# PSC-CUNY Research Awards (Enhanced)

| | |
|---|---|
| Control No : | ENHC-52-41 |
| Rank : | Associate Professor |
| Tenured : | Yes |
| College : | NEW YORK COLLEGE OF TECHNOLOGY |
| Panel : | Physics & Engineering |
| Discipline : | Physics |
| Co-PI : | |

| | |
|---|---|
| Name : | Acquaviva, Viviana |
| Address : | <br>, |
| Telephone : | |
| Email : | vacquaviva@citytech.cuny.edu |

| | |
|---|---|
| Human Subject Use | No |
| Animal Subject Use | No |
| Supplementary Materials | No |
| List of Supplementary Material | |
| Will Interviews be Conducted? | No |
| Department | Physics |
| List of Undesirable Reviewers | No |

Title of Proposed Project:

**Yes, we can: Inferring galaxy properties from cosmological simulations using machine learning**

Brief Abstract

Machine learning techniques have been used successfully in Astrophysics for a variety of problems, from recognizing galaxy morphologies to identifying outliers. One significant limitation of supervised learning techniques is that they require a training set for which the ground truth is known, which is not readily available because most astrophysical processes happen on unobservable timescales. One possible approach to this problem is to train machine learning models on state-of-the-art cosmological simulations; however, it is unclear how models will perform once applied to real data. We are carrying out an innovative study with the goal to model the generalization error of a machine learning algorithm as a function of an appropriate measure of distance between the source domain and the application domain. Our preliminary results have shown great promise when applied to the predictions of stellar masses. Our goal is to extend and improve our framework, and ultimately obtain a reliable estimate of how a model trained on simulations might behave on data.

Relevant Publications & Scholarship

Papers published between 2015 and 2020:

V. Acquaviva, C. Lovell, and E. Ishida, "Debunking Generalization Error or: How I Learned to Stop Worrying and Love My Training Set", accepted to the NeurIPS 2020 workshop "Machine Learning and the Physical Sciences" (2020), https://arxiv.org/abs/2012.00066. (Resulted from PSC-CUNY funding, cycle 51).

A. Lawler and V. Acquaviva, "Detecting episodes of star formation using Bayesian model selection", submitted for publication in Monthly Notices of the Royal Astronomical Society, Volume 491 (2020). Resulted from PSC-CUNY funding, cycle 50).

S. Sherman, S. Jogee, J. Florez, M. L Stevans, L. Kawinwanichakij, I. Wold, S. L Finkelstein, C. Papovich, V. Acquaviva, R. Ciardullo, C. Gronwall, Z. Escalante, "Exploring the high-mass end of the stellar mass function of star-forming galaxies at cosmic noon", Monthly Notices of the

Royal Astronomical Society, Volume 491 (2020).

C. Lovell, V. Acquaviva, P. Thomas, K. Iyer, E. Gawiser, and S. Wilkins, “Learning the Relationship between Galaxies Spectra and their Star Formation Histories using Convolutional Neural Networks and Cos- mological Simulations”, Monthly Notices of the Royal Astronomical Society, Volume 490 (2019) (Resulted from PSC-CUNY funding, cycle 49).

V. Acquaviva, Pushing the Technical Frontier: From Overwhelmingly Large Data Sets to Machine Learning, Invited Review, Proceedings of the International Astronomical Union, IAU Symposium, Volume 341 (2019).

J. Fang, S. Faber, D. Koo, A. Rodriguez-Puebla, Y. Guo, G. Barro, et al (including V. Acquaviva), Demographics of Star-forming Galaxies since $z \sim 2.5$. I. The UVJ Diagram in CANDELS, The Astrophysical Journal, Volume 858 (2018).

B. Lee, M. Giavalisco, K. Whitaker, C. Williams, H. Ferguson, V. Acquaviva, et al, “The Intrinsic Char- acteristics of Galaxies on the SFR-M* Plane at $1.2 < z < 4$: I. The Correlation between Stellar Age, Central Density, and Position Relative to the Main Sequence”, The Astrophysical Journal, Volume 853 (2018).

A. Leung, V. Acquaviva, E. Gawiser, R. Ciardullo, E. Komatsu, A. Malz, et al, “Bayesian Redshift Classification of Emission-line Galaxies with Photometric Equivalent Widths”, The Astrophysical Journal, Volume 843 (2017). (my contribution stemming from PSC-CUNY funding, cycle 44)

P. Kurczynski, E. Gawiser, V. Acquaviva, et al “Evolution of Intrinsic Scatter in the SFR-Stellar Mass Correlation at $0.5 < z < 3$, Astrophys. Journal Lett. 820, 1 (2016). (my contribution stemming from PSC-CUNY funding, cycle 43)

V. Acquaviva, "How to measure metallicity from five-band photometry with supervised machine learning algorithms," Monthly Notices of the Royal Astronomical Society, 456, 2, 1618 (2016). (Resulted from PSC-CUNY funding, cycle 46)

B. Mobasher, T. Dahlen, H. Ferguson, V. Acquaviva, et al, "A Critical Assessment of Stellar Mass Measurement Methods", The Astrophysical Journal 808, 1 (2015).

V. Acquaviva, A. Raichoor and E. Gawiser, "Simultaneous Estimation of Photometric Redshifts and SED Parameters: Improved Techniques and a Realistic Error Budget," The Astrophysical Journal 804, 8 (2015).

J. Bridge, C. Gronwall, R. Ciardullo, A. Hagen, G. Zeimann, A. Malz, V. Acquaviva, D. Schneider, N. Drory, K. Gebhardt, S. Jogee, "Physical and Morphological Properties of [O II] Emitting Galaxies in the HETDEX Pilot Survey," ApJ 799, 205 (2015).

## Education

| **Institution** | **Degree** | **Year(s)** | **Discipline** |
|---|---|---|---|
| SISSA/ISAS, Trieste | Ph. D. | 2006 | Astrophysics |
| University of Pisa | B. Sc. | 2002 | Physics |

## Other Current & Past Funding (last 5 years)

| **Period** | **Role** | **Title** | **Amount** | **Funding Source** |
|---|---|---|---|---|

|  |  |  |  |  |
|---|---|---|---|---|
| 2019-2024 | co-I | Engaging, Empowering, and Retaining New Scholars in Science, Technology, Engineering and Mathematics | $12,620.00 | NSF S-STEM |
| 7/1/2019-06 /30/2020 | PI | Measuring Galaxy Star formation Histories with Machine Learning and Bayesian Model Selection | $3,500.00 | PSC-CUNY |
| 7/1/2019-06 /30/2020 | PI | Research in the classroom: Estimating the physical properties of galaxies using Machine Learning | $7,380.00 | CUNY Research Foundation |
| 7/1/2018-06 /30/2019 | PI | Measuring galaxy star formation histories with machine learning | $3,500.00 | PSC-CUNY |
| 7/1/2017-06 /30/2018 | PI | Measuring the metallicity of galaxies through strong line indicators | $6,000.00 | PSC-CUNY |
| 9/1/2016-06 /30/2017 | PI (Collaborative Research) | Reconstructing Star Formation Histories to Reveal the Origin and Evolution of the SFR-M* Correlation | $8,789.00 | Space Telescope Science Institute |
| 7/1/2016-06 /30/2017 | PI | Measuring the metallicity of high-redshift galaxies with Machine Learning and SED fitting | $3,500.00 | PSC-CUNY |
| 7/1/2015-06 /30/2016 | PI | A novel approach to measuring metallicity in galaxies | $3,499.00 | PSC-CUNY |
| 7/1/2014-06 /30/2015 | PI | Galaxy Classification in HETDEX using Machine Learning | $3,491.00 | PSC-CUNY |

## Attachments

| Description | File Name | File Size | Date Attached |
|---|---|---|---|
| Budget Justification | PSC_CUNY_2020_Budget_Justification. | 43177 | 12/8/2020 3:46:23 PM |
| Project Description | PSC_CUNY_2020.pdf | 441908 | 12/6/2020 8:55:01 PM |

## Budgets

| Description |  |  | Requested Amount |
|---|---|---|---|
| Laboratory Fees |  | 0.00<br>0.00 | 0.00 |
| Equipment |  | 0.00<br>0.00 | 0.00 |
| Manuscript Preparation/ Publication Costs |  | 0.00<br>0.00 | 0.00 |
| Research Staff | Fringe Benefit Expense | 0.00<br>0.00 | 0.00 |

|  |  |  |  |
|---|---|---|---|
|  | MTA Payroll Tax |  |  |
| Clerical Staff | Fringe Benefit Expense<br>MTA Payroll Tax | 0.00<br>0.00 | 0.00 |
| Summer Salary | Fringe Benefit Expense | 0.00<br>0.00 | 0.00 |
| General Office Supplies/Xeroxing |  | 0.00<br>0.00 | 0.00 |
| Research Supplies |  | 0.00<br>0.00 | 0.00 |
| Domestic Travel |  | 0.00<br>0.00 | 0.00 |
| Independent Contractors |  | 0.00<br>0.00 | 0.00 |
| Subject Payments |  | 0.00<br>0.00 | 0.00 |
| Released Time | Fringe Benefit Expense | 3000.00<br>1530.00 | 4530.00 |
| Foreign Travel<br>2,550 (PI to travel to conference). 4,600 (two 4-weeks collaborator visits). Additional details in the budget justification. |  | 7150.00<br>0.00 | 7150.00 |
|  |  | **Total** | **11680.00** |

# Yes, we can:
# Inferring galaxy properties from cosmological simulations using machine learning

Machine learning techniques have been found to be increasingly useful in a variety of data-intensive sciences, and were particularly well received in Astrophysics, because of the predominance of large data sets and of the need to model complex physical processes without the possibility of laboratory testing. As a result, they have been successfully used in many research problems, from recognizing galaxy morphologies, to identifying gravitational lenses, to evaluating distances of astrophysical sources, to enhancing the effective volume of surveys (*e.g.,* [1]). In many cases, science questions can be set up as a supervised learning problem, in which the goal is for the algorithm to model an input/output relationship (for example, to infer a certain physical property on the basis of other observed ones) after being shown a set of examples, known as a *learning set*, for which both the observed quantities and the desired one are known. A supervised learning algorithm is only as good as its learning set: only if the examples shown are both representative and complete of the application domain, it is possible to obtain correct inference.

In the problem we consider here, we aim to determine some physical properties of distant galaxies (for example, stellar masses or stellar ages) from their observed spectra (charts of their luminosity versus wavelength of light emission), using machine learning methods. We know that different astrophysical processes leave their imprint in various regions of the spectra with characteristic signatures, so we are confident that the information is present in the data. However, identifying a learning set for this problem (i.e., a set of galaxies whose properties are known) is very hard: the ground truth is not available, because we have no way of knowing the true history of how galaxies have formed.

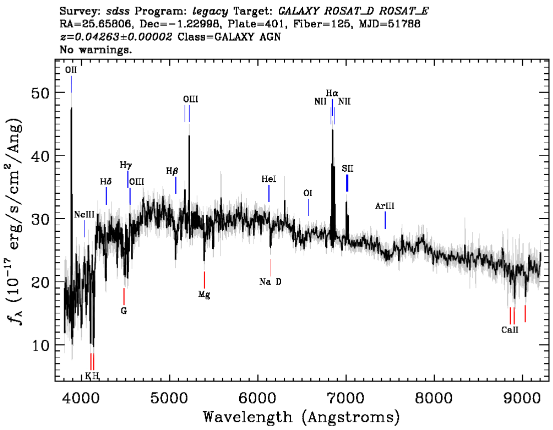

**Figure 1:** An example spectrum for a galaxy in one of our target data surveys, the Sloan Digital Sky Survey. The input features of our ML problem are the observed brightness at each wavelength; the target will be properties such as stellar mass or stellar age.

To circumvent the difficulty of observing astrophysical processes directly, in the last decade there has been an enormous community effort running large, computationally expensive numerical simulations of galaxy formation and evolution. The results of the most sophisticated of these (*e.g.,* the recent Illustris TNG[1]) resemble quite well the Universe that we presently observe, confirming that such state-of-the-art simulations are able to capture the most important processes in galaxy formation and evolution. In [2], we leveraged this remarkable progress by carrying out pioneering work (funded by a PSC-CUNY award) in determining the star formation history of galaxies by training several algorithms, most notably a Convolutional Neural Network, on spectra obtained through the cosmological simulations, with very promising results. We were also able to show that the results were relatively

[1] `http://www.tng-project.org/`

robust to training models using one cosmological simulations and then applying it to galaxies from a different simulation.

However, a potential weakness of these methods remained unexplored: **the algorithms are trained on the simulations, and we want to apply them to real data**. How can we make sure that our method does not break down when extended to objects that are significantly different from those in our training set? In our most recent work [3], also funded by a PSC-CUNY award, we started to scratch the surface of this important problem, as described in the next section. **The final goal of this project is to provide a quantitative assessment of the generalization error, *i.e.,* the average deviation from the true values incurred when applying the model to previously unseen objects, of these models when applied to observed spectra from real galaxies.**

## Framework and Methodology

Our project is based on a very simple idea: if the simulations are realistic, the spectra of galaxies that they produce will be similar to those of real observed galaxies. We aim to leverage the idea of similarity in the space of observed spectra and check whether it can be translated into a measure of similarity in the space of physical properties.

The problem setup is the following. We know that in nature, a set of physical properties for galaxies (for example, stellar mass, star formation history, chemical enrichment history...), plus many latent variables, will lead to the spectra that we observe; we could call this "mapping" function $f(x)$. *We are interested in learning the "inverse" function $f^{-1}(x)$*, which would teach us to go from observed quantities (spectra) to the quantities that we'd like to measure (physical properties).

In simulations, a set of chosen input physical properties is transformed into a set of simulated spectra by some known modeling function, let's say $g(x)$. Hopefully $g(x)$ is a good approximation of $f(x)$; we can verify this because if our variables are meaningful and our modeling is correct, then the simulated spectra generated by applying $g(x)$ will resemble the observed spectra. In this sense, the distance in spectral space traces the similarity between $f(x)$ and $g(x)$.

Now let us consider the other direction. The function $g^{-1}(x)$ can be learned by, for example, training a machine learning model. The learned function will, of course, have its own generalization error, which breaks down as usual in a model-independent noise term, a bias term, and a variance term. This will cause some difference between the "true" input physical parameters and the inferred physical parameters.

What happens if we apply the learned function $g^{-1}(x)$ *to the observed spectra* (in other words, when we use it a proxy for the function we want, $f^{-1}(x)$)? Our hypotheses are the following:

1. There is an additional term in the mean square error, which comes from the fact that we learned the "wrong" function, $g^{-1}(x)$ instead of $f^{-1}(x)$;
2. The additional error will depend in a *predictable* way on an appropriate distance metric describing the similarity (or lack of it) between the observed spectra and the simulated spectra.

**If we can show that 2. is true, we would be able to predict the generalization error on data.** Our scheme is described in Fig. 2.

Our strategy for (empirically) proving our hypotheses consist of these steps:

1. Generate several sets of simulations, changing the modeling assumptions (in this case, we work with 20 of them, $i = 1 - 20$). The simulations are chosen to represent the full range of models that are compatible with current observations.

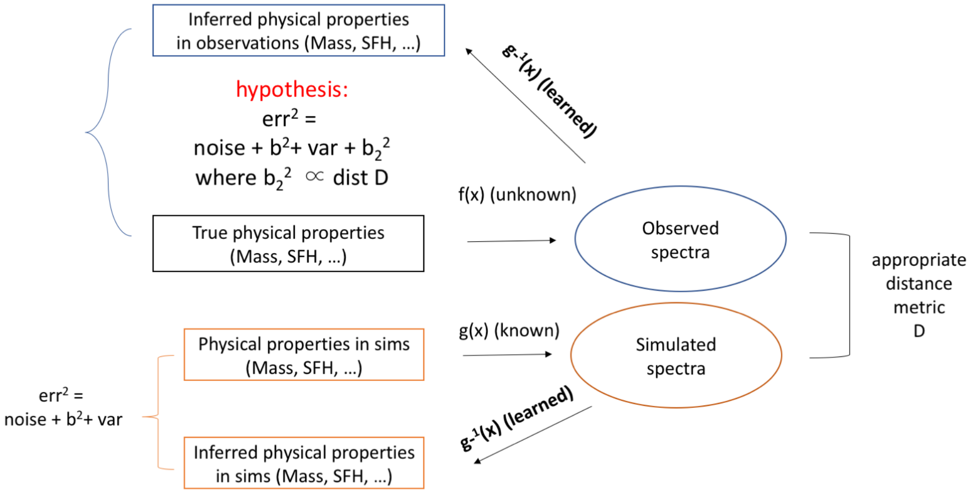
Inferred physical properties in observations (Mass, SFH, …)
hypothesis:
err$^2$ =
noise + b$^2$+ var + $b_2^2$
where $b_2^2 \propto$ dist D
True physical properties (Mass, SFH, …)
$g^{-1}(x)$ (learned)
f(x) (unknown)
Observed spectra
appropriate distance metric D
g(x) (known)
Simulated spectra
Physical properties in sims (Mass, SFH, …)
err$^2$ =
noise + b$^2$+ var
$g^{-1}(x)$ (learned)
Inferred physical properties in sims (Mass, SFH, …)

**Figure 2:** Scheme of the main hypothesis we are testing: if we can learn an imperfect function $g^{-1}(x)$ that gives us physical properties of galaxies (such as stellar mass, or star formation history) starting from spectra, can we estimate how much the parameters inferred through $g^{-1}(x)$ deviate from the true ones, based on some distance D measured in the space of spectra, which acts as a proxy for the distance between $f(x)$ and $g(x)$?

2. Find a suitable representation feature space for all the simulations sets, and optimize machine learning models in this space;
3. Identify an appropriate measure of distance between data sets ($D_{i,j}$ where i, j $\in [1, 20]$);
4. Train 20 models, one per simulated set of spectra, excluding the objects who participated to the feature selection process in step 2, to learn as many inverse modeling functions, indicated as $g_1^{-1}(x), g_2^{-1}(x), ...g_{20}^{-1}(x)$;
5. Apply each of the learned functions to each of the simulated sets of spectra;
6. Generate and analyze 20 scatter plots, one for each simulation set $i$, plotting the distance metric $D_{i,j}$ (where $j = 1, ...20$) versus the generalization error obtained by applying the functions $g_1^{-1}(x), g_2^{-1}(x), ...g_{20}^{-1}(x)$ to learn the parameters of simulation $i$;
7. Use these 20 examples to infer a robust regression between the distance metric $D_{i,j}$ mentioned above and the generalization error incurred.
8. Use the regression model to predict the generalization error on data, based on the distance between data and simulations.

## Preliminary results

We have generated 20 simulations, each of which contains $\sim$ 6,500 galaxies, with a representative range of physical properties. The differences between the simulations arise from using varying modeling assumptions, for example by changing stellar libraries and dust properties. The feature set of this problem is the vector of measured brightness at each wavelength, in the range between 3,000 and 9,000 Å; this matches the Sloan Digital Sky Survey data.

For this exploratory step, we have chosen to predict a simple quantity, the total mass held in stars.At zero order, the stellar mass of a galaxy is proportional to its luminosity, and in particu-

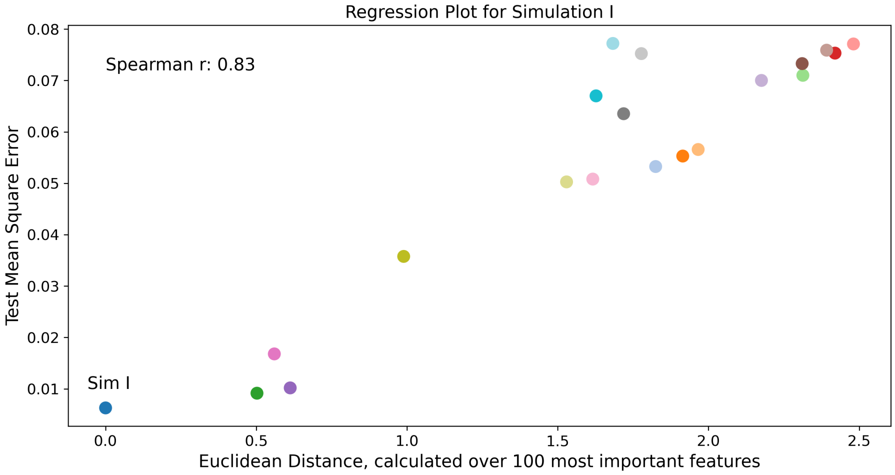

**Figure 3:** An example scatter plot of Mean Square Error versus pairwise distance, for Simulation 1. The MSE is calculated by applying the inverse mapping function learned on each simulation, from 1 to 20, to the data of simulation 1; each color corresponds to a different simulation. There is a clear correlation trend, which suggest the possibility of successfully fitting a regression model. This would allow us to predict the MSE when the model is applied to data. Plots for the other 19 simulations show similar trends. From [3].

lar, to the luminosity in the near-infrared region of the EM spectrum, where fewer confounding effects/degeneracies with other parameters exist. Therefore, this problem is a feasibility pilot for more complicated parameter estimation tasks. Our measure of the generalization error is the Mean Square Error (MSE) on the stellar mass.

After an extensive exploration of domain adaption techniques, we reached the conclusion that a similarity measure that correlates with the MSE in stellar mass need to derive from a supervised feature selection process, as opposed to an unsupervised dimensionality reduction process.

Our preliminary results are shown for a simplistic feature selection process. We assembled a super-data set by compiling together a random selection of 1000 objects from each of the 20 simulation sets, and fitting a Random Forest Regressor to predict Stellar Mass. We then ranked the features according to their importance, selected the first 100, and calculated the pairwise distance between simulated data sets as the mean Euclidean distance in this 100-dimensional space. The objects that participated to the feature selection process are excluded from further processing, so that the performance we report is a true generalization error, obtained for *previously unseen* data.

The results are quite promising. We show one example plot where the "target" set of spectra is simulation 1, and we show the MSE (again from a Random Forest model) when we apply the 20 functions $g_1^{-1}, g_2^{-1}, ...g_{20}^{-1}$ to recover the stellar mass. There is a clear trend that suggest the possibility of fitting the regression successfully. The trends seen here are similar to what we observe in the other 19 plots, where we apply the learned inverse-modeling functions to the other 19 sets of simulations.

## Plan of Work and Anticipated Impact

Our preliminary results confirmed that this technique is promising, but are really just a first step towards a proper assessment of the generalization error on data. Our next steps include improving the techniques as well as extending the method to different parameter estimation tasks:

- Feature selection processes based on ranking can be quite misleading when the features are highly collinear, which is the case here. Therefore, we expect that by using a more sophisticated feature selection technique, for example by clustering highly correlated features, selecting one per cluster, and adding weights proportional to the feature importance in the calculation of distances, we can obtain tighter correlations.
- The generalization error obtained from tree-based ensemble algorithms, such as that shown in Fig. 3, can deviate from the expected behavior because of the lack of extrapolation capability of these methods. We expect that using Convolutional Neural Networks, which are already part of our existing framework, will lead to improved results and stability.
- Further understanding of the applicability domain of our technique will come from understanding "failing" cases, such as outliers in our distance/generalization error regressions, as well as investigating those simulations that have poorer generalization properties.
- After optimizing our method for the Stellar Mass estimation, we will be ready to extend it to the problem of determining the star formation history, the dust abundance and profile, and the chemical enrichment histories of different galaxies. Each of those will require the activation of different features, but the core of the pipeline will remain unchanged.
- We will compare our results to those obtained through traditional (templated-based) Spectral Energy Distribution fitting methods, such as those obtained with the software BAGPIPES[2], which our group is already familiar with (e.g., [4]).
- The final step of this analysis will consist of using the above tools to estimate the generalization error achievable on real galaxies, for different publicly available data sets. This includes, for example, SDSS and 3D-HST[3].

The potential impact of our work goes well beyond enabling the use cosmological simulations to train algorithms that can be applied to data, and can in principle be adapted to any problem where the development of a machine learning model happens on simulations. As it is customary in our research group, we will make all code available on github.

We expect that this project will result in at least one publication in a high-impact peer reviewed journal, and in several conference presentations, as well as support the scholarship of an early-career researcher (Chris Lovell). As a reference, the last two projects in collaboration with him resulted in two publications in high-rated journals and five presentations in international conferences.

## References

[1] V. Acquaviva, Pushing the Technical Frontier: From Overwhelmingly Large Data Sets to Machine Learning, invited review article, Proceedings of the IAU symposium 341, Cambridge University Press, 2019.

[2] C. Lovell, V. Acquaviva, P, Thomas, K. Iyer, E. Gawiser, and S. Wilkins. Learning the relationship between galaxies spectra and their star formation histories using convolutional neural networks and cosmological simulations, *Monthly Notices of the Royal Astronomical Society*, Vol. 449, Issue 4, 2019.

[3] V. Acquaviva, C. Lovell, and E. O. Ishida, Debunking Generalization Error or: How I Learned to Stop Worrying and Love My Training Set, accepted for the NeurIPS 2020 workshop "Machine Learning and the Physical Sciences", arXiv:2012.00066.

[4] A. Lawler and V. Acquaviva, Detecting episodes of star formation using Bayesian model selection, submitted for publication in *Monthly Notices of the Royal Astronomical Society*, arXiv:2012.02285.

[2] https://bagpipes.readthedocs.io/

[3] https://3dhst.research.yale.edu/