# Deep Learning for Reintegrating Biology

Rolf Müller[1]*, Jin Ping Han[2], Sriram Chandrasekaran[3], Paul Bogdan[4]

[1]Department of Mechanical Engineering, Virginia Tech
[2]IBM TJ Watson Research Center
[3]Department of Biomedical Engineering, University of Michigan, Ann Arbor
[4]Department of Electrical and Computer Engineering, University of Southern California

**Summary:** The goal of this vision paper is to investigate the possible role that advanced machine learning techniques, especially deep learning could play in the reintegration of various biological disciplines. To achieve this goal, a series of operational, but admittedly very simplistic, conceptualizations have been introduced: Life has been taken as a multidimensional phenomenon that inhabits three physical dimensions (time, space, and scale) and biological research as establishing connection between different points in the domain of life. Using this conceptualizations, fragmentation of biology can be seen as the result of too few and especially too short-ranged connections. Reintegrating biology could then be accomplished by establishing more, longer ranged connections. Deep learning methods appear to be very well suited for addressing this particular need at this particular time. Not withstanding the numerous unsubstantiated claims regarding the capabilities of AI, deep learning networks represent a major advance in the ability to find complex relationships inside large data sets that would have not been accessible with traditional data analytic methods or to a human observer. In addition, ongoing advances in the automation of taking measurements from phenomena on all levels of biological organization, continue to increase the number of large quantitative data sets that are available. These increasingly common data sets could serve as anchor points for making long-range connections by virtue of deep learning. However, connections within the domain of life are likely to be structured in a highly nonuniform fashion and hence it is necessary to develop methods, e.g., theoretical, computational, and experimental, to determine linkage of biological data sets most likely to provide useful insights on a biological problem using deep learning. Finally, specific deep learning approaches and architectures should be developed to match the needs of reintegrating biology.

## 1 Purpose

The challenge of reintegrating biology arises because biology has grown into a large, heterogeneous field of science that contains many distinct subdisciplines, each with a its own specific research questions, methodology, and terminology. This makes it hard to find common ground among the different biological subdisciplines. Nevertheless, all areas of biology are connected by common themes such as the rules of life and evolution. The tension between these strong common themes on the one side and the current state of fragmentation of the field raises the question whether it would be possible to reintegrate biology into a single coherent science again. The goal of this vision paper is to investigate the potential role of deep learning (DL) approaches in the reintegration of biology. To do so, the paper presents operational conceptualizations for the current fragmentation of biology and how reintegration of biology could be phrased as a deep learning problem.

## 2 Background

Life on earth is a multi-dimensional phenomenon. Its dimensions could be defined in several ways [19], in terms of basic physical dimensions one could pick time, space, and scale to organize biological structures and processes (Fig. 1a). For the sake of simplicity, these dimensions will be used in the current paper to illustrate the points discussed, but choosing a different set of dimensions should not affect the basic argument presented here. In addition to the multidemensionality of life, it is important to note that life on earth has spanned a very large range of values along each of the physical dimensions

*rolf.mueller@vt.edu

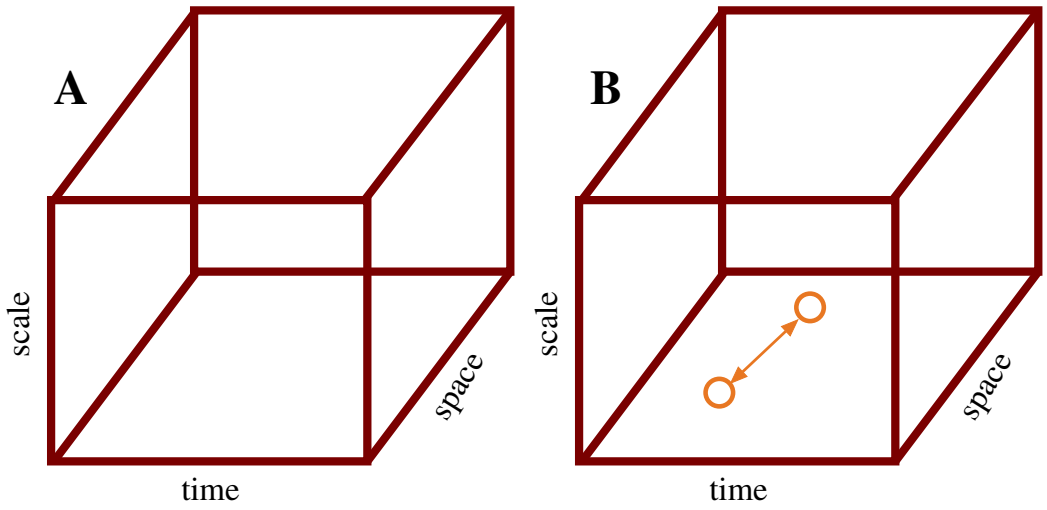


Figure 1: Conceptualization used for life and biological research: a) life as a multidimensional phenomenon spanning the physical dimensions of time, space, and scale; b) biological research as a way to establish connections between different points in the time-space-scale domain of life.

adopted here and this is likely to hold for other dimensions as well. In time, life has extended from its origin at least 3.5 billion years ago [24] to the presence and the future. In space, it covers the entire surface of the earth and the depths of the oceans [3]. Finally, the scales of life range from individual atoms to the earth's entire biosphere.

Based on the notion of life as a multi-dimensional phenomenon, the process of gaining new biological knowledge could be regarded as establishing linkages between different points in the joint time-space-scale domain of life (Fig. 1b). For example, one may want to predict if the population of a certain biological species will become (functionally) extinct in the future based on its current demographic composition and other circumstances [17]. Making such a prediction could be conceptualized as establishing a linkage along the time dimension (between the current and future status of the population) while maintaining fixed positions along the space and scale dimensions. Similarly explaining the clinical symptoms of a disease in terms of the underlying molecular mechanisms would be conceptualized as establishing a linkage along the scale dimension, but it could also involve changes along the time and space dimensions as the disease may develop over time and spread within the body of an individual or across a population of individuals.

## 3 Problem Statement: Fragmentation of Biology

In general, the difficulty of establishing relationships between points in the time-space-scale domain of life can be expected to increase with the distance between the points to be connected and the multiscale spatiotemporal complexity of interactions. For example, short-time effects tend to be more readily predictable than long-term changes and localized effects tend to be easier to understand than distributed effects. However, even steps that are small compared to the extent of the entire domain spanned by life can be very hard. For example, predicting the structure of a protein from its sequence of amino acids [9] bridges only a comparatively small distance along the dimensions of scale (molecular to macromolecular) and time (tenths of microseconds to seconds [2]). Nevertheless, making these connections is very computationally expensive, requires deep domain-specific knowledge, and still poses unresolved challenges [9].Moreover, due to existing bias in biological datasets towards manipulating only certain biological knobs and not all degrees of freedom, when designing the deep learning based discovery we also need to quantify the uncertainty and trustiness of these predictions.

Using the conceptualization of biological research as making connections between points across the time-space-scale domain of life, the current fragmentation of biology could be seen as a consequence of the difficulty associated with making such connections, especially long-ranging ones. Different biology research communities are often bogged down by trying to cross local boundaries in their respective subsections of life's time-space-scale domain and hence have neither the tools nor the capacity to reach out to biological phenomena and research communities at more distant points in the domain. The current fragmentation of biology can hence be conceptualized as a shortage of connections – especially long-range

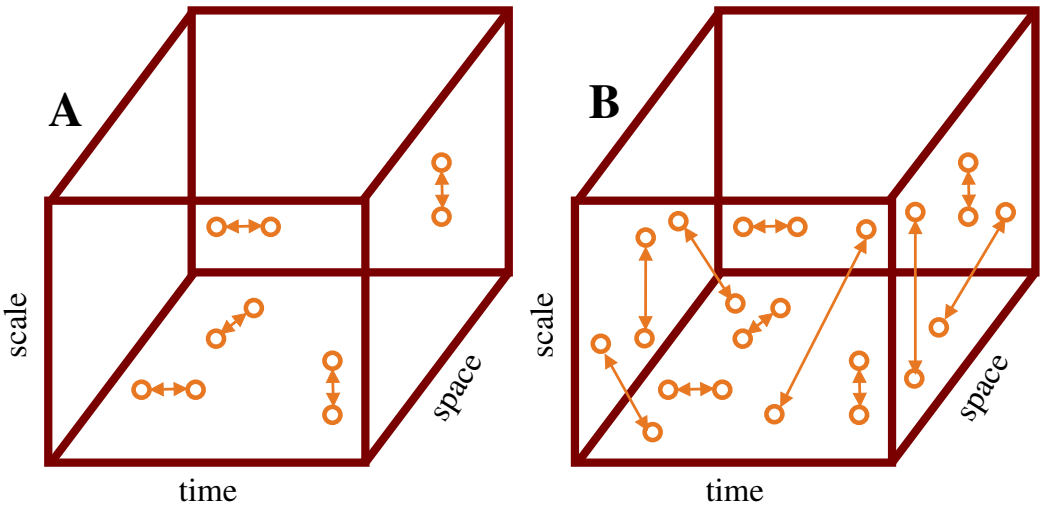


Figure 2: Conceptualization for the fragmentation and reintegration of biology: a) The current fragmented state of biology is the result of few and local connections; b) Reintegrating biology would be achieved by establishing more longer-ranging connections.

connections – within the time-space-scale domain of life (Fig. 2a). Using the same conceptualization, reintegrating biology could be achieved by establishing more and especially longer-ranging connections (Fig. 2b).

Finally it should be pointed out that in this view, the current fragmentation of the biological research community is not a purely social problem that is, e.g., due to the tendency of researchers to stay within a community of like-minded peers and use a specific terminology that isolates them from other life scientists. Instead, the root cause of this fragmentation can be traced back to current fundamental scientific limitations on making connections between more distant points in the domain of life. This does not preclude, of course, that various social and organizational factors also play important, amplifying roles in maintaining the current fragmentation of biology.

## 4 Transformative Opportunity for Reintegrating Biology

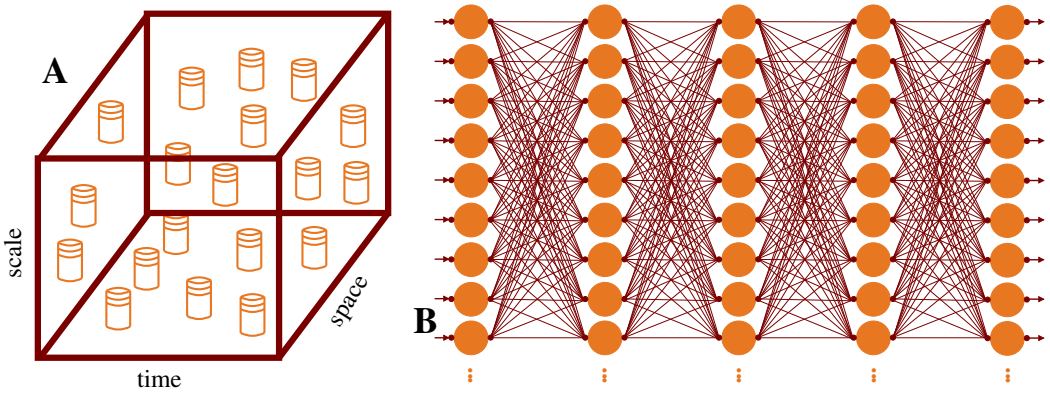


Figure 3: Convergence of "big data" and deep learning that could be critical for the reintegration of biology: a) The ongoing automation of biological data acquisition leads to data sets across the domain of life that can act as anchor points for learning connections between these points; b) Deep neural networks have the ability to learn complex relationships between their inputs and outputs and are hence well suited for establishing links between data sets across the domain of life.

With a dedicated effort, it should be possible to harness two important ongoing and developments in science and engi-

neering to create a transformative impact on the reintegration of biology. These two developments are (i) automation and (ii) the deep learning revolution [25]. Achieving a convergence between these two broader scientific developments within the life sciences could have a transformative impact on the goal of reintegrating biology.

Automation has found its way into many areas of biology that range from the smallest to largest scales of biological organization. For example, at the molecular scale, obtaining genetic sequence data is now largely automated and highly efficient [18]. At the organismal levels, there are many efforts to digitize the specimens in natural history collections [4]. At the population level, the advent of electronic health records and genotype-phenotype assays [6] provide millions of data points on variation between individuals in a population. At the ecosystems level, camera trapping [21] and automated acoustic monitoring devices [26, 13] are producing large amounts of image and audio data from ecosystems around the globe. As a result of all these efforts, "big data" sets have begun to appear at many different locations across the time-space-scale domain of life.

Where successful, such automated methods enable the collection of quantitative data sets that are much larger than would have been possible with their respective manual equivalents. However, such large data sets also pose a challenge since obtaining insights from large amounts of data requires sophisticated analysis methods. The DL revolution could address this challenge. It is driven by methodological progress [25] that has made it possible to design a large and diverse family of learning networks that are able to learn complex, i.e., nonlinear relationships between their respective inputs and outputs [16]. By achieving this, the DL research community has managed to vastly expand the ability of scientists in any area to discover patterns and find functional relationships in data that were not accessible by the much more restrictive classes of the previously existing linear and nonlinear learning methods.

One of the remaining key issues with DL methods is that they typically require very large amounts of data for training. Although there are many research efforts to remedy this situation and realize deep networks that can learn from small data, this need for training on large data sets remains a restriction that is unlikely to be completely overcome in the foreseeable future, Moreover, in order to overcome the variability and stochasticity that intrinsically characterizes biological systems and the existing bias or even errors that exist in the biological data acquisition, we need to endow the deep learning framework with capabilities for quantifying uncertainty and providing a degree of trust and confidence for each prediction. Hence, the convergence between data automation and the use of DL in biology could alleviate the shortcomings of each of the two methods and hence provide a powerful approach for addressing the root causes of the current fragmentation problem in biology. In this scenario, the increasing number of "big data" sets produced by virtue of automated methods could be seen as potential "anchor points" for efforts to learn new connections across the domain of life. DL methods could then be used to exploit these anchor points and test whether a link between any given pair of points can be established.

# 5 The Need for Guiding Insight

Besides automation and DL methods, a third component could be critical to enable the reintegration of biology. This additional component is insight that could be used to guide attempts to identify linkages across the time-space-scale domain of life. Such guiding insight is likely to be critical because of two factors: (i) the enormous size of the domain of life and (ii) the likelihood of a large variability in the strength of the linkages between different point pairs in this space. The latter point is based on the assumption that the domain of life has very pronounced inhomogeneities in its connectivity structure. This is to be expected because any linkages across the time-space-scale domain of life are the results of flows of matter or energy and with it information. Points that are connected – either directly or indirectly – by strong flows of matter or energy are likely to have likewise strong connections whereas those that do not will also not exhibit such linkages. In the spatial domain, for example, the estuary of a river can be expected to be linked strongly to its upstream regions because matter is being transported downstream. At the other end of these spatial examples would be ecosystems that are separated by strong geographical barriers such as large bodies of water, mountain ranges, or deserts. For example, one may wonder whether it would make sense to attempt a linkage between ancient DNA sequences from Siberia and recent camera trap data from the Amazon since the objects described by the two data sets are widely separated in time, space, and scale.

Given this situation, a successful reintegration of biology using anchor points provided by quantitative data sets and linkages established through DL will critically require guiding insight. The domain of life is likely way too large for a selection of possible anchor points that is solely based on trial and error. This situation could be handled by virtue of existing *a priori* knowledge and common sense, the efficient use of adaptive sampling strategies, and – perhaps – a specific theory that could be developed for this purpose. These points call for the development of a specific theory and methodology that could result in a better understanding of the "connectivity of life", i.e., the linkage structure of the time-space-scale domain of life. At present, it is not clear what such a theory could look like and developing it would most likely require an iterative

process in which trial-and-error experimental attempts would alternate with research efforts dedicated to formulating such a theory.

## 6 Customized DL for Reintegrating Biology

Research in DL has already produced a diverse set of different network architectures with proven advantages for specific problems [16]. Some well-known examples are convolutional neural networks for processing of images [15], recurrent neural networks for time-variant data (e.g., speech, [11]), reinforcement learning for establishing feedback control structures [14], networks with residual layers for dealing with vanishing gradients [12], and transfer learning to enable crossings between different learning and testing domains [22]. Based on the successes of these customized network architectures, it appears to be worthwhile to conduct research to determine whether DL architectures can be customized to match structure in the space-time-scale domain of life. For example, would it be possible/advantageous to come up with a network architecture that would take in data from an orderly set of points along the scale dimension into subsequent network layers? Of particular interest in this context would be deep multimodal learning paradigms [23] that could be a good fit to the multi-modal nature of many biological data sets.

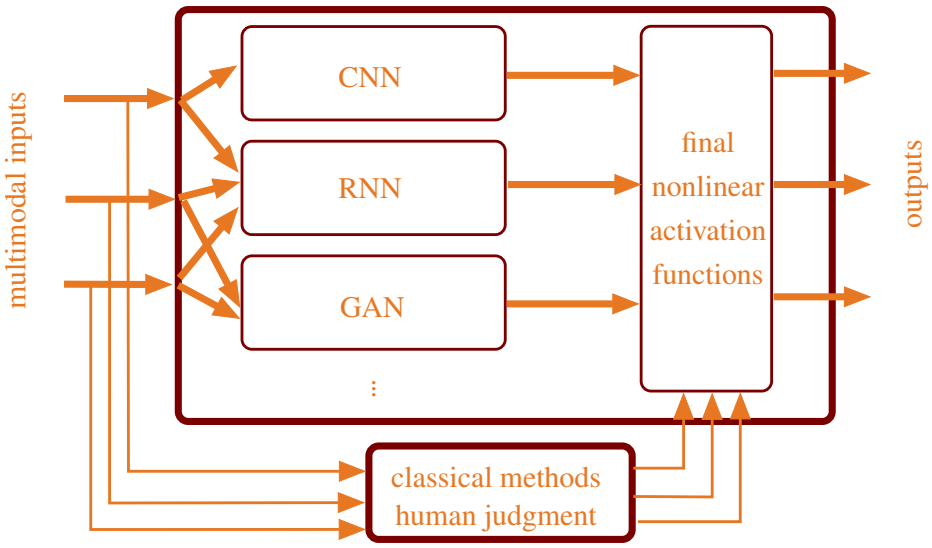


Figure 4: Schematic outline of a deep multimodal learning paradigm to deal with different biological data that also takes into account results from classical methods and human judgment.

Besides specific network architectures, another major thrust for developing DL for the purpose of reintegrating biology could be the adaptation of DL approaches that can provide insights into the nature of the relationships that the network has learned. In the DL community, these methods are often subsumed under the term "transparent AI" [7] and the basic idea is to analyze a deep network that has learned a useful relationship to make the nature of the learned relationship explicit. A common approach to improve interpretability of machine learning models is by creating hybrid models that incorporate mechanistic detail based on relevant physical and biological principles. For example, a 'white-box' machine-learning approach that integrates mechanistic metabolic modeling with generic machine learning has been used to understand the complex response of microbes to antibiotics [27]. Such hybrid models have also been used for integration of diverse data types. A probabilistic regulation of metabolism method, for example, has combined mechanistic biochemical modeling with data-driven probabilistic modeling to integrate thousands of transcriptomics and metabolic data set with growth phenotype measurements in microbes [8]. This integration of modeling methods creates a 'constrained' optimization problem where in the boundaries of AI/machine-learning are set forth by mechanistic rules driven by biophysical and biochemical laws and prior knowledge. Another strategy for increased interpretability of deep learning models involves using regularized autoencoders that provide a lower dimensional representation of complex datasets. This approach was used to de-noise and interpret single cell transcriptomics data generated using various technologies [1].

Given the large gaps that still remain in the coverage of the time-space-scale domain of life by scientific exploration, e.g., the large number of biological species that are not yet described scientifically [20], it would also be important to develop DL concepts that can deal with novelty and fill in gaps in data. Species that are not yet known to science could be discovered with the help of DL methods such as semi-supervised novelty detection [5]. In this case, the known data that surrounds the novel data point could be seen as a distributed cloud of anchor points that are to be connected to the novelty data point. Biological data sets with gaps due to missing data or sample sizes that are too small could be augemented using generative DL methods such as generative adversarial networks [10].

# 7 Conclusions and Possible Broader Impacts

Understanding the multidimensional, highly connected nature of life poses not only a key scientific challenge but is also likely of prime importance to ensure that the earth's biosphere can continue to support the survival of mankind. Despite all the impressive progress in understanding biological phenomena at various levels, the current fragmentation of biology poses the risk that critical relationships – especially long-ranging connections – will go unnoticed for too long. A pervasive use of the best available techniques for the discovery of more, long-ranging connections across the entire domain of life could mitigate this risk considerably and also lead to a much deeper understanding of life through time, space, and scale.

# References


[1] M. Amodio, D. Van Dijk, K. Srinivasan, W.S. Chen, H. Mohsen, K.R. Moon, A. Campbell, Y. Zhao, X. Wang, M. Venkataswamy, A. Desai, Ravi V., Priti K., Montgomery R., Wolf G., and Smita K. Exploring single-cell data with deep multitasking neural networks. *Nature Methods*, 16:1139–1145, November 2019.

[2] R. L. Baldwin. Why is protein folding so fast? *Proc. Natl. Acad. Sci. USA*, 93(7):2627–2628, 1996.

[3] D.H. Bartlett. Microbial life in the trenches. *Marine Technology Society Journal*, 43(5):128–131, 2009.

[4] R.S. Beaman and N. Cellinese. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys*, (209):7, 2012.

[5] G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(Nov):2973–3009, 2010.

[6] B.R. Bochner. Innovations: New technologies to assess genotype–phenotype relationships. *Nature Reviews Genetics*, 4(4):309, 2003.

[7] D. Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

[8] S. Chandrasekaran and N.D. Price. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. USA*, 107(41):17845–17850, 2010.

[9] K.A. Dill and J.L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[11] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] A.P. Hill, P. Prince, E. Piña Covarrubias, C.P. Doncaster, J.L. Snaddon, and A. Rogers. Audiomoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. *Methods in Ecology and Evolution*, 9(5):1199–1211, 2018.

[14] L.P. Kaelbling, M.L. Littman, and A.W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[15] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[17] T.M. Lee and W/ Jetz. Unravelling the structure of species extinction risk for predictive conservation science. *Proceedings of the Royal Society B: Biological Sciences*, 278(1710):1329–1338, 2010.

[18] E.R. Mardis. Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008.

[19] R.M. May. The dimensions of life on earth. *Nature and Human Society*, pages 30–45, 2000.

[20] C. Mora, D.P. Tittensor, S. Adl, A.G.B. Simpson, and B. Worm. How many species are there on earth and in the ocean? *PLoS Biology*, 9(8):e1001127, 2011.

[21] A.F. O'Connell, J.D. Nichols, and K.U. Karanth. *Camera traps in animal ecology: methods and analyses*. Springer Science & Business Media, 2010.

[22] S.J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[23] D. Ramachandram and G.W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[24] J.W. Schopf, A.B. Kudryavtsev, A.D. Czaja, and A.B. Tripathi. Evidence of archean life: stromatolites and microfossils. *Precambrian Research*, 158(3-4):141–155, 2007.

[25] T.J. Sejnowski. *The deep learning revolution*. MIT Press, 2018.

[26] R.S. Sousa-Lima, T.F. Norris, J.N. Oswald, and D.P. Fernandes. A review and inventory of fixed autonomous recorders for passive acoustic monitoring of marine mammals. *Aquatic Mammals*, 39(1), 2013.

[27] J.H. Yang, S.N. Wright, M. Hamblin, D. McCloskey, M.A. Alcantar, L. Schrübbers, A.J. Lopatkin, S. Satish, A. Nili, B.O. Palsson, Walker G.C., and J.J. Collins. A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell*, 177(6):1649–1661, 2019.