are rooted in symmetry [4]. The authors also introduced the concept of geometric deep learning and demonstrated how group theory, along with function invariance and equivariance principles, can serve as foundation for composing and describing different deep learning algorithms. Along these lines, the present manuscript explains the structure of ANNs and the core principles of machine learning algorithms. Additionally, it offers a review of the mathematical and statistical foundations pertinent to the development of artificial intelligence applications using bioinformatics data.

THE STRUCTURE OF ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS

We will begin by describing the structures and functions of deep learning and Artificial Neural Networks (ANNs), which form the foundation of artificial intelligence [5]. We use a dataset consisting of n pairs of $(x_i, y_i)_n$, where x_i represents n data points and y_i their corresponding labels. Each x_i data point can take the form of a number, a vector (an array of numbers), or a matrix (a grid of numbers), storing various types of bioinformatics data. The labels can assume different formats, such as binary (two-options), like $y_i = 1$ "inhibits cancer growth", or $y_i = 0$ "does not inhibit cancer". The labels can also be continuous numbers, for instance, $y_i = 0.3$ indicating 30% inhibition, or a composite label such as $y_i = (0, 1, 0)$, which signifies drug attributes like '0 - no inhibition', '1 - yes for toxicity', '0 - not metabolized', respectively. Similarly, the input data points can also be composite, for example, $x_i = (50, 100)$ representing two measuments for a single biological entity. Regardless of the label structure, the primary objective of deep learning algorithms and the overarching goal of artificial intelligence

applications in bioinformatics is to first train the ANN using data with known labels. Subsequently, the ANN is utilized to classify newly generated data by predicting their labels [6].

The simplest structure of an artificial neural network, as depicted in **Fig.1,** is "fully connected". In this structure, each neuron k within the ANN possesses a specific number of incoming and outgoing connections. These connections correspond to the quantity of neurons present in the previous and next layers within the neural network [6]. For example, the neuron $k_1^{(1)}$ of the *First Layer (1)* on **Fig.1**, which has $n = 2$ incoming and $n = 3$ outgoing connections. These connections align with the "input layer", which comprises two neurons, and the three connections extend to the neurons of the internal ("hidden layer") denoted as *Second Layer (2)* in the figure. The designation “hidden” is attributed to the internal layers because they do not directly receive input data.

This concept parallels the behavior of neurons engaged in cognition within animal brains, in contrast to sensory neurons. While the number of neurons in the hidden layers can vary based on the complexity of the label classification problem that the ANN is intended to address [7], the input layer must have a precise number of neurons that align with the structure of the input data. In **Fig. 1,** for instance, there are two input neurons, and the data can take the form $x_i = (50, 100)$. Lastly, the output layer consists of a number of neurons corresponding to the count of labels y_i associated with each input data point in the dataset. In **Fig. 1,** a single label is presented.

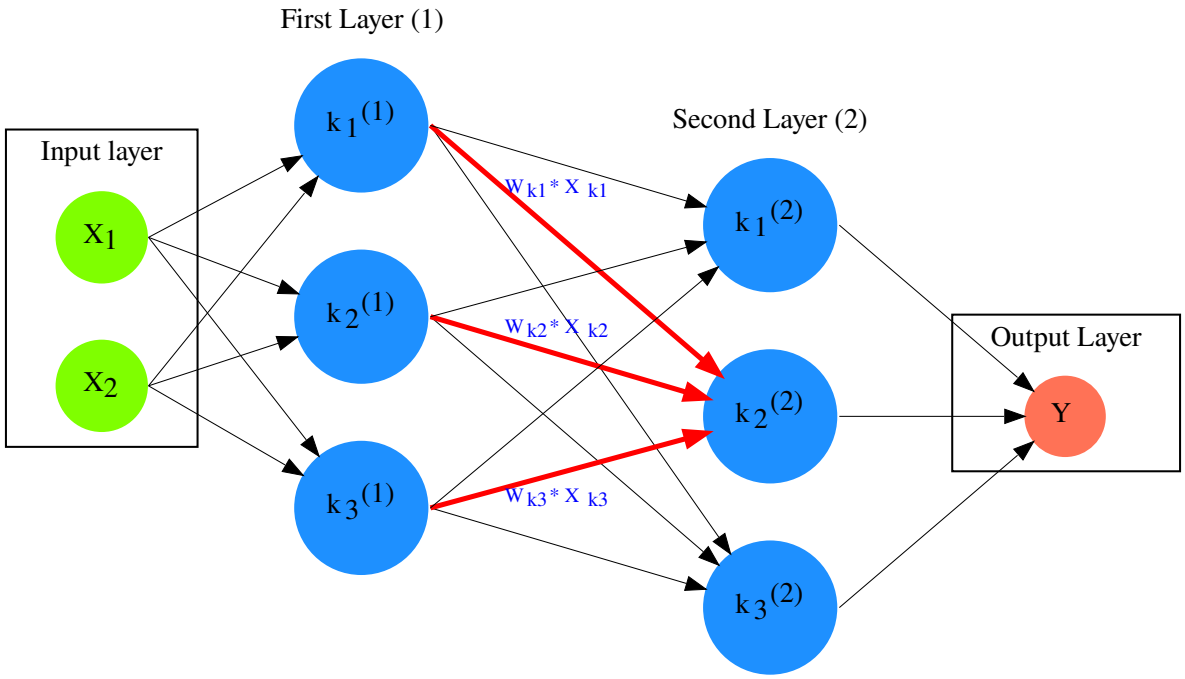


Figure 1. An example **Artificial Neural Network (ANN)**. The signal aggregation taking place on the second neuron $\sigma_{k_2^{(2)}}$ of the second hidden layer, can be expressed with the formula $\sigma_{k_2^{(2)}} = \sum_{k_{1,2,3}}^{(1)} w_{k1} * x_{k1} + w_{k2} * x_{k2} + w_{k3} * x_{k3} - b$, which is the aggregation of neuron signals from the first layer, shown as red arrows in the figure. b represents the threshold that needs to be overcome by the aggregation sum in order for the neuron to fire, and then the neuron will transmit a signal along the line shown towards the output on the final layer of the figure. The reader should refer to the text for more details. ”’

Similar to neural networks in animal brains, the computational abstractions used in machine learning

and artificial intelligence model neurons as computational units that execute signal summation and threshold activation [8]. Specifically, each artificial neuron performs a summation of incoming signals from its connected neighbooring neurons in the preceding layer on the network, shown for example as red arrows on **Fig.1** for $\sigma_{k_2^{(2)}}$. The signal processing throughout the ANN transitions from the input data x_i on the leftmost layer (**Fig.1**) to the output of data labels y_i on the rightmost end. Within each neuron, when the aggregated input reaches a certain threshold, the neuron "fires" and transmits a signal to the subsequent layer.

The signals entering the neuron can either be the data directly from the input layer or signals generated by the activation of neurons in the intermediate "hidden" layers. The summation and thresholding computation within each neuron is represented with the function $\sigma_k = \sum_1^k w_k * x_k - b$, where w_k represents the connection weights of the preceding neurons. Each connection arrow in **Fig.1** has a distinct weight, such as, for example, x_{k1} which is the incoming signal from the neuron $\sigma_{k_1^{(1)}}$ to neuron $\sigma_{k_2^{(2)}}$, multiplied by the weight w_{k1}, which symbolizes the strength of the connection between these two artificial neurons.

The weights in artificial neural networks embody the strength of connections between neurons. They determine the impact of input signals on the final output of the network. Throughout the training process, these weights are adjusted to minimize the difference between the network's predicted and intended output. The weights govern the information flow within the network, enabling it to learn and generate precise predictions. Accurately calibrated weights are crucial for the network to effectively learn patterns and extrapolate its knowledge to novel input data [8].

For the majority of applications, the weight values w_k constitute the only elements in the ANN structure that are variable and adjusted by the algorithms during training using the input data. This process is similar to the biological brain, where learning takes place by strengthening connections among neurons [9]. However, unlike the biological brain, the ANNs used for practical data analysis have fixed connections between neurons and the structure of the neural network remains unaltered during the process of training and learning to recognize and classify new data. The last term b in the summation signifies a threshold that must be surpassed, as in $\sum_1^k w_k * x_k > b$, to trigger the activation of a neuron.

A final step prior to transmitting the neuron's output value involves the application of a "logit" function to the summation value that is represented as $\varphi(\sigma_k)$. φ can be selected from a range of non-linear functions contingent on the type of input data and the specific analysis and data classification domain for which the ANN will be used [5]. The value of the logit function is the output of the neuron, which is transmitted to its interconnected neurons in the subsequent layer through outgoing connections, illustrated as an arrow in **Fig.1** and corresponding to the brain cell axons in the biological analogy. Multiple layers of interconnected neurons (**Fig.1**), along with multiple connections per layer, each having its own weight w_k, together form the framework of the Artificial Neural Network (ANN).

From a mathematical formalism perspective, a trained ANN is a function f that predicts labels y_{pred_i}, which can include categories such as 'no inhibition', 'yes for toxicity' etc., for different types of input data x_i, ranging from histology images to drug molecules represented as graph data structures. Therefore, the ANN undertakes data classification by operating as a mapping function $f(x_i) = y_{\text{pred}_i}$, that connects the input data to the respective labels. Furthermore, the $f(x_i)$ is a non-linear function, since it is an aggregate composition of the non-linear functions $\varphi(\sigma_k)$ of the individual interconnected neurons within the network [5]. As a result, the $f(x_i)$ can successfully classify labels for data inputs originating from complex data distributions. This fact enables ANNs to attain heightened analytical capability compared to conventional statistical learning algorithms [10]. The $f(x_i)$ estimation is carried out by fitting a training dataset, which establishes correlations between labels y_i and data points x_i. With hundreds of papers and monographs that were written on the technical details of training ANNs, we will next attempt to briefly summarize the process and direct the reader to provided citations for further details [11].

As mentioned earlier, the only variable elements in the ANN structure are the weights w_k of neuron connections. Therefore, training an ANN to classify data involves the estimation of these weights. Furthermore, the training process entails minimizing the error E, which is the difference between the labels y_{pred_i} predicted by the function f and the true labels y_i. This error metric is akin to true/false positive and negatives (precision and recall) used in statistics, however, different formulas are used for its estimation when dealing with multi-label or complex input data for the ANN (for further details, refer to [12]). The estimation of neuron connection weights w_k is executed by the algorithm through fitting the network function f to a large training dataset of $\{x_i, y_i\}_i^n$ pairs of input data and labels, while the error E is calculated by using a subset of the data for testing and validation purposes. The training algorithm starts with an initial value of the weights, and then performs multiple cycles, referred to as "epochs", to estimate the function f. This is achieved by fitting the data x_i to the network and calculating the error E by comparison between the predicted y_{pred_i} and the true labels y_i. At the end of each cycle, a process called "backpropagation" is performed [10], which involves a gradient descent optimization algorithm, which fine-tunes the weights of individual neurons to minimize E.

The gradient descent [13] optimization examines a large subset of all possible combinations of weight values, yet as a heuristic algorithm, it minimizes E, but cannot reach zero error. Upon the completion of multiple training cycles, the training algorithm identifies a set of weights that best fit the data with minimal error. The ANN settles on the optimal values that estimate each $\varphi(\sigma_k)$ function for $\sigma_k = \sum_1^k w_k * x_k - b$, where w_k is the weight in each interconnected neuron. Consequently, the overall function f represented by the network is also estimated, as it comprises the composition of the individual $\varphi(\sigma_k)$ neuron functions, as mentioned earlier. Following the completion of the artificial neural network training, where the most optimal set of weights is determined, the network is ready to be used for label prediction with new, unknown x_i data.

ARTIFICIAL INTELLIGENCE, GROUP THEORY, SYMMETRY AND INVARIANCE

Data domains in relation to group theory and symmetry

In the remaining sections, we will examine how the principles of group theory, symmetry, and invariance provide a foundational framework for comprehending the function of machine learning algorwthms. Furthermore, the classifying power of ANNs, particularly in relation to statistical variance, transformations, and non-homogeneity in the input data. In broad terms, symmetry entails the analysis of geometric and algebraic mathematical structures and finds applications across different research fields, including physics, molecular biology, and machine learning. A core concept in symmetry is invariance, which, in our context, is changing data coordinates, such as relocating a drug molecule in space or shifting the position of a cancer histology tissue sample, while maintaining the shape of the object unchanged [3]. Following such an alteration, which will be formally defined later in this text as an *invariant transformation*, it becomes imperative for the machine learning algorithms and ANNs to be capable of identifying a drug molecule even after rotation or recognizing cancerous tissue from a shifted histology image.

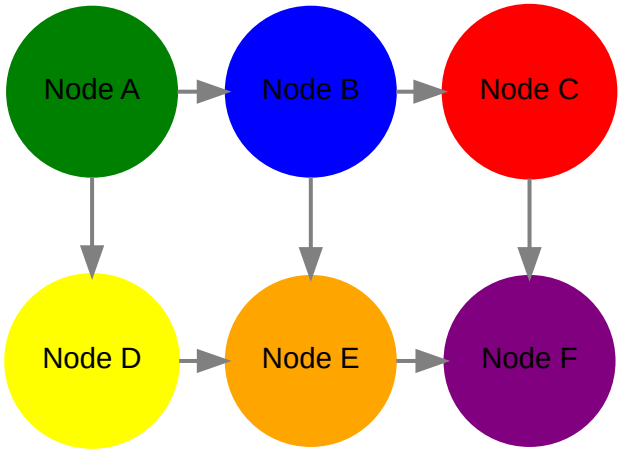
In order to link the abstract symmetry concepts with data classification in machine learning, as per the terminology of Bronstein et al., we consider the input data x_i to originate from a symmetry domain denoted as Ω. This Ω serves as the foundational structure upon which the data are based, and it is upon this domain structure that we train artificial neural networks to undertake classification, employing the label prediction function f as mentioned in the earlier section. For example, microscopy images are essentially 2-dimensional numerical grids of n x n pixels (**Fig.2a**), with each pixel having an assigned value corresponding to the light intensity captured when the image was taken.

In this scenario, the data domain is a grid of integers ($\mathbb{Z}$), represented as $\Omega : \mathbb{Z}_n \times \mathbb{Z}_n$. Similarly, for color images, the data domain is $x_i : \Omega \to \mathbb{Z}_n^3 \times \mathbb{Z}_n^3$, encompassing three overlaid integer grids that individually represent the green, bluem and red layers composing the color image [14]. In either case,

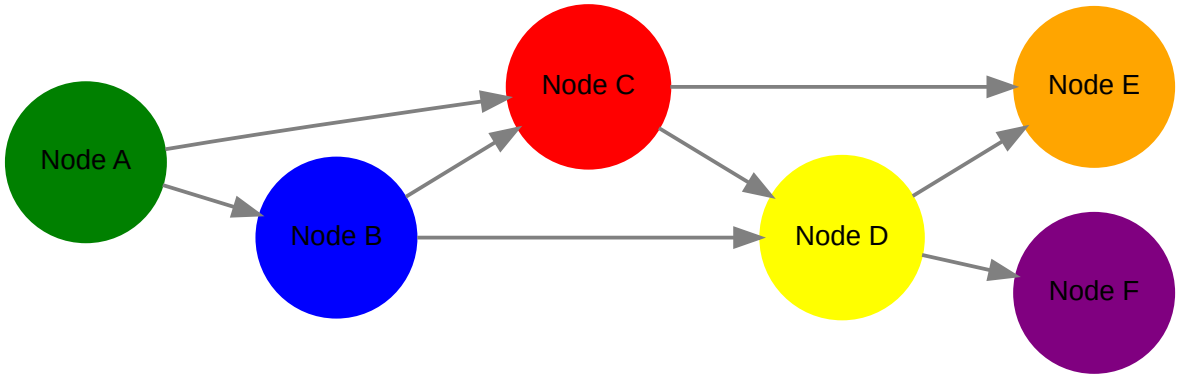
the Ω contains all possible combinations of pixel intensities, while the specific pixel value combinations of the images in the input data x_i are a "signal" $\mathrm{X}(\Omega)$ from the domain. The ANN's data classification and label prediction function $y_{\mathrm{pred}_i} = f(x_i)$ is applied upon the signal $\mathrm{X}(\Omega)$, which fundamentally constitutes a subset of the domain Ω.

A *symmetry group* G contains all possible transformations of the input signal $\mathrm{X}(\Omega)$, referred to as symmetries g or *group actions*. A symmetry transformation g preserves the properties of the data; for instance, it ensures that objects within an image remain undistorted during rotation. The constituents of the symmetry group, denoted as $g \in G$, are the associations of two or more coordinate points $u, v \in \Omega$ on the data domain (grid in our image example). Between these coordinates, the image can undergo rotation, shifting or other transformations without any distortion.

Consequently, the key aspect of the formal mathematical definition of the group lies in its capacity to safeguard data attributes during object distortions that frequently occur during the experimental acquisition of bioinformatics data. The concept of symmetry groups is important for modeling the performance of machine learning algorithms, particularly for classifying the data patterns despite the variability inherently present within the input data.



a. grid data for image pixels



b. graph data structure for a protein or other molecule

Figure 2. (a). A *grid* data structure representing image pixels, is formally a *graph* **(b).** A *graph* $G = (V, E)$, is composed of *nodes* V shown as circles, and *edges* connecting the nodes and shown as arrows. It can represent a protein, where the amino acids are the nodes and the peptide bonds between amino acids are the edges.

Another important data structure within bioinformatics is a *graph* denoted as $G = (V, E)$, composed of *nodes* V that signify biological entities, and *edges* representing connections between pairs of nodes (**Fig. 2b**). In a specific instance of a graph corresponding to a real-world object, the edges are a subset of all possible links between nodes. An example graph data structure for a biological molecule such a protein or a drug would portray the amino acids or atoms as node entities, while the chemical bonds between each of these entities are captured as edges. These edges could signify the carbonyl-amino (C-N) peptide bonds between amino acids and molecular interactions across the peptide chain on the protein structure, or the chemical bonds between atoms in a drug molecule [15].

Furthermore, attributes in the molecular data such as, for example, polarity, amino acid weight, or drug binding properties can be depicted as s - dimensional node attributes, where s represents the attributes assigned to each node. Similarly, edges or even entire graphs can have attributes, for experimental data measured on the molecular interactions represented by the edges, and measurements of the properties of the complete protein or drug. Finally, from an algorithmic perspective, images can be viewed as a special case of graphs in which the pixels serve as nodes, interconnected by edges following a structured pattern that generates a grid formation (**Fig.2a**) representing the adjacent positions of the pixels.

Group theory and symmetry principles applied to machine learning

Having established the mathematical and algorithmic parallels between graphs and images, we will now utilize the principles of the *symmetry group* G to examine the analytical and classification power of machine learning ANNs, with respect to data variability and transformations. Whether it involves data types like input images or molecules represented as graphs, which may undergo shifts or rotations, we introduce the concept of invariance guided by the principles of group theory and symmetry. These foundational mathematical and algorithmic formalisms serve as the basis for modeling the performance and output of machine learning algorithms, specifically ANNs, with regard to the diversity present in the dataset.

Consecutively, these principles can be extrapolated and generalized to encompass other types of data beyond graphs and images, for which ANNs are trained to predict and categorize. While we present the group and symmetry definitions following a data-centric approach, we will remain consistent with the mathematical framework, while describing how the group operations can effect transformations on the input data. Furthermore, different types of data may have the same symmetry group, and different transformations could be performed through identical group operations. For example, an image featuring a triangle, which essentially is a graph with three nodes, might possess the same rotational symmetry group as a graph with three nodes or a numerical sequence of three elements.

When chemical and biological molecules are represented as graphs as described earlier, the nodes V can be in any order depending on how the data were measured during the experiment. However, this variation does not change the underlying information contained in the data. As long as the edges **E,** which represent the connections between molecules, remain unchanged, we maintain an accurate representation of the molecular entity, irrespective of the sequence of nodes in $\mathbf{V}$. In cases where two graphs portraying the same molecule have identical edges but differ in node arrangement, they are called *isomorphic*. It is crucial that any machine learning algorithm designed for pattern recognition on graphs, should not depend on the ordering of nodes. This ensures that classification using ANNs and artificial intelligence remain robust against variations in experiment measurement encountered in real-world data [16]. This is something that is taken for granted with human intelligence, where, for example, we can recognize an object even when a photograph is rotated at an angle.

Invariance and the classification power of artificial neural networks

Returning to our earlier formal definitions of ANNs as function estimators fitted to the data, in order for ANNs algorithms to equivalently recognize *isomorphic* graphs, the functions $\varphi(\sigma_k)$ and overall $f(x_i)$ of the ANN acting on graph data should be *permutation invariant*. This implies that for any permutation

of the input dataset, the output values of these functions remain unchanged, regardless of the ordering of the nodes $\mathbf{V}$. This concept can be similarly applied to images, which, as previously mentioned, are specialized instances of fully connected graphs. Furthermore, these principles can also be generalized for other data types beyond images or graphs.

To further formalize the concept of invariance, and considering that both image and graph examples are essentially points on a grids on a two-dimemensional plane, we can use linear algebra. Specifically, by using a matrix we can represent the data transformations as group actions, denoted by g, within the symmetry group G. The use of matrices enables us to connect the group symmetries with the actual data by performing matrix multiplications that modify the coordinates of the object and consecutively represent the data transformations through the multiplication. The dimensions of the matrix, $n \times n$, typically are similar to these of the signal space $\mathrm{X}(\Omega)$ for the data (e.g., $\mathbb{Z}_n \times \mathbb{Z}_n$ images). The matrix dimensions not depend on the size of the group (i.e. the number of possible symmetries) or the dimensionality of the underlying data domain Ω. With this definition in place, we can formalize symmetries and group actions for modifying data objects, employing matrix and linear transformations as the foundation for connecting invariance in relation to variability in the data.

We will now conclude by establishing the mathematical and linear algebra formalisms that underlie the resilience of ANNs and machine learning algorithms in pattern recognition, considering transformations in the data. While our framework is based on a two-dimensional grid data domain Ω, the formalisms developed here can also be extrapolated to any number of dimensions or data formats without loss of generality. First, we will connect matrices to group actions g (such as rotations, shifts) within the symmetry group $g \in G$ by defining a function θ that maps the group to a matrix as $\theta : G \rightarrow \mathbf{M}$. As mentioned earlier, a matrix $\mathbf{M} \in R^{n \times n}$ consisting of numerical values (integers, fractions, positive and negative), when multiplied by the coordinate values of an object on the plane Ω, results in rotation or shifts of the object's coordinates for the exact amount corresponding to the group action within the symmetry group.

With these definitions in place, we will now connect the matrix formalisms with the neural network estimator function $y_{\mathrm{pred}_i} = f(x_i)$, which is identified by adjusting neuron connection weights during multiple training cycles with the input data. Our goal is to leverage the mathematical formalisms of group symmetry and invariance to establish the resilience of ANNs in classifying and assigning labels to new data points [17]. These data points originate from real-world data that might contain tranformations and distortions. First, we define the estimator function of the ANN to be *invariant* if the condition for the input data holds, i.e. $f(\mathbf{M} \times x_i) = f(x_i)$ for all matrices $\mathbf{M}$ representing the actions $g \in G$ within the symmetry group.

This formula encapsulates the requirement for the neural network function to be invariant: its output value remains the same whether the input data x_i are transformed or not (e.g., an image or graph is not rotated on the plane), as represented by the matrix multiplication $\mathbf{M} \times x_i$. Therefore, the output values $y_{\mathrm{pred}_i} = f(x_i)$ produced by the ANN, which essentially represent predicted output labels (e.g., y_{pred_i} = potent drug / not potent), based on the input data, exhibit resilience to noisy and deformed real-world data when the network estimator function is invariant. In a different case, the estimator function approximated by the ANN can be *equivariant* and defined as $f(\mathbf{M} \times x_i) = \mathbf{M} \times f(x_i)$. This signifies that the output of the ANN will be modified, but the label prediction result will shift equally alongside the shift in the input data.

Neural networks and group theory in relation to continuous data transformations

Up to this point, we have exclusively discussed discrete tranformations in linear algebra terms, utilizing matrix multiplications that lead to coordinate shifts and rigid transformations of the data, like rotating an image or graph by a specific angle on the grid Ω. However, in real-world data scenarios, we often also encounter continuous, more fine-grained shifts. In such cases, ANNs algorithms should be able to recognize patterns, classify, and label the data without any loss of performance [18]. Mathematically, the

continuous transformations follow equally with the invariant and equivariant functions described earlier. For instance, if the domain Ω contains data with smooth transformations and shifts, such as moving images (videos) or shifts of molecules and graphs that maintain *continuity* in a topological definition [19], in this case we deal with a concept known as *homeomorphism* instead of *invariance*.

Finally, if the rate of continuous transformation of the data is quantifiable, meaning that the function θ, which maps the group to a matrix, is *differentiable*, then the members of the symmetry groups will be part of a *diffeomorphism*. As it follows from the principles of calculus, in this case, infinitely multiple matrices $f((M)$ will be needed to be produced by θ for the continuous change of the data coordinates at every point. These differentiable data structures are common with manifolds, which, for example, could be used to represent proteins in fine detail. In this case, the molecule would be represented as a cloud with all atomic forces surrounding the structure, as opposed to the discrete data structure of nodes and edges in a graph. Finally, if the manifold structure also includes a metric of *distance* between its points to further quantify the data transformations, in this case, we will have an *isometry* during the transformation due to a group action from the symmetry group.

APPLICATIONS OF ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS IN BIOINFORMATICS

Artificial Intelligence (AI) and Deep Learning have emerged as powerful tools with diverse applications in the field of bioinformatics, and multiple research studies have been reported in the literature [20], [21], [22], highlighting the potential of the technology to revolutionize healthcare and life sciences. One of the significant applications is drug discovery, as AI algorithms facilitate the analysis of large datasets of chemical compounds, predicting their effectiveness and safety [23], [24], [25]. These studies have demonstrated that AI can accelerate the drug discovery process by screening potential candidates and optimizing their properties, resulting in substantial cost and time savings.

In the field of genomics, AI algorithms have been applied to the analysis of DNA sequencing and gene expression data, facilitating, for example, the identification of disease-causing mutations and enhancing our understanding of genetic variations [26], [27], [28], [29]. Moreover, in these studies, genomic data analysis with AI algorithms has provided critical insights, which can assist in the development of personalized medicine approaches and as result tailor treatments to individual patients. Consecutively, the use of AI algorithms in bioinformatics can contribute to the advancement of precision medicine. By integratively analyzing also other omics data (e.g., transcriptomics, proteomics, metabolomics), patient data, encompassing genetic information, medical history, and lifestyle factors, AI-driven insights can lead to improved predictions of drug responses, identification of potential side effects, and the recommendation of optimal treatment options for individual patients.

This personalized medicine approach can also involve enhancing patient care and treatment outcomes, through disease diagnosis improved by machine learning analysis of medical images, including computed tomography (CT) and magnetic resonance imaging (MRI) scans, X-rays, and histopathology images, of diseases like cancer [30], [31], [32], [33]. The AI algorithms can assist pathologists and radiologists in rendering precise diagnoses, enabling early detection and diagnosis, and ultimately contributing to overall improvements in patient outcomes.

AI can also play a significant role in assisting the development of bioinformatics tools and software accelerating the process of code development for the analysis and interpretation of biological data, such as sequence alignment, protein structure prediction, and functional annotation [34], [35], [36]. Furthermore, AI-powered natural language processing techniques have been employed to analyze scientific literature, patents, and clinical trial reports. This capability enables researchers to stay updated about the latest discoveries and facilitates knowledge discovery in the field.

Finally, in the area of clinical trials, machine learning algorithms have been appplied to mine vast amounts of data from clinical trials. As a result, the rates of success for new drugs and treatment strategies have improved for patients partipating in the trials [37], [38]. Additional studies have also

demonstrated that machine learning algorithms can result in enhanced optimization of clinical trial designs, reduction in costs, and an overall acceleration of the drug development pipelines [23], [24].

CONCLUSION

The rapid advancements in the fields of Machine Learning and Artificial Intelligence in recent years have exerted a substantial influence in the field of Bioinformatics. With these accelerated developements, the chance to systematically categorize algorithms and their corresponding applications, along with their perfomance across various types of bioinformatics data, has diminished. By harnessing the mathematical formalisms of symmetry and group theory, we can establish the operational principles of Artificial Intelligence algorithms concerning bioinformatics data. This not only paves the way for a deeper understanding of their functionality but also provides insights into the directions for future development in the field.

Funding Information: This work has been supported by Award Number U54 CA221704(5) from The National Cancer Institute.

Author Contributions: K. Krampis wrote the manuscript and performed the research. C. Wultsch provided overview during the development of the research and the manuscript. E. Ross, O. Ogunwobi, G. Ma and R. Mazumder contributed to the development of the research and provided feedback during the development of the manuscript.

Conflict of Interest: The authors declare no conflicts of interest.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: No data were generated as part of the present review paper.

Acknowledgments: The authors would like to thank their respective institutions for supporting their scholarly work.

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

[1] K. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister, and C. O'Sullivan, "The sequence read archive: a decade more of explosive growth," *Nucleic acids research*, vol. 50, no. D1, pp. D387–D390, 2022.

[2] L. Clissa, "Survey of Big Data sizes in 2021." 2022.

[3] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv preprint arXiv:2104.13478*, 2021.

[4] E. Noether, "Invariante variationsprobleme, math-phys," *Klasse, pp235-257*, 1918.

[5] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan, and X. Gao, "Deep learning in bioinformatics: Introduction, application, and perspective in the big data era," *Methods*, vol. 166, pp. 4–21, 2019.

[6] T. M. Nair, "Building and Interpreting Artificial Neural Network Models for Biological Systems.," *Methods in molecular biology (Clifton, N.J.)*, vol. 2190, pp. 185–194, 2021, doi: 10.1007/978-1-0716-0826-5_8.

[7] M. Uzair and N. Jamil, "Effects of hidden layers on the efficiency of neural networks," in *2020 IEEE 23rd international multitopic conference (INMIC)*, 2020, pp. 1–6.

[8] V. Renganathan, "Overview of artificial neural network models in the biomedical domain.," *Bratislavske lekarske listy*, vol. 120, no. 7, pp. 536–540, 2019, doi: 10.4149/BLL_2019_087.

[9] M. Wainberg, D. Merico, A. Delong, and B. J. Frey, "Deep learning in biomedicine," *Nature biotechnology*, vol. 36, no. 9, pp. 829–838, 2018.

[10] B. Tang, Z. Pan, K. Yin, and A. Khateeb, "Recent advances of deep learning in bioinformatics and computational biology," *Frontiers in genetics*, vol. 10, p. 214, 2019.

[11] J. Zou, Y. Han, and S.-S. So, "Overview of artificial neural networks.," *Methods in molecular biology (Clifton, N.J.)*, vol. 458, pp. 15–23, 2008, doi: 10.1007/978-1-60327-101-1_2.

[12] N. Kriegeskorte and T. Golan, "Neural network models and deep learning," *Current Biology*, vol. 29, no. 7, pp. R231–R236, 2019.

[13] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[14] G. Chartrand *et al.*, "Deep Learning: A Primer for Radiologists.," *Radiographics : a review publication of the Radiological Society of North America, Inc*, vol. 37, no. 7, pp. 2113–2131, 2017, doi: 10.1148/rg.2017170077.

[15] N. Kriegeskorte and T. Golan, "Neural network models and deep learning.," *Current biology : CB*, vol. 29, no. 7, pp. R231–R236, Apr. 2019, doi: 10.1016/j.cub.2019.02.034.

[16] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research.," *Journal of pharmaceutical and biomedical analysis*, vol. 22, no. 5, pp. 717–727, Jun. 2000, doi: 10.1016/s0731-7085(99)00272-1.

[17] A. Eetemadi and I. Tagkopoulos, "Genetic Neural Networks: an artificial neural network architecture for capturing gene expression relationships.," *Bioinformatics (Oxford, England)*, vol. 35, no. 13, pp. 2226–2234, Jul. 2019, doi: 10.1093/bioinformatics/bty945.

[18] L. G. Wright *et al.*, "Deep physical neural networks trained with backpropagation.," *Nature*, vol. 601, no. 7894, pp. 549–555, Jan. 2022, doi: 10.1038/s41586-021-04223-6.

[19] W. A. Sutherland, *Introduction to metric and topological spaces*. Oxford University Press, 2009.

[20] M. Lee, "Recent Advances in Deep Learning for Protein-Protein Interaction Analysis: A Comprehensive Review," *Molecules*, vol. 28, no. 13, Jul. 2023.

[21] M. Wysocka, O. Wysocki, M. Zufferey, D. Landers, and A. Freitas, "A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data," *BMC Bioinformatics*, vol. 24, no. 1, p. 198, May 2023.

[22] B. Jahanyar, H. Tabatabaee, and A. Rowhanimanesh, "Harnessing Deep Learning for Omics in an Era of COVID-19," *OMICS*, vol. 27, no. 4, pp. 141–152, Apr. 2023.

[23] F. W. Pun, I. V. Ozerov, and A. Zhavoronkov, "AI-powered therapeutic target discovery," *Trends Pharmacol Sci*, Jul. 2023.

[24] G. Floresta, C. Zagni, V. Patamia, and A. Rescifina, "How can artificial intelligence be utilized for de novo drug design against COVID-19 (SARS-CoV-2)?," *Expert Opin Drug Discov*, pp. 1–4, Jul. 2023.

[25] Y. Zhou *et al.*, "Deep learning in preclinical antibody drug discovery and development," *Methods*, Jul. 2023.

[26] A. rez-Mena, E. n, M. J. Alvarez-Cubero, A. Anguita-Ruiz, L. J. Martinez-Gonzalez, and J. Alcala-Fdez, "Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression," *Comput Methods Programs Biomed*, vol. 240, p. 107719, Jul. 2023.

[27] W. Wei, Y. Li, and T. Huang, "Using Machine Learning Methods to Study Colorectal Cancer Tumor Micro-Environment and Its Biomarkers," *Int J Mol Sci*, vol. 24, no. 13, Jul. 2023.

[28] D. Shigemizu *et al.*, "Classification and deep-learning-based prediction of Alzheimer disease subtypes by using genomic data," *Transl Psychiatry*, vol. 13, no. 1, p. 232, Jun. 2023.

[29] Z. Mirza *et al.*, "Identification of Novel Diagnostic and Prognostic Gene Signature Biomarkers for Breast Cancer Using Artificial Intelligence and Machine Learning Assisted Transcriptomics Analysis," *Cancers (Basel)*, vol. 15, no. 12, Jun. 2023.

[30] R. Adam, K. Dell'Aquila, L. Hodges, T. Maldjian, and T. Q. Duong, "Deep learning applications to breast cancer detection by magnetic resonance imaging: a literature review," *Breast Cancer Res*, vol. 25, no. 1, p. 87, Jul. 2023.

[31] Y. Tong *et al.*, "Prediction of lymphoma response to CAR T cells by deep learning-based image analysis," *PLoS One*, vol. 18, no. 7, p. e0282573, 2023.

[32] L. R. Archila *et al.*, "Performance of an Artificial Intelligence Model for Recognition and Quantitation of Histologic Features of Eosinophilic Esophagitis on Biopsy Samples," *Mod Pathol*, p. 100285, Jul. 2023.

[33] Q. Li, A. Sandoval, and B. Chen, "Advancing spinal cord injury research with optical clearing, light sheet microscopy, and artificial intelligence-based image analysis," *Neural Regen Res*, vol. 18, no. 12, pp. 2661–2662, Dec. 2023.

[34] M. Santorsola and F. Lescai, "The promise of explainable deep learning for omics data analysis: Adding new discovery tools to AI," *N Biotechnol*, vol. 77, pp. 1–11, Jun. 2023.

[35] B. Waissengrin *et al.*, "Artificial intelligence (AI) molecular analysis tool assists in rapid treatment decision in lung cancer: a case report," *J Clin Pathol*, Jul. 2023.

[36] F. Hosseini, F. Asadi, H. Emami, and M. Ebnali, "Machine learning applications for early detection of esophageal cancer: a systematic review," *BMC Med Inform Decis Mak*, vol. 23, no. 1, p. 124, Jul. 2023.

[37] S. M. Ahmed, R. V. Shivnaraine, and J. C. Wu, "FDA Modernization Act 2.0 Paves the Way to Computational Biology and Clinical Trials in a Dish," *Circulation*, vol. 148, no. 4, pp. 309–311, Jul. 2023.

[38] A. Aliper *et al.*, "Prediction of clinical trials outcomes based on target choice and clinical trial design with multi-modal artificial intelligence," *Clin Pharmacol Ther*, Jul. 2023.