מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Davidson Institute
The Educational Arm of the Weizmann Institute

The Dr. Bessie F. Lawrence 53rd International Summer Science Institute 2022

Journal of Scientific Reports

Dear graduates of the 2022 ISSI,

A few months ago, we stood in front of a difficult decision to hold the ISSI in as a digital program for a second year. Covid-19 has changed our social and cultural behaviors, affected our habits, and our means to interact, study and socialize. Following the successful program last summer, we had a drive to further improve the program this year.

It was important to us to preserve the spirit of the ISSI, to bring the atmosphere of the Weizmann, the open doors and open minds approach which are at the heart of the Weizmann Institute. Through diverse scientific talks, discussing the harmony and boundaries of arts and science, asking questions and raising theories- all these provide the foundations to enable the freedom to think differently and the courage to dare, to test, to go where no one has gone before and to pave your own road with partners, collaborators and friends from any field and discipline, even out of scientific fields.

We have put together the structure, but you students made it your own meaningful experience. You have shown your commitment and overcame the obstacles along the way: The different time zones, maximizing your interactions with your mentors and peers, and came forward to assist where needed. The outcomes of your hard work were highly commended by your mentors and creatively crafted into your final presentations, demonstrating your dedication and enthusiasm in your scientific tasks. The mini-papers presented ahead are yet another part of your high achievements in this short period of research, and an introductory step into the endless world of scientific discoveries.

We hope that by participating in this program you were able to acquire new sets of skills and tools for your future academic studies, career, and as a general life habit no less. That the broad outlook on science as a collaborative field has encouraged you to create a community of shared interests and passion, and we wish for you that these friendships will last for life as you are now members of the ISSI community.

It was our great pleasure to experience the ISSI program of July 2022 with every one of you. We wish you all the best and continued success in your future academic studies and career.

Sincerely,

ISSI coordinators

Dr. Dorit Granot, Dr. Aya Shkedy & Ms. Nirit Alon

Table of Contents

Bioimaging Data Generation for Machine Learning in Blender: An Unorthodox Approach

Avraham Balsam[3], Jonas Kuehne[2], Victoria Rodríguez de León[1], Sara Sánchez Vargas[1]

Mexico[1], Switzerland[2], United States[3]

Mentored by Dr. Vyacheslav Kalchenko
Department of Veterinary Resources
Weizmann Institute of Science, Rehovot, Israel

Abstract

In Vivo Bioluminescence Imaging (BLI) is an increasingly popular method which has been used in modern biological research applications such as gene activation tracking, tumorigenesis, cancer treatment, disease progression and drug development. Nevertheless, it comes with the ethical cost of animal testing, as it requires the genetic modification and direct injection of many laboratory animals, generally mice, to achieve ideal experimental results. In this report, we propose a system that tackles this problem efficiently through a synthetic data generation program that is able to simulate accurately through the use of Blender (a 3D computer graphics software) the behavior of the different luciferin-luciferase kinetic reactions that occur in the test subjects. To represent reality, we also developed a *Python* module capable of producing various noise models that mimic image sensor behavior at various lighting conditions, representing common problems that inevitably happen during imaging. The presented approach is efficient, respects animal life and opens possibilities for experimentation in the field of optical imaging, such as using the previously generated data for training an Artificial Neural Network (ANN) in order to improve the way in which the real light signal is captured.

Introduction

Bioluminescence is a type of chemiluminescence where a living organism has the capacity of emitting light. This happens as a result of a chemical reaction between luciferins (bioluminescent substrates) and luciferases, which are enzymes that catalyze the oxidation of the luciferins producing luminous energy [1]. Some organisms synthesize luciferin on their own, others obtain it from their medium, and, for experimental purposes, some can be changed by

humans through genetic modification or by direct injection methods so they can have it in their system [2]. For the latter case, the luciferase enzymes are obtained from insects (e.g., FFLuc from the firefly), marine species (e.g., RLuc from the anthozoan sea pansy) or bacteria (e.g., from the symbiotic bacteria *Vibrio fischeri*) [3, 4, Fig. 1]. Also, there are synthetic luciferins made for achieving a more efficient performance, such as the case of CycLuc1 [5, 6]. Each luciferin has a different bioluminescent color effect as a result of their individual molecular arrangement.

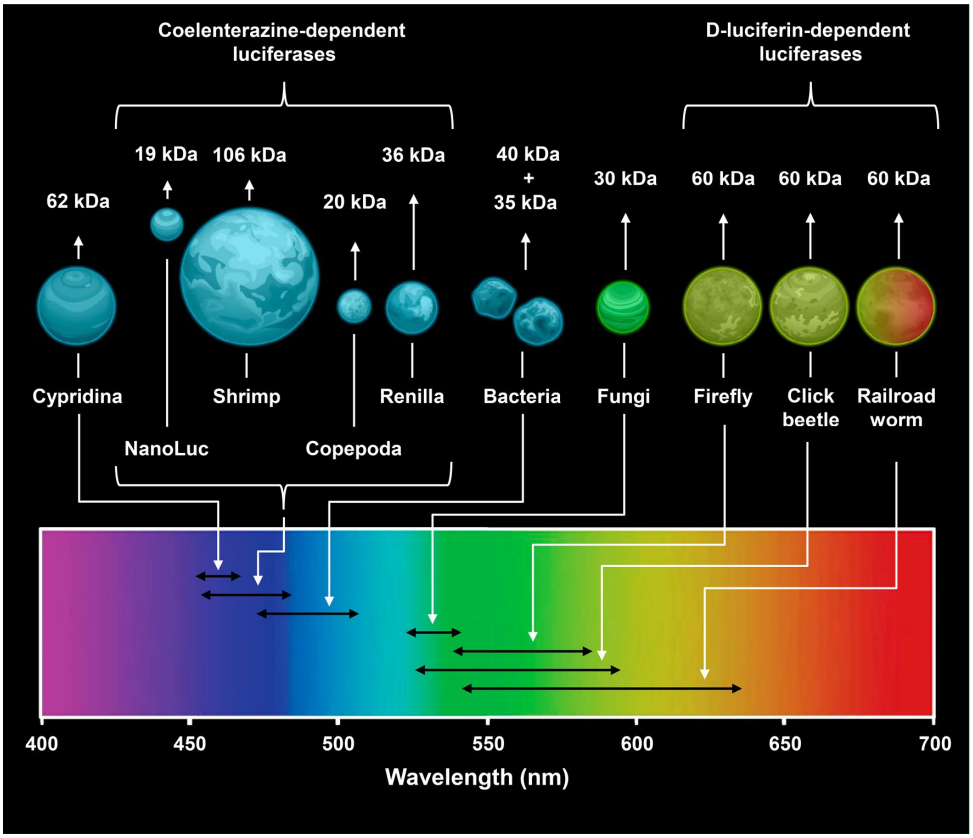


Figure 1. Luciferase wavelength palette for bioluminescent species.

BLI (*In vivo* Bioluminescence Imaging) is a developed technique used for detecting the visible photons released at specific wavelengths after the chemical process of bioluminescence [7, Fig. 2]. This technology consists of specialized cameras with highly sensitive detection systems (housed inside compartments sealed against external light) that reveal the signal coming from the sites and levels of luciferase expression activity in the test subject [8]. The strength of the signal is limited by the properties of light passing through tissue before reaching the detector, in which diffraction and absorption phenomena make imaging only possible up to a

depth of approximately 1 or 2 cm; implying that the deeper the tissue, the lower is the acquired signal [9].

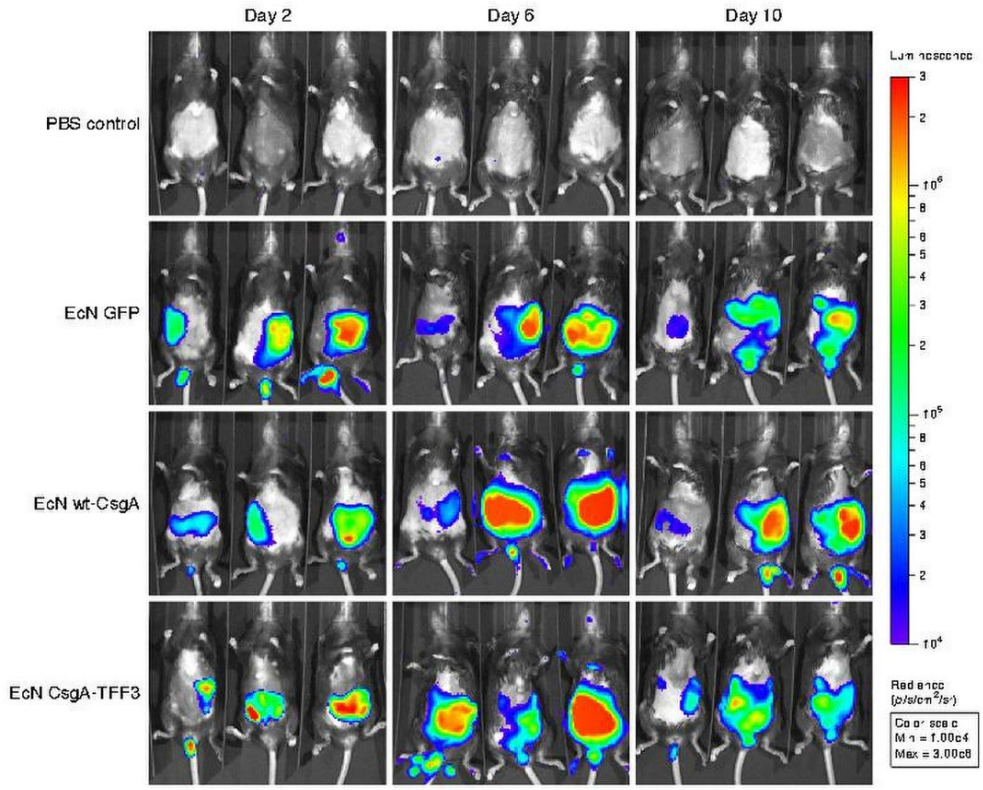


Figure 2. Representative bioluminescence imaging of engineered E. coli in the mouse gut from P. Praveschotinunt et al. [7]

The general mechanism of photon transport in biological tissue can be represented by the *Beer-Lambert Law,* which is used to describe the behavior of light when passing through a biological material where absorbance depends on variables such as molar absorptivity, the length of the light path and the density of the medium [10, Fig. 3].

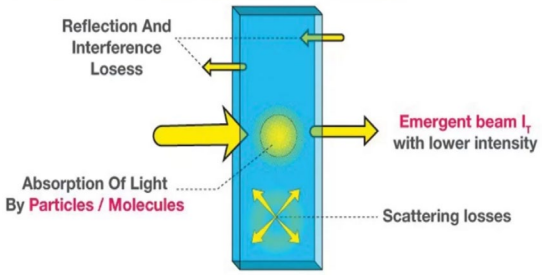


Figure 3. Beer-Lambert Law diagram.

The following model, extracted from the Digimouse mesh [11, Fig. 4] in MCX, portrays the optical properties of bioluminescence in 21 different types of tissue in a mouse:

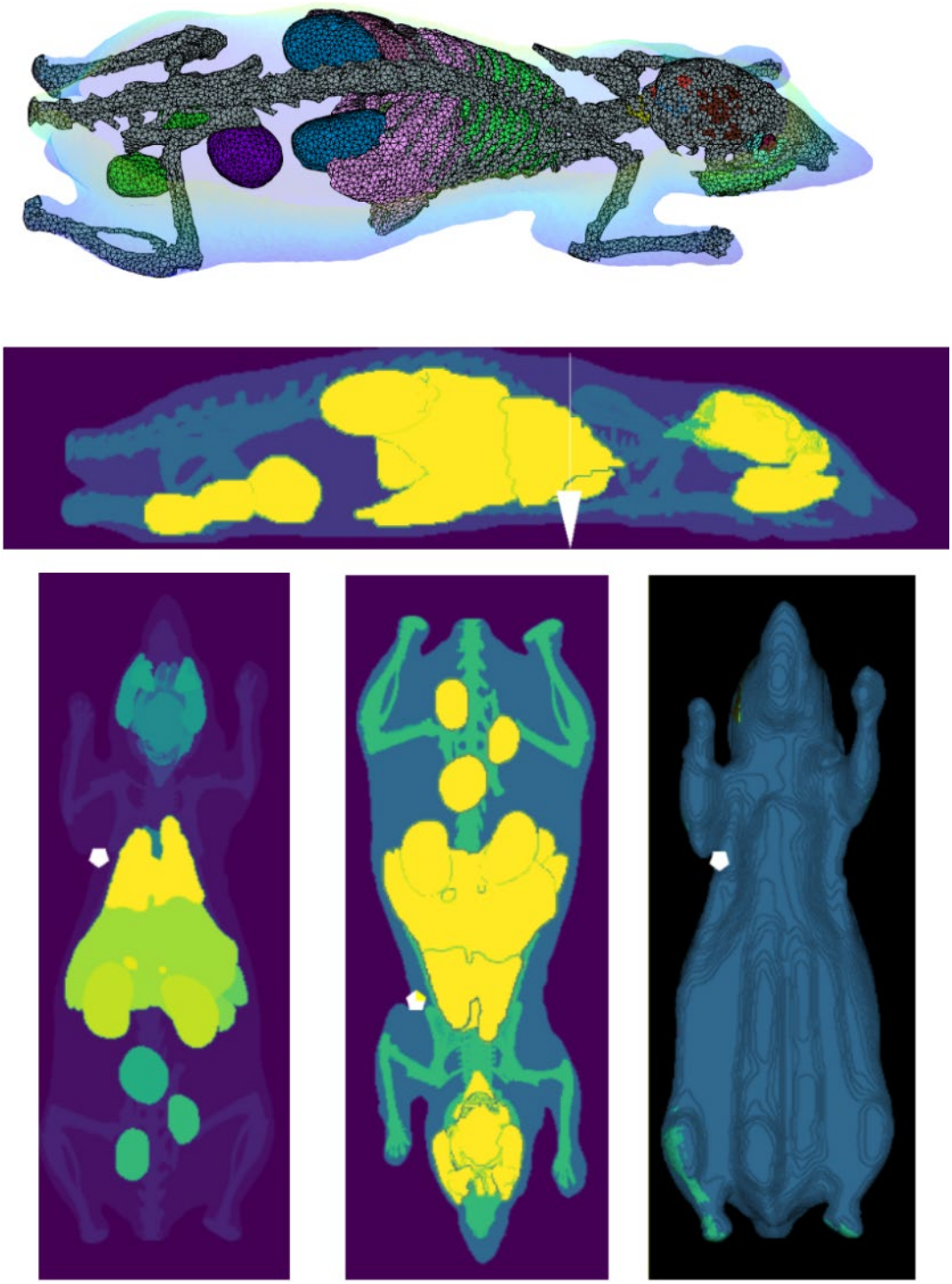


Figure 4. Bioluminescence in the MCX Digimouse.

Rapid advances in biophotonics require different efficient and model-based computational approaches to obtain useful information about biological systems that have different complexities and scales. For simulating photon transport in tissue, there is already existing software such as Monte Carlo eXtreme (MCX) and Blender Photonics (a MCX add-on for Blender use). Both work in principle with the Monte Carlo method (MC), which is considered as the gold standard in biophotonic applications due to its efficacy in simulating light propagation within heterogeneous tissues due to the random, yet accurate, approach in describing photon propagation trajectories [12].

By definition, MCX is a photon transport simulator, powered by GPU-based parallel computing, capable of performing MC photon simulations faster than a fully optimized CPU-based implementation by saving photon partial path lengths, optimizing random number generators and improving accuracy with atomic operations [13]. Blender Photonics is an integrated open-source software environment for tetrahedral 3D meshing in Blender's GUI (Graphical User Interface) in order to create complex tissue models and launch MMC light simulations [14].

In BLI, through an ordered series of scans, the luciferin-luciferase chemical reaction during a specific frame of time is commonly modeled after with a kinetic curve based on time-signal intensity. These kinetic curves are helpful to determine the best point after substrate injection to image the experimental subjects as well as analyzing the efficiency of different substances in order to determine the best choice. It is very important to model them because each behavior changes depending on the unique characteristics of the living organism that is tested such as size, therapeutic intervention, luciferin delivery method and concentration, tissue depth and the presence of scar tissue (see Fig. 5). For the case of tumor detection, factors such as location, size and geometry are considerably relevant [15].

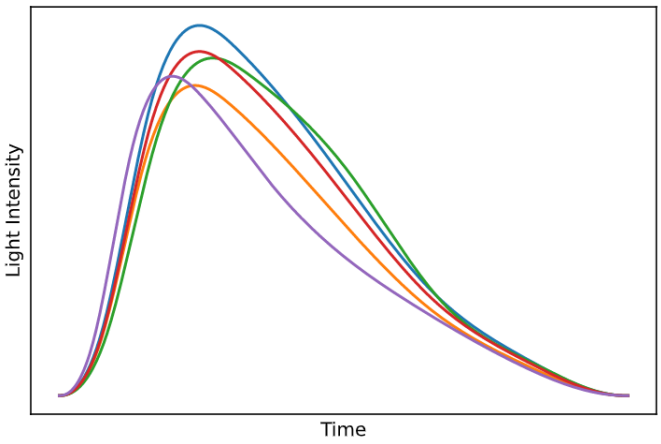


Figure 5. Qualitative kinetic curve modeling of Luciferin-Luciferase reactions.

For reproducibility and accuracy, peak signaling time must be identified on a wide number of experiments in order to be able to determine an average. This implies the constant experimentation on laboratory animals, which is in many ways unnecessary and could be avoided and can be unsustainable when there is a need to collect a large amount of data. Considering this, we seek to offer an alternative, making it possible to do effective science while avoiding as much as possible interference on animals' lives.

For our research on synthetic data generation, we decided to use mice as the base test subject model. This is because mice are the most common choice for BLI due to their physical constitution (photons can travel several millimeters through their tissue or thin bones and still emit a detectable signal) and, since they share biological traits as well as diseases with humans, they are very useful for biomedical research [16]. The existing simulations for BLI are based on the Monte Carlo method, a robust mathematical technique that offers both generality and accuracy in modeling photon transport inside tissue, nevertheless, it consumes too much time and computational resources. Therefore, for our model, we propose an innovative approach that tackles both problems through the Blender software. It consists of a parameterizable program capable of representing visually, through realistic mice models, different cases of kinetic curves describing the bioluminescence reaction in mice after luciferin injection. This allows visualization of a wide number of mice behaviors in a simple and efficient way (see Fig. 6).

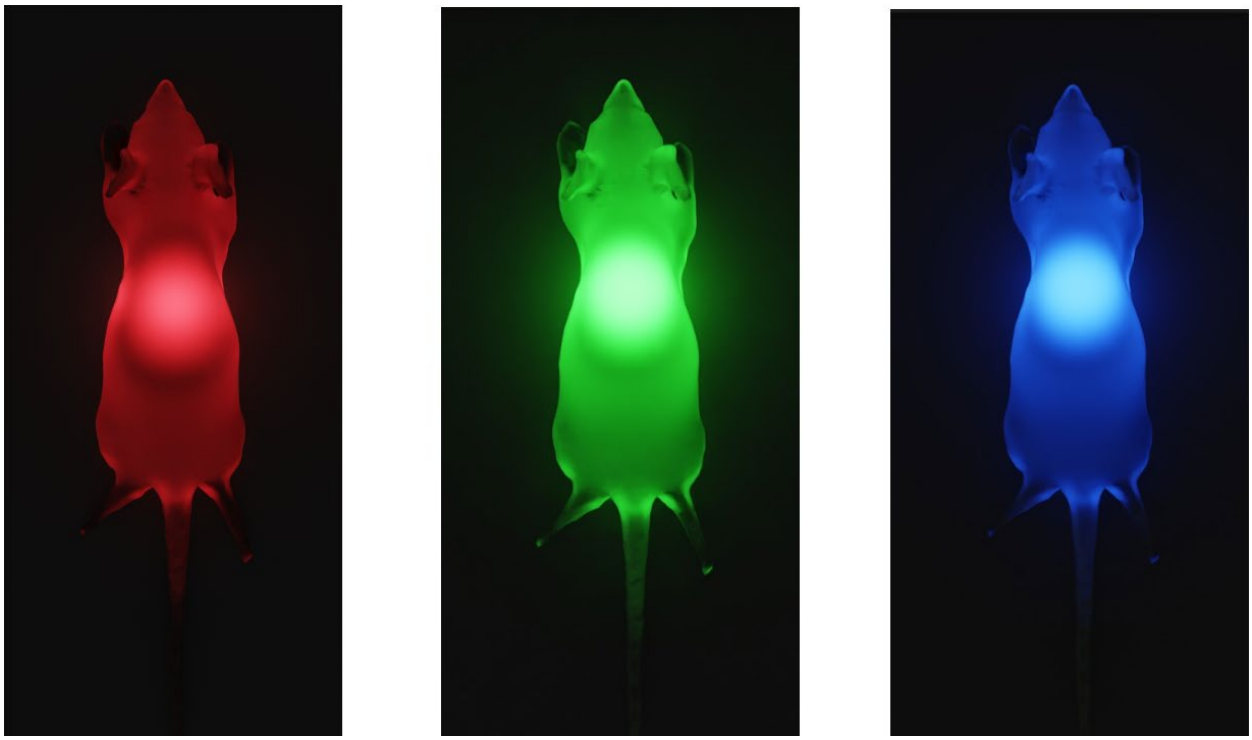


Figure 6. Sample mice created in Blender.

In BLI, a common problem that happens during imaging is noise, either thermal noise (coming from the camera) or cosmic noise, which appears in the form of completely white random pixels. For the purpose of adding realism to our simulation, we created a complementary *Python* module that creates this effect in the rendered models. We did this envisioning the great opportunities it can represent for Machine Learning applications, where the modified data has the potential of being used for training an ANN with the capacity of extracting the real signal. From both approaches, we picture how transformative bioimaging can be for areas such as modern healthcare and general research in organisms.

Materials and Methods

We decided to use Blender because it is a free and easy to use alternative even for users without a strong programming background. For initial purposes of illustration, we did multiple 3D animations to represent the different cases of BLI in mice, see Fig. 6, 10.

Our data generation system works through an automated algorithm based on Blender's *Python* scripting library. After setting initial parameters, for example the light intensity, the size and location of the mouse, the program uses an approximation of the kinetics function for luciferin to set the light emitted by the blob, with this term we refer to the light source, for each frame and approximates the mouse's natural bioluminescence using exponential decay. Additionally, we developed a robust object-oriented structure to allow for changes and modifications to the parameters of the simulation. The end user can modify the number of blobs, the minimum and maximum intensities of the blobs, the scale of the mouse, the light emitted by the mouse, the position of the blobs, and a host of other parameters. It is important to note the user can replace the mouse model or the blob provided with another model of their own choosing or add more objects into the mix and thus generate data for specific applications of bioluminescence, for example tracking of tumors. The official name of our method is *BlenderBioImaging.*

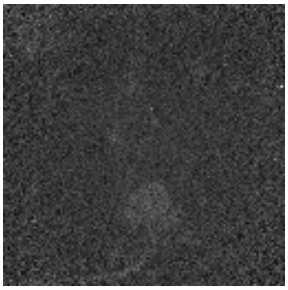
For the addition of noise to the previously generated data, we built a separate python script. To simulate the thermal noise to the camera and the cosmic noise in our synthetic data, we combined two already existing noise types. First we applied Poisson noise, also known as shot noise. This type of noise is commonly used to simulate noise occurring in photon counting in optical devices [17]. As Poisson noise is signal dependent, we scaled and rescaled the image to be able to apply different strengths of noise. On top of this we added Gaussian noise. This kind of noise occurs often in digital images when the environment has, for example, a high temperature [18]. Thus, it is a suitable choice to simulate the noise caused by the thermal radiation emitted by the camera. Gaussian noise is, in contrast to Poisson noise, not signal dependent. For this we can simply change the strength of it by changing mean and standard deviation.

To address the problem of overfitting in machine learning applications, we randomized the noise for each frame. As the raw data produced with the blender simulation consists of single frames, the last step of the script is merging them into one sequence to simulate the kinetic curve produced by the decaying luciferin.

Results

Using our script in Blender, we were able to generate large quantities of data in a short time. The source images we created do not have any noise applied yet, see Fig. 8 and 9 on the left.

For better visualization we show the peak intensity frame, meaning the frame in which the light intensity of the light source, the blob, is at its maximum.



With our noise-generation script, we were able to generate different noise strengths efficiently. The combination of the two different kinds of noise enabled us to greatly resemble real mouse data taken using LAGOX, see Fig. 7, 8, 9. LAGOX is an optical imaging system.

Figure 7. Original sample.

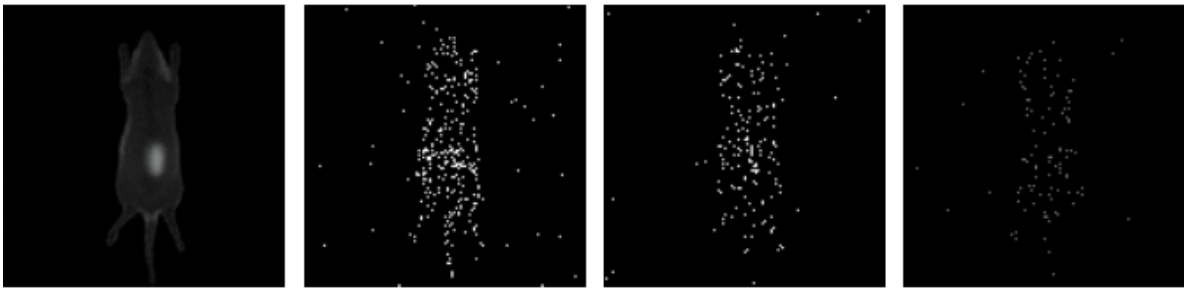


Figure 8. Poisson noise of strength 0.0, 0.3, 0.5, 1 (left to right) at peak intensity frame.

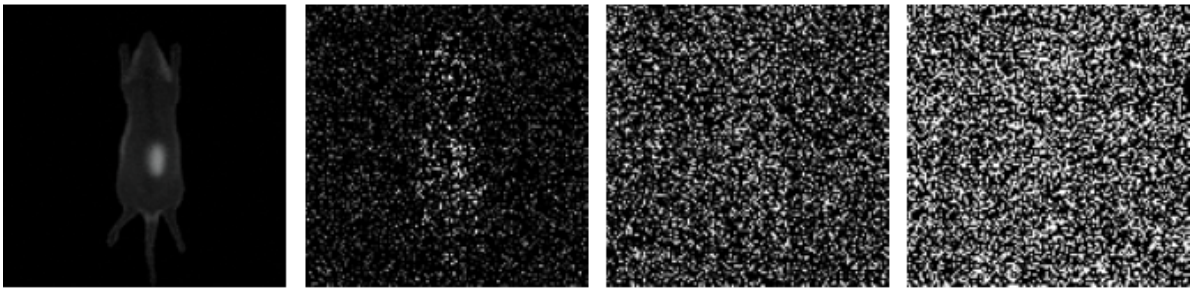


Figure 9. Combined noise of strength 0.0, 0.1, 0.9, 1.9 (left to right) at peak intensity frame.

We could also generate in Blender 3D realistic models for representing BLI applications, the following pictures show one of them, which is tumorigenesis, or tumor development tracking, see Fig. 10. We show on the top the development of lung cancer over time and at the bottom snapshots of brain and bone cancer. The choice of color is purely for visual purposes.

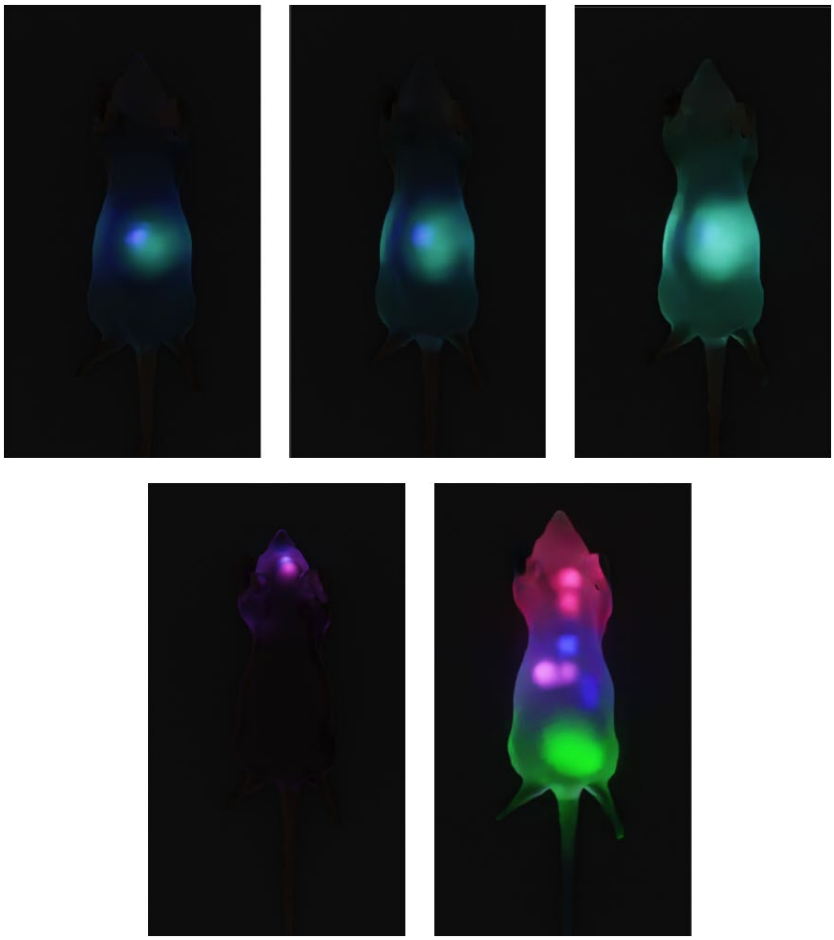


Figure 10. Tumorigenesis in the lung, brain, and bones (top to bottom, left to right) of the mouse.

Discussion

We decided to create our own approach from scratch in Blender because MCX and Blender Photonics weren't suitable enough for data generation purposes. In detail, for MCX there were many issues related to correct software installation and general interaction with the interface, which made the simulator not practical for our research purpose. For the case of Blender Photonics, we were able to install the necessary software for its use, but we noted it was very time consuming for us to get the simulation running, in addition to the fact that it wasn't

user friendly. As MCX is, computationally speaking, significantly more complex than a native Blender approach, we hypothesize the computation time per frame would very likely be much longer. The computation time per frame was critical for us, as we needed to generate large amounts of data in a short time span. Thus, our approach, BlenderBioImaging, has the advantage that it works out of the box. There is only one Python library which has to be installed for the python installation of blender.

Our discoveries may open the world of synthetic data generation to those without the resources to run some of the more computationally expensive methods other researchers have developed. The system we have developed may be useful to scientists inside and outside the biophotonics community who do not have time to scale the steep learning curve Monte Carlo simulations present. Additionally, Blender as a tool for data generation is computationally inexpensive, making it useful for generating large amounts of synthetic data for machine learning algorithm training or multi-frame data for kinetic analysis.

Because of this synthetic data generation could be a viable option and has the additional benefit of potentially reducing animal usage in clinical trials, which is a strong approach for doing ethical science in general bioimaging experimental contexts.

Nevertheless, there are also some limitations to our methods, as without the MC the simulation one cannot simulate real photon propagation in Blender. Also, we were able to resemble the real noise in the images to a large extent, but not completely. So possible usage depends on the needs regarding accuracy of the depiction of reality.

Acknowledgements

We would like to express our gratitude to the research group of Dr. Slava Kalchenko and the Department of Veterinary Resources for providing us with the necessary resources and mentoring through the course of this project as well as Aya Shkedy, Dorit Granot and Nirit Alon for coordinating the Virtual 2022 ISSI, allowing us the opportunity to participate in this amazing program.

References

[1] I. Navizet, Y.-J. Liu, N. Ferré, D. Roca-Sanjuán, and R. Lindh, “The Chemistry of Bioluminescence: An Analysis of chemical functionalities,” *ChemPhysChem*, vol. 12, no. 17, pp. 3064–3076, 2011.

[2] A. Van Praagh, “Optical Imaging Webinar: Scientific Principals and Applications,” *YouTube*, 29-Jul-2020. [Online]. Available: https://www.youtube.com/watch?v=IxQVAmLpmT4&t=1035s.

[3] D. Heymann, “Chapter 46 - Pre-clinical molecular imaging of ‘the seed and the soil’ in bone metastasis,” in *Bone cancer primary bone cancers and bone metastases*, Saint Louis: Elsevier Science & Technology, 2014.

[4] A. A. Kotlobay, Z. M. Kaskova, and I. V. Yampolsky, “Palette of luciferases: Natural Biotools for new applications in biomedicine,” *Acta Naturae*, vol. 12, no. 2, pp. 15–27, 2020.

[5] X. Ji, S. T. Adams, and S. C. Miller, “Bioluminescence imaging in mice with synthetic luciferin analogues,” *Methods in Enzymology*, pp. 165–183, May 2020.

[6] H. Simonyan, C. Hurr, and C. N. Young, “A synthetic luciferin improves in vivo bioluminescence imaging of gene expression in cardiovascular brain regions,” *Physiological Genomics*, vol. 48, no. 10, pp. 762–770, 2016.

[7] P. Praveschotinunt, A. M. Duraj-Thatte, I. Gelfat, F. Bahl, D. B. Chou, and N. S. Joshi, “Engineered E. coli Nissle 1917 for the delivery of matrix-tethered therapeutic domains to the gut,” *Nature Communications*, vol. 10, no. 1, Dec. 2019.

[8] N. D. H. B., “Chapter 36 - Neuroimaging of Brain Tumors in Animal Models of Central Nervous System Cancer,” in *Handbook of Neuro-oncology neuroimaging*, Elsevier Science Publishing Co, 2016.

[9] E. C. Lattime and S. L. Gerson, “Chapter 31 - Imaging of Oncolytic Virus Gene Expression,” in *Gene therapy of cancer: Translational Approaches from preclinical studies to clinical implementation*, Amsterdam: Elservier / Academic Press, 2014.

[10] I. Oshina and J. Spigulis, “Beer–lambert law for Optical Tissue Diagnostics: Current state of the art and the main limitations,” *Journal of Biomedical Optics*, vol. 26, no. 10, 2021.

[11] Q. Fang, “Digimouse Atlas FEM mesh,” *Monte Carlo eXtreme: GPU-based Monte Carlo Simulations*. [Online]. Available: http://mcx.space/wiki/index.cgi?MMC%2FDigimouseMesh.

[12] MN Editors, 2022. *The Beer–Lambert law*. [online] Microbiology Note. Available at: https://microbiologynote.com/beer-lambert-law.

[13] Fang, Q., n.d. Monte Carlo eXtreme. [online] Mcx.space. Available at: http://mcx.space/wiki.

[14] Zhang, Y. and Fang, Q., 2022. BlenderPhotonics: an integrated open-source software environment for three-dimensional meshing and photon simulations in complex tissues. *Journal of Biomedical Optics*, 27(08).

[15] K. Martin, “The importance of kinetic curves in bioluminescence imaging,” *GoldBio*. [Online]. Available: https://www.goldbio.com/articles/article/Importance-of-Kinetic-Curves-in-Bioluminescence-Imaging.

[16] E. C. Bryda, “The Mighty Mouse: The Impact of Rodents on Advances in Biomedical Research,” *Missouri Medicine*, 2013.

[17] T. Tyson and A. Bradshaw, 2013. Signal & Noise, UC Davis. [online] Available at: https://123.physics.ucdavis.edu/shot_files/ShotNoise.pdf.

[18] Cattin, P., 2016. *Image Restoration - Prof. Dr. Philippe Cattin - MIAC, University of Basel*. [online] Miac.unibas.ch. Available at: https://miac.unibas.ch/SIP/06-Restoration.html.

Making the Invisible Visible: Regression and Artificial Neural Networks in Low Light Signal Detection

Avraham Balsam[4], Jonas Kuehne[3], Anna Noyvert[3], Jansen Wong[4], Xinyue Yu[1]

China (Hong Kong)[1], Switzerland[2], United Kingdom[3], United States[4]

Mentored by Dr. Vyacheslav Kalchenko
Department of Veterinary Resources
Weizmann Institute of Science, Rehovot, Israel

Abstract

Data analysis is crucial for biological experiments, and often involves substantial manpower investments, high time cost, and strict requirements of accuracy. Moreover, manual detection of low-light signals is sometimes complicated and tedious. Consequently, finding efficient solutions for data generation and analysis is critical for the continued progress of scientific research. Many researchers have used machine learning, supervised and unsupervised, to detect low-light signals in biological experiments. In our study, we used some of those methods on kinetic frame sequences to recognize anomalous patterns which may not be discernible from single frame data. We used a variety of supervised and unsupervised machine learning algorithms, and eventually concluded that Truncated Singular Value Decomposition and Non-Negative Matrix Factorization are the most effective. Additionally, we explored a novel use case of the autoencoder as an effective tool for feature analysis and denoising. To make our research available to the general public, we developed a Flask application with the capacity to enhance user-submitted images. We hope that our research will open the world of low-light image analysis to a wider audience, and thereby increase the efficiency of cross-disciplinary research intersecting with biophotonics.

Motivations

Since scientists began conducting biological experiments on mice, ethical controversies were raised regarding animal experimentation. According to a new study, more than 111 million

mice and rats are experimented on annually in United States biomedical research, and the number of mice and rats used for experiments all over the world may be much larger [1]. Additionally, machine learning data generation and analysis algorithms are often more cost effective than traditional methods. Mainly based on these two factors, machine learning may be useful as a complement to traditional low-light image analysis solutions.

Machine learning

Machine learning (ML) is the subset of artificial intelligence (AI) that focuses on building systems that learn from input data [2]. In short, machine learning is performed by networks or systems that adapt to rules or regulations when given input data to match the output data. The process of machine learning involves discovering patterns in the data, making predictions, and then offering suggestions or solutions for future improvement. Moreover, once the networks or systems are built, they can be used to improve the efficiency of data analysis. Machine learning algorithms can be categorized into two types: supervised learning and unsupervised learning. In essence, supervised learning takes labeled data, where in unsupervised learning the data is unlabeled.

Materials and Methods

The data used in this study were taken from existing physical bioluminescence images, and this limited dataset was then augmented with artificially generated mouse bioluminescent images generated in Blender that accurately simulate the behavior of different luciferin-luciferase kinetic reactions that occur in test subjects.

For the supervised learning training, which requires both an underexposed input image and a well-exposed target image, we used three frames from the time sequence of both the artificial and real bioluminescent data: the first frame, the 30th frame which was determined to consistently have the highest bioluminescent signal, and the last frame. Using these frames, we set the first frame as the red channel of the resulting image, the 30th frame as the green channel, and the last frame as the blue channel. To create the underexposed input image, we multiplied the green channel with the brightest signal by 0.4 to dim the signal, and to create the well-exposed target image, we multiplied the green channel with the brightest signal by 2 to enhance the signal, see Fig. 1. Then, using these image pairs, we trained the supervised learning model.

Supervised Learning Process

In supervised learning, labeled datasets are used to train models to classify data or predict possible outcomes accurately. After weights of input data in neural networks are adjusted according to the cross-validation process, they can be applied to execute further data classification or analysis [3]. In this study, we used an existing neural network called MIRNet [4, Fig. 2], which was pre-trained to enhance low-light images, followed by transfer learning to transfer the knowledge learned from the source dataset to our target dataset. MIRNet uses the LOL dataset, composed of 500 low-light and normal-light image pairs and divided into 485 training pairs and 15 testing pairs. Each image pair in the dataset consists of a low-light input image and its corresponding, well-exposed reference image, which acts as a label and allows the network to learn how to recover high-quality image content from its degraded content [5]. During the fine-tuning process, we froze some layers in the network to ensure that the weights in these layers were not changed when the model was partially re-trained on our data. We then used our RGB image pairs to re-train the neural network, which updates the weights in specific unfrozen layers, to make it better at enhancing the light sources in our bioluminescent images. However, training of MIRNet with our dataset was accomplished with an unsatisfactory outcome (MIRNet contains over 12 million parameters and over 1000 layers). We were not able to fully finish the fine-tuning process to get the desired results, but with some more adjustments, it should be able to work as expected.

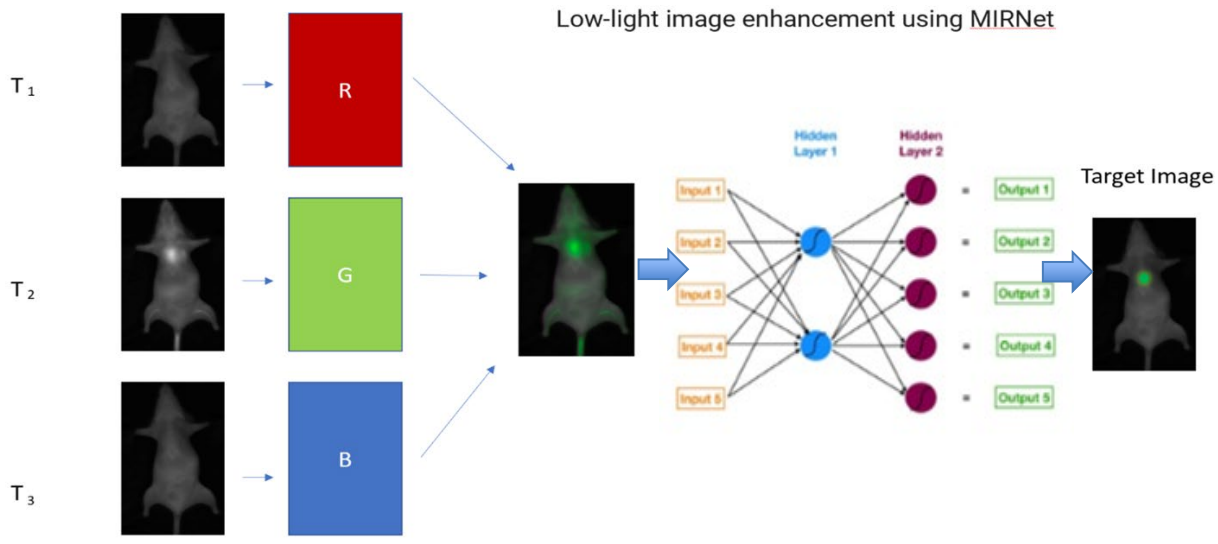


Figure 1. Process of supervised learning in this study.

Figure 2. MIRNet enhanced image examples.

Unsupervised Learning Methods

There exist many methods for unsupervised learning. One of them is dimensionality reduction. One can imagine this process like expressing a data set with much fewer data. The context of images one could imagine an object from multiple perspectives. We can, for example, imagine an object with just one isometric view. The object is not represented accurately with, for example, just the view from top. Our goal was to represent the whole sequence of images, picturing the luciferin decay, with one single image.

This, in essence, is the idea of dimensionality reduction. There exist multiple algorithms to do so. We tried a number of them. Those included principle component analysis (PCA), sparse pda, factor analysis, independent component analysis, truncated singular value decomposition (TSVD) and non-negative matrix factorization (NMF)[6]. To quantify our results, we defined an error function, which shows the distance between the right and the detected location of the brightest spot in the image. This was done for multiple noise strengths.

Autoencoder

Another approach used an autoencoder to detect anomalies in individual images. Autoencoders are a specialized type of artificial neural network, sometimes called “semi-supervised” since they lie on the border between supervised and unsupervised learning. Autoencoders are a composite of two neural networks: the encoder, which compresses the input data into a lower dimensionality vector, and the decoder, which decodes that vector to reconstruct the original image [7]. This approach has been used for noise reduction [8] and anomaly detection [9], both of which have fundamental relevance to low-light image anomaly detection.

To minimize the risk of overfitting and streamline the data analysis process, we chose not to combine the training data we generated into a single dataset. Instead, each individual frame sequence was segmented into one hundred individual images which were passed to an autoencoder as training data. This approach allowed our model to make full use of the kinetics incorporated into the training data, as it could more accurately analyze patterns in blob activation and noise intensity. The network was trained for two hundred epochs, after which it was able to successfully denoise and detect anomalies in the frame sequence it was passed. Because the autoencoder is trained on an individual frame sequence, training is relatively short (55 seconds) when compared to other machine learning algorithms.

Results

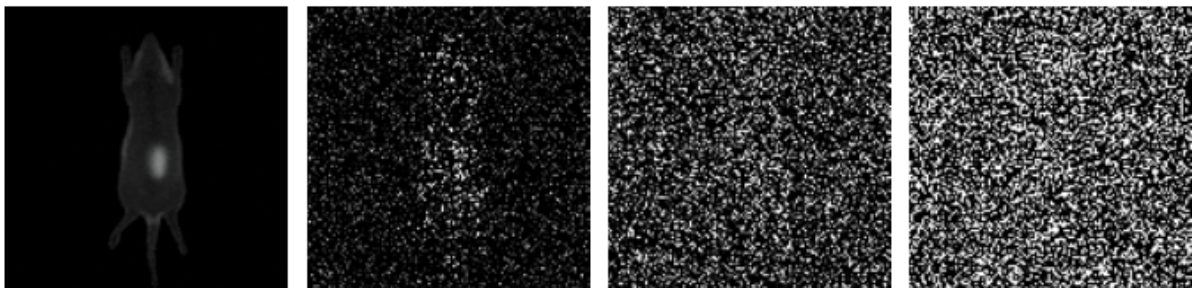


Figure 3. Combined noise of strength 0.0, 0.1, 0.9, 1.9 (left to right) at peak intensity frame

To compare performance, we required images with different noise levels. The noise in our data is a combination of Poisson- and Gaussian noise [10, Fig. 3].

Unfortunately, we did not have the time to get results with the approach using supervised learning, thus it will not be mentioned in this section.

With the unsupervised methods we used, we were able to compute the above-mentioned error measure and averaged it over all 500 samples, see Fig. 4.

It becomes clear that the TSVD and NMF algorithms outperform the rest. If computation time is taken into account, we suggest the use of TSVD, as it takes on average 0.9 seconds per sample while NMF takes 2.2 seconds. Thus, only results generated with this algorithm will be shown, see Fig. 5.

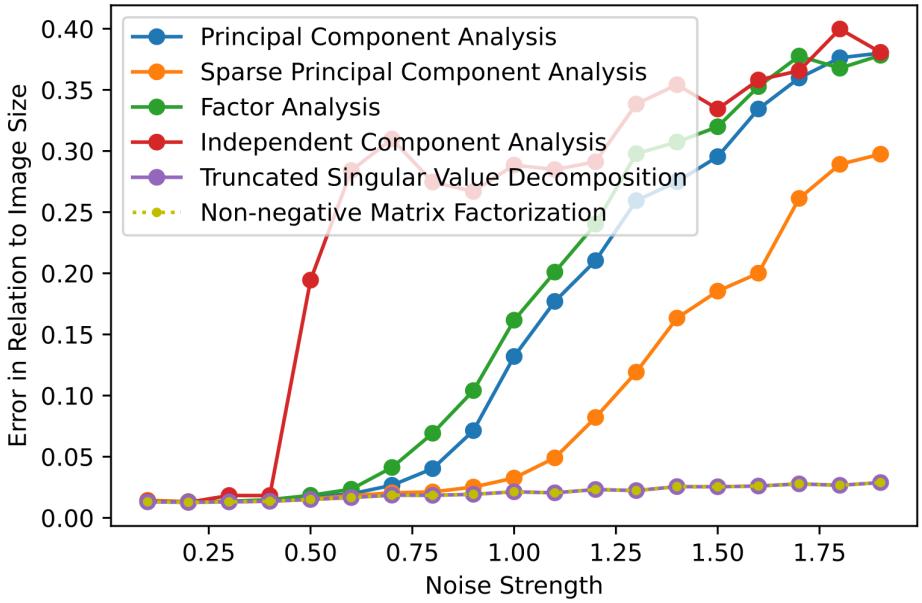


Figure 4. Performance of the different dimensionality reduction / regression algorithms.

For the auto-encoder, we were also able to create a working structure, see Fig. 5.

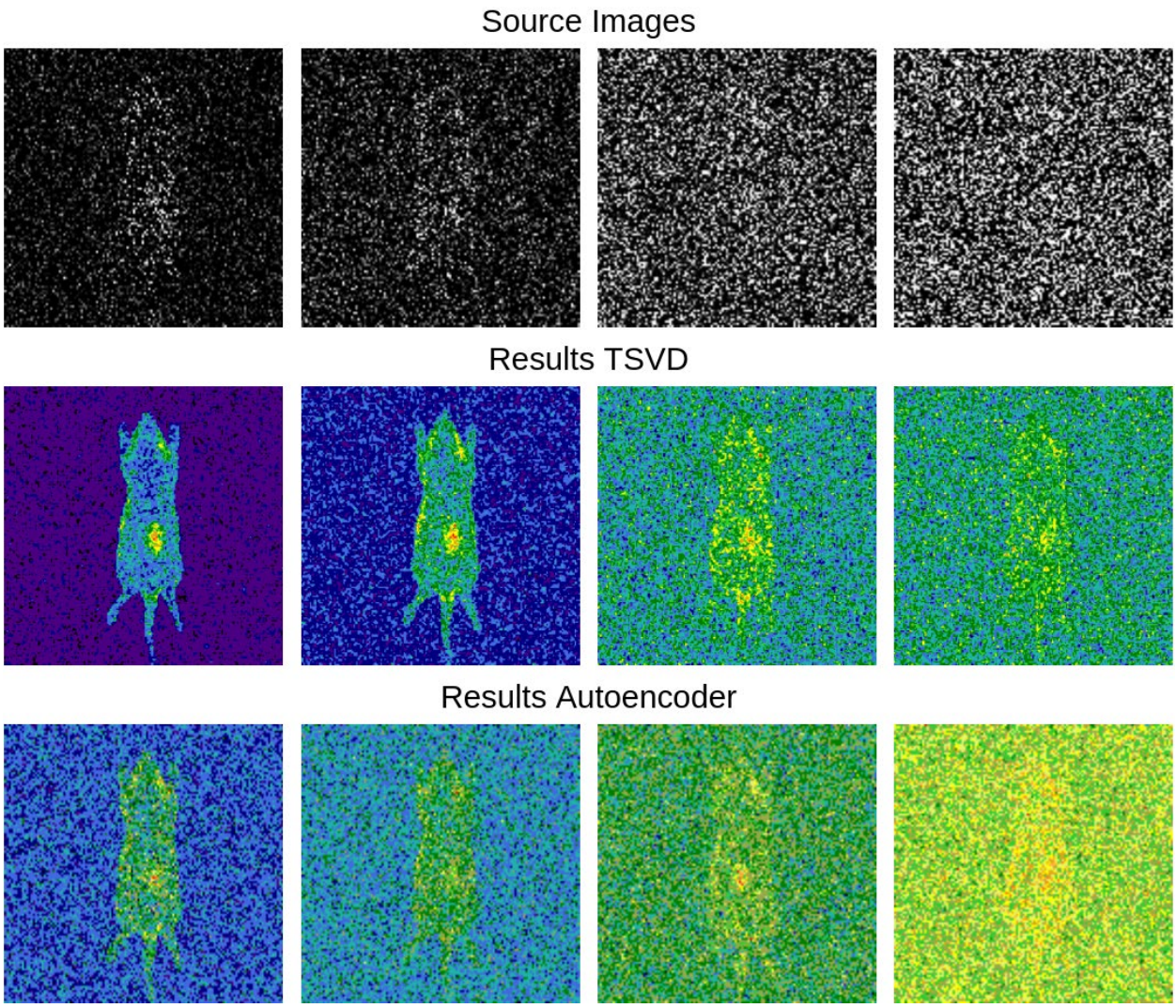


Figure 5. Results of TSVD and autoencoder on noise strengths 0.0, 0.1, 0.9, 1.9 (left to right).

To further enhance the images, we applied some minor denoising to the results generated with TSVD, see Fig. 6.

Comparison

As shown in the previous section, dimensionality reduction is marginally more effective at denoising and feature analysis than the autoencoder, although both were able to detect low-signal images at high noise intensities. With further refinement, we believe that both approaches can be improved upon. These results may be useful in the field of bioluminescence imaging as well as any other study which involves kinetic data analysis.

With the additional denoising, which could also be applied to the results of the autoencoder, the signal can be detected and shown clearly, even at high noise strength.

If we compare computation time, then the processing of each sample takes with the autoencoder (55s) about 55 times longer than with TSVD (0.9s).

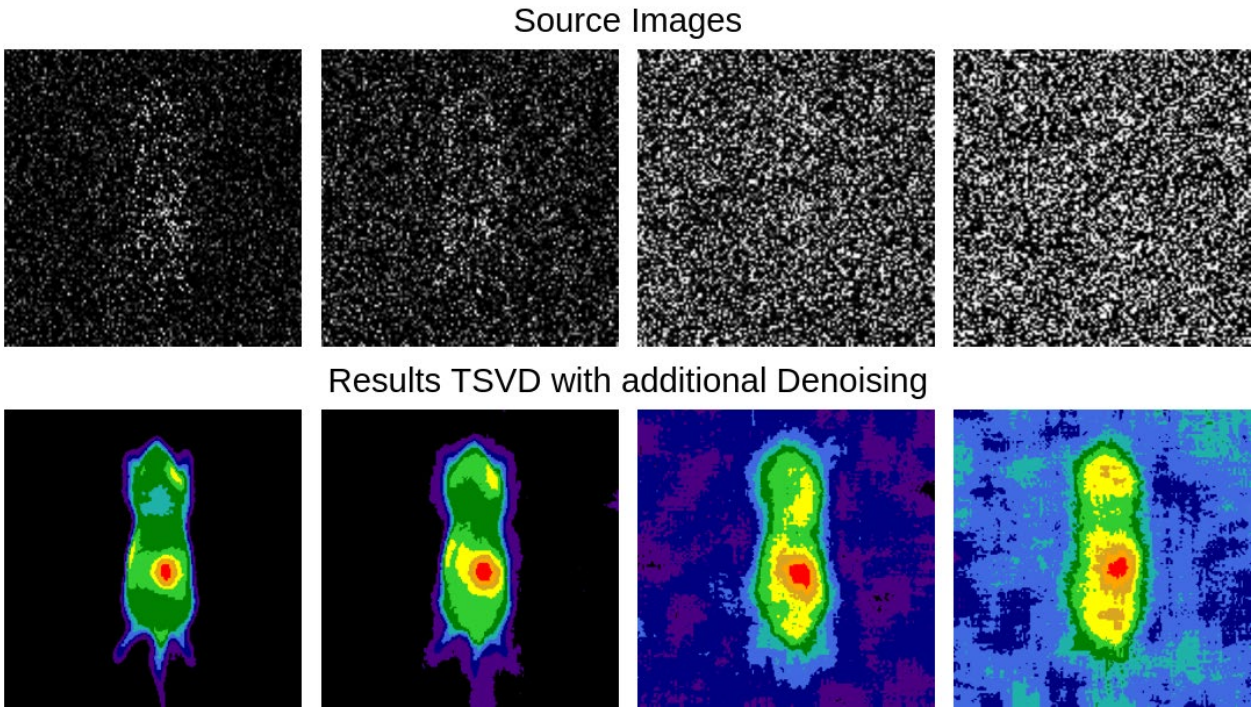


Figure 6. Results of TSVD with additional denoising on noise strengths 0.0, 0.1, 0.9, 1.9 (left to right).

Flask

To make the machine learning algorithms we used more accessible, we developed a Flask application which incorporates all of the unsupervised learning algorithms discussed in this paper, see Fig. 7. The user uploads a noisy image frame sequence, and the server enhances it and detects any anomalies that may be present in the image. We implemented a customizable form template to allow the end user to adapt our software to their own needs; the user can toggle the colormap in which images will be displayed, decide which algorithms to use, and set some low-level parameters of the autoencoder. The application was deployed using Heroku and can be accessed by anybody with an internet connection. We hope that this Flask application will open our methods to the broader scientific world and make clear the benefits of kinetic data analysis.

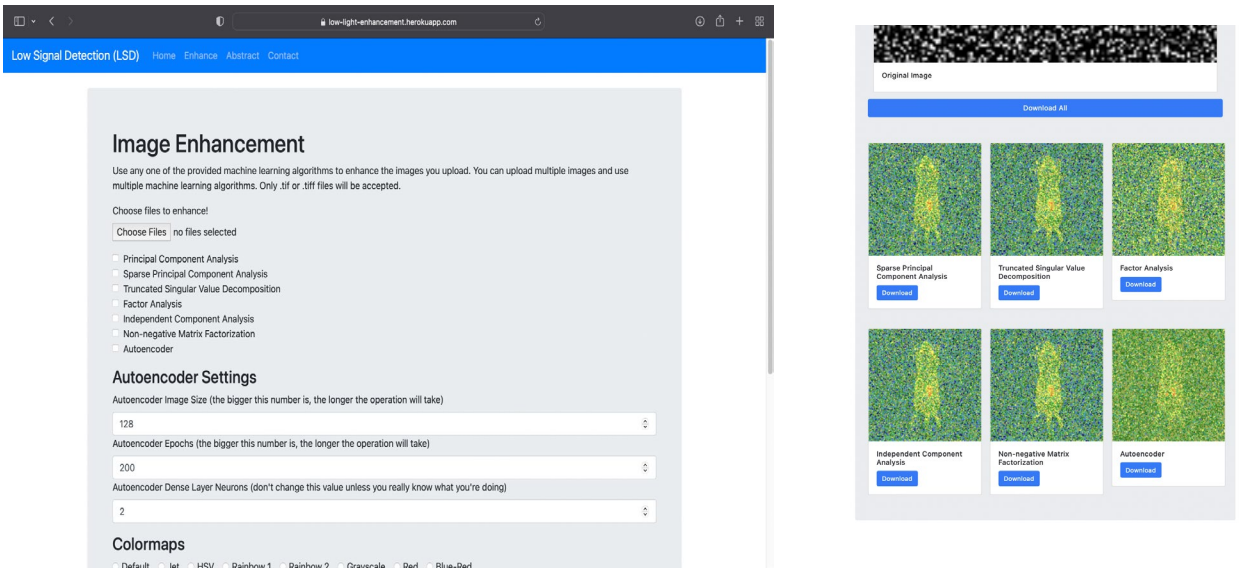


Fig. 7. User interface of Flask application

Discussion

We believe we have conclusively demonstrated the effectiveness of unsupervised machine learning techniques in analyzing noisy and low-quality images using kinetics. Using dimensionality reduction techniques, we were able to detect anomalies in an image which, to the naked eye, were indiscernible. We also have shown that for our examined data, these techniques have achieved improved results in comparison to supervised and semi-supervised machine learning techniques, both of which require long training periods and a wealth of training data. To an extent, the fact that these images behaved in a similar fashion to real-world data when subjected to analysis indicates that our synthetic data generation system accurately mimics low-light imaging conditions. However, it is still important to recognize the limitations of some of the approaches we have used in this research. In order to conserve computational resources and demonstrate that data generation is possible even without high-end software, we used Blender to generate our data. In order to rigorously test the efficacy of the algorithms we used, we must test them on real bioluminescence imaging data. Additionally, although we accounted for both gaussian noise and poisson noise, we have not tested the noise our system generates against natural noise. In the future, we hope to quantify the similarity of our data to random noise and continue to improve the accuracy of our synthetic data. Also, we will incorporate more sophisticated Monte Carlo simulations such as Blender Photonics and MCX into our model and use them to generate more accurate data.

As discussed earlier, the systems we have developed can aid the scientific community in quickening the pace and humanity of preclinical studies. Our models require no specialized knowledge of Monte Carlo systems nor a deep understanding of computer programming.

Lowering the threshold for participation and engagement with the biophotonics community will allow for more efficient and productive cross-disciplinary research, and for greater exposure of the biophotonics community to the greater scientific world.

However, while our methods are effective as a tool for kinetic data analysis, the tools we developed will not work on still images. Thus, researchers without access to multi-frame images cannot use our software. Additionally, training the autoencoder is much more time-consuming than implementing one of the other unsupervised methods, which perform marginally better. Hyperparameter optimization of the autoencoder is also more difficult than optimization of the regression methods. But these weaknesses are also strengths. The autoencoder is a complicated and powerful instrument which, when applied in the correct circumstances, can solve complicated problems. We hope that the novel approach we have developed will be used by other scientists to increase the power and effectiveness of the autoencoder in kinetic imaging. Some possible avenues of exploration include convolutional [11], variational [12], and LSTM autoencoders [13].

Acknowledgements

We would like to express our gratitude to the research group of Dr. Slava Kalchenko and the Department of Veterinary Resources for providing us with the necessary resources and mentoring through the course of this project as well as Aya Shkedy, Dorit Granot and Nirit Alon for coordinating the Virtual 2022 ISSI, allowing us the opportunity to participate in this mind-opening program.

References

[1] Grimm, D., 2021. *How many mice and rats are used in U.S. labs? Controversial study says more than 100 million*. [online] Science.org. Available at: https://www.science.org/content/article/how-many-mice-and-rats-are-used-us-labs-controversial-study-says-more-100-million.

[2] Oracle.com. n.d. *What is Machine Learning?*. [online] Available at: https://www.oracle.com/hk/data-science/machine-learning/what-is-machine-learning.

[3] Ibm.com. 2020. *What is Supervised Learning?*. [online] Available at: https://www.ibm.com/cloud/learn/supervised-learning.

[4] Zamir, S., Arora, A., Khan, S., Munawar, H., Khan, F., Yang, M. and Shao, L., 2022. Learning Enriched Features for Fast Image Restoration and Enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.1-1.

[5] Wei, C., Wang, W., Yang, W. and Liu, J., 2018. *Deep Retinex Decomposition for Low-Light Enhancement*. [online] Paperswithcode.com. Available at: https://paperswithcode.com/paper/deep-retinex-decomposition-for-low-light.

[6] scikit-learn. n.d. *Decomposing signals in components (matrix factorization problems)*. [online] Available at: https://scikit-learn.org/stable/modules/decomposition.html#decompositions.

[7] Tschannen, M., Bachem, O. and Lucic, M., 2018. Recent Advances in Autoencoder-Based Representation Learning. [online] Available at: https://arxiv.org/abs/1812.05069.

[8] L. Yasenko, Y. Klyatchenko and O. Tarasenko-Klyatchenko, "Image noise reduction by denoising autoencoder," 2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT), 2020, pp. 351-355, doi: 10.1109/DESSERT50317.2020.9125027.

[9] Z. Chen, C. K. Yeo, B. S. Lee and C. T. Lau, "Autoencoder-based network anomaly detection," 2018 Wireless Telecommunications Symposium (WTS), 2018, pp. 1-5, doi: 10.1109/WTS.2018.8363930.

[13] Balsam, A., Kuehne, J., Rodríguez de León, V., Sánchez Vargas S. and Kalchenko V., 2022. Bioimaging Data Generation for Machine Learning in Blender: an Unorthodox Approach.

[10] Guo, X., Liu, X., Zhu, E. and Yin, J., 2017. Deep Clustering with Convolutional Autoencoders. *Neural Information Processing*, pp.373-382.

[11] An, J. and Cho, S., 2015. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. [online] Available at: http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf.

[12] H.D. Nguyen, K.P. Tran, S. Thomassey, M. Hamad,
Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management, International Journal of Information Management, Volume 57, 2021, 102282, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2020.102282.

Enabling Practical Alternatives for Tumor Profiling: Spatial Transcriptomics Preserves the Cellular Resolution of Single-Cell RNA Sequencing

Adithi Adusumilli[1], Maria Fernanda Argote de la Torre[2], Sinan Arif Aramaz[3], Svenja Heß[4], Selin Kocalar[1], Federica Maestri[5], Jinho Aron Moon[4], Rohan Raghavan[3]

USA[1], Mexico[2], United Kingdom[3], Germany[4], Luxembourg[5]

Mentored by Dr. Michael Tyler, Lab of Dr. Itay Tirosh
Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel

Abstract

RNA sequencing (RNA-seq) is an extremely useful tool to study tumors and has been widely used for cellular state characterization. While in the past RNA-seq profiling could only be done by sequencing together the RNA from the whole tumor, '*in bulk*', it is now possible to profile tumors at since cell resolution with single-cell RNA-seq (scRNA-seq) by isolating and sequencing tumor cells separately. Nevertheless, this process ignores the tumor cell's spatial organization and morphology. In our research, we investigated and evaluated the role of spatial transcriptomics, where all cells and their spatial organization are retained but single-cell resolution is lost. To do this we investigated the ability to distinguish intra-tumor heterogeneity at lower resolutions by using scRNA-seq data to simulate bulk expression profiles with different sizes and spatial relationships. Our findings showed that while increasing the number of cells analyzed together does result in less distinct cell classes, spatial transcriptomics still allows for intra-tumor heterogeneity characterization.

Introduction

Transcriptomics refers to the study of RNA within biological samples and has been used extensively to study tumors [1]. In biological systems, RNA is the key molecular intermediate between genomic DNA and the proteins that DNA encodes for, which define cellular phenotypes

[2]. Transcriptomic analysis allows for phenotypic trends to be defined by levels of DNA transcription, which in cancer biology, enables the determination of various cell classes based on gene expression profiles, even within highly heterogeneous tumor microenvironments [3]. Understanding the transcriptomic activity of various cell subtypes within tumors is valuable as it can enable a deeper understanding of cancer cell biology, ultimately aiding efforts in the development of new and effective cancer therapies.

While bulk RNA sequencing of whole tumors has traditionally been used in cancer research, this approach is not suitable as it fails to provide clarity into the varying gene expression profiles of the diverse cell classes [10]. The development of single-cell RNA sequencing has partially addressed this need by allowing for the screening of individual cells [5], but the sample preparation process is both costly and biases the survival of certain cell types. As a result, single-cell RNA sequencing does not allow for accurate characterization of the tumor microenvironment and is not a widely accessible methodology due to its high costs [6].

Spatial transcriptomics, a method in which affixed tissue samples are screened in small sections of ~10 cells, poses a more promising alternative to bulk and single-cell RNA sequencing [7]. Not only does this method preserve the spatial organization and architecture of glioblastoma, but it also allows for the characterization of small clusters of cells comprising the tumor [8]. This advantage allows for further analysis of the cell classes within the tumor and the distribution of these cell classes, a key insight into the understanding of this tumor and the optimal methods of treatment. However, single-cell resolution is not achieved. Here, we seek to probe whether the loss of single-cell resolution detracts from the ability of spatial transcriptomics to elucidate the diverse cell classes within tumors.

As a test case, we use scRNA-seq data for glioblastoma multiforme (GBM), a particularly aggressive high-grade glioma and the most prevalent brain tumor. This disease's high mortality rate can be attributed to the heterogeneity of cell classes within the glioblastoma microenvironment, which makes it challenging to characterize the tumor, identify optimal treatments, and provide patients with the timely care they require. Such characteristics make it highly difficult to prevent this tumor from quickly infiltrating other parts of the brain, contributing to the low success rates of common treatments including surgery [9].

Previously, Neftel and coworkers identified 8 major cell classes through the analysis of single-cell RNA sequencing data from glioblastoma patient samples [10]. In this study, we use these results to compare cell subtype classifications between data obtained from single-cell RNA sequencing as well as simulated bulk profiles using matrix factorization while also investigating the effects of cell type proportion and spatial organization of tumors on our analysis. This dataset

was selected because of both the aforementioned heterogeneity within GBM and the well-defined cell classes Neftel and coworkers identified, which makes it ideal when it comes to comparing the scRNA-seq and simulated bulk profile results to determine the feasibility of spatial transcriptomics in preserving the resolution of scRNA-seq while efficiently characterizing glioblastoma heterogeneity.

Materials and Methods

In order to assess how cellular resolution affects detection in glioblastoma tumor RNA sequencing, the scRNA-seq gene expression dataset taken from the study conducted by Neftel et al. was investigated [10]. This study focused exclusively on the 209 cells belonging to sample MGH-100. This sample only contained malignant cells from an adult patient and contained all 4 cell states characterized by Neftel et al.

Regardless of whether it would be used as single-cell or bulk data, all sample information was normalized to TPM (Transcript Per Million) and centered. Additionally, only the 100 most expressed genes in each cell were considered during the single-cell analysis, while the top 5,000 were used in all bulk profiles. The entirety of this prep process as well as the data analysis was performed using the R Programming Language version 4.2.1.

The *Non-negative Matrix Factorization* (*NMF)* package (Version 0.24.0) was essential. *NMF* is a standard and well-trusted dimensionality reduction method that has been used in numerous publications related to gene expression analysis. This process clusters the RNAseq data in accordance with eight well-defined cellular states the glioblastoma cells exist in. As classified by Neftel et al., these are the two MES-like, AC-like, OPC-like, and two NPC-like as well as cell cycle G1 and G2 ‘expression program’ behaviors. During the course of this study, we also referenced the meta-modules provided in the Neftel paper: a consensus set of gene lists characterizing each of these programs [10]. Using NMF, we first decomposed the Neftel et al. tumor scRNA-seq expression matrix into multiple lower-rank matrices, running it with ranks three through six in order to carry out a more in-depth analysis.

Moreover, *the Jaccard index* was used to compare different meta-modules and gene sets, processing their intersections. Using this similarity index, overlaps between gene sets could be quantified, whereby a matrix of Jaccard values could be tabulated for each pairwise intersection of gene lists. We could subsequently use *hierarchical clustering with complete linkage*, which grouped together gene sets representing the same cellular state within matrices of Jaccard values. This allowed similarities between gene sets to be visualized. Using this method, we determined which gene sets were representative of the same cell classes and which individual

gene sets represented cell states not accounted for by other sets, using this data to redefine the Neftel et al. meta-modules. *Random sampling* was initially employed in our modeling of cell aggregates, drawing from a uniform probability distribution to generally account for the high levels of heterogeneity in glioblastoma.

To simulate the random samples more realistically incorporating the cellular proportions and spatial organization of the tumor, data from Neftel et al. [10] in which they reported the frequency of the six cell classes in adult glioblastomas was used. To reduce to our four cell classes the frequencies of the subtypes were summed, leading to the probability distribution OPC: 12.5; NPC: 30; MES: 21.5; AC: 36. In our classification of cells into each expression program, we calculated the average gene expression in each cell's gene-set. The maximum was used to classify each cell in accordance with the redefined gene lists.

To simulate a clustering pattern, the probability of each preceding cell type was proportionally increased by a base multiplicative factor of 2.5, which would change with each iteration according to the user-specified function parameters and cell types selected in the previous iteration. This was representative of the noted scenario whereby cells of a certain state are more likely to occur next to other cells in the same state. More precise empirical information on the spatial relationship of glioblastoma is not yet available (07.22).

Results

Single-cell data

We initially set out to accurately represent the glioblastoma microenvironment at single-cell resolution, creating a 'ground truth' model from which comparisons could be drawn in accordance with the aim of our experiment. Having run the NMF algorithm and created a single matrix composed of the 18 resulting columns or gene sets, we mapped the intersection between all 18 of them (Fig 1A). Afterward, we re-ran this same intersection analysis, only this time comparing the gene sets to the previously mentioned meta-modules defined by Neftel et al. to determine which cells (and as such, which genes) within the MGH100 sample most closely resemble each cell class (Fig 1B).

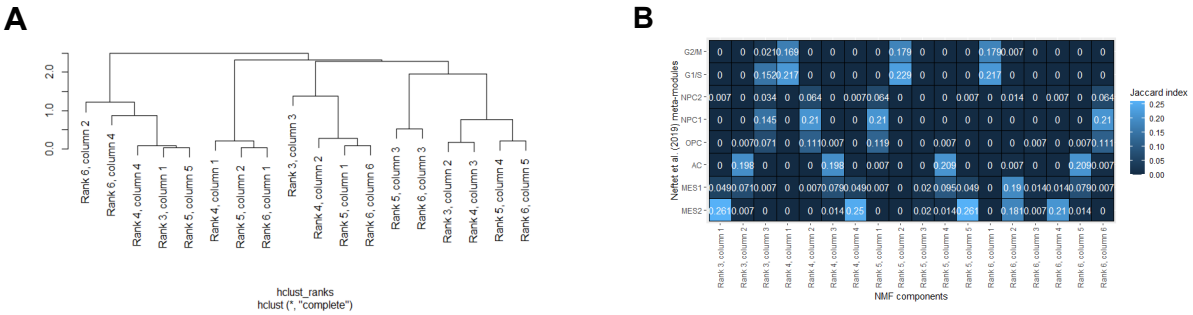


Fig 1- Intersections between MGH100 samples A) Intersection between MGH100 samples ranks 3 through 6 using hierarchical clustering. B) Intersection between MGH100 samples and Neftet et al. (2019) meta-modules

Based on these two figures, we were able to determine our 'ground truth' and define four gene lists: NPC, OPC, Mes, and AC as described in the Materials and Method section. The purpose of redefining the Neftel et al. cell classes and tailoring them to the MGH100 sample was to discard the genes that were not relevant to this study and thus minimize noise within the data. An important thing to note in Figure 1B is the relationship between the NPC and the OPC lists. As we can see, when a gene set intersects with either of them it will most likely also intersect with the other, albeit at a different level. As such, for the purposes of this study, while these two lists will be represented independently, during data analysis they will be examined together.

Simulating bulk profiles with a uniform distribution

To gauge the glioblastoma microenvironment at a lower resolution like the one observed through spatial transcriptomics, we simulated aggregates of 10, 100, and 1000 cells from the data and carried out the same process we did for our single-cell data. We initially did so by sampling from a uniform distribution, whereby 500 aggregates of random cells in the sample could be generated for each varying aggregate size. These could then be tabulated against the overall expression of each gene per aggregate. After normalizing and processing this data through NMF as before, we could then compare the resultant gene sets to our ground truth model once again using the Jaccard index. Hierarchical clustering was used, as above, to define and classify expression signatures using different cell resolutions for each expression program. This data subsequently examined the concordance between aggregate expression signatures and our ground truth model. Hence, we could determine the limiting cellular resolution for accurate tumor representation and thus detection. We also decided to discard the G1/S and G2/M classes due to the fact that they are not tumor-specific and that Neftel et al. Figure 3B shows cells in any other

state can also be cycling. Figure 2A and Figure 2B show the intersection between our single-cell and bulk data and the new gene lists.

The following two figures (Fig 2C and Fig 2D) show the comparison between these three bulk profiles, our single-cell data, and the defined gene lists from our 'ground truth' with slight variations in how the intersection means were calculated. Figure 2C took into account both hierarchical clustering and the intersection between each gene set and the four gene lists, resulting in a mean that considered all non-zero values for each gene list. Figure 2D, on the other hand, only takes the cell class each gene list or column most intersects with into account, with only OPC and NPC allowed to occur simultaneously. All the data used to plot these figures was obtained from heatmaps such as Figures 2A and 2B. These plots show that the ability to capture the meta-modules indeed decreases with larger cluster sizes, but the rate of decrease seems to slow or even plateau, suggesting that even relatively large cluster sizes are still sufficient to capture these meta-modules. In some cases, the intersection appears to increase along with cluster size, which is likely due to the randomness in cell sampling and may not truly reflect the improved performance of the NMF method or that larger clusters are better at capturing the meta-modules.

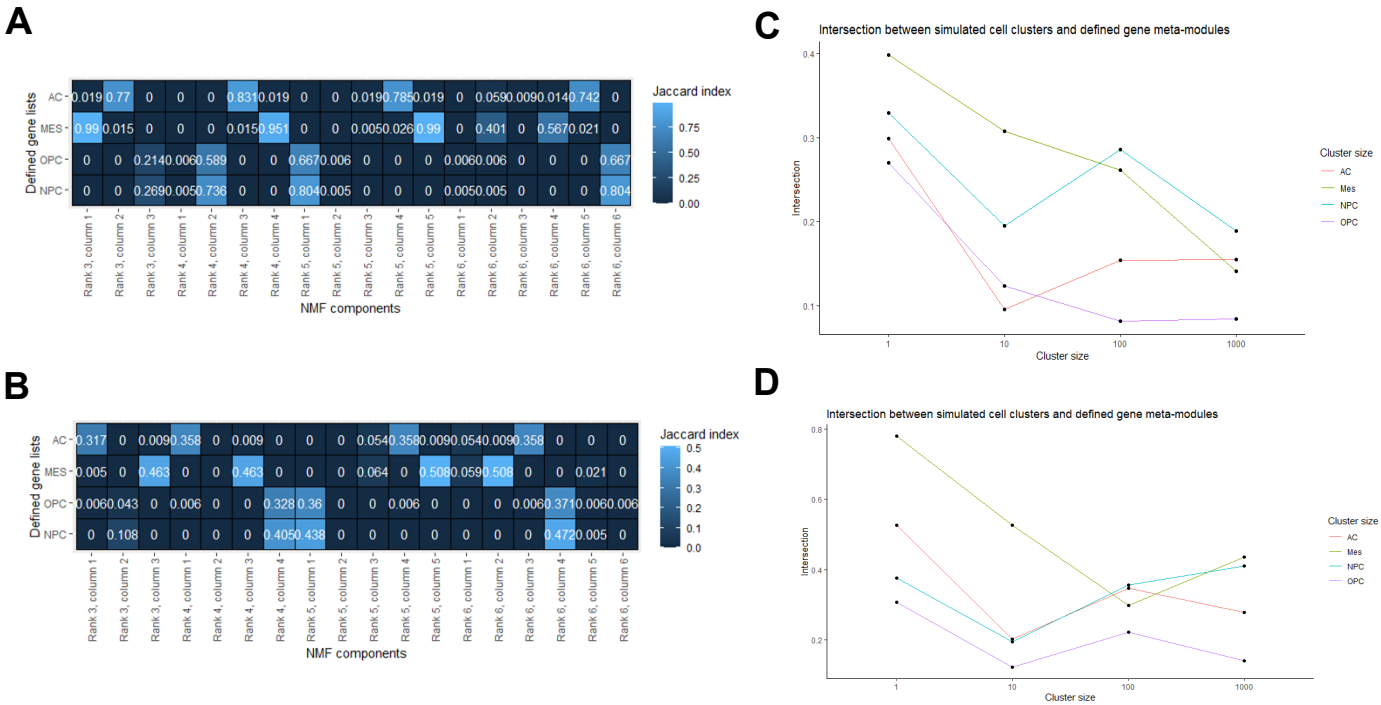


Figure 2- Intersections of single-cell and bulk data with newly defined gene lists A) Intersection between MGH100 samples and the defined gene lists. B) Intersection between MGH100 uniformly sampled 100-cell clusters and the defined gene lists. C) Intersection between MGH100 randomly sampled cell clusters and the defined gene lists considering all intersection values above zero. D) Intersection between MGH100 randomly sampled cell clusters and the defined gene lists only considering each column's highest expressed cell class.

Inclusion of spatial relationships while sampling

It is important to outline that there are many potential spatial relationships and consequently, to simplify things, we considered only the scenario where cells of each type are likely to occur next to other cells of the same type, which should be sufficient to simulate the overall degree of spatial segregation or "*compartmentalization"*. In order to simulate spatial relationships, and therefore investigate cell type proportion and spatial organization, we introduced a control multiplicative factor to avoid unrealistic cluster sizes. We used the already defined gene lists also used in the uniformly sampled analysis.

Through the algorithms we designed, we sample each cell sequentially, with the probability distribution changing depending on the preceding cell's type. If the first cell is chosen, the standard probabilities (outlined by Neftel et al.) will be chosen.

There are three different parameters: the base and control factors already explained above and the multiplicative factor. With each sampled cell, if the cell previously selected is the same type as the one currently selected, the probability for that cell type is multiplied by a multiplicative factor. However, if a different cell is chosen, all the probabilities are reset.

By keeping the cluster size constant (100 cells) and varying the multiplicative factor (0.1 and 1), we were able to clearly outline a difference, having the 1 multiplier give the best clustering. The results (Figures 3A and 3B) suggest that a higher degree of spatial segregation increases the ability to distinguish the expression programs, even at a low cellular resolution (100 cells). This suggests that lower resolutions could be tolerated for more spatially organized tumors.

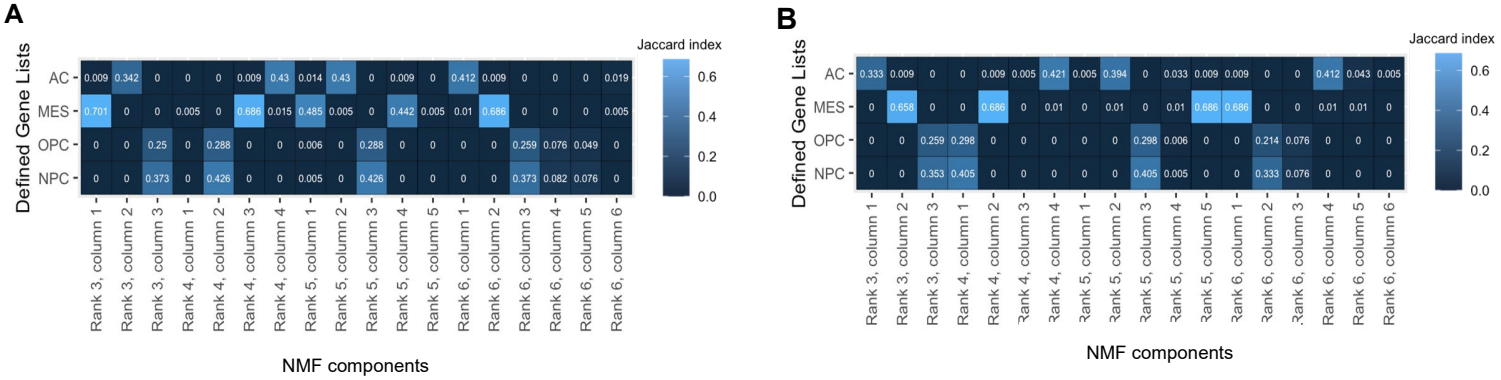


Fig 3- Cell clustering in consideration of spatial relationships within the samples A) Intersection of NMF gene signatures with meta-modules using a multiplicative factor of 0.1. B) Intersection of NMF gene signatures with meta-modules using a multiplicative factor of 1.

Discussion

Ultimately, through this study, we validated that spatial transcriptomics preserves the cellular resolution of single-cell RNA sequencing, and, thus, that spatial transcriptomics can be used as an alternative method for efficient determination of glioblastoma heterogeneity. Not only does this finding enable a more insightful approach to RNA sequencing as spatial transcriptomics provides insight into the architecture of tumors, but it also demonstrates that these additional insights can be achieved while maintaining the cost-effectiveness of the methodology itself. In the future, spatial transcriptomics may become the standard technique for studying different cancer types, more specifically in our case glioblastomas, over single-cell RNA sequencing given these benefits. It is also worth mentioning that spatial relationships could be relevant for decision-making in the clinic. For instance, a clinician can assess under a microscope how spatially segregated a tumor is and decide accordingly what resolution is needed for RNA-sequencing. Our results also suggest even lower resolutions (100 or 1000 cells) to be sufficient to characterize tumors. Hence, we could imagine more cost-effective sequencing methods such as overloading droplets as in scifi-RNA-seq or a simple dissection approach.

Due to limitations in experimental parameters, it would be essential to also look at non-malignant cells in addition to malignant cells during the clustering process to most accurately recapitulate the native tumor environment. Additionally, the computational findings should be compared to and repeated with experimental data obtained from spatial transcriptomic analyses to verify that the *in-situ* results corroborate with what is achieved experimentally. Moreover, our findings could be used to optimize a computational sampling method in order for spatial relationships to be inferred using only single-cell or bulk RNA-seq data, without *in situ* methods. This program could be employed on large scales to gain further insight into pre-existing single-cell RNA sequencing data. In fact, by applying this method to study other cancer types, we can better understand its applicability to different scenarios of varying heterogeneity. Finally, other cancer types should be investigated to assess the generality of the results.

Understanding the complex tumor microenvironments that result in large numbers of mortalities each year necessitates the availability of inexpensive, practical, and easy-to-analyze laboratory techniques. By demonstrating the high precision of spatial transcriptomics, we pave the way for such methodologies to be harnessed and applied toward ensuring safer prospects for cancer patients.

Acknowledgments

We would like to thank our mentor, Dr. Michael Tyler, for inspiring us, providing continuous feedback, support, encouragement, and constant motivation to dig deeper into our research. A special thanks to the Tirosh Lab and the Department of Molecular Cell Biology at the Weizmann Institute of Science for virtually having us and providing the opportunity of investigating such interesting topics. Lastly, we would like to thank the coordinators of the ISSI, Dr. Dorit Granot, Dr. Aya Shkedy, and Ms. Nirit Alon, for organizing all the different activities and giving us the possibility of meeting such talented and brilliant people.

References

1. Liang, Kung-Hao. “Transcriptomics.” *Bioinformatics for Biomedical Science and Clinical Applications*, Woodhead Publishing, 27 Mar. 2014, https://www.sciencedirect.com/science/article/pii/B9781907568442500036.
2. Milward, E.A., et al. “Transcriptomics.” *Encyclopedia of Cell Biology*, Academic Press, 20 Aug. 2015, https://www.sciencedirect.com/science/article/pii/B9780123944474400295.
3. Crick, Francis. “Central Dogma of Molecular Biology.” *Nature News*, Nature Publishing Group, 8 Aug. 1970, https://www.nature.com/articles/227561a0.
4. Jiang, Zhihua, et al. “Whole Transcriptome Analysis with Sequencing: Methods, Challenges and Potential Solutions.” *Cellular and Molecular Life Sciences: CMLS*, U.S. National Library of Medicine, Sept. 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6233721/.
5. Zhang, Yijie, et al. “Single Cell RNA Sequencing in Cancer Research - Journal of Experimental & Clinical Cancer Research.” *BioMed Central*, BioMed Central, 1 Mar. 2021, https://jeccr.biomedcentral.com/articles/10.1186/s13046-021-01874-1.
6. Hong, Mingye, et al. “RNA Sequencing: New Technologies and Applications in Cancer Research - Journal of Hematology & Oncology.” *BioMed Central*, BioMed Central, 4 Dec. 2020, https://jhoonline.biomedcentral.com/articles/10.1186/s13045-020-01005-x.
7. Burgess, Darren J. “Spatial Transcriptomics Coming of Age.” *Nature News*, Nature Publishing Group, 12 Apr. 2019, https://www.nature.com/articles/s41576-019-0129-z.
8. Yoosuf, Niyaz, et al. “Identification and Transfer of Spatial Transcriptomics Signatures for Cancer Diagnosis - Breast Cancer Research.” *SpringerLink*, BioMed Central, 13 Jan. 2020, https://link.springer.com/article/10.1186/s13058-019-1242-9.
9. Davis, Mary Elizabeth. “Glioblastoma: Overview of Disease and Treatment.” *Clinical journal of oncology nursing* vol. 20,5 Suppl (2016): S2-8. doi:10.1188/16.CJON.S1.2-8.

10. Neftel, Cyril, et al. “An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma.” *Cell*, ScienceDirect, 8 Aug. 2019, https://www.sciencedirect.com/science/article/pii/S0092867419306877#mmc2.

Analysing Galaxy Rotation Curve Data to Investigate and Compare the Dark Matter Theory and the MOND Theory

José Andrés Cepeda Santiago[1], Christian Dancker[2], Shashank Kalyanaraman[3], Nimrod Boshi Levine[4], Paul Philip Obernolte[2], Aamod Paudel[5], Yin Lam Wong[6]

Mexico[1], Germany[2], India[3], Israel[4], Nepal[5], and Hong Kong[6]

Mentored by Abhishek Banerjee, Lab of Prof. Gilad Perez
Department of Particle Physics & Astrophysics
Weizmann Institute of Science, Rehovot, Israel

Abstract

At the galactic scale, standard Newtonian models fail to explain the high observed velocities of stellar systems. Specifically, the observed stellar mass only accounts for around 10% of the mass required to explain this high velocity trend [1]. There are two popular solutions to this 'missing mass' problem - the Dark Matter (DM) Theory and the Modified Newtonian Dynamics (MOND) Theory. We use rotation curve data from the Spitzer Photometry and Accurate Rotation Curves (SPARC) database to test different models of MOND. In addition, we fit a one particular DM profile to the same SPARC dataset to compare with our MOND fit. Finally, we present a comparative discussion of both approaches.

1. Introduction

The rotation curve of a given galaxy is the plot of the orbital speeds of the visible stars and gas in that galaxy against the radial distance of those stars/gases from the centre of the galaxy. Fig. 3 shows the rotation curve of a disc galaxy from the Spitzer Photometry and Accurate Rotation Curves (SPARC) database [2]. Next to the observed velocities, the plot also contains the expected velocities if Newtonian dynamics were always right. These expected velocities are based on the observed mass distribution of visible matter within the galaxy. While the velocity is generally expected to decrease with an increasing distance from the galaxy's centre, it actually stays constant. We investigated two possible solutions to this problem, Modified Newtonian Dynamics (MOND) and dark matter (DM). MOND is the idea that our Newtonian Dynamics are

not complete, but need modifications in order to explain the observed rotation curves with the observed matter distribution. The other is DM, the idea that our understanding of dynamics is correct, but that the observed rotation curve is the result of thus far unobserved matter (hence called DM), that would make up 90 % of the mass of a galaxy [1]. The details are explained in 2.1 and 2.2.

2. Theoretical Background

2.1 MOND

The MOND theory has proposed a modification of Newtonian dynamics at accelerations significantly lower than those that can be found within the solar system. MOND has only one parameter a_0 (which can be seen as a new fundamental constant), and it marks the transition between the Newtonian regime and the MOND regime. Mathematically, the MOND theory can be expressed by

$$F_N = a \cdot m \cdot \mu\left(\frac{a}{a_0}\right) \tag{1}$$

where F_N is the Newtonian force and $\mu(x)$ is an interpolation function that is needed besides the parameter a_0. In order to not change the dynamics within the solar system, for $x >> 1$, the function $\mu(x)$ should approach 1. Meanwhile, for $x << 1$, the function $\mu(x)$ should approach x. For $x \approx 1$, the graph of $\mu(x)$ takes the form of a curve-fitting function connecting both asymptotes [3]. Naturally, there are an infinite amount of possible interpolation functions. We introduce six common interpolation functions in section 3.3.

A direct result from the MOND theory is the fact that the acceleration experienced by a star in the MOND regime is not inversely proportional to the square of its distance from the galaxy's centre anymore, but inversely proportional to the distance itself. Our area of research centres around two aspects of the MOND theory:

(1) In previous publications, the calculated value of the independent parameter a_0 was found to be around $1.2 \cdot 10^{-10}\ m/s^2$. We find the best value of a_0 independently and test it across 175 different galaxies provided in the SPARC database, wherein our only free parameter is a_0.

(2) We find a_0 using six common interpolation functions $\mu(x)$, and examine which of them gives the smallest value of deviation for the galaxy rotation curve fit.

2.2 DM fit

DM is matter that does not emit light, so it cannot be observed through traditional means. One way of observing it is through its gravitational pull. DM is predicted to make up around 80% of matter of our universe, while the remaining 20% is made up by baryonic (observable) matter [4].

In addition to our investigations of the MOND theory, we also adopt the DM approach to fit the galaxy rotation curves. For this purpose, we used the Navarro–Frenk–White (NFW) profile which describes the spatial distribution of potential DM within a galaxy and is based on N-body simulations. The NFW profile relates the density of DM with its radius from the galactic centre as:

$$\rho_{DM}(r) = \frac{\rho_0}{\frac{r}{R_S}\left(1+\frac{r}{R_S}\right)^2} \tag{2}$$

with ρ_0 is some density parameter and the scale radius R_S vary according to the given DM halo [5].

3. Materials and methods

3.1 Observed data

We used velocity data provided by SPARC which is an open source database created by Federico Lelli, Stacy S. McGaugh and James Schombert, using observation data mainly from the Spitzer Space Telescope. It contains the observed velocities at several radii and mass distribution data for 175 different galaxies.

On observing the increase in light wavelength (called a 'redshift') in the HI-Light emitted by interstellar nebulae, the observed velocity v_{obs} can be calculated. In 56 of the 175 galaxies, there is also observational data used from the redshift in the Hα-light, emitted by the stars in the inner parts of the galaxy, hence improving the velocity data in the inner parts of the galaxy. Additionally, SPARC gives error bars for the velocity data due to uncertainties in velocity measurement. Furthermore, SPARC contains luminosity data from the observation of the 3.6 μm hydrogen line. SPARC is used to calculate the distribution of the visible mass from this luminosity data using Mass-to-Luminosity ratios [2].

3.2 Expected Newtonian velocity

We shall consider an object of mass m with distance r from the galaxy's centre. There is a Newtonian gravitational force F_{total} acting on this object that is made up of different components:

$$F_{total} = F_{gas} + F_{disk} + F_{bul} \tag{3}$$

In this equation, F_k with $k \in \{gas, disk, bul\}$ denotes the gravitational force applied on the considered object by the galaxy gas, disk, and bulge, respectively. All forces point towards the galaxy's centre or away from it. If we assume that the considered object rotates in a perfectly circular orbit around the galaxy's centre, we have

$$F_k = \frac{|v_k| \cdot v_k}{r} \cdot m \tag{4}$$

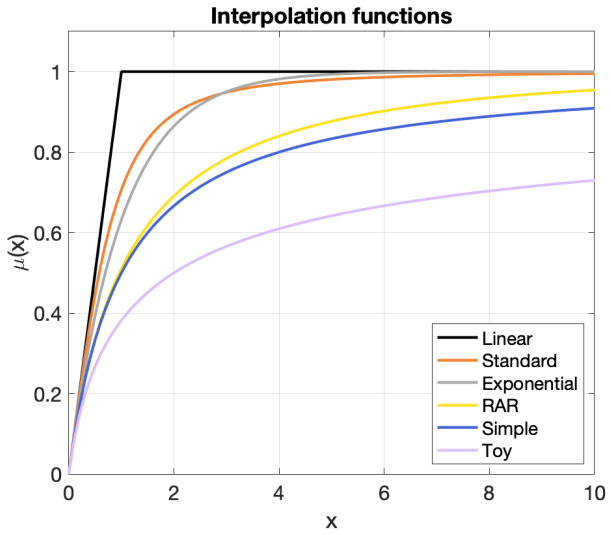
with $k \in \{total, gas, disk, bul\}$. Note, that F_k and v_k are positive or negative if they point towards the galaxy's centre or away from it, respectively. From equation (3) and (4) we can derive a formula for the expected velocity of the considered object if Newtonian dynamics are true:

$$v_{Newt,bar} = \sqrt{|v_{gas}|v_{gas} + \gamma_{disk}|v_{disk}|v_{disk} + \gamma_{bul}|v_{bul}|v_{bul}} \tag{5}$$

The velocities v_k with $k \in \{gas, disk, bul\}$ are given in the SPARC database for different radii for every galaxy [6]. They are theoretical velocities calculated from the visible mass distribution in each galaxy. These mass distributions were calculated with the assumption that the mass-to-luminosity ratio of each part of each galaxy is equal to the mass-to-luminosity ratio $\gamma_\odot$ of our sun. Thus, we need γ_{disk} and γ_{bul}, which are dimensionless corrections for the galaxy disk and bulge, respectively, compensating for deviations from $\gamma_\odot$. These deviations are partly due to the usage of different wavebands in surface photometric measurement [6, 7]. In our analysis, γ_{disk} = 0.5 $M_\odot/L_\odot$(where $M_\odot$ and $L_\odot$ are solar mass and solar luminosity respectively) and γ_{bul} = 0.7 $M_\odot/L_\odot$ were fixed as suggested by previous research [8].

3.3 MOND fit

We used MATLAB in order to calculate the expected MOND velocity v_{MOND} from the expected Newtonian velocity $v_{Newt,bar}$ (see Appendix A). We did this using six different interpolation functions – linear, simple, standard, exponential, toy, and the radial acceleration relation (RAR), respectively [9]:



$$\mu_{linear}(x) = min(x, 1) \tag{6}$$

$$\mu_{simple}(x) = \frac{x}{1+x} \tag{7}$$

$$\mu_{standard}(x) = \frac{x}{\sqrt{1+x^2}} \tag{8}$$

$$\mu_{exp}(x) = 1 - e^{-x} \tag{9}$$

$$\mu_{toy}(x) = \frac{\sqrt{1+4x} - 1}{\sqrt{1+4x} + 1} \tag{10}$$

Fig. 1. Graphs of all used interpolation functions.

$$v_{MOND,RAR}(r, v_{Newt,bar}) = \sqrt{\frac{{v_{Newt,bar}}^2}{1 - e^{-\sqrt{{v_{Newt,bar}}^2 \,/\, (r \cdot a_0)}}}} \tag{11}$$

Note that the radial acceleration relation is not expressed as an interpolation function $\mu(x)$, but as an equation directly relating v_{MOND} to $v_{Newt,bar}$. However, an interpolation function can be derived from this equation using numerical methods [9]. Fig. 1 shows a plot of the different interpolation functions. As can be observed, their main difference is how quickly they transit between both regimes.

For the interpolation functions other than RAR, the following equation can be used to convert $v_{Newt,bar}$ to v_{MOND}:

$$F_N = \frac{{v_{Newt,bar}}^2}{r} \cdot m = \frac{{v_{MOND}}^2}{r} \cdot m \cdot \mu\left(\frac{a}{a_0}\right) \Leftrightarrow \frac{{v_{Newt,bar}}^2}{{v_{MOND}}^2} = \mu\left(\frac{a}{a_0}\right) \text{ with } a = \frac{{v_{MOND}}^2}{r} \tag{12}$$

For the linear, simple, and standard interpolation function, this equation was solved for v_{MOND} explicitly. For the other interpolation functions, a simple numerical implicit equation solver was applied in MATLAB.

For our analysis, we used $N = 3607$ datapoints across 175 galaxies.[1] For these data points, we calculated v_{MOND} for different values of a_0 within a reasonable range ($a_0 \in [0.01 \cdot 10^{-10} m/s^2, 2.50 \cdot 10^{-10} m/s^2]$) with a step of $0.01 \cdot 10^{-10} m/s^2$. For each value of a_0, we calculated the mean square weighted deviation (MSWD) per degree of freedom [10]:

$$\chi_\nu{}^2 = \frac{1}{\nu} \cdot \sum_{i=1}^{N} \left(\frac{v_{obs,i} - v_{MOND}(r_i, v_{Newt,bar,i})}{\delta v_{obs,i}} \right)^2 \text{ with } \nu = N - p - 1 \qquad (13)$$

Here, p is the number of free parameters. In our case, we only vary a_0, so that $p = 1$. Our MATLAB algorithm finds the value of a_0 for which $\chi_\nu{}^2$ is minimal. The described process was repeated for each of the six interpolation functions.

3.4 NFW fit

The NFW profile gives us a density-radius relation $\rho_{DM}(r)$ for DM (see equation (2)), using which, by integration, one can determine the enclosed mass M_{DM} within a radius R_{max} of the galaxy's centre [5]:

$$M_{DM}(R_{max}) = \int_0^{R_{max}} 4\pi r^2 \, \rho_{DM}(r) \, dr = 4\pi \rho_0 {R_S}^3 \left[ln\left(\frac{R_S + R_{max}}{R_S} \right) + \frac{R_S}{R_S + R_{max}} - 1 \right] \qquad (14)$$

Using this mass, one can determine the gravitational force applied by the DM on an object of mass m at distance r from the galaxy's centre. By equating this force with the centripetal force, one can determine the velocity v_{DM}, at which the considered object is expected to rotate[2] around the galaxy's centre if the galaxy was *only* made up of DM:

$$F_{DM} = G \cdot \frac{M_{DM}(r) \cdot m}{r^2} = \frac{{v_{DM}}2}{r} \cdot m \Leftrightarrow {v_{DM}}^2 = \frac{G \cdot M_{DM}(r)}{r} \qquad (15)$$

Similar to equation (3) to (5), we then calculate the expected velocity of the object considering both DM and visible matter. We use the equation

[1] We excluded those data points from the SPARC database where F_{total} is zero or negative (see equation (2)).

[2] Again, we assume a perfectly circular rotation.

$$v_{NFW} = \sqrt{{v_{Newt,bar}}^2 + {v_{DM}}^2} \tag{16}$$

where $v_{Newt,bar}$ is calculated using equation (5). For each galaxy and for each pair of values ρ_0 and R_S, one can calculate the MSWD of the NFW fit using the following equation [10]:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{v_{obs,i} - v_{NFW}(r_i,\, v_{Newt,bar,i})}{\delta v_{obs,i}} \right)^2 \tag{17}$$

This way, the MSWD can be defined as a function of ρ_0 and R_S for each galaxy. Using the built-in MATLAB function "fminsearch" [11], we were able to find the global minimum of $\chi^2(\rho_0, R_S)$, i. e. the best values of ρ_0 and R_S for the considered galaxy (see Appendix A). This way, we found the best values of ρ_0 and R_S for each of the 175 galaxies.

Finally, we calculated the overall MSWD per degree of freedom for all galaxies. We have $\nu = N - p - 1$ degrees of freedom, where $N = 3607$ is the number of datapoints across all galaxies and where $p = 2 \cdot 175$ is the number of free parameters. Note that there are two free parameters, ρ_0 and R_S, per galaxy.

4. Results

Fig. 2 plots the MSWD per degree of freedom against all tried values of a_0 for each interpolation function. Each minimum marks the best value of a_0 for the considered function. Both the simple and RAR function as well as the standard and exponential function produce similar results. This is very natural as those interpolation functions look very similar, respectively (see Fig. 1). As one can see in Fig. 2 as well as in Table 1, simple and RAR produce the best fits.

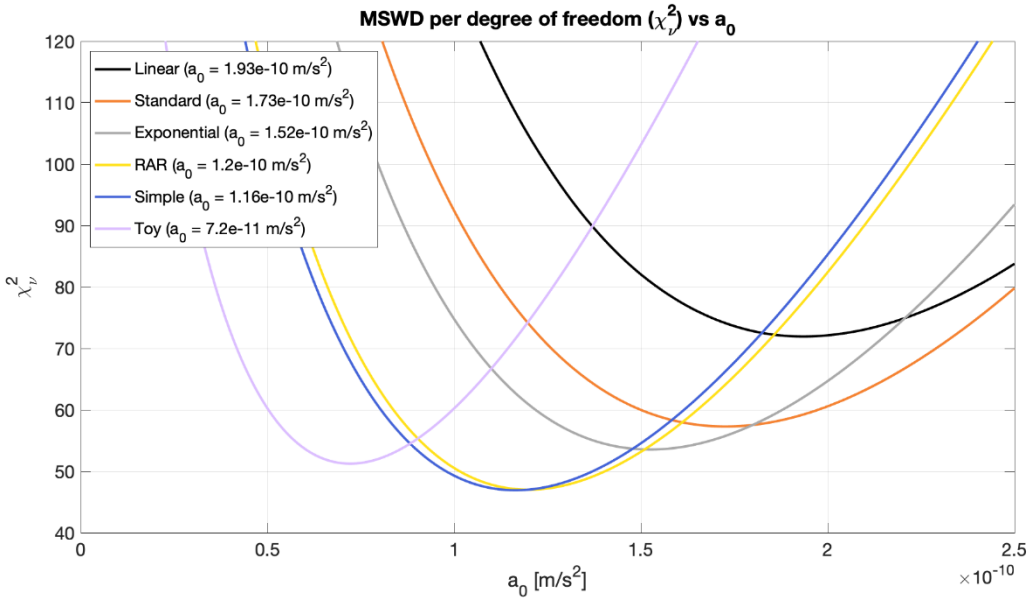


Fig. 2. The overall MSWD per degree of freedom plotted against a_0 for each interpolation function.

The best a_0 value reaches from $0.72 \cdot 10^{-10}\ m/s^2$ for the toy function to $2.17 \cdot 10^{-10}\ m/s^2$ for the linear approach. Thus, the spread is very high. However, the a_0 values obtained by RAR ($1.20 \cdot 10^{-10}\ m/s^2$) and simple ($1.16 \cdot 10^{-10}\ m/s^2$) should be valued most as the MSWD per degree of freedom is least for these functions.

Notably, the MSWD per degree of freedom is substantially smaller for the NFW fit (see Table 1). Additionally, there are more galaxies for the NFW fit which have a MSWD per degree of freedom that is smaller than 1, so that the fit can be considered good. These can be attributed to the fact that each interpolation function uses one best a_0 for fitting all galaxies while NFW fits each galaxy with the respective best pairs of ρ_0 and R_S. Still, for some galaxies, both the MOND and the NFW fits are not satisfactory[3], partly due to the fact that the error bars on the observed velocities are sometimes very small (see Appendix B), maybe incorrectly so [6]. For detailed information on the MSWD and the NFW parameters ρ_0 and R_S obtained for each galaxy, please refer to Appendix C and D. Note that the extreme values of ρ_0 and R_S shown in Appendix C (e. g. for the galaxy PGC51017) are known as the Core-Cusp problem of the NFW profile [12].

[3] If the MSWD per degree of freedom of a galaxy for a particular fit is smaller than 1, the result is considered highly satisfactory. However, if it is much larger than 1, the result is considered not satisfactory.

Table 1. The best value of a_0 for each interpolation function and the overall MSWD per degree of freedom for each interpolation function and the NFW fit for comparison. The MSWD per degree of freedom is calculated using the best value of a_0, respectively.

Fitting approach	Best a_0 $[m/s^2]$	${\chi_v}^2$ for all galaxies	Num. of galaxies with ${\chi_v}^2 < 1$
Linear	1.93E-10	71.94	5
Standard	1.73E-10	57.29	10
Exponential	1.52E-10	53.55	9
RAR	1.20E-10	47.01	13
Simple	1.16E-10	46.92	11
Toy	7.20E-11	51.25	10
NFW fit	-	8.86	78

The made observations are supported by Fig. 3 and Appendix A. The rotation curves obtained by RAR and simple are very similar again.

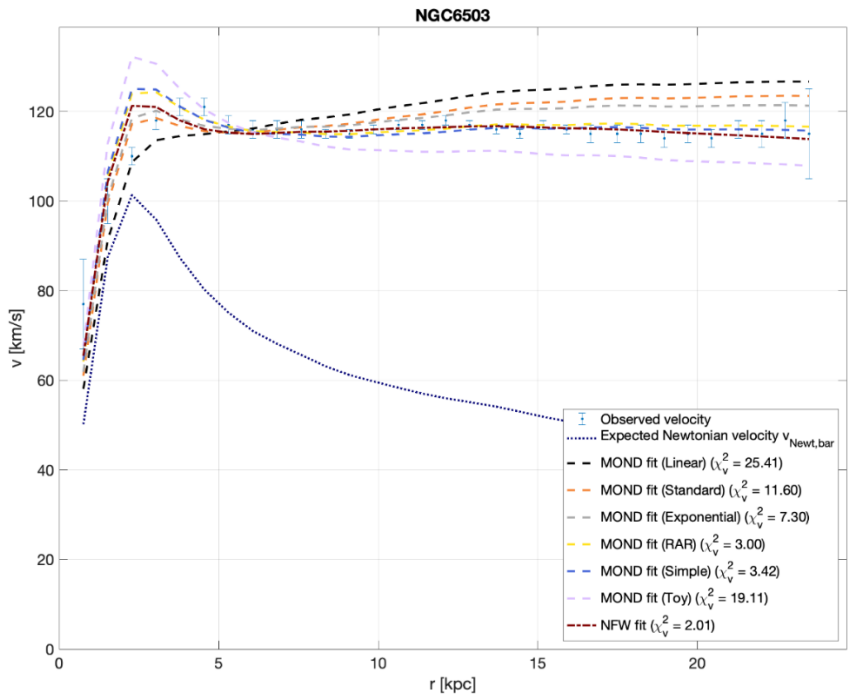


Fig. 3. Rotation curve for the galaxy NGC6503 containing the observed velocity and expected velocities for different MOND interpolation functions as well as the NFW fit. Note that each MOND fit relies on the best a_0 value overall of the respective interpolation function.

5. Discussion

The method used was similar to that employed in previous research [6]; however, the comparing research made use of a Markov Chain Monte Carlo simulation (MCMC) in order to optimise the MOND fit by adjusting several measured parameters (γ_{disk}, γ_{bul}, galaxy distance, and disk inclination) within their margin of error. Naturally, given the use of more than a single parameter, the fits they presented were much better in terms of MSWD per degree of freedom. However, the values of a_0 obtained by our method using the RAR and simple functions are still consistent with other research [6-9, 13]. We understand that the mass-to-luminosity ratios γ_{disk} and γ_{bul} are not entirely constant for each galaxy and have to be varied around their expected values $0.5\ M_\odot/L_\odot$ and $0.7\ M_\odot/L_\odot$ [8], respectively, in order to improve the quality of our results. With our method, we currently only consider the error bars on the observed velocities. However, our calculations of MOND and NFW velocities rely on other observed quantities, e. g. luminosity data. Because we do not consider the error bars on these quantities, our calculated deviations are unreasonably high. Hence, this is a possible area of improvement of our research that could be realised by using a MCMC simulation, for instance.

The meaning of a_0 and its physical relation to the universe is unknown. Previous research has established a relationship between the value of a_0 and the Hubble-Lemaitre constant, the unit of measurement to describe the universe expansion [14]. The relation has been expressed by previous research for a_0 to be equal to $cH_0/2\pi$, where H_0 represents the value of the Hubble-Lemaitre constant of the present universe [9]. Whether this relationship could be coincidental or not is still unresolved, but should not be disregarded for that matter and should be taken into consideration in future research to identify the physical meaning behind the value of a_0.

Furthermore, the scope of MOND remains unclear, towards whether it is limited to Newtonian gravitation or Newton's Second Law to its full extension. While MOND very well explains the behavior inside a single galaxy, currently it does not completely explain the dynamics of galaxy clusters [15]. There are theories towards how MOND could explain the dynamics of galaxy clusters, but these theories still need further examination. This is the same with the bullet cluster [16], which is viewed by many as a proof for the existence of DM. One theory exists, which explains the bullet cluster with MOND [17], but is viewed widely as the less plausible option than DM. Additionally, there is no accepted relativistic theory of MOND and no way, how it can be derived from general relativity. There is one theory to generalise MOND in a relativistic manner, called TeVeS [18]; however, TeVes itself shows not to be consistent with measurements done by

LIGO [19]. On the other hand, there is no observation or any direct indication of the existence of DM. The existence of DM is only assumed to explain the behavior of galaxies, galaxy clusters and the early universe. As with MOND, there is no convincing evidence so far that DM exists. DM is a very convenient theory as it allows for many free parameters, so one can assume its presence whenever it is needed to explain a certain behavior. Therefore, is appears quite obvious, that DM will always provide nearly perfect fit, while the MOND fits, although remaining in an acceptable range, have a bigger $\chi_v{}^2$ than DM, given that MOND has only one free parameter. Hence, it remains impossible to say whether MOND or DM are the more plausible explanation as none of them is completely convincing.

References

[1] S. Trippe, Walter de Gruyter, **69**, 173-187 (2014) https://arxiv.org/abs/1401.5904v1

[2] F. Lelli, S. S. McGaugh, J. M. Schombert, Astrophys.J., **152**, 157 (2016) https://arxiv.org/abs/1606.09251

[3] M. Milgrom, Astrophys.J., **270**, 365-370 (1983) 10.1086/161130

[4] D. Bolles, Dark Energy, Dark Matter, 2022, https://science.nasa.gov/astrophysics/focus-areas/what-is-dark-energy

[5] J. F. Navarro, C. S. Frenk, S. D. M. White, Astrophys.J., **462**, 563-575 (1996) https://arxiv.org/abs/astro-ph/9508025

[6] P. Li, F. Lelli, S. S. McGaugh, J. Schombert, A&A, **615**, A3 (2018) https://arxiv.org/abs/1803.00022

[7] F. Lelli, S. S. McGaugh, J. M. Schombert, M. S. Pawlowski, Astrophys.J., **836,** 152 (2017) https://arxiv.org/abs/1610.08981

[8] S. S. McGaugh, F. Lelli, J. Schombert, Phys. Rev. Lett. **117**, 201101 (2016) https://arxiv.org/abs/1609.05917

[9] X. Li, S. Zhao, H. Lin, Y. Zhou, Chinese Physics C, **45**, 025107 (2021) https://iopscience.iop.org/article/10.1088/1674-1137/abce53

[10] P. Young, (2012) https://uol.de/f/5/inst/physik/ag/compphys/download/Alexander/dpg_school/bh_notes_peter3.pdf

[11] MathWorks, fminsearch, https://la.mathworks.com/help/matlab/ref/fminsearch.html?s_tid=mwa_osa_a

[12] W.J.G. De Blok, Advances in Astronomy, **2010**, 789293 (2009) https://arxiv.org/abs/0910.3538v1

[13] S. S. McGaugh, Astrophys.J., **143**, 40 (2012) https://arxiv.org/abs/1107.2934

[14] M. Milgrom, Phys. Rev. D **100**, 084039 (2019) https://arxiv.org/abs/1908.01691

[15] M. Milgrom, Phys. Rev. D **98**, 104036 (2018) https://arxiv.org/abs/1810.03089

[16] D. Clowe, M. Bradač, A. H. Gonzalez, M. Markevitch, S. W. Randall, C. Jones, D. Zaritsky, Astrophys.J., **648**, L109-L113 (2006) https://arxiv.org/abs/astro-ph/0608407

[17] G. W. Angus, B. Famaey, H. S. Zhao, MNRAS, **000**, 1-14 (2006) https://arxiv.org/abs/astro-ph/0606216v1

[18] J. D. Bekenstein, Phys. Rev. D **71**, 069901 (2005) https://arxiv.org/abs/astro-ph/0403694v6

[19] S. Boran, S. Desai, E. O. Kahya, R. P. Woodard, Phys. Rev. D **97**, 041501 (2018) https://arxiv.org/abs/1710.06168

Acknowledgments

We would like to thank Abhishek Banerjee for mentoring through the course of this project. Also, we would like to thank Prof. Mordehai Milgrom for discussing his MOND theory with us and providing further insight. Finally, thanks to Dr. Aya Shkedy, Dr. Dorit Granot, and Nirit Alon for their organisation of and assistance throughout the whole ISSI virtual camp.

Appendix A: Open-source MATLAB code

In order to do the necessary curve fitting computations for the MOND theory with different interpolation functions as well as for the NFW profile we created an extensive package of MATLAB functions. These functions were also used to generate all plots and tables in this paper including the appendix. The code is completely open source and can be found under https://github.com/paul019/issi2022-mond. Using this code, one can experiment with MOND and NFW using the SPARC database. Of course, the code also includes a full documentation.

Appendix B: More galaxy rotation curves

Rotation curves for the galaxies UGC01281, NGC5055 and NGC2366 containing the observed velocity and expected velocities for different MOND interpolation functions as well as the NFW fit. Note that the three galaxies (including the galaxy from Fig. 3) were chosen to represent different qualities of fit.

Figures of the three galaxies can be seen at
https://drive.google.com/drive/folders/1WYOf5WPf31YYy_4hSwhxW-joIfb-rZff?usp=sharing
as AppendixB.1, AppendixB.2 and AppendixB.3, respectively.

Appendix C: Detailed information on the MOND and NFW fit for each galaxy

Note that the table is ordered by $\chi_\nu{}^2$ for the interpolation function “simple”. The quality flag is a metric from the SPARC database that represents the respective quality of measured data (1/2/3 means high/medium/low quality, respectively).

The table can be seen at: https://docs.google.com/spreadsheets/d/1U1ngcq-K9cf5-VtZtJfIMjX2dbLFp_jL/edit?usp=sharing&ouid=112495751641776499570&rtpof=true&sd=true

Appendix D: Histogram of χ_ν^2 for each interpolation function and the NFW profile