# MACHINE LEARNING FOR 21CM STUDIES AND COSMOLOGY

Michelle Gurevich

***Abstract.*** — Precise measurement of cosmological and astrophysical parameters is fundamental to a comprehensive theory of cosmology. This has traditionally required computationally-intensive numerical simulations to be run on scarce telescope resources, prompting researchers to seek new methods for their study. Recently, machine learning has emerged as a useful tool for constraining parameter space and dimensionality while allowing for a high degree of accuracy. Specifically, the methods of emulation and parameter estimation have proven particularly suited to studies of the 21cm signal. We investigate the benefits and shortcomings of both methods in this review, and suggest refinements as well as prospective applications.

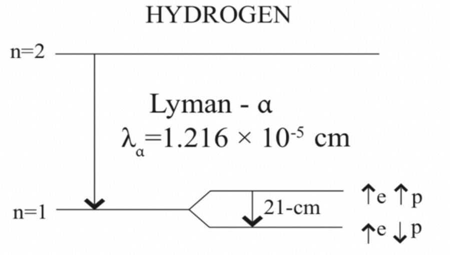
## Contents

***Supervisor.*** — Dr. Jonathan Pritchard.

## 1. Background on 21cm Cosmology

The first stars to form populated a bleak, cold universe, but their ultra-violet radiation ionized the surrounding gas medium (Loeb, 2010) and made it possible for new generations of stars to follow. Local perturbations in mass density led to halos where gravitational collapse coalesced matter into galaxies (Pritchard, 2012). At this point the universe had expanded and cooled from its previously opaque state (Peebles, 1994), and so as photons interacted with electrons and protons, they produced emission lines that could be observed. Though the early stars could theoretically be detected with powerful enough telescopes (Windhorst, 2018), their low relative luminosities make detection challenging. In order to learn about this epoch, known as "Cosmic Dawn," it is helpful to study instead the effect of such early luminous matter on its surrounding material.

Figure 1. Spin flip diagram of the 21cm line of neutral hydrogen (Loeb, 2010).



Due to its ubiquity as well as its sensitivity to the ionization level of the IGM (Gillet, 2019), hydrogen arises as a natural candidate. In fact, by studying the spectral lines produced when photons interact with neutral hydrogen, it is possible to create snapshots of the universe over the course of its evolution (Pritchard, 2012). The excitation of a neutral hydrogen atom by a photon releases energy in the form of a spectral line, which is seen as absorption or emission depending on its temperature relative to that of the local background. In Figure 1 this is illustrated in the $n = 1$ section where the spin flip of a proton releases a spectral line corresponding to a wavelength of 21cm.

As the universe expands, the regions from which these spectral lines emerged recede farther away from us, and the wavelength appears redshifted. This is described by:

$$z + 1 = \frac{\lambda_{observed}}{\lambda_{emitted}}, \tag{1}$$

where $z$ is the redshift (Loeb, 2010). From Eq. 1 it is clear to see higher redshift corresponds to the true wavelength being much smaller than the one observed, which is analogous to saying the wavelength was stretched while it travelled toward us. It is then possible to partition the universe at different times by comparing sections corresponding to wavelengths of $21cm \cdot (1 + z)$ (Loeb, 2010), to effectively map the history of the evolution of the early universe.

Variation in atomic hydrogen at these selected times provides insight into changes in the concentration and distribution of ionized regions, which in turn informs us about the composition and even the temperature of the intergalactic medium (Pritchard, 2012). From this, it is possible to better understand the conditions under which the first stars and galaxies formed. However, these distributions which resemble Gaussian fields during the inflation period, are subject to gravitational effects such as galaxy clustering (Loeb, 2010), and devolve over time into non-Gaussian fields (Ramanah, 2020). The cumulative effects of these perturbations in the gradient potential can be measured, but the modeling of these matter distributions is rendered computationally intensive (Kern, 2017). This is especially true where images of faraway, early galaxies are further distorted by weak gravitational lensing (Mootoovaloo, 2020). Small scales are also nonlinear; whereas density fields appear Gaussian at large scales ($\delta(x) \ll 1$ in Eq.3 below) , and can be statistically described by the power spectrum,

$$P(k) = (2\pi)^{-3}\langle| \delta_{\mathbf{k}}^2 |\rangle \\ \text{for } \delta_{\mathbf{k}} = \int d^3x\delta(x)e^{-i\mathbf{kx}}, \tag{2}$$

and over density,

$$\delta(x) = \frac{\rho(x) - \bar{\rho}}{\bar{\rho}}, \tag{3}$$

such is not the case for small scales ($\delta(x) \sim 1$ in Eq. 3) (Loeb, 2010). Consequentially, numerical simulation of small scales again prove costly.

There is an ongoing effort to use scarce telescope time as efficiently as possible, and to generate reasonable estimates of parameter values for when new radio telescopes capable of reading as-of-yet undetected signals are operational. Numerical simulations are very costly to run, even for optimized algorithms (Villaescusa-Navarro et al., 2020). The non-linear structure of these distributions often means tools such as traditional power spectrum analysis, while useful (Ribli, 2019; Schmit, 2017), are incomplete (Gillet, 2019). Thus, researchers in recent years have turned to alternative methods for computational and statistical modeling of the parameters describing the early Universe.

## 2. Machine Learning

21cm cosmology is a natural candidate for machine learning due to the high dimensionality and limited constraints of its parameter spaces (Kern, 2017), as well as the many opportunities for optimizing simulation run time. Furthermore, the usual concerns of machine learning, such as the black box nature of many algorithms (Buhrmester, 2019), are not inherently problematic in the context of 21cm.

### 2.1. Definition and basic architecture

Machine learning is the iterative application of an algorithm to a training data set such that predictions are generated without having been programmed explicitly, such that these predictions tend toward increased accuracy (Nichols, 2019). The figure provided in Appendix 1 illustrates the layers of a Convolutional Neural Network (CNN), a type of machine learning algorithm, used by Gillet et al. (2019) to predict the values of several astrophysical parameters (shown at the bottom of the figure). This is useful for conceptualizing the overarching structure of a machine learning algorithm, and emphasizes the repeated application of hidden layers responsible for feature extraction while also showing the interconnectedness between these and the outside layers.

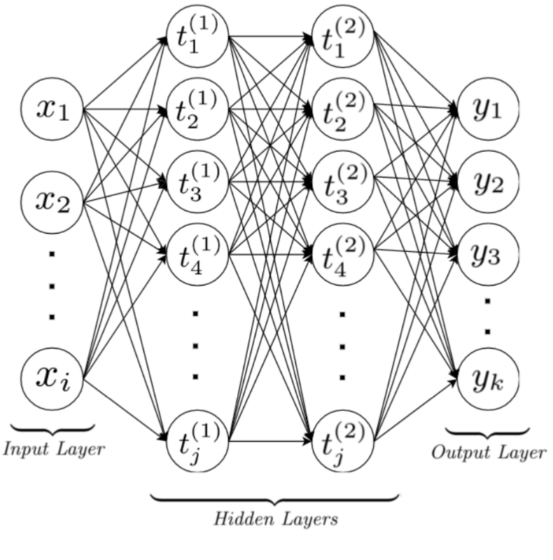


FIGURE 2. Schema for multilayer perceptron (MLP) network (Schmit 2017).

Additionally, Figure 2 visualizes the three layers in a neural network, these consisting of the input layer which defines some data nodes, the hidden layers which are adjusted iteratively and which describe the relationships between sets of nodes (Nichols, 2019) as weighted linear combinations, and the output layer which returns output values (Schmit, 2017). The weights associated to each vertical slice of nodes are adjusted by minimizing the mean square error distance of values predicted by the algorithm from those seen in the training data set. The training algorithm for the multilayer perceptron (MLP) design described in this paper is useful in situations where the training data set is sparse and parameter space has low dimension (Schmit, 2017).

There are two fundamental ways for the learning to transpire: supervised, which is suited to classification problems, and unsupervised, which has the model infer outcomes from data (Nichols, 2019). In order for the model to be trained, there must be some idea of minimizing the error at each stage of its run. This is defined by the loss function (Nichols, 2019), and in the case of 21cm studies tends to be some form of mean square error minimization. A notable variation on the loss function is discussed in subsection 3.1. The evaluation of models is centered in their ability to recreate results established by other, non machine learning, methods as well as to predict results from observational data. In addition to this, computational cost and application to problem statements where traditional approaches fail, are useful metrics for evaluating machine learning approaches. Machine learning in the context presented here is a tool with the purpose of advancing 21cm studies.

### 2.2. Motivation

Both the direct and inverse approaches discussed in this review have as their central objective the measurement of parameters (Pritchard, 2012) which in turn describe the large-scale structures of interest at specific points in the early universe's evolution. Researchers implementing machine learning in 21cm cosmology seek to statistically map the distribution of matter density over time (Villaescusa-Navarro et al., 2020), and in particular during the Cosmic Dawn, by refining those parameters' constraints. This is accomplished via a variety of approaches, and methods must be tailored to the specific context of each experiment. The result is an amalgam of previously established constraints and new adjustments, which is used to iteratively reduce error bars and further constrain parameter spaces. As such, it makes sense to adopt a piecewise approach when evaluating 21cm machine learning research.

There is an intrinsic pairing between emulation and parameter estimation, and many papers pursue either a forward-modeling (emulation) or inverse-modeling (parameter estimation) approach; the techniques can also be combined, though the scope would be tremendous. As such, the majority of papers discussed in this review pursue either emulation or parameter estimation, and these will be handled in separate sections. The results are then summarized in the conclusion.

### 2.3. Limitations

Though machine learning presents many exciting opportunities for 21cm, these are not without their own associated drawbacks.

The problem is not defined over a known structure that can be consistently, adequately described by the power spectrum (Loeb, 2010). The machine learning algorithms can be understood to skip this definition stage and proceed immediately to parameter estimation. However, this reduces the problem to a dependence on a black box system (Buhrmester, 2019). In such situations, the latent layer, used for identifying the features the network picks out, may not have a clear motivation for making the selections it does (Buhrmester, 2019). Additionally, the problem of isolating bases and accounting for their respective influences on the general parameter space depends on successful interpretation of feature engineering (Hortua, 2020).

As the goal is to refine both the parameter space and the values of the parameter elements themselves, this is clearly problematic. It may still be possible to identify the physical implications of dependencies, but since much of the actual learning is intrinsically based on feature engineering (Gillet, 2019), it is unlikely that the structure particulars can be recovered or explained *ex post facto*. Again, this is often ignored as the primary goal is recovery of parameter estimates. A network may output some set of parameter estimates that can be evaluated against the degree to which they reflect reality. Assuming observed values are in agreement with the algorithm output for a significant number of numerical analyses, it may not be necessary to provide further justification. In that case, the structure may not be of foremost interest.

This may not always be possible, however, as the 21cm signal has not yet been directly measured for periods such as the cosmic dark ages. Machine learning is not solely a means for speeding up computation; it is used also where traditional methods fail entirely. For example, machine learning may successfully circumvent the need to explicitly state a likelihood function, in situations where one cannot be easily defined (Villaescusa-Navarro et al., 2020), or skip over the formulation of the nonlinear structure the problem is defined in. And particularly in such simulations where machine learning outputs cannot be verified against traditional numerical methods, it is of utmost importance to compare outputs to observational data, and for situations in which such data has not yet been recorded, even the best estimates remain unverifiable. This problem is discussed at length in Section 5 of this review, which focuses on instrumentation which is slated for completion in a matter of years, and which will prove essential in validating learned outputs.

## 3. Model Emulation

The principle objective of model emulation is to use deep generative models to simulate Cosmic Dawn for 21cm by assuming parameter values are known, with some reasonable

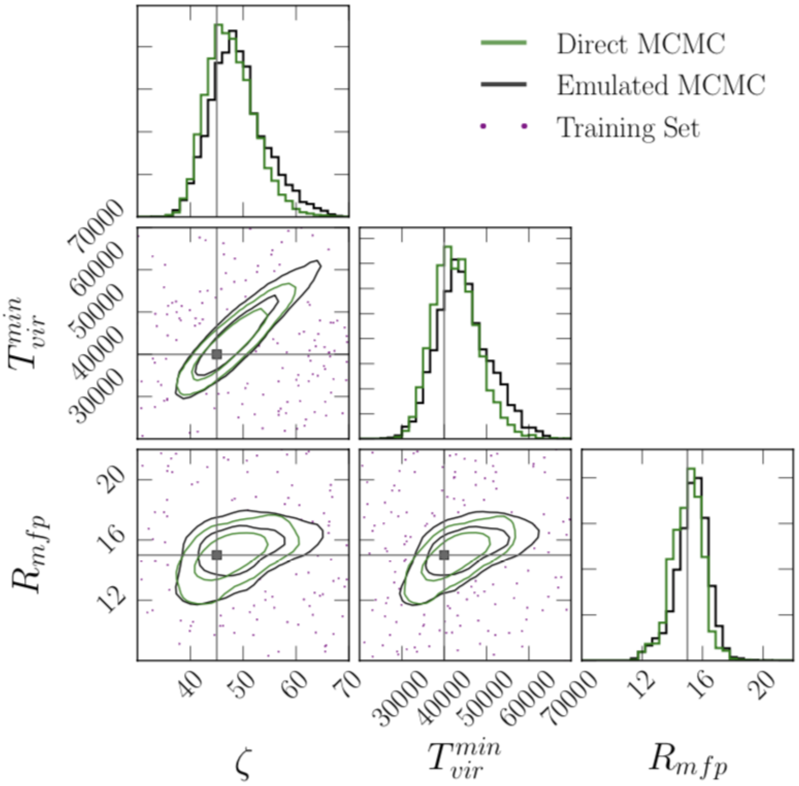


FIGURE 3. Contours representing constraints derived from direct MCMC (`21CMMC`) and embedded-emulator methods (Kern, 2017).

variance, and verifying the resulting distributions. The goal of such models is to map from the lower dimension spaces, such as those of density fields, to ones of higher dimension, such as small-scale structures, more efficiently (Ramanah, 2020). This is required by the prohibitive computational intensity associated with traditional methods of modeling systems of many non-linearly behaved components and tracking their interactions and interdependencies.

Model emulation generally refers to the collection of surrogate models which are used to represent a simulation defined by its parameter space (Kern, 2017). Fiducial values are varied individually for a given set of parameters and from these a model is developed to emulate, rather than directly reconstruct, a distribution, e.g. the 21cm power spectrum (Schmit, 2017). Since the 21cm distribution has previously been computed for different parameter values, it is possible to draw lines of best fit for the outputs of emulators in order to judge these models' effectiveness. In situations where emulators are able to recreate such distributions, they often supersede traditional methods given their computational superiority. Several examples of emulators built on neural networks and their comparative computation requirements are discussed below.

### 3.1. Embedded versus standalone neural networks for MCMC refinement

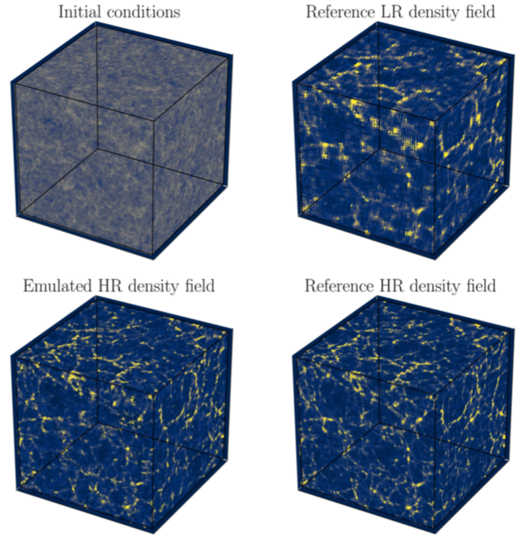
In the case of the Kern et al. (2017), an emulator is embedded within the `21CMMC` model developed by Greig and Mesinger (Greig, 2015) to optimize computation time while maintaining a high degree of accuracy. `21CMMC` is a Monte Carlo Markov Chain (MCMC) analysis tool built on `21cmFAST` (Greig, 2015), and until recently was the only means of simulating the parameter space from the Epoch of Reioniziation (Kern, 2017). As shown by the contours in Figure 3, parameter constraints computed by the emulator were very similar to those computed by brute-force, and perhaps most importantly the former were more conservative (Kern, 2017). Stated differently, the constrained space of the emulator was by and large enclosed inside that of the brute-force method, implying the estimates were useful and trustworthy. One shortcoming explored was with respect to the dependence of emulators on the size of their training data sets (Kern, 2017), especially when dealing with highly-unconstrained parameters. This is potentially a temporary problem as new, large data sets are expected from instruments which will be operational in coming years (Villaescusa-Navarro et al., 2020), and as new methods are developed for even the most ill-behaved spaces.

A neural network can also be implemented as a standalone design, and the formulation discussed here had significant computational advantages to a comparable MCMC model. Schmit and Pritchard (2017) show artificial neural networks (ANN) are an effective tool for accelerating computation for models in which output is continuous and can be mapped with few points, as is the case for the power spectrum of the 21cm wavelength (Schmit, 2017). As a result, they were able to faithfully recreate the 21cm power spectrum from a set of parameters using an ANN while varying $\zeta$, $T_{vir}^{min}$, and $R_{mfp}$, and error was largely confined to runs where parameters values were near interval bounds. Emulators are particularly useful for narrowing constraints on reionization and heating parameters (Kern, 2017), as is reflected here. The emulator was run for multiple subsets of the training data set and training durations to establish the importance of both on the parameter best fit values and it was shown that so long as training data is representative of the space it need not exactly follow its true distributions (Schmit, 2017). A comparable process which used MCMC had a runtime of 2.5 days with 6 cores per redshift (Greig, 2015); the ANN discussed here was able to perform the computation in 4 minutes.

### 3.2. Restricted neural networks and Wasserstein loss

Alternatively, a restricted neural network is best for emulation scenarios in which one wishes to map from low resolution density field models to high resolution small-scale structures (Ramanah, 2020). This method is discussed by Ramanah et al. (2020) and uses an approximation of the Wasserstein distance as its loss function. The application is not specific to 21cm but worth exploring for its implications to that field. The principal assumption made here that is not immediately obvious for 21cm and may indeed pose issues, is rotational symmetry. It is possible this can be remedied as outlined in subsection 4.2, where convolutional kernel symmetries are discussed, but this remains to be demonstrated.

Figure 4. Depiction of density field emulated from initial conditions and low-resolution version; a simulated high resolution density field depicted for comparison.(Ramanah, 2020).


Initial conditions
Reference LR density field
Emulated HR density field
Reference HR density field

The Wasserstein approach is a variation on a generative adversarial network (GAN) that defines an approximation to a Wasserstein loss function [(1)] as the distance between the generated distribution and the target Whereas a GAN typically uses a discriminator to classify outputs of the generator as either close to that of the target or otherwise, this version instantiates a loss function directly correlated to the output image resolution. As can be seen in Figure 4, the results visually closely resemble those of a traditional high-resolution simulation forgoing the associated computation intensity.

The effect on computation time is, as hoped, substantial: the emulator, which clocks 45 CPU hours, features a speed up of a factor of 11 when compared to running the normal high-resolution simulation for 500 CPU hours (Ramanah, 2020). This is

[(1)] Computation of a Wasserstein distance is, in practical terms, intractable. Therefore, an approximation may be used provided that certain constraints are met; otherwise, a gradient penalty may be imposed.

achieved with accuracy bounded from below by that of other deep generative models. Ramanah (2020) finds also that this particular set up is largely unaffected by small variations in the mass density, $\Omega_m$, which contrasts greatly with the results of Villaescusa-Navarro et al. (2020) shown in Figure 5, where adjusting $\Omega_m$ yields high fluctuation of matter (specifically, gas metallicity) distribution in output images.

### 3.3. Lagrangian deep learning

Lagrangian deep learning (LDL) exploits translations and rotational symmetries (Dai 2020) to constrain the simulation and minimize computational intensity. It is useful for emulating hydrodynamic simulations (Dai 2020) such as those discussed in the CAMELS project, and is modeled on effective theory (Dai 2020): the Lagrangian is rewritten to encapsulate the most general form with the restriction that symmetries are satisfied, and its free coefficients represent unresolved small scales (Dai 2020). By comparing the predicted and target distributions of the power spectrum, researchers have shown LDL methods are better predictors of baryonic distributions than full hydrodynamic simulations (Dai 2020), a powerful result given how much more costly [(2)] those simulations are than their machine learning counterparts. Therefore, LDL methods are both more accurate and cheaper to run than the full simulations traditionally used for fitting observational data. LDL offers great potential for speeding up computation, just as the other machine learning algorithms discussed in prior subsections, but with the added benefit that now results are of higher resolution than those of non-ML simulations.

### 3.4. Discussion

The nature of research discussing emulation techniques is adaptive but myopic, in the sense that a given paper focuses largely on the particulars of a specific computation or the adjustment and refinement of a preceding approach. This is no accident, as the improvement on previously-established methods is the most effective way machine learning can benefit 21cm studies. One means of achieving such results is speeding up the time it takes to perform computations, and can be accomplished by embedding an emulator within a more general model (Kern, 2017; Ramanah, 2020), or replacing the model entirely (Schmit, 2017). Emulation offers many advantages for efficiently computing distributions given large training data sets are available. Selecting the right type of model depends largely on what is to be achieved: in the case of establishing connections and interdependencies of parameters, a convolutional neural network is best.

Where the goal is to map from low resolution density field models to high resolution small-scale structures, it makes sense to pursue a strategy that directly related output to a loss function, and the Wasserstein optimized GAN is a reasonable approach. In every paper discussed, incorporating an emulator into an existing algorithm or replacing a brute-force approach with an emulator led to orders of magnitude of savings in computation time, and sometimes even lowered storage space requirements (Ramanah, 2020). The most effective uses of emulation were those that used it to extend the scope of analysis such as further constraint of parameter spaces (Kern, 2017) and insight into loss calculation (Ramanah, 2020). Emulation scales well and speeds up model evaluation while maintaining a high degree of accuracy (Schmit, 2017), and as such will be a crucial component of 21cm models as it becomes necessary to process more and more experiment data in coming years.

## 4. Parameter Estimation

Parameter estimation is the inverse approach of the one described in the previous section; now, parameters are extracted from a given data set. As discussed, data sets are projected to grow with new experimental and observational results coming in, and the need to examine these both effectively and efficiently will benefit from the computational advantages offered by machine learning methods.

[(2)] Two LDL solvers are set up, one that is `FastPM`-based and the other N-body-based. These are used to generate maps that require 7 and 4 orders of magnitude less computation time respectively when compared to hydrodynamic simulations (Dai 2020).

### 4.1. The CAMELS Project

The problem of estimating parameter values is approached by constraining the possible values of a given parameter or the space itself. The CAMELS Project (2020) is a recent, monumental effort consisting of 4233 cosmological simulations that compares the full power spectrum distributions to those of baryonic matter, demonstrating that it is necessary to marginalize over the latter in order to properly interpret cosmological surveys (Villaescusa-Navarro et al., 2020). CAMELS uses two suits, "IllustrisTNG" and "SIMBA," to perform this marginalization. One set of SIMBA data examines the effect of varying cosmological parameters on a region of $25 \times 25 \times 5(h^{-1}Mpc)^3$. In Figure 5 it is clear to see varying each parameter even somewhat can have large effects on the metallicity of the gas, and thus on the conditions that region would have for star formation. In order to draw conclusions about the independent effects of a given parameter, it would be necessary to decouple its effects from that of the others. Each parameter explores a different basis direction, but often two parameters might be degenerate, and overall the model is not robust to even small parameter variance.

The suites are found to be both fast and accurate, and can be used to estimate parameter values such that these estimates are in close agreement with observational data (Villaescusa-Navarro et al., 2020). This is accomplished via non-Bayesian modeling, i.e. the posterior,

$$P(\theta|d) = \frac{P(d|\theta) \times P(\theta)}{P(d)}, \tag{4}$$

where $\theta$ represents the set of parameter values to be extracted and $d$ the data set used, is ignored and instead the likelihood function, $P(d|\theta)$ is optimized. Unfortunately, constraints are not as rigid as the researchers had hoped, but this is justified by the scatter from cosmic variance resulting from a different initial random seed being used in each run (Villaescusa-Navarro et al., 2020). This illustrates CAMELS is very useful in instances where it is not possible to write a likelihood function directly, as it provides a way to still find parameter values that would maximize it

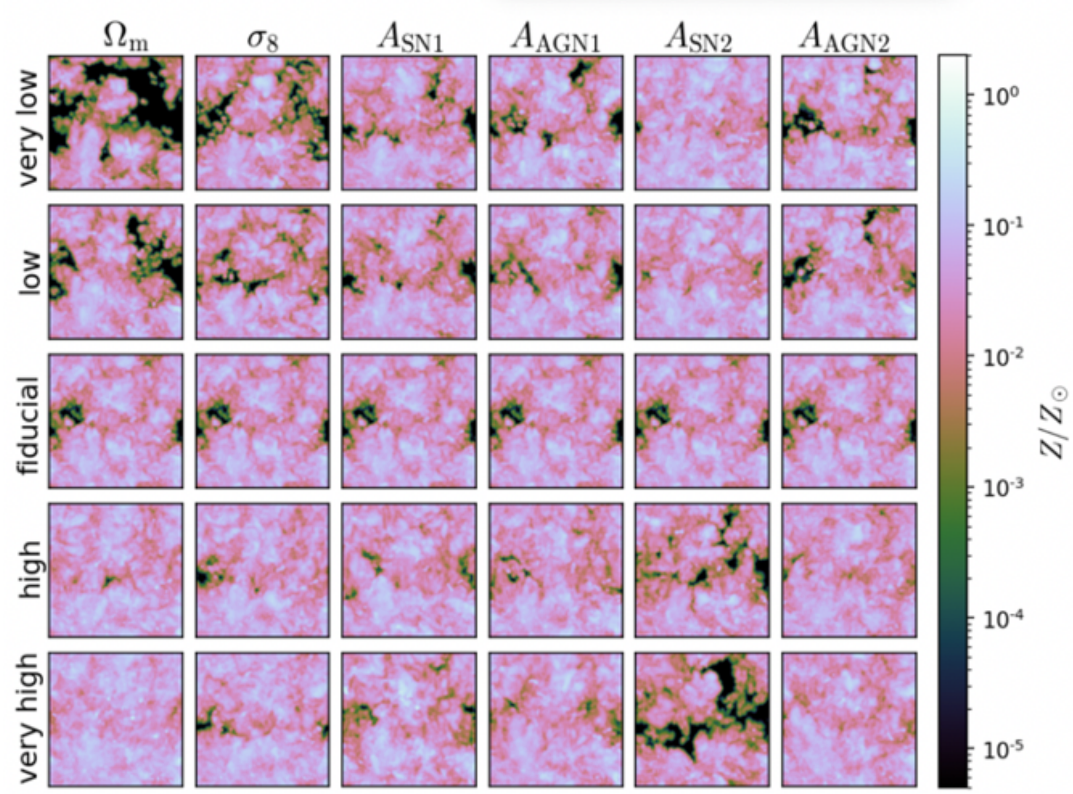


FIGURE 5. Gas metallicity concentrations of a region as a result of individually varying paramaters $\Omega_m$, $\sigma_8$, $A_{SN1}$, $A_{AGN1}$, $A_{SN2}$, $A_{AGN2}$,. (Villaescusa-Navarro et al., 2020).

(Villaescusa-Navarro et al., 2020). It may prove more efficient also where the covariance matrix must be computed from a large number of simulations by skipping that step entirely.

One limitation of CAMELS is the volume of simulation is insufficient to account for scales such as galaxy clusters. Also, and perhaps most importantly, by varying parameters individually, it fails to establish the effects that a change in ratio of two or more would have on the outputs. This limitation is the one regarding isolation of bases discussed in Section 2.1. Future work by Villaescusa-Navarro et al. (2020) is expected to entail perturbing cosmological parameters to better explain the affects of combinations of parameters on parameter space constraints, as well as varying parameters more, and increasing volume of simulation sizes.

Overall, despite these drawbacks, the significant reduction in computation time makes it a serious candidate for learning about underlying structures on these scales. The training, which is responsible for the majority of computational cost, can be run in parallel on 150 GPU hours (Villaescusa-Navarro et al., 2020).

### 4.2. Convolution neural networks and symmetries

CNNs are traditionally popular in image recognition technology and Gillet et al. (2019) claim they have been adapted in cosmology for finding and analyzing gravitational lensing in images (Gillet, 2019). The convolution layer is responsible for extracting features from the input image, and this is accomplished by weighted filtering matrices. The resulting series of convolutions is seen in the output image, which is then pooled (shrunk while maintaining a maximal value defined for the original image) and flattened (rows are concatenated into a one dimensional array), in order to produce the input for a classical neural network, which runs a regression to continuously fit the parameter values.

CNNs are particularly useful in this instance because they allow for the convolution kernel to be learned by the network, and thus do not depend on it being known prior to the running of the algorithm (Gillet, 2019). Due to the translationally invariant nature of convolutional kernels, it is possible to simulate arbitrarily large box sizes (Ramanah, 2020). This is useful for generating large, high-resolution models (Ramanah, 2020) at much faster speeds than would be possible without the implementation of machine learning tools. It may be possible to exploit additional symmetries such as rotation, if properties of parameter spaces can be shown to obey spherical harmonics, as has been done in the case of inferring dark matter halo distributions with neural engines (Charnock, 2020). CNNs have been shown to be particularly useful in situations of large weak lensing surveys despite susceptibility to noise (Ribli, 2019), which was excluded in the proof of concept paper from Gillet et al. (2019) and which they intend to investigate in future work.

### 4.3. Discussion

Parameter estimation is concerned with how to most efficiently infer parameters, and though location-specific precision is of course important, it may not be paramount. The statistical analysis from a given simulation may not depend on parameters of a particular region being correctly identified, but rather on the probability that given many regions, some percent of them would be expected to behave in a certain way (Nichols, 2019). The likelihood function in this context would be learned from performing many simulations and then applied to data in order to return the most probable sets of parameters. If the intention is to derive summary statistics, say for finding that some percentage of regions have metallicities needed for star formation, it may be sufficient to summarize distribution probabilities without ascertaining which of the locations would fall into that percentage.

Autoencoders used in the training of neural networks are able to reduce dimensionality of data for nonlinear spaces, similarly to how Principle Component Analysis is used in generic constructions. The purpose of the autoencoder is to find a lowest dimensional representation from which the original can be reconstructed, and to then use this representation in lieu of the costly original to generate data (Villaescusa-Navarro et al., 2020). Consequently, the information loss is on the same order as the loss in accuracy from original to reconstruction, and thus error is behaves both predictably and consistently. Because the success of parameter estimation methods is related to how well a method qualitates underlying relationships and patterns, PCA-analogous reduction is quite useful, and has been applied to both simulated and real 21cm data (Makinen, 2020). It is not discussed whether autoencoders discriminate between loss of dimension due to innate sparsity, e.g. if the original image is of lower rank than its size, versus loss of information. It would certainly be valuable to determine whether such a separation is possible to make, and could perhaps be inferred from error variances (since the first case would presumably not contribute substantially to error). One consideration could be to develop an autoencoder to mimick Singular Value Decomposition, which would also make it possible to *a prior* specify allowed error and compute the representation to be within those bounds (Trefethen 1997). Overall, autoencoders are indeed a valuable feature of generative models and offer exciting opportunities for exploiting the dimensionality reduction capabilities of numerical methods.

## 5. Telescopes and Signal Detection

Ewen and Purcell first observed the 21cm signal in 1951 using a microwave radiometer. Since then, a variety of telescopes have been built and more commissioned with the hopes of detecting the signal directly (Schmit 2018) and in nonlocal galaxies (Davies, 2020). The Square Kilometer Array (SKA) will be the most sensitive radio instrument built for this purpose and is projected to be completed in 2027. Since the Rayleigh criterion relates the minimum angular resolution of a telescope to its size as follows:

$$\theta_{min} = 1.22 \times \frac{\lambda}{D_{telescope}}, \tag{5}$$

where $\lambda$ is the observed wavelength and $D_{telescope}$ the diameter of the telescope mirror (Loeb 2010), it makes sense to build large radio telescopes as arrays. Then, measurements taken at individual antennae can be combined to provide meaningful results. The SKA will be built in South Africa and Australia, and away from population centers so as to limit the interference of everyday broadcasts which are transmitted over a similar frequency. Other noteworthy radio arrays include the Hydrogen Epoch of Reionization Array (HERA), which has as its primary purpose the 3D mapping of hydrogen gas distribution during the epoch of reionization (DeBoer, 2016), as well as its precursors: the Low-Frequency Array (LOFAR) and Murchison Widefield Array (MWA).

Early detection of the 21cm signal was noisy and higher signal measurements are still required for the accurate mapping of the evolution of luminous matter distributions throughout the universe. Measuring the signal is relatively straightforward for the epoch of reionization but becomes a more ambitious task for the cosmic dark ages, as the progression toward lower frequency ranges means foregrounds are brighter, there are contributions from Earth's ionosphere, and the atmosphere becomes reflective. For this reason, more precise instruments offer a significant contribution to the task of mapping these distributions, and as always are crucial for the collection of observational data which will either support or contradict the models discussed here, but in either case will help to further advance the field of 21cm cosmology. Recent developments in 21cm modeling described in this review as well as the slated completion of the SKA make this a truly exciting time to study the Cosmic Dawn period.

## 6. Conclusion

Researchers engaged in 21cm study find themselves at a crossroads. Due to the non linear nature of the scales being resolved, traditional tools such as power spectrum analysis fail to describe matter distributions completely, and numerical simulation accuracy carries a prohibitive operational cost. In order to generate more stringent bounds on the parameters of interest or better simulate small scale distributions, such that telescopes can better search for the 21cm signal in even the most faraway stars and galaxies, it becomes necessary to engage new methods.

In this context, machine learning offers both savings on computational time and possibilities for generating boundary estimates in situations where it were previously not possible. This review has described the two principal techniques used by cosmologists incorporating machine learning algorithms in their work, model emulation and parameter estimation. Both methods are able to significantly improve on more established numerical analyses and overcome their associated shortcomings. We find that, though neural networks and deep generative models are exciting in their own right, they are better able to constrain parameter spaces when combined with numerical methods such as MCMC, into hybrid models.

The possibility of reducing parameter space dimensionality offered by autoencoders used in neural networks offers an as of yet untapped area for improvement over traditional methods. This topic, explored by the CAMELS project, among others, is a likely direction for future work. It is to be expected that machine learning algorithms, which must minimize loss functions in the training stage, should be concerned with exploiting sparsity. Similarly, it may be possible to use additional symmetries in the problem statement to more efficiently encode information and reduce computational intensity.

On the other hand, the question of how best to emulate existing numerical simulations depends entirely on the scope and intent of the problem definition. The many different types of emulators available are successful for specific conditions outlined in the section where they are described. As such, there is no best emulator in general, though all of the ones discussed in this review offered orders of magnitude of reduction in computation time.

As further constraints are determined for cosmological and astrophysical parameters, better models are formulated for describing early universe matter distributions, and more precise instrumentation comes into operation, it will become possible to significantly advance our understanding of fundamental physics and the state of the early universe.

## Citations

Buhrmester, V., Münch, D., Arens, M. (2019). Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey.

Charnock, T., et al. (2020). Neural physical engines for inferring the halo mass distribution function. *Monthly Notices of the Royal Astronomical Society.* **494**(1), 50-61. Available from doi: 10.1093/mnras/staa682.

Dai, B., and Seljak, U. (2020). Learning effective physical laws for generating cosmological hydrodynamics with Lagrangian Deep Learning. *Monthly Notices of the Royal Astronomical Society.* **501**(1), 146-156. Available from doi: 10.1093/mnras/staa3531.

Davies, J. et al. (2020). Stacking Redshifted 21cm Images of HII Regions Around High Redshift Galaxies as a Probe of Early Reionization.

DeBoer, D. R., Parsons, A. R., et al. (2016). Hydrogen Epoch of Reionization Array (HERA) *Publications of the Astronomical Society of the Pacific*, **129**(974). Available from doi: 10.1088/1538-3873/129/974/045001.

Gillet, N., Mesinger, A., Greig, B., Liu, A., Ucci, G. (2019). Deep learning from 21-cm tomography of the Cosmic Dawn and Reionization. *Monthly Notices of the Royal Astronomical Society.* **484**(1), 282-293. Available from doi: 10.1093/mnras/stz010.

Greig, B. and Mesinger, A. (2015). 21CMMC: an MCMC analysis tool enabling astrophysical parameter studies of the cosmic 21cm signal. *Monthly Notices of the Royal Astronomical Society.* **449**(4), 4246–4263. Available from doi: 10.1093/mnras/stv571.

Hortúa, H. J., Volpi, R., Malagó, L. (2020). Parameters Estimation from the 21 cm signal using Variational Inference.

Kern, N. S., Liu, A., Parsons, A. R., Mesinger, A., Greig, B. (2017). Emulating Simulations of Cosmic Dawn for 21cm Power Spectrum Constraints on Cosmology, Reionization, and X-ray Heating. *The Astrophysical Journal.* **848**(1). Available from doi: 10.3847/1538-4357/aa8bb4.

Loeb, A. (2010). *How Did the First Stars and Galaxies Form?.* Princeton University Press.

Makinen, T. L., et al. (2020). deep21: a Deep Learning Method for 21cm Foreground Removal.

Mootoovaloo, A., Heavens, A. F., Jaffe, A. H., Leclercq, F. (2020). Parameter Inference for Weak Lensing using Gaussian Processes and MOPED. *Monthly Notices of the Royal Astronomical Society.* **497**(2), 2213–2226. Available from doi: 10.1093/mnras/staa2102.

Nichols, J. A., Chan, H., Baker, M. (2019) Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews.* **11**, 111-118. Available from doi: 10.1007/s12551-018-0449-9.

Peebles, J. E., et al. (1994). The Evolution of the Universe. *Scientific American.* **271**(4), 52-57. Available from doi: 10.1038/scientificamerican1094-52.

Pritchard, J. R., Loeb, A. (2012). 21cm cosmology in the 21st century. *Reports on Progress in Physics*, **75**(8).

Ramanah, D. K., Charnock, T., Villaescusa-Navarro, F., Wandelt, B. D. (2020). Super-resolution emulator of cosmological simulations using deep physical models. *Monthly Notices of the Royal Astronomical Society.* **495**(4), 4227-4236. Available from doi: 10.1093/mnras/staa1428.

Ribli, D., Pataki, B. A., Zorilla Matilla, J. M., Hsu, D., Haiman, Z., Csabai, I. (2019). Weak lensing cosmology with convolutional neural networks on noisy data. *Monthly Notices of the Royal Astronomical Society.* **490**(2), 1843–1860. Available from doi: 10.1093/mnras/stz2610.

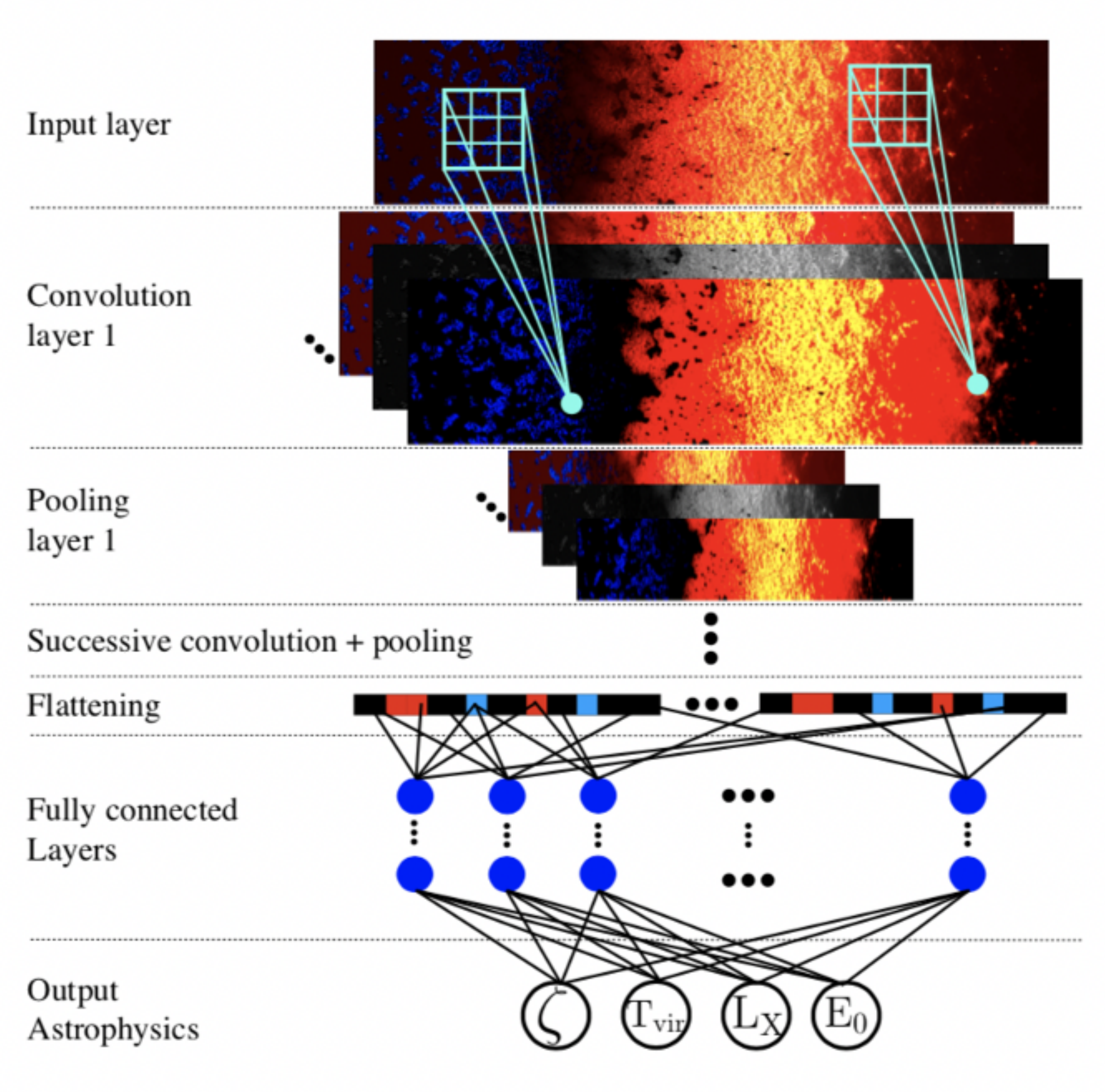
Schmit, C. J. and Pritchard, J. R. (2017). Emulation of reionization simulations for Bayesian inference of astrophysics parameters using neural networks. *Monthly Notices of the Royal Astronomical Society.***475**(1), 1212-1223. Available from doi: 10.1093/mnras/stx3292.

Trefethen, L. N. and Bau, D. (1997). *Numerical Linear Algebra* SIAM.

Villaescusa-Navarro, F., Daniel Anglés-Alcázar , D., Genel, S., Spergel, D. N., Somerville, R. S., Dave, R., Pilepich, A., Hernquist, L., Nelson, D., Torrey, P., Narayanan, De., Li, Y., Philcox, O., La Torre, V., Delgado, A. M., Ho, S., Hassan, S., Burkhard, B., Wadekar,. D., Battaglia, N., Contardo, G. (2020). The CAMELS project: Cosmology and Astrophysics with MachinE Learning Simulations.

Windhorst, R. A., et al. (2018). On the observability of individual Population III stars and their stellar-mass black hole accretion disks through cluster caustic transits. *The Astrophysical Journal Supplement Series.* **234**(2).

## A. Appendix 1



CNN diagram taken from Gillet et al. (2019). An input image is filtered by successive applications of convolution and pooling layers which produce a lower resolution representation of the original. This is further reduced by a flattening layer and is in turn fed into a neural network to extract parameter values.

*Self study project, dated January 2020*

Michelle Gurevich, Physics Department, Imperial College London, Exhibition Rd, South Kensington, London SW7 2BU, United Kingdom