Goodness-of-Fit Testing using Wasserstein Distance to Validate Model of DEC Background

REU Program at Columbia University - Nevis Labs

Eleni Mandarakas[1]

[1]Bucknell University

January 12, 2026

Abstract

Dark matter is the leading explanation for the observed discrepancy between visible matter and the total gravitational mass in the universe. Detecting a dark matter particle would not only confirm its existence but also deepen our understanding of the fundamental constituents of matter and the structure of the universe. The XENON Collaboration is dedicated to advancing direct detection efforts, targeting dark matter candidates known as Weakly Interacting Massive Particles (WIMPs) in the mass range GeV - TeV. Achieving this goal requires effective background-signal discrimination. One such background arises from double electron capture (DEC) events. Models have been developed to describe the DEC background, and this study investigates whether the Wasserstein distance as a Goodness-of-Fit test is sensitive enough to evaluate model agreement with observed data. This study shows that the Wasserstein test has the most power to detect differences in hypotheses when the LL parameter is positive and the LM parameter is negative and the least power when the parameters are close to the null hypothesis or when the parameters are close to each other. Overall, it has shown to be a powerful test in some regions, but future work in a two tail test would be suggested to gain power specifically in the regions of greatest interest.

Contents

1 Introduction

Astrophysical and cosmological observations, including measurements of the cosmic microwave background, gravitational lensing, galaxy rotation curves, and the dynamics of the Bullet Cluster, provide compelling evidence for the existence of a non-luminous, non-baryonic form of matter known as dark matter. This hypothetical substance does not appear to interact via the electromagnetic force, making it extraordinarily difficult to detect directly. Its presence is inferred solely through gravitational effects on visible matter, radiation, and the large-scale structure of the universe.

2 XENON Collaboration

To search for direct evidence of dark matter, the XENON collaboration utilizes a dual-phase liquid xenon time projection chamber (TPC) designed to detect rare interactions between dark matter particles, particularly weakly interacting massive particles (WIMPs), and xenon nuclei. The detector consists of a cylindrical volume filled with liquid xenon, with a thin layer of gaseous xenon above it. Photomultiplier tubes (PMTs) are positioned at both the top and bottom of the chamber, while a cathode and anode establish an electric field throughout the detector.

When a particle interacts with a xenon atom in the liquid phase, the collision produces prompt scintillation photons (referred to as S1) and ionization electrons. The electrons drift upward under the influence of the electric field and enter the gas phase, where they are accelerated by a stronger field and produce a secondary scintillation signal (S2) via electroluminescence. The signals are then reconstructed during data analysis and grouped together as events. Electronic recoil (ER), which constitutes most of the background, refers to collisions of particles with the electrons in the xenon atom, whereas nuclear recoil (NR) refers to collisions of a WIMP with the nucleus of a xenon atom. The ratio of S2 to S1 provides a means of discriminating between ER and NR, as ER events typically exhibit a higher ionization yield and thus a larger S2/S1 ratio.

Despite this discrimination, certain rare background processes can mimic NR signals. Notably, the double electron capture (DEC) decay of xenon isotopes such as ^{124}Xe can produce signals with S2/S1 ratios consistent with those of nuclear recoils. These events pose a challenge to background rejection efforts, as they may be misidentified as potential dark matter candidates.

To evaluate the performance of models designed to identify double electron capture (DEC) events, the Wasserstein distance is employed as a goodness-of-fit test statistic. The statistical power of the test is then computed to quantify its sensitivity in distinguishing between competing hypotheses.

3 Double Electron Capture

Double electron capture (DEC) events deposit energy through X-ray and Auger cascades, resulting in a more spatially localized energy deposition than β decays of comparable energy. This increased localization leads to higher ionization densities and enhanced electron-ion recombination, which causes DEC signals to more closely resemble nuclear recoil (NR) events than typical β decays.

Two-neutrino double electron capture is an extremely rare nuclear process in which two orbital electrons are simultaneously captured by the nucleus, converting two protons into neutrons and emitting two electron neutrinos.

$$^{124}\text{Xe} + 2e^- \rightarrow^{124} \text{Te} + 2\nu_e + (\text{X-rays and Auger electrons})$$

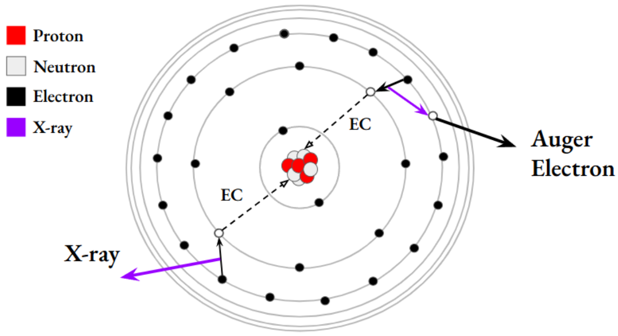


Figure 1: ^{124}Xe *captures two electrons and releases two neutrinos, x-rays, and auger electrons.*

Following the capture, the resulting vacancies in the atomic shell—most commonly in the innermost (K) shell—are filled through a cascade of X-rays and Auger electrons, producing a low-energy, localized signature in the detector. While K-shell captures are more probable and produce more energetic, easily distinguishable signals, decays involving two electrons from the L shell (LL) or one each from the L and M shells (LM) result in lower-energy signatures. These events are of particular interest because their more localized energy deposition leads to higher ionization densities and enhanced recombination, causing them to resemble nuclear recoil (NR) events more closely.

Because the energy deposited by LL and LM modes lies within the WIMP search region of interest (ROI), they represent an important background, with this analysis predicting approximately 13.7 DEC events in total, occurring in an estimated LL:LM ratio of roughly 2:1, as predicted in the XENONnT DEC hypothesis study [7].

4 Wasserstein Distance

The Wasserstein distance, also known as the Earth Mover's Distance (EMD), is a metric used to quantify the difference between two probability distributions. Intuitively, it represents the minimum amount of "work" required to transform one distribution into another, where "work" is defined as the product of the amount of probability mass moved and the distance it is transported.

The p-Wasserstein distance between two probability distributions μ and ν on a metric space $(\mathbb{R}, d)$ is defined as:

$$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} d(x,y)^p \, d\gamma(x,y) \right)^{1/p}$$

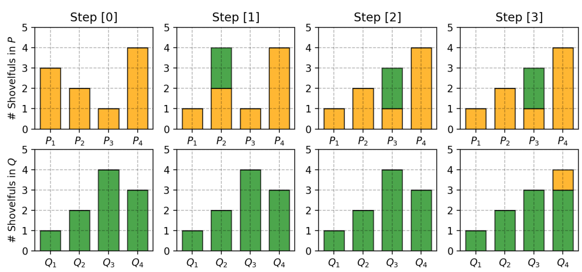


Figure 2: *Discrete 1D example of EMD[10].*

Here, $\Gamma(\mu,\nu)$ denotes the set of all joint distributions (couplings) on $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν, and $d(x,y)$ is the distance between points x and y.

In this work, we focus on the cases $p = 1$ and $p = 2$. For $p = 1$, the Wasserstein distance corresponds to the Earth Mover's Distance with cost measured by the Euclidean distance between points, $d(x,y) = \|x-y\|_2$. For $p = 2$, the cost function is given by the squared Euclidean distance, $d(x,y)^2 = \|x-y\|_2^2$, which places greater emphasis on larger displacements. The choice to use $p = 2$ increases the sensitivity of the Wasserstein distance to variations.

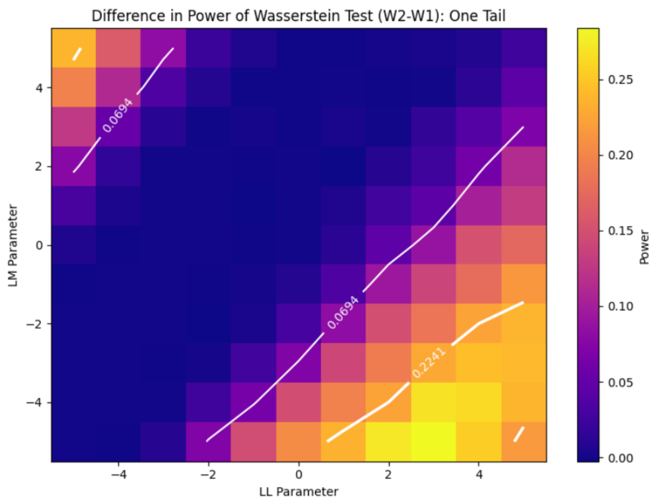
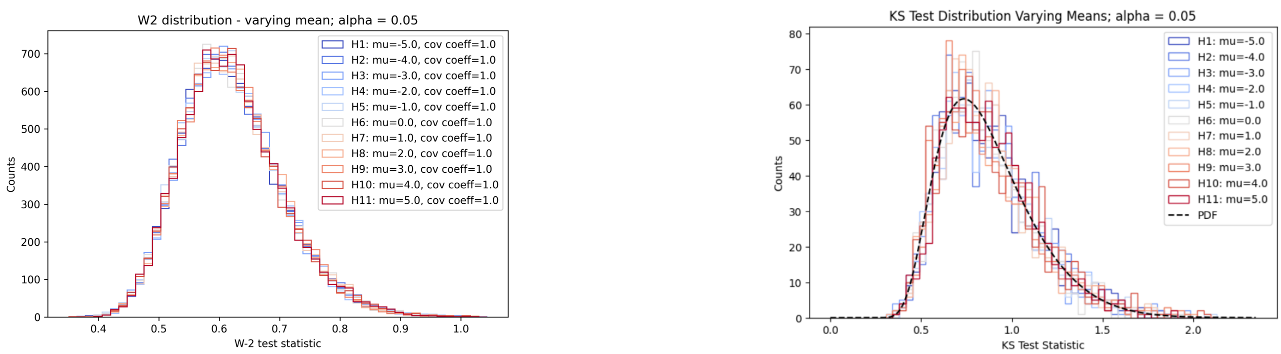


Figure 3: *Power values are positive showing that using $p = 2$ is more powerful than $p = 1$. 2-Wasserstein Distance will be used moving forward as it is a more sensitive test.*

We apply the Wasserstein distance as a two-dimensional, unbinned goodness-of-fit test to compare observed data to a background model of DEC events. The test operates on empirical samples in two-dimensional space ($\mathbb{R}^2$), thereby preserving the full resolution of the data.

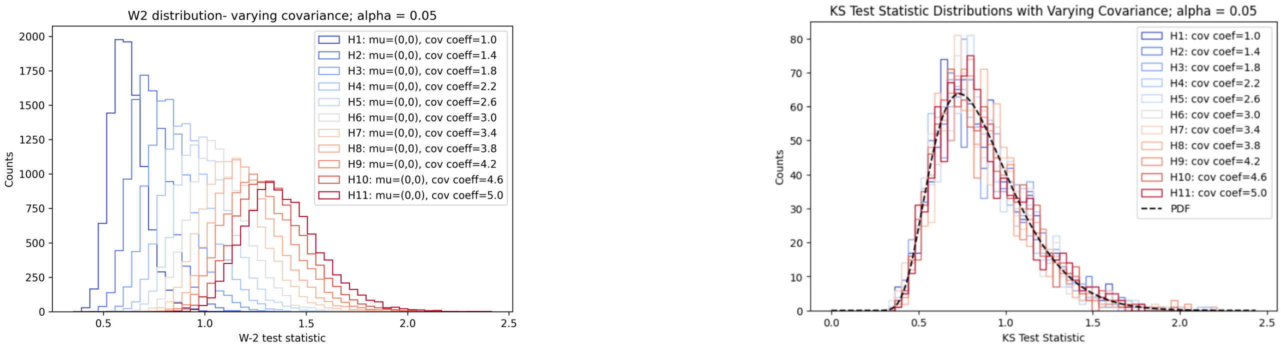
Unlike classical goodness-of-fit tests such as the Kolmogorov–Smirnov or χ^2 tests; its sampling distribution under the null hypothesis depends on the specific shape and variance of the underlying distribution. To investigate this, sample pairs were drawn from a 2D Gaussian distribution with fixed covariance, and the Wasserstein distance was computed for each pair. This process was repeated 10,000 times to approximate the distribution of the test statistic under the null hypothesis. The procedure was then repeated with Gaussians having different means and covariances to assess whether the statistic depends on the absolute location and spread of the distribution. The same methodology was applied using the KS test to validate the approach and confirm its distribution-free behavior under analogous conditions.



(a) ***(b)***

Figure 4: *Wasserstein (a) and Kolmogorov–Smirnov (b) test statistic distributions for 10,000 sample pairs drawn from identical 2D Gaussian distributions with fixed covariance and varying mean. In each case, samples were drawn from the same distribution to evaluate the null distribution of the test statistic. The results confirm that both tests are insensitive to shifts in the absolute location of the mean under* H_0.

Consequently, we employ a toy Monte Carlo approach to numerically approximate the distribution of the test statistic under the null hypothesis. This procedure enables the calculation of empirical p-values and allows us to assess the statistical significance of deviations between the observed data and the background model.



(a) *(b)*

Figure 5: *Wasserstein (a) and Kolmogorov–Smirnov (b) test statistic distributions for 10,000 sample pairs drawn from identical 2D Gaussian distributions with fixed mean and varying covariance. In each case, samples were drawn from the same distribution to evaluate the null distribution of the test statistic. The results confirm that while the KS test is insensitive to changes in the spread of* H_0*, as expected, the Wasserstein distribution is easily influenced. This confirms the Wasserstein distance in not distribution-free.*

To quantify the sensitivity of the Wasserstein distance as a test statistic (W), we perform a one-tailed hypothesis test to compute the statistical power of detecting deviations from the background model. Under the null hypothesis H_0, the observed data are assumed to follow the DEC distribution. The alternative hypothesis H_1 corresponds to the mis-modeling of DEC events.

The power of the test is the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true. Mathematically, it is defined as:

$$\text{Power} = 1 - \beta$$

$$\beta = \mathrm{P}_{H_1}(W < T)$$

where P_{H_1} denotes the probability evaluated under the alternative hypothesis.

At a 95% confidence level, the threshold T is determined from the upper tail of the null distribution, obtained via toy Monte Carlo simulations, such that 0.05 of the area of the pdf lies beyond the threshold. The power of the test is then calculated as the fraction of MC pseudo-experiments, or samples, generated under the alternative hypothesis for which the Wasserstein distance exceeds T. This approach allows us to determine the probability of correctly identifying a model deviation when it is present, thereby characterizing the test's discriminating ability; it provides a quantitative measure of any mis-modeling of the data.

Earth Mover's Distance Function

To compute the 2-Wasserstein (Earth Mover's) distance between two empirical distributions, we consider two samples: a source distribution and a target distribution. Each distribution consists of a collection of events, where each event is characterized by a location and an associated probability mass.

Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^2$ be the locations in the source distribution with corresponding mass vector $\mathbf{a} \in \mathbb{R}^n$, and $\{y_j\}_{j=1}^m \subset \mathbb{R}^2$ be the locations in the target distribution with mass vector $\mathbf{b} \in \mathbb{R}^m$, such that:

$$\sum_{i=1}^{n} a_i = 1, \quad \sum_{j=1}^{m} b_j = 1, \quad a_i, b_j \geq 0$$

Each location is defined by its cs1 and cs2 values (i.e., $x_i = (\mathrm{cs1}_i, \mathrm{cs2}_i)$), and the mass represents the normalized fraction of events at that location in the dataset.

To quantify the cost of transporting mass between the source and target distributions, we define a cost matrix $C \in \mathbb{R}^{n \times m}$, where each entry is the squared Euclidean distance between locations:

$$C_{ij} = \|x_i - y_j\|_2^2$$

The goal is to find a transport plan $\gamma \in \Gamma(\mathbf{a}, \mathbf{b}) \subset \mathbb{R}^{n \times m}$, which specifies how much mass to move from x_i to y_j, that minimizes the total transport cost subject to the marginal constraints, where $\Gamma(\mathbf{a}, \mathbf{b})$ denotes the set of all valid transport plans with marginals a and b. In other words, each row of γ sums to the source mass $\mathrm{a_i}$, each column of γ sums to the target mass $\mathrm{b_j}$, all γ_{ij} must be non-negative, and γ must represent the least costly path, or the optimal transport plan. The optimal plan is found using a linear programming algorithm, or more specifically the network simplex algorithm.

$$\begin{array}{ll} \textbf{Mass Constraints} & \textbf{Non-negativity Constraint} \\ \sum_{j=1}^{m} \gamma_{ij} = a_i \quad \forall i \quad \text{(row sums)} & \gamma_{ij} \geq 0 \quad \forall i, j \\ \sum_{i=1}^{n} \gamma_{ij} = b_j \quad \forall j \quad \text{(column sums)} & \end{array}$$

The 2-Wasserstein distance (W), used as the test statistic, is computed by summing the product of the optimal and the corresponding cost matrix C_{ij} over all source and target locations.

$$W = (\sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_{ij} C_{ij})^{1/2}$$

This formulation enables a rigorous computation of the distance between two empirical distributions, providing a measure of mis-modeling, which serves as a sensitive test statistic in our analysis.

5 Methodology

5.1 Templates

The templates used to model DEC (double electron capture) events were constructed from simulations in which the light and charge yields of the DEC events were varied. The LL and

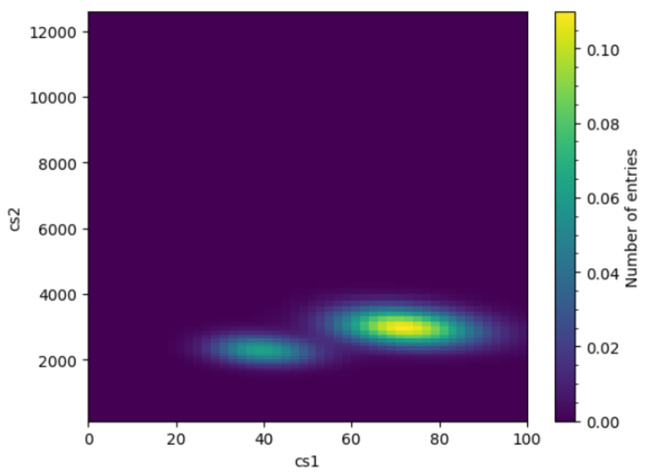


Figure 6: *Example Template Plotted*

LM parameters define the energy deposited by events resulting from electron captures originating from the L shell and from both the L and M shells, respectively. In total, 11 distinct LL parameter values and 11 distinct LM parameter values were used across the simulations, yielding 121 unique templates spanning the combined range of these two parameters. Adjusting these parameters shifts the bulk of the simulated distribution along energy contours. Simulated events are subsequently binned and transformed into multi-dimensional histograms for further analysis.

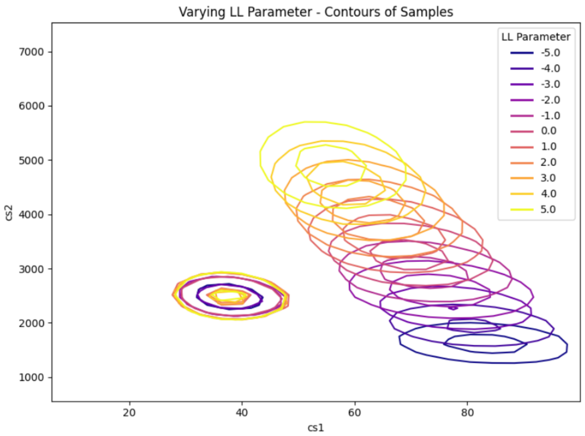


Figure 7: *Varying LL parameter changes energy deposition of DEC events from LL shell.*

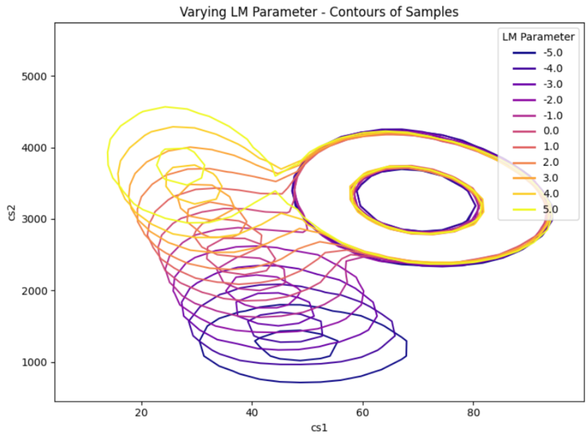


Figure 8: *Varying LM parameter changes energy deposition of DEC events from L and M shells.*

5.2 Sampling

A total of 13 events were randomly drawn using Monte Carlo methods for each test, corresponding to the expectation of 13 DEC events [7]. Each Monte Carlo sample was directly compared to the template from which it was generated using the Wasserstein distance. To approximate the distribution under the null hypothesis, this sampling procedure was repeated multiple times, generating a distribution of Wasserstein distances for each template. The templates are binned representations of simulated data and exhibit edges where the distribution is effectively truncated, limiting their ability to capture the full spread of the data as the underlying parameters shift. However, given the small size of each data set, the influence of edge behavior on the computed Wasserstein distance is minimal and does not significantly affect the results. Additionally, excessively large reference samples were avoided to prevent overemphasizing the discrete structure of the template. Conversely, very small samples were also avoided, as they reduce sensitivity to differences between competing hypotheses. For these reasons, comparisons were made directly between the test samples and their respective templates.

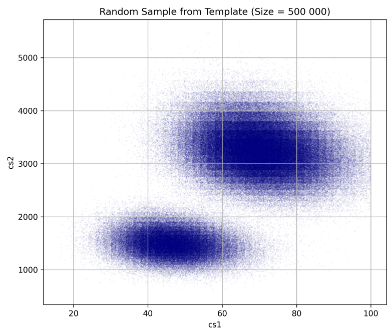


Figure 9: *Large Sample (500 000) reveals binning in template and edge.*

6 Results and Discussion

Figure 10 shows the statistical power of the 2-Wasserstein distance test evaluated over a grid of LL and LM parameter values, where the null hypothesis corresponds to the baseline DEC model

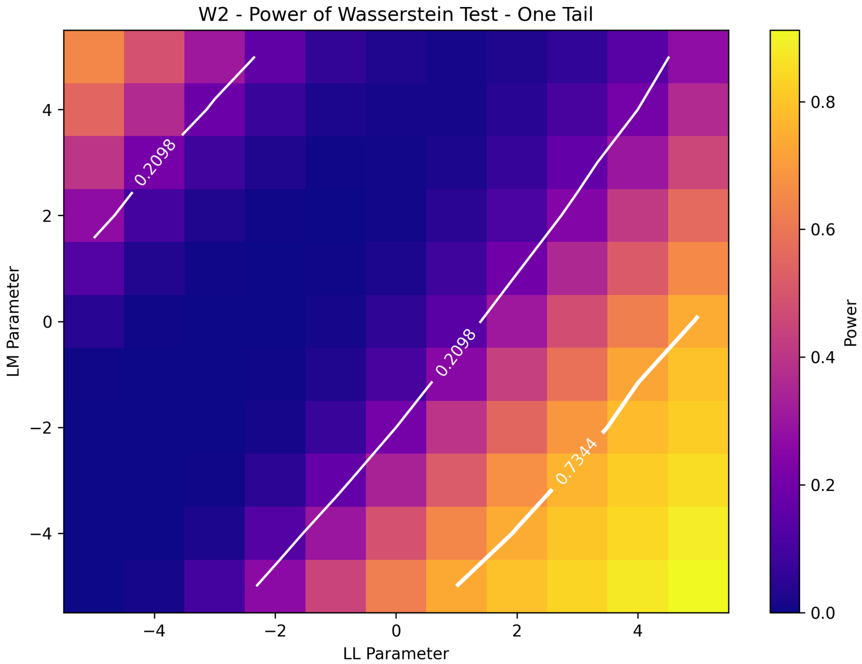


Figure 10: *Power Heat Map - Most powerful when (LL,LM) = (-5,-5) where power = X.*

at (LL, LM) = (0,0). For each parameter combination, the power represents the probability of correctly rejecting the null hypothesis when the alternative is true, as estimated via Monte Carlo simulations. The heatmap indicates that the Wasserstein test is most sensitive when either the LL or LM parameter takes on negative values, corresponding to hypotheses that deviate substantially from the null model. In these regions, the test achieves near-maximal power, demonstrating a strong ability to detect significant differences in the energy deposition characteristics of DEC events. Conversely, the power decreases markedly as the parameters approach zero or deviate only slightly from the null, suggesting that the test is less effective at distinguishing subtle model variations with the available sample size. Moreover, the power distribution is asymmetric, implying that deviations in certain directions in parameter space produce more detectable changes in the underlying distribution than others. These findings indicate that while the 2-Wasserstein distance is an effective goodness-of-fit metric for identifying pronounced mis-modeling of the DEC background, its sensitivity diminishes for small parameter shifts. This highlights the importance of sample size and suggests that complementary methods or additional data features may be necessary to improve discrimination in cases of minor deviations. Overall, this analysis confirms the utility of the Wasserstein distance for validating DEC background models within the XENONnT experiment and clarifies the parameter regimes where it is most informative.

7 Summary and Conclusions

This study demonstrated the effectiveness of the 2-Wasserstein distance as a two-dimensional, unbinned goodness-of-fit test for validating the DEC background model in the XENONnT experiment. By comparing Monte Carlo random samples to simulated templates across a range of LL and LM parameters, we evaluated the test's statistical power in detecting deviations from the baseline model. The results showed that the test is highly sensitive to large mis-modelings, particularly when the energy deposition parameters deviate significantly from the null hypothesis. However, sensitivity declines for small parameter shifts, limiting its ability to detect subtle discrepancies under the current sampling conditions. These findings establish the 2-Wasserstein distance as a powerful diagnostic tool for model validation while also revealing its limitations in low-contrast scenarios.

Further analysis revealed that using a two-tailed test significantly improved the ability to detect deviations within regions of low sensitivity, particularly within the central purple band of the power heatmap. Unlike the one-tailed test, which only captures deviations in a single direction, the two-tailed approach accounts for bidirectional differences and was able to identify mismatches that were previously undetectable. Future work will focus on integrating this approach more systematically, optimizing the sensitivity across a larger parameter space, and exploring additional test statistics or dimensional representations to enhance detection capability.

8 Acknowledgements

I am grateful to Prof. Elena Aprile for her constant support, Dr. Pueh Leng Tan for her mentorship, direction, and impactful discussions, and XENON Collaboration and Laboratori Nazionali del Gran Sasso for hosting and supporting the project. Thank you also to the Nevis Labs REU Program. This material is based upon work supported by the National Science Foundation under Grant No. PHY-2349438.

References

[1] Jelle Aalbers. *Dark Matter Search with XENON1T*. PhD thesis, University of Amsterdam, 2018. PhD Thesis.

[2] E. Aprile et al. Observation of two-neutrino double electron capture in ^{124}xe with xenon1t. *Nature*, 568(7753):532–535, 2019.

[3] E. Aprile et al. Wimp dark matter search using a 3.1 tonne-year exposure of the xenonnt experiment. *arXiv preprint*, 2025. Submitted February 2025.

[4] Laura Baudis. Dual-phase xenon time projection chambers for rare-event searches. *Philosophical Transactions of the Royal Society A*, 382(20230083), 2023.

[5] Glen D. Cowan. *Statistical Data Analysis*. Oxford Science Publications. Oxford University Press, Oxford, UK, 1998. Clarendon Press edition.

[6] Marc Hallin, Gilles Mordant, and Johan Segers. Multivariate goodness-of-Fit tests based on Wasserstein distance. *arXiv preprint*, 2021. Revised version (v3) submitted 27 Jan 2021.

[7] Robert Hammann. Dec hypothesis test and wimp roi in xenonnt, 2024. XENON1T Internal Note, accessible via XENON Wiki.

[8] Tudor Manole. Pot: Python optimal transport code (contributed by t. manole among others). `https://github.com/PythonOT/POT`, 2025. Open-source optimal transport library, MIT license.

[9] Tudor Manole. *Statistical Inference for Optimal Transport*. PhD thesis, Carnegie Mellon University, 2025. PhD Thesis, Department of Statistics Data Science.

[10] Lilian Weng. Gan by example (part 1), 2017. Accessed: 2025-07-22.