Waveform simulation in PandaX-4T*

Jiafu Li (李家富)[14] Abdusalam Abdukerim (阿卜杜萨拉姆)[1] Zihao Bo (薄子豪)[1] Wei Chen (陈葳)[1]
Xun Chen (谌勋)[1,19] Chen Cheng (程晨)[14] Zhaokan Cheng (成赵侃)[15] Xiangyi Cui (崔祥仪)[12]
Yingjie Fan (樊英杰)[21] Deqing Fang (方德清)[18] Lisheng Geng (耿立升)[4,5,6] Karl Giboni[1]
Xuyuan Guo (郭绪元)[8] Chencheng Han (韩晨成)[12] Ke Han (韩柯)[1] Changda He (何昶达)[1] Jinrong He (何金荣)[8]
Di Huang (黄迪)[1] Junting Huang (黄俊挺)[1] Zhou Huang (黄周)[1] Ruquan Hou (侯汝全)[19] Yu Hou (侯钰)[13]
Xiangdong Ji (季向东)[11] Yonglin Ju (巨永林)[13] Chenxiang Li (李晨翔)[1] Mingchuan Li (李名川)[8]
Shuaijie Li (李帅杰)[8,1] Tao Li (李涛)[15] Qing Lin (林箐)[3,2†] Jianglai Liu (刘江来)[1,12,19‡] Congcong Lu (陆聪聪)[13]
Xiaoying Lu (芦晓盈)[9,10] Lingyin Luo (罗棱尹)[7] Yunyang Luo (罗云阳)[3] Wenbo Ma (马文博)[1§]
Yugang Ma (马余刚)[18] Yajun Mao (冒亚军)[7] Yue Meng (孟月)[1,19] Xuyang Ning (宁旭阳)[1] Binyu Pang (庞彬宇)[9,10]
Ningchun Qi (祁宁春)[8] Zhicheng Qian (钱志成)[1] Xiangxiang Ren (任祥祥)[9,10] Nasir Shaheed[9,10]
Xiaofeng Shang (尚晓凤)[1] Xiyuan Shao (邵熙元)[17] Guofang Shen (申国防)[4] Lin Si (司琳)[1]
Wenliang Sun (孙文良)[8] Andi Tan (谈安迪)[11] Yi Tao (陶奕)[1,19] Anqing Wang (王安庆)[9,10] Meng Wang (王萌)[9,10]
Qiuhong Wang (王秋红)[18] Shaobo Wang (王少博)[1,20] Siguang Wang (王思广)[7] Wei Wang (王为)[14,15]
Xiuli Wang (王秀丽)[13] Xu Wang (王旭)[1] Zhou Wang (王舟)[1,12,19] Yuehuan Wei (魏月环)[15]
Mengmeng Wu (武蒙蒙)[14] Weihao Wu (邬维浩)[1] Yuan Wu (吴渊)[1] Mengjiao Xiao (肖梦姣)[1] Xiang Xiao (肖翔)[14]
Binbin Yan (燕斌斌)[1] Xiyu Yan (阎玺宇)[16] Yong Yang (杨勇)[1] Chunxu Yu (喻纯旭)[17] Ying Yuan (袁影)[1]
Zhe Yuan (苑哲)[18] Youhui Yun (运有辉)[1] Xinning Zeng (曾鑫宁)[1] Minzhen Zhang (张敏桢)[1] Peng Zhang (张鹏)[8]
Shibo Zhang (张世博)[1] Shu Zhang (张澍)[14] Tao Zhang (张涛)[1] Wei Zhang (张威)[12] Yang Zhang (张洋)[9,10]
Yingxin Zhang (张瀛心)[9,10] Yuanyuan Zhang (张园园)[12] Li Zhao (赵力)[1] Jifang Zhou (周济芳)[8]
Ning Zhou (周宁)[1,19] Xiaopeng Zhou (周小朋)[4] Yong Zhou (周永)[8] Yubo Zhou (周钰博)[1] Zhizhen Zhou (周知震)[1]

[1]School of Physics and Astronomy, Shanghai Jiao Tong University, Key Laboratory for Particle Astrophysics and Cosmology (MoE), Shanghai Key Laboratory for Particle Physics and Cosmology, Shanghai 200240, China
[2]State Key Laboratory of Particle Detection and Electronics, University of Science and Technology of China, Hefei 230026, China
[3]Department of Modern Physics, University of Science and Technology of China, Hefei 230026, China
[4]School of Physics, Beihang University, Beijing 102206, China
[5]International Research Center for Nuclei and Particles in the Cosmos & Beijing Key Laboratory of Advanced Nuclear Materials and Physics, Beihang University, Beijing 100191, China
[6]School of Physics and Microelectronics, Zhengzhou University, Zhengzhou 450001, China
[7]School of Physics, Peking University, Beijing 100871, China
[8]Yalong River Hydropower Development Company, Ltd., 288 Shuanglin Road, Chengdu 610051, China
[9]Research Center for Particle Science and Technology, Institute of Frontier and Interdisciplinary Science, Shandong University, Qingdao 266237, China
[10]Key Laboratory of Particle Physics and Particle Irradiation of Ministry of Education, Shandong University, Qingdao 266237, China
[11]Department of Physics, University of Maryland, College Park, Maryland 20742, USA
[12]Tsung-Dao Lee Institute, Shanghai Jiao Tong University, Shanghai, 200240, China
[13]School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[14]School of Physics, Sun Yat-Sen University, Guangzhou 510275, China
[15]Sino-French Institute of Nuclear Engineering and Technology, Sun Yat-Sen University, Zhuhai 519082, China
[16]School of Physics and Astronomy, Sun Yat-Sen University, Zhuhai 519082, China
[17]School of Physics, Nankai University, Tianjin 300071, China

Received 18 December 2023; Accepted 26 March 2024; Published online 27 March 2024

* This work was supported in part by the National Science Foundation of China (12090060, 12090061), Ministry of Science and Technology of China (2023YFA1606200), and Office of Science and Technology, Shanghai Municipal Government (22JC1410100). We are thankful for the support provided by the Double First Class Plan of the Shanghai Jiao Tong University and Guangzhou Municipal Science and Technology Project (202201010991)

† E-mail: qinglin@ustc.edu.cn
‡ E-mail: jianglia.liu@sjtu.edu.cn (Spokesperson)
§ E-mail: wenboma@sjtu.edu.cn

[18]Key Laboratory of Nuclear Physics and Ion-beam Application (MOE), Institute of Modern Physics, Fudan University, Shanghai 200433, China
[19]Shanghai Jiao Tong University Sichuan Research Institute, Chengdu 610213, China
[20]SJTU Paris Elite Institute of Technology, Shanghai Jiao Tong University, Shanghai 200240, China
[21]Department of Physics,Yantai University, Yantai 264005, China

Abstract: Signal reconstruction through software processing is a crucial component of the background and signal models in the PandaX-4T experiment, which is a multi-tonne dark matter direct search experiment. The accuracy of signal reconstruction is influenced by various detector artifacts, including noise, dark count of photomultiplier, photoionization of impurities in the detector, and other relevant considerations. In this study, we presented a detailed description of a semi-data-driven approach designed to simulate a signal waveform. This work provides a reliable model for the efficiency and bias of the signal reconstruction in the data analysis of PandaX-4T. By comparing critical variables that relate to the temporal shape and hit pattern of the signals, we found good agreement between the simulation and data.

Keywords: dark matter, time projection chamber, simulation

DOI: 10.1088/1674-1137/ad380f

I. INTRODUCTION

The search for dark matter (DM) [1−3] is a highly active research area in astroparticle physics. Among the various experimental approaches [4−12], dual-phase xenon time projection chamber (TPC) based experiments have been recognized as highly sensitive for detecting cold DM particles within the mass range of approximately 10 GeV/c^2 to TeV/c^2. PandaX-4T [10], located in the China Jinping Underground Laboratory (CJPL) [13, 14], is a typical experiment of this kind. PandaX-4T employs a cylindrical dual-phase xenon TPC with a sensitive volume measuring approximately 1.2 meters in diameter and 1.2 meters in height. The TPC is equipped with 368 3-inch photomultipliers (PMTs), 169 and 199 of which are distributed at the top and bottom of the PMT array, respectively. The TPC detects and measures the prompt scintillation signal ($S1$) and the subsequent amplified proportional scintillation signal ($S2$), which arises from the delayed ionization signals. By analyzing the time difference between the $S1$ and $S2$ signals, as well as the spatial distribution of the PMT hits associated with the $S2$ signal, we are capable of reconstructing the vertical and horizontal positions of the interaction vertex, respectively. Accurate 3-D position reconstruction and the determination of the $S2/S1$ ratio are crucial in particle discrimination within the TPC detector. This discrimination capability is important as it greatly reduces the effective background for DM direct searches.

To interpret the DM search results in PandaX-4T, precise models for the low-energy background and DM signal are crucial. A critical component of the low-energy models is the reconstruction of the $S1$ and $S2$ signals. This reconstruction process involves various steps such as peak identification, pulse classification, and clustering. To accurately evaluate the efficiency and potential biases inherent in the software reconstruction, it is essential to obtain pure events that faithfully represent the desired physical signal. However, acquiring pure samples from the recorded reconstructed data is challenging because the recorded data have already been influenced by the effects of the software signal reconstruction inefficiency and biases. To address this challenge, a dedicated waveform simulation (WS) framework that generates synthetic data waveforms has been developed for PandaX-4T. The WS framework incorporates our best understanding of the processes involved in the generation, collection, and reconstruction of the $S1$ and $S2$ signals, as well as the accompanying sources of noise, such as dark counts, PMT after-pulsing, and impurity photoionization. In addition, the extensive samples generated by the WS can be utilized to train machine learning and deep neural network algorithms, which can further enhance background rejection capabilities.

This manuscript presents a detailed account of WSs employed in various scientific studies conducted using the PandaX-4T data to search for weakly interacting massive particles (WIMPs) [10], detect solar ^{8}B neutrinos [15], investigate light dark matter particles [16], etc [17, 18]. Section II of the manuscript elucidates the details of simulating the $S1$ and $S2$ pulses, as well as incorporating various sources of noise into the simulation process. Section III presents a comparison between simulation and real data, wherein critical variables such as signal width, pattern, and charge are evaluated. Finally, Section IV summarizes the work and discusses future perspectives. Similar works of other experiments can be found in Ref. [19].

II. WAVEFORM SIMULATION

The simulated event waveforms in a WS further undergo the software processing and reconstruction chain that is used for data processing and analysis in PandaX-4T to ensure that the efficiency and bias of the simulated waveforms align with the observed data. WSs employ a

semi-data-driven approach to simulate the $S1$ and $S2$ waveforms. The most fundamental segments of the $S1$ and $S2$ signals (hit and single-electron waveforms) are randomly selected from real data and then randomly reassembled to form simulated signals. Additionally, WSs take into account various spurious pulses, including noise, dark counts, PMT after-pulsing, and delayed afterglow after a large $S2$ pulse arising from impurity ionization and delayed electron effects. The high level characteristics of the simulated and real data are then compared to validate the modeling and prediction efficacies. To illustrate, two simulated waveforms are presented in the middle and bottom panels of Fig. 1. The bottom panel demonstrates successful identification of the $S1$ and $S2$ signals, albeit with a slight difference in the reconstructed charge values with respect to the true charge due to fluctuations in the reconstruction steps and interference from spurious pulses. In contrast, the middle panel exhibits complete misidentification of the $S1$signal, indicating a potential inefficiency in the software signal reconstruction process.

A. *S1* pulse

The $S1$ signal represents the prompt scintillation signals detected in the TPC detector, arising from the interaction of particles with the xenon shell electrons or nucleus. The $S1$ signal exhibits a relatively short time scale, typically ranging from a few tens of nanoseconds to approximately 100 nanoseconds, which originates from light propagation in the TPC. The shape of the $S1$ pulse is influenced by the PMT signal shaping, decay time profile of singlet and triplet xenon dimers, and propagation of photons within the TPC. The ratio between the singlet and triplet dimer decays differs slightly between electronic recoils (ERs) and nuclear recoils (NRs). In ER events, where the incoming particle interacts with the xenon shell electrons, this ratio tends to be slightly higher compared to NR events [20, 21], where the interaction occurs with the xenon nucleus. However, the difference in pulse shapes between ER and NR events is obscured as a consequence of photon propagation effects, such as photon Rayleigh scattering and the reflection on the liquid surface and TPC wall.

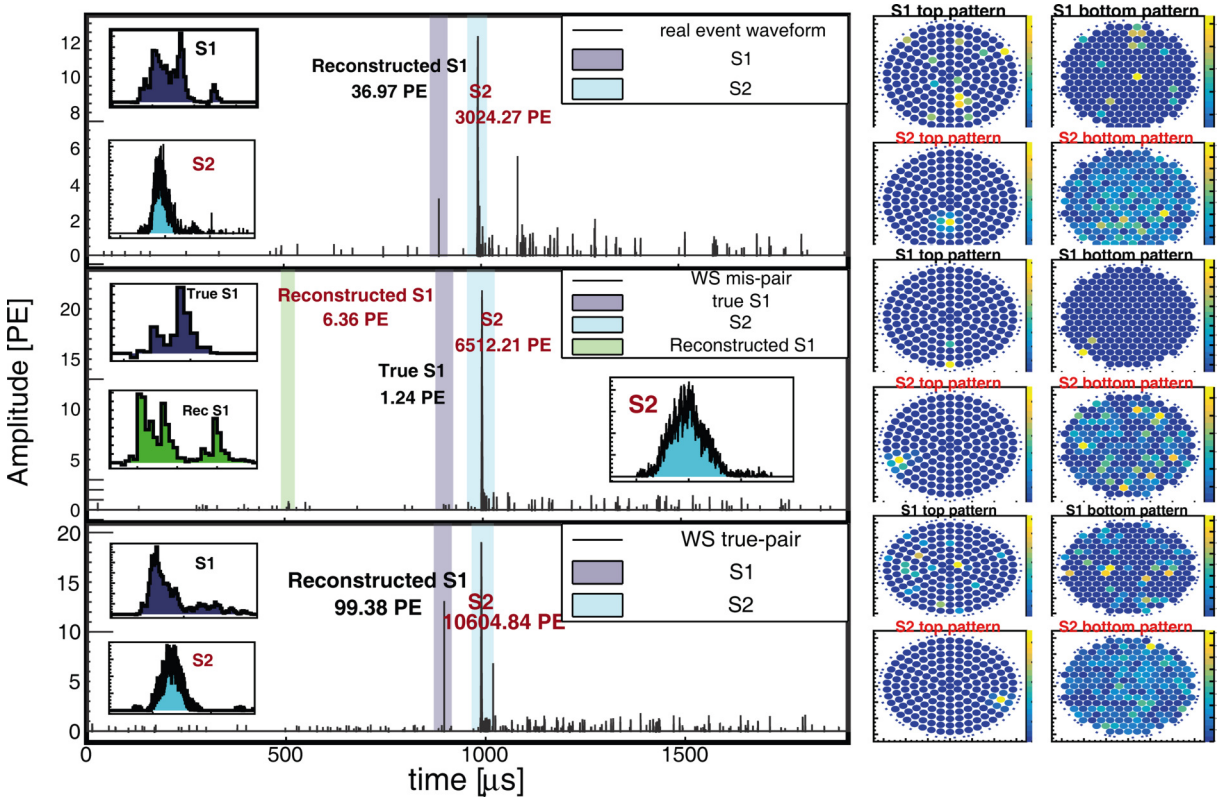

Fig. 1. (color online) Three waveforms corresponding to events with similar drift times. Top panel represents a recorded data waveform. Middle panel shows a simulated waveform, where the true $S1$ signal is overshadowed by a noise $S1$ signal with a slightly higher charge. Bottom panel displays a simulated waveform, where the $S1$ and $S2$ signals are correctly identified. Gray, green, and cyan shaded regions represent time windows associated with true $S1$, misidentified $S1$, and true $S2$ signals, respectively. Reconstructed charge of identified $S1$ and $S2$ signals are displayed in panels. For enhanced detail, inset panels provide zoomed-in views of$S1$ and $S2$ signals. Hit patterns of main $S1$ and $S2$ signals are shown in right columns.

To simulate the *S*1 waveform, a data-driven approach is adopted. In this approach, the *S*1 signals from ^{220}Rn [22] and neutron calibration data, including a ^{241}AmBe radioactive source and Deuteron-Deuteron (DD) neutron generator, are utilized to simulate the *S*1 waveforms corresponding to the ERs and NRs, respectively. The *S*1 signals with a charge of 2–200 photoelectrons (PE) and 2–150 PE were selected as the ER and NR *S*1 pools, respectively. It was observed that the hit distribution among the PMTs for the majority of the *S*1 signals that occurred in the fiducial volume was quite uniform owing to Rayleigh scattering and reflection on PTFE. Within the charge levels of the *S*1 pools, the *S*1 waveforms consisted of single-photon hits that were distributed among multiple PMTs. The probability of two hits occurring on same PMT was negligible. Additionally, the time response of the PMTs to a single hit was within appoximately 10 ns, while the time spread due to photon propagation was approximately 40 ns. Two hits occurring on same PMT were likely distinguishable by time. The single-photon hit waveform on a PMT is depicted in the left panel of Fig. 2. Based on the above facts, a simulation of the *S*1 waveform was performed by sampling the hits of an *S*1 pool. The *S*1 pool was required to have the same event position as the simulated *S*1 signal (transverse distance <5 cm and vertical distance <8 cm). Figure 3 shows an example of a simulation. An *S*1 pool with 19 hits was used to simulate an *S*1 signal with 13 hits, and 68% of the pooled hits were sampled to form the simulated *S*1 signal. This data-driven approach naturally takes into account the position-dependent characteristics of the *S*1 pulse time profile resulting from light propagation and variations of the PMT characters, such as the gains and SPE resolution. The right panel of Fig. 2 displays the hit time distributions of the *S*1 signals in the ER and NR calibration data from various *Z* positions. The differences in the shape of the waveforms arises from the differences in the spatial position distributions in the ER and NR calibration data.

B. *S*2 pulse

The *S*2 signal corresponds to the detection of proportional scintillation light emitted by the drifted electrons when they reach the gaseous xenon layer between the anode and gate electrodes in the TPC detector. The shape of the *S*2 pulse is predominantly determined by two factors: the diffusion of electrons during their drift from the interaction vertex to the gaseous xenon layer, and the subsequent travel within the gaseous layer. The longitudinal diffusion coefficient in liquid xenon has been experimentally measured to be approximately 30–40 cm^2/s at an electric field of 100 V/cm [23]. To obtain accurate *in-situ* values for the longitudinal diffusion coefficient, measurements are performed using different calibration sources. Specifically, the 41 keV gamma line from the ^{83m}Kr cal-

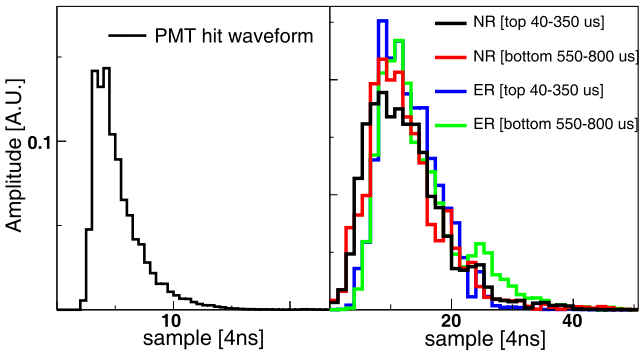

Fig. 2. (color online) Single PE waveform (left) and time distributions (right) of *S*1 hits. Blue, orange, and red solid lines in right panel represent distributions from ^{220}Rn, ^{241}AmBe, and DD calibration data, respectively.

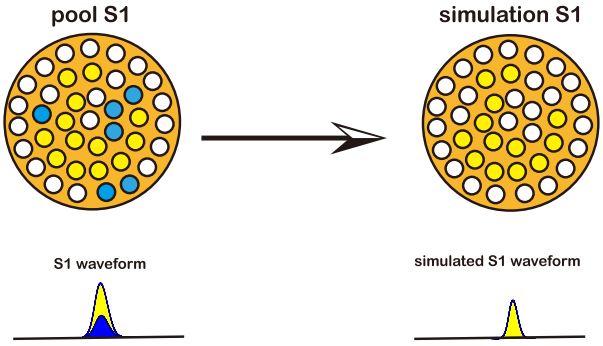

Fig. 3. (color online) Example of *S*1 waveform simulation. *S*1 pool has 19 hits and simulated *S*1 signal is required to have 13 hits. Six hits from *S*1 pool are randomly selected for dropping, and the rest are taken as simulated *S*1 signal. Pattern and waveform of dropped hits are in blue in the left panel.

ibration data, and the α decay events from ^{222}Rn are utilized. Figure 4 illustrates the 80%-CDF *S*2 width (defined as the length of time window that covers central 80% of the *S*2 charge) over drift time distributions for these events, and the mean 80%-CDF *S*2 widths as a function of drift time were used to calculate the respective longitudinal diffusion coefficients. The longitudinal diffusion coefficient D_L can be obtained by fitting the mean 80%-CDF *S*2 widths (denoted as $\langle W_{S2}^{80}\rangle$) as a function of drift time by the relation:

$$\langle W_{S2}^{80}\rangle = \frac{f^{80}}{v_d}\sqrt{2D_L T+\sigma_0^2}, \tag{1}$$

where T is the drift time and σ_0 is the standard deviation of the photon hit time in a single electron (SE) waveform. σ_0 is affected by the travel time of the electron in the gaseous xenon layer, and depends on the gas gap and electric field strength in the gap. σ_0 is calibrated to be 0.2 mm in Run0 of PandaX-4T. v_d is the drift velocity in LXe, which was set to 1.44 mm/μs. The factor $f^{80} = 2.56$ is the conversion factor from the standard deviation to

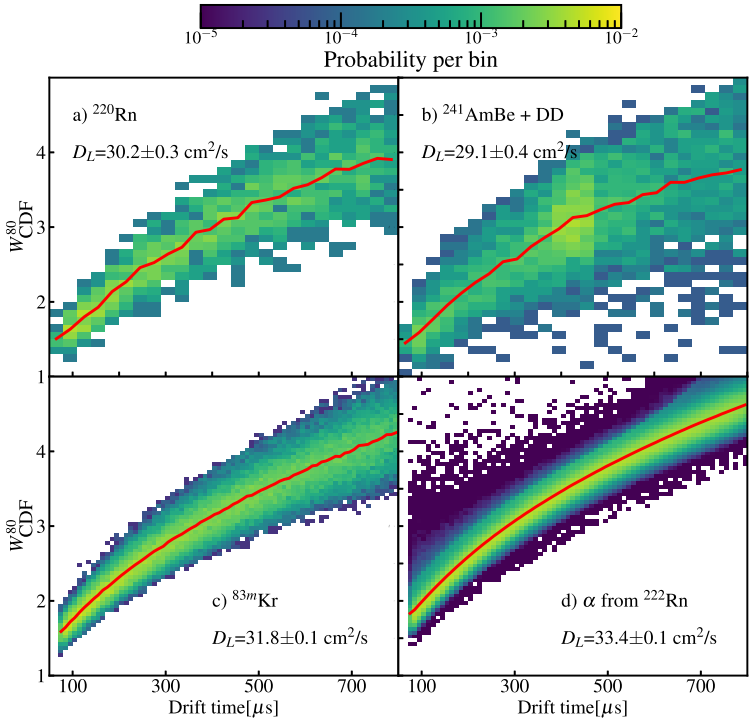

Fig. 4. (color online) Normalized distributions of $W^{80}_{\rm CDF}$ over drift time for four types of data: a) ^{220}Rn calibration, b) neutron calibration using ^{241}AmBe and DD, c) ^{83m}Kr calibration, and d) α events originating from ^{222}Rn impurities. Red solid lines represent distribution means. D_L values obtained from fitting the means using Eq. (1) are given in each panel.

80%-CDF width that assumes the *S*2 pulse shape is Gaussian. The best-fit D_L using low-energy data (^{220}Rn, ^{241}AmBe, and DD) was systematically lower than the results obtained using the high-energy data (^{83m}Kr and α from ^{222}Rn). This is due to the statistical bias of the *S*2 pulse width when the *S*2 charge is only caused by a few electrons. The variations observed in the best-fit diffusion coefficients obtained from high-energy α also suggest the potential presence of systematic effects in high-energy samples, such as PMT saturation. Consequently, the WS employs the diffusion coefficient from ^{83m}Kr calibration data. This coefficient agrees well with the width versus drift time distributions of the low-energy data, discussed later in the text. The time profile of electron travel within the gaseous xenon can be determined by analyzing the SE waveforms, as depicted in Fig. 5. The average SE waveforms shown in Fig. 5 are obtained by stacking the SE waveforms with the arrival time of their first hit set to zero. The electric field and decay characteristics of xenon dimers in the gaseous phase influenced the observed time profile.

The SE waveforms from the data were reassembled to generate simulated *S*2 signals in the WS. To ensure compatibility of the (*X*, *Y*) positions between the simulated *S*2 signals and data, SE waveforms with reconstructed positions R_0 within a certain range, specifically within 4 cm from the simulated position, were selected for the assembly process. This value was determined to efficiently

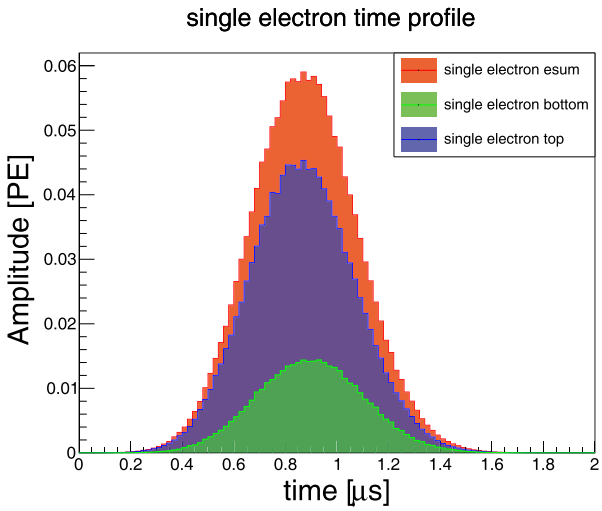

Fig. 5. (color online) Average single electron waveforms.

match the pattern of the *S*2 signals (see Section III for more detail). Each selected SE pulse waveform was given a time shift before reassembly, to ensure that the simulated SE arrival times adhered to the diffusion principle in Eq. (1). Figure 6 shows an example of a waveform simulation for an *S*2 signal from a 2-electron signal. Two SE waveforms were sampled from the SE pool, with charges of 31.56 and 14.61 PE, respectively. We required the simulated *S*2 signal to have a *z* position of 47.2 cm, which

results in a *S*2 CDF width of 2.56 μs. The two SE waveforms were given time shifts of –0.87 and 0.57 μs, respectively.

C. PMT after-pulsing

When a PMT detects light signals, residual gases in the PMT can become ionized owing to the acceleration of photoelectrons (PEs) [24−27]. These positively charged ions then drift back towards the photocathode of the PMT, causing the emission of additional PEs. Consequently, small delayed signals may appear after the main pulse. The majority of residual gases tend to be concentrated on the first dynode of the PMT, leading to characteristic and constant time delays for certain ions in the PMT response due to after-pulsing. The magnitude of the time delay is primarily proportional to the square root of the atomic mass over the ion charge of the gas molecule. Furthermore, the number of PEs that each residual gas ion can generate is dependent on the type of residual gas. In WSs, the number of PMT after-pulsing hits and the charge of each hit are sampled according to the average differential probability and mean charge of the PMT after-pulsing, respectively, as a function of the delay time, as shown in Fig. 7. The average probability of after-pulsing after each PE in the 3-inch PMTs utilized in the PandaX-4T experiment was approximately 3%.

D. Photoionization and delayed electrons

In several LXe-TPC detectors, it has been observed that additional *S*2 signals appear subsequent to a large *S*2 signal [28−30], commonly referred to as delayed *S*2 signals or delayed electrons. The appearance of these delayed signals is attributed to various factors. One significant factor is the photoionization of the electrode metal. The electrons resulting from this photoionization process exhibit distinct drift times, as depicted in Fig. 8. Delayed electrons can also arise from the photoionization of electro-negative impurities (mainly oxygen) uniformly distributed within the detector. It is also speculated that impurities present in the LXe may "capture" and subsequently release drifting electrons [30], leading to delayed *S*2 signals. Other processes, such as electron trapping at the liquid-gas interface [31] and spontaneous electron emission from the electrode [32, 33], have also been proposed as potential contributors to the delayed *S*2 phenomenon. In the PandaX-4T experiment, a data-driven approach was employed to model the probability of delayed electron generation. By stacking selected *S*2 waveforms with a fixed reference time (e.g., the start of the waveform), the resulting stacked waveform was analyzed to give the production probability, as shown in Fig. 8. The correlation between the probability of the delayed electron production and the corresponding delay times was modeled empirically using a combination of two Gaussians and two exponential distributions. The Gaussian components represent delayed electrons originating from the gate and cathode electrodes. The mean delay

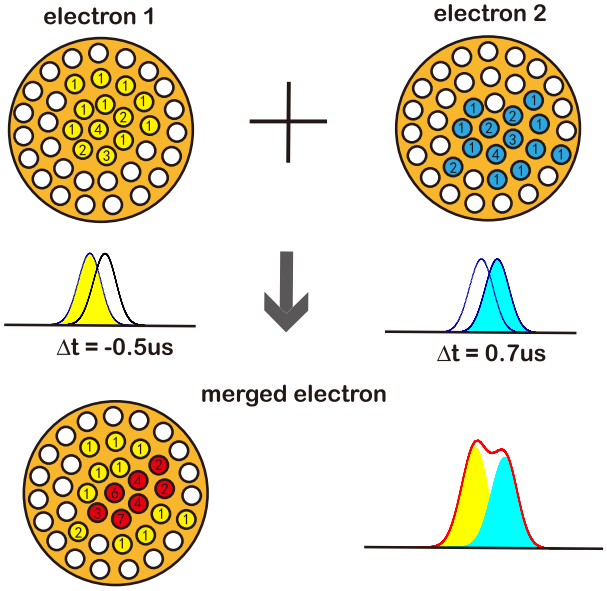

Fig. 6. (color online) Example of *S*2 simulation. Two SE events were selected from SE pools. Their patterns and waveforms are shown in the upper panels. Time shifts were given to each SE waveforms and followed a Gaussian distribution, which is compatible with the diffusion of an event from Z = 47.2 cm. The pattern and waveform of the then merged simulated *S*2 signal are shown in the lower panel. The number given in the pattern plot represents the hit number seen by the PMT.

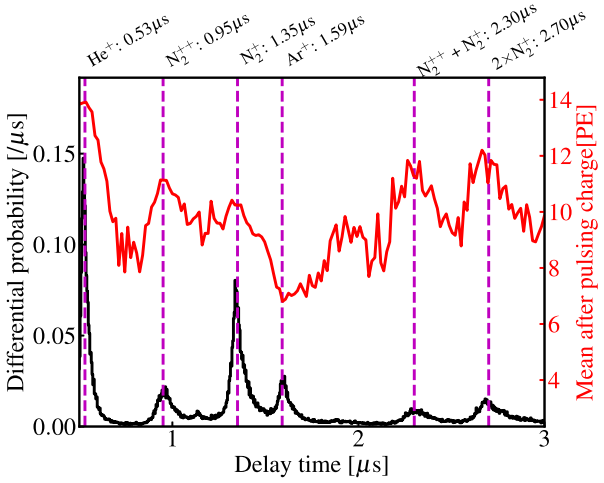

Fig. 7. (color online) Differential probability of PMT after-pulsing as a function of delay time. After-pulsings caused by helium, nitrogen, and argon residual gases are most visible, and highlighted by the magenta dashed lines. Secondary after-pulsing of nitrogen is also indicated by magenta dashed lines. Red solid line shows the mean charge of after-pulsing after each single-PE.

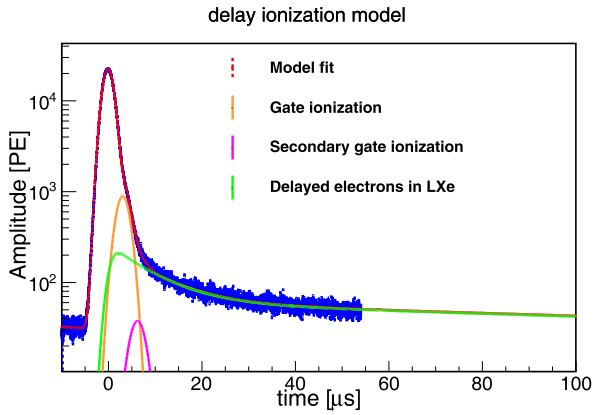

Fig. 8. (color online) Average $S2$ waveform (blue solid line). Red solid line shows a fit to the average $S2$ waveform, consisting of gate ionization (orange solid line) and delayed electron contribution from LXe (green solid line). The gate photoionization was generated from a fixed Z position and was modeled as Gaussian. The delayed electrons in LXe were modeled as the sum of two exponentials convoluted by a Gaussian. The Gaussian convolution represents time spread due to electrons traveling in the gas gap during proportional amplification. The $S2$ light of gate photoionization (magenta solid line) also further generated photoionization (secondary gate photoionzation).

time for the gate delayed electrons, extracted from the fit to the stacked waveform, was 3.2 μs. These values align with the expected behavior, assuming the liquid-gas interface is 5–8 mm above the gate. The delayed electrons from the cathode were negligible compared to other components. The two exponential distributions were found to adequately model delayed electrons resulting from impurity photoionization and electron delays caused by impurity or liquid surface trapping. The probabilities of generating one delayed electron per detected $S2$ photon were determined as 0.24 and 0.15%, respectively, for the gate photoionization and other aforementioned effects. In the WS, the number of delayed electrons were sampled based on these probabilities and the delay time distributions are shown in Fig. 8. The secondary photoionization caused by the primary photoionization was negligible and not implemented in the WS. Further adjustments to the photoionization probability were made to account for the presence of small $S2$ signals in the $S2$ waveform samples obtained from real data.

E. Noise and dark counts

Apart from the aforementioned effects, it is important to consider the presence of noise and dark counts from the PMTs, as they can potentially overshadow the small $S1$ signals and lead to incorrect pairing. Noise mainly refers to spurious lights present in the TPC that are observed to follow a large $S2$ signal [34]. As the $S2$ rates increase in the calibration data, the corresponding noise rate also increase. To account for these sources of noise and dark counts, they were incorporated into the WS by stacking real segmented waveforms on top of the simulated waveform. These noise segments were randomly sampled from the 2-ms window preceding the identified $S1$ signal. The noise level appeared to have a correlation with the type of run conducted. Figure 9 displays selected segmented waveforms from scientific runs for the DM search, ER calibration, and NR calibration.

III. COMPARISON TO DATA

To evaluate the performance of the WS, a comparative analysis was performed by examining the distributions of key variables between the simulated samples and experimental data. These key variables were derived by subjecting the simulated waveforms to the same data processing algorithm employed in the PandaX-4T experiment. The variables selected for comparison can be classified into four distinct categories, with respect to $S1$ pulse shape, $S2$ pulse shape, pattern, and waveform "dirtiness". The description of these variables can be found in Table 1.

The comparison was conducted using the ^{220}Rn and neutron calibration data for ER and NR, respectively. The data selections that were used for DM search [35] were applied. To mitigate the potential influence arising from correlations between the key variables and parameters such as $S1$ charge, $S2$ charge, and event position, it is necessary to ensure that the distributions of $S1$, $S2$, and position in the WS and data samples are compatible.

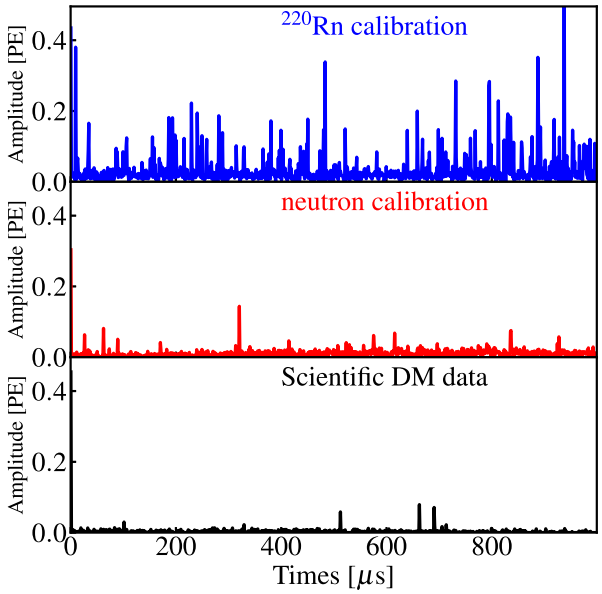

Fig. 9. (color online) Sample noise waveforms from ^{220}Rn calibration data (top), neutron calibration data (middle), and DM data (bottom).

Table 1. List of key variables used to compare the data and WS.

Category	Symbol	Description	reference
$S1$-related	$N_{S1}^{\rm peak}$	Number of peaks of major $S1$.	Figs. 10 and 12
	Q_{S1}	Total charge of major $S1$.	
	h_{S1}	Pulse maximum height of major $S1$.	
	W_{S1}	Difference of end and start times of major $S1$ pulse.	
	$W_{S1}^{\rm FWHM}$	Half-height width of major $S1$.	
	$W_{S1}^{\rm ten}$	10%-height width of major $S1$.	
	$N_{S1}^{\rm hit}$	Number of hits for major $S1$.	
	$N_{S1}^{\rm cand}$	Number of candidate $S1$ signals in the event.	
	$M_{S1}^{\rm bot}$	Charge on most-fired bottom PMT of major $S1$.	
$S2$-related	$N_{S2}^{\rm peak}$	Number of peaks of major $S2$.	Figs. 11 and 13
	Q_{S2}	Total charge of major $S2$.	
	$W_{S2}^{\rm FWHM}$	Half-height width of major $S2$.	
	$W_{S2}^{\rm ten}$	10%-height width of major $S2$.	
	h_{S2}	Pulse maximum height of major $S2$.	
	$R_{\rm preS2}$	Fraction of charge in pre-maximum-height window to total of major $S2$.	
	$N_{S2}^{\rm hit}$	Number of hits of major $S2$.	
	W_{S2}^{80}	$S2$ width containing 80% charge.	
	$\sigma_{S2}^{\rm hit}$	Standard deviation of charges of hits among PMTs of major $S2$.	
Pattern-related	A_{S1}	Top-bottom asymmetry of major $S1$.	Figs. 14 and 15
	A_{S2}	Top-bottom asymmetry of major $S2$.	
	$F_{S1}^{\rm maxq}$	Fraction of charge on most-fired PMT over total charge of major $S1$.	
	$F_{S2}^{\rm maxq}$	Fraction of charge on most-fired PMT over total charge of major $S2$.	
	σ_{S1}^{ch}	Standard deviation of charges of fired PMTs of major $S1$.	
	$\sigma_{S2_b}^{\rm CoG}$	Root-mean-square of hit PMT positions to center-of-gravity reconstructed position of major $S2$.	
Waveform "dirtiness"	$\rho_{\rm preS1}$	Charge density before major $S1$.	Figs. 16 and 17
	ρ_{S1-S2}	Charge density inbetween major $S1$ and $S2$.	
	$\rho_{\rm postS2}$	Charge density after major $S2$.	
	$F_{\rm S1-S2}$	Charge fraction of major $S1$ plus $S2$ to total charge in event waveform.	

A. Comparison of *S*1 and *S*2 pulse shape related variables

Because it is crucial to verify the fidelity of the WS by closely matching the pulse shapes of the simulated waveforms to those observed in real data, a comparison was conducted between the WS and data based on a set of selected variables, including pulse height, pulse width, number of hits, and hit charge variations. The comparisons were performed using the ER and NR calibration data, and the chi-square values were used as a measure of agreement. Figs. 10−13 illustrate the comparisons for the $S1$ and $S2$ pulses in the ER and NR calibration data. The majority of the variables demonstrated good agreement between the WS and data, with only a few variables exhibiting noticeable differences, including h_{S2}, $\sigma_{S2}^{\rm hit}$, and $N_{S2}^{\rm peak}$. This is speculated to arise from the limited statistical significance of the calibration data comprising the $S1$ and SE pools in the WS. In addition, a relaxed selection criterion was employed for these $S1$ and SE events to augment the statistics, potentially introducing impurities into the collected data.

B. Comparison of pattern related variables

In addition to the temporal features of the simulated pulses, the pattern on the PMTs plays a crucial role. Figures 14 and 15 show the comparison of the $S1$ and $S2$ pattern related variables between the WS and calibration data, including the top-bottom asymmetries for the $S1$ and $S2$ signals, standard deviation of the hit charges among the fired PMTs, and RMS of the hit PMT positions. Most variables show good agreement, except the fraction of the

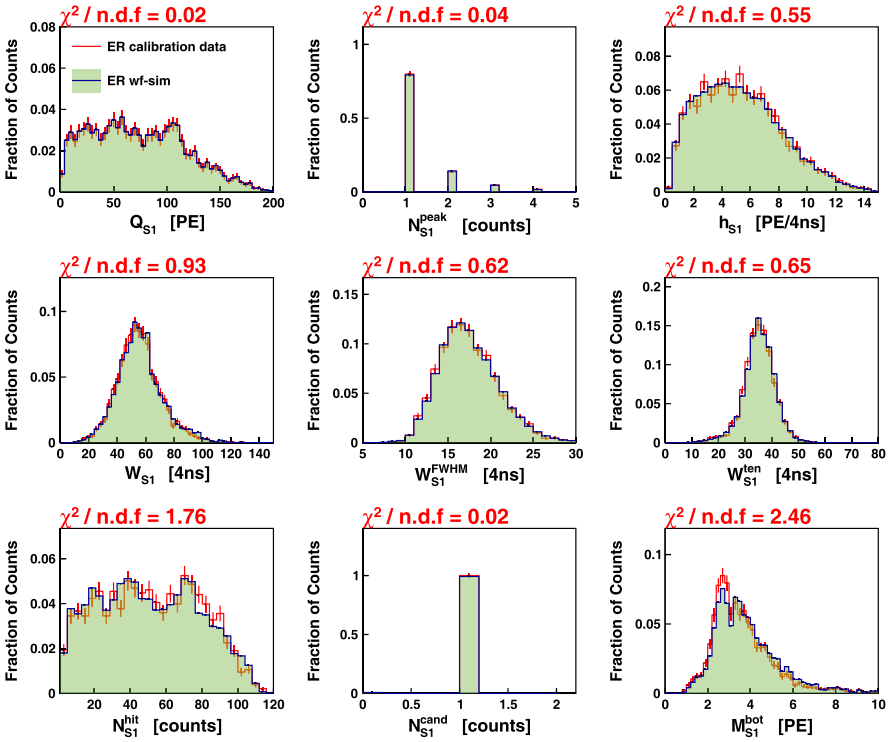

Fig. 10. (color online) Comparisons of $S1$ key variables related to pulse shapes using ER calibration data. Red error bars represent data distribution, and shaded green histogram shows distribution from WS. χ^2 values divided by the degree of freedom are shown on top of each panel.

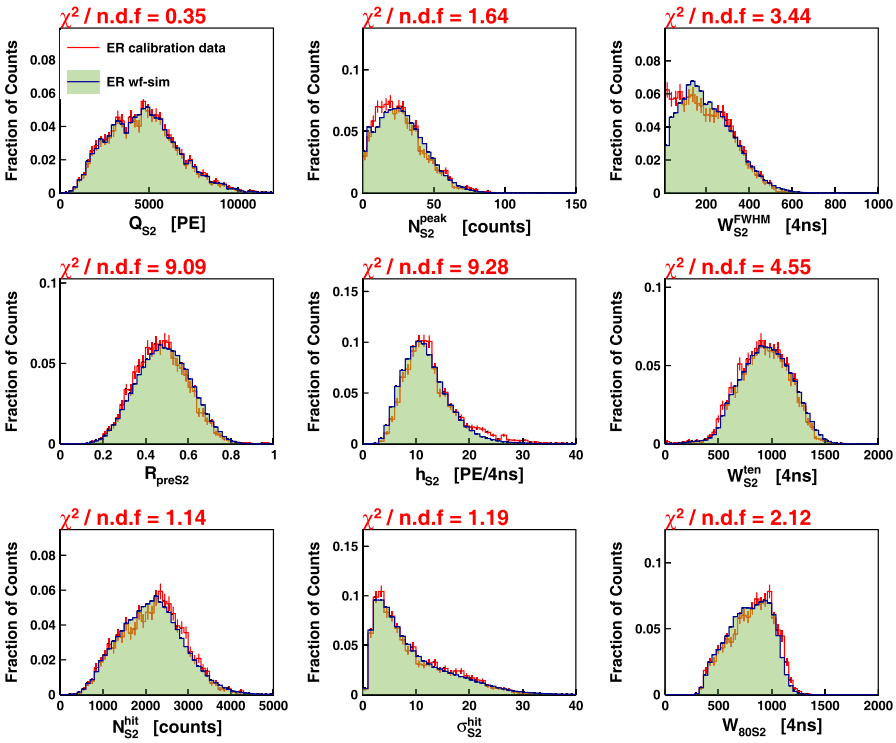

Fig. 11. (color online) Comparisons of $S2$ key variables related to pulse shapes using ER calibration data. Red error bars represent data distribution, and shaded green histogram shows distribution from WS. χ^2 values divided by degree of freedom are shown on top of each panel.

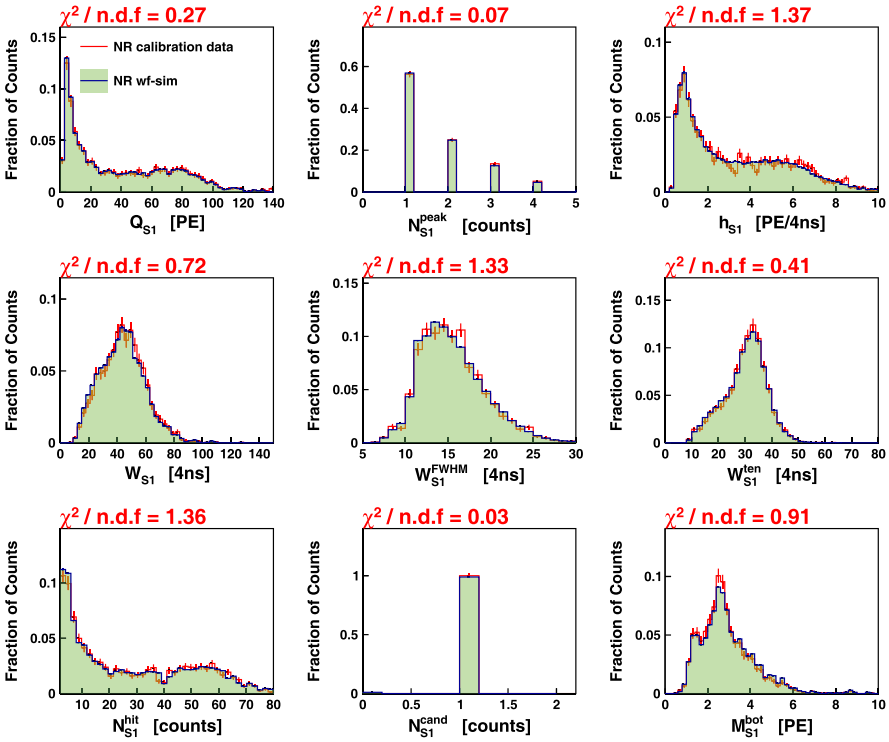

Fig. 12. (color online) Comparisons of $S1$ key variables related to pulse shapes using NR calibration data. Red error bars represent data distribution, and shaded green histogram shows distribution from WS. χ^2 values divided by degree of freedom are shown on top of each panel.

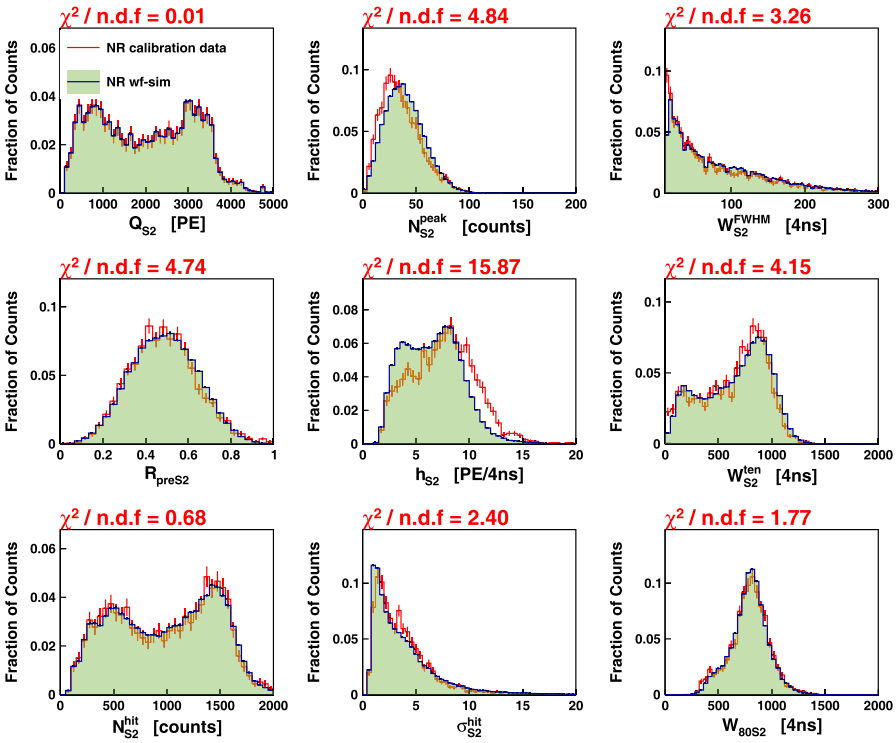

Fig. 13. (color online) Comparisons of $S2$ key variables related to pulse shapes using NR calibration data. Red error bars represent data distribution, and shaded green histogram shows distribution from WS. χ^2 values divided by degree of freedom are shown on top of each panel.

$S2$ charge on the most-fired PMT over the total $S2$ charge F_{S2}^{maxq}. This is attributed to the inherent uncertainty in the reconstructed positions of the SEs within the simulation pool. Notably, the reconstructed positions of the SEs exhibit considerable statistical uncertainty due to the low $S2$ charge, particularly in proximity to the edge of the TPC. Given that the actual positions of the SE blocks constituting the simulated $S2$ may not align with the expected position, discrepancies in the $S2$ pattern-related variables between the data and WS may arise.

C. Comparison of waveform "dirtiness" related variables

The presence of spurious noise in the waveforms can introduce incompatibility between the WS and data. To evaluate this effect, direct comparisons were performed for several selected variables, as summarized in Table 1 and depicted in Figs. 16 and 17. The initial results indicate slight deviations between the WS and data, suggesting the need for adjustments in the production probabil-

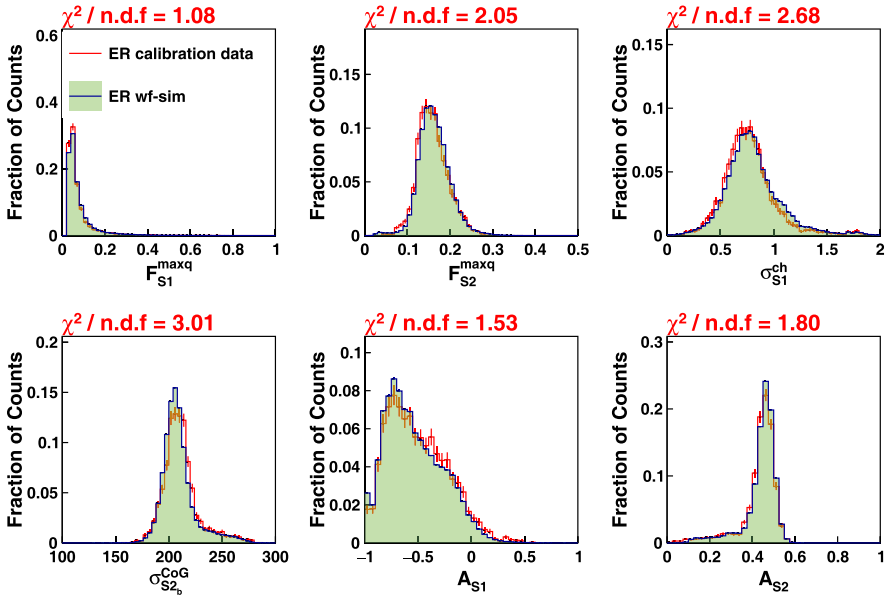

Fig. 14. (color online) Comparisons of pattern-related key variables using ER calibration data. Red error bars represent data distribution, and shaded green histogram shows distribution from WS. χ^2 values divided by degree of freedom are shown on top of each panel.

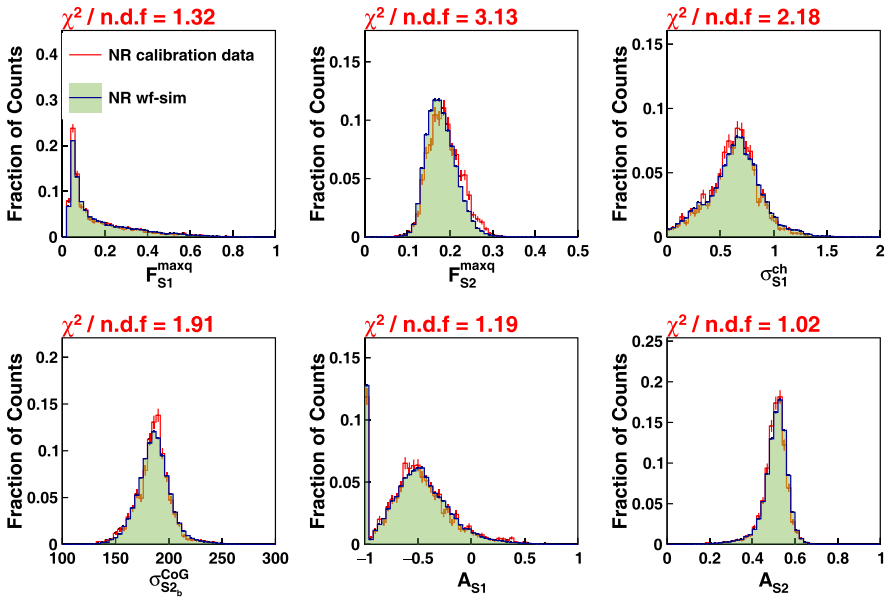

Fig. 15. (color online) Comparisons of pattern-related key variables using NR calibration data. Red error bars represent data distribution. Shaded green histogram shows distribution from WS. χ^2 values divided by degree of freedom are shown on top of each panel.

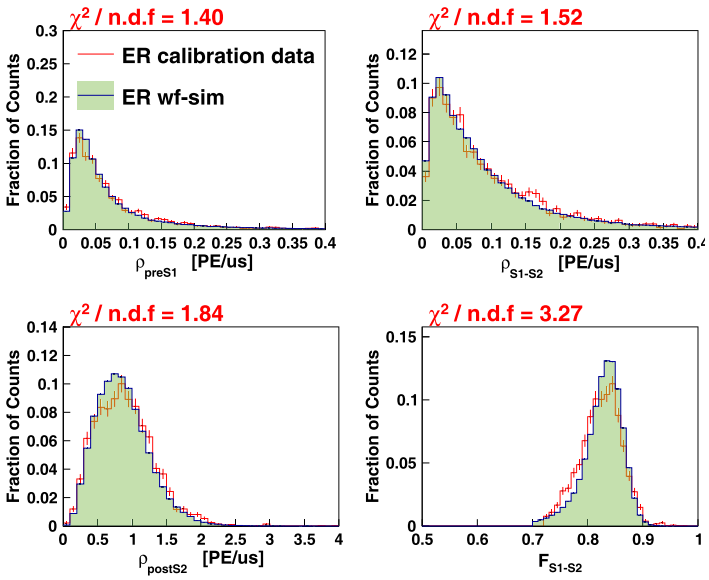

Fig. 16. (color online) Comparisons of waveform "dirtiness"-related key variables using ER calibration data. Red error bars represent data distribution, and shaded green histogram shows distribution from WS. Chi square values divided by degree of freedom are shown on top of each panel.

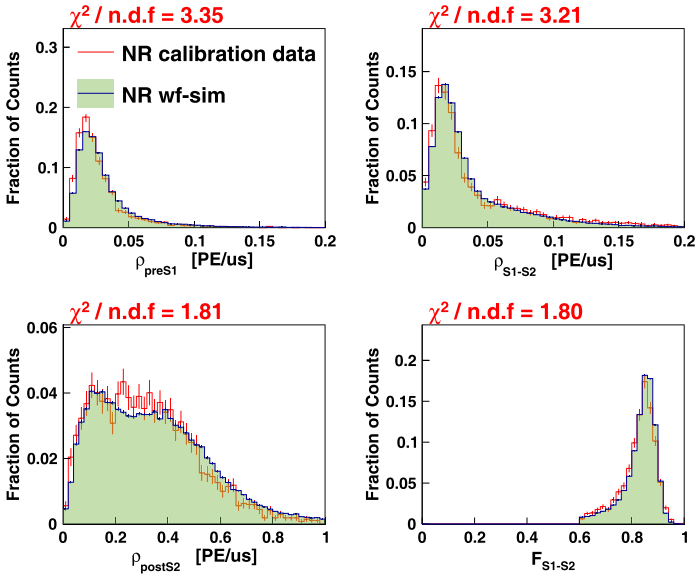

Fig. 17. (color online) Comparisons of waveform "dirtiness"-related key variables using NR calibration data. Red error bars represent data distribution, and shaded green histogram shows distribution from WS. Chi square values divided by degree of freedom are shown on top of each panel.

ity of the photoionization model within the WS. Subsequent modifications were made based on these findings, leading to a satisfactory improvement in the matching between the WS and data.

IV. SUMMARY AND DISCUSSION

We presented a detailed overview of the WS framework developed for the PandaX-4T experiment. This WS framework has been extensively utilized to address various research objectives that require a sufficient statistical sample size, which may not be readily available in the experimental data. The primary application of the WS framework is to generate synthetic samples for conducting studies related to reconstruction efficiency and bias, as referenced in [35]. Additionally, the WS framework serves as a valuable resource for generating training samples for boosted decision tree algorithms, as mentioned in [15]. Through our analysis, we demonstrated that the data-driven waveform simulation provides a com-

parable description of the experimental waveforms, particularly in terms of the pulse shape, pulse pattern, and presence of spurious noise. This highlights the effectiveness and reliability of the WS framework in capturing important features of the observed data. With more calibration data and understanding of the detector, we expect to further improve the WS framework, especially in terms of the correlation between the pulse shape and pattern, as well as its efficacy in high-energy regimes. Additionally, we envision broadening the application of the WS framework to encompass the interpretation of a wider range of physical models.

ACKNOWLEDGEMENTS

We would like to thank the sponsorship from the Chinese Academy of Sciences Center for Excellence in Particle Physics (CCEPP), Hongwen Foundation in Hong Kong, Tencent Foundation in China, and Yangyang Development Fund. We would also like to thank the CJPL administration and Yalong River Hydropower Development Company Ltd. for their indispensable logistical support and help.

References

[1] Gianfranco Bertone, Dan Hooper, and Joseph Silk, Physics Reports **405**(5-6), 279 (2005)
[2] Gerard Jungman, Marc Kamionkowski, and Kim Griest, Physics Reports **267**(5-6), 195 (1996)
[3] Jonathan L Feng, Annual Review of Astronomy and Astrophysics **48**, 495 (2010)
[4] Xiang Xiao *et al*., Phys. Rev. D **92**(5), 052004 (2015)
[5] Andi Tan, Mengjiao Xiao, Xiangyi Cui *et al*., Phys. Rev. Lett. **117**(12), 121303 (2016)
[6] E Aprile *et al*. (XENON Collaboration), Phys. Rev. Lett. **121**(11), 111302 (2018)
[7] DS Akerib, S Alsum, HM Araújo *et al*., Phys. Rev. lett. **118**(2), 021303 (2017)
[8] R Agnese, T Aramaki, IJ Arnquist *et al*., Phys. Rev. Lett. **120**(6), 061802 (2018)
[9] F Petricca, G Angloher, P Bauer *et al*., First results on low-mass dark matter from the cresst-iii experiment. In *Journal of Physics: Conference Series*, volume 1342, page 012076. IOP Publishing, 2020.
[10] Y. Meng, Z. Wang, Y. Tao *et al*., Phys. Rev. Lett. **127**(26), 261802 (2021)
[11] E Aprile, K Abe, F Agostini *et al*., Phys. Rev. Lett. **131**(4), 041003 (2023)
[12] LUX-ZEPLIN Collaboration, Phys. Rev. Lett. **131**(4), 1 (2023)
[13] Ke-Jun Kang, Jian-Ping Cheng *et al*., Status and prospects of a deep underground laboratory in china. In *Journal of Physics: Conference Series*, volume 203, page 012028. IOP Publishing, 2010.
[14] Jian-Ping Cheng, Ke-Jun Kang, Jian-Min Li *et al*., Annual Review of Nuclear and Particle Science **67**, 231 (2017)
[15] Wenbo Ma, Abdusalam Abdukerim, Chen Cheng *et al*., Phys. Rev. Lett. **130**(2), 021802 (2023)
[16] Shuaijie Li, Mengmeng Wu, Abdusalam Abdukerim *et al*., Phys. Rev. Lett. **130**(26), 261001 (2023)
[17] Linhui Gu, Abdusalam Abdukerim, Zihao Bo *et al*., Phys. Rev. Lett. **129**(16), 161803 (2022)
[18] Dan Zhang, Ab dusalam Abdukerim, Zihao Bo *et al*., Phys. Rev. Lett. **129**(16), 161804 (2022)
[19] Peter Gaemers, R. Angevaare Joran Zhu, Tianyu *et al*. Xenonnt/wfsim: v1.0.2, 2022. https://zenodo.org/records/7216324.
[20] Shinzou Kubota, Masahiko Hishida, and Jian-zhi Raun, Journal of Physics C: Solid State Physics **11**(12), 2645 (1978)
[21] Akira Hitachi, Tan Takahashi, Nobutaka Funayama *et al*., Phys. Rev. B **27**(9), 5279 (1983)
[22] Wenbo Ma, Abdusalam Abdukerim, Zihao Bo *et al*., J. Instrum. **15**(12), P12038 (2020)
[23] Oumarou Njoya, T Tsang, M Tarka *et al*., Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **972**, 163965 (2020)
[24] Shaoli Li, Xun Chen, Karl L Giboni *et al*., J. Instrum. **11**(02), T02005 (2016)
[25] Kevin Lung, K Arisaka, A Bargetzi *et al*., Characterization of the hamamatsu r11410-10 3-in. photomultiplier tube for liquid xenon dark matter direct detection experiments. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 696: 32–39, 2012.
[26] Laura Baudis, Annika Behrens, Alfredo Ferella *et al*., J. Instrum. **8**(04), P04026 (2013)
[27] D Yu Akimov, AI Bolozdynya, Yu V Efremenko *et al*., Instruments and Experimental Techniques **58**, 406 (2015)
[28] E Aprile, M Alfonsi, K Arisaka *et al*., Journal of Physics G: Nuclear and Particle Physics **41**(3), 035201 (2014)
[29] Elena Aprile, K Abe, F Agostini *et al*., Phys. Rev. D **106**(2), 022001 (2022)
[30] DS Akerib, S Alsum, HM Araújo *et al*., Phys. Rev. D **102**(9), 092004 (2020)
[31] P Sorensen and K. Kamdin, J. Instrum. **13**(02), P02032 (2018)
[32] A Tomás, HM Araújo, AJ Bailey *et al*., Astroparticle Physics **103**, 49 (2018)
[33] Adam Bailey. *Dark Matter Searches and Study of Electrode Design in LUX and LZ.* PhD thesis, Imperial College London, 2016.
[34] Elena Aprile, Jelle Aalbers, F Agostini *et al*., Phys. Rev. Lett. **126**(9), 091301 (2021)
[35] Yunyang Luo, Zihao Bo, Shibo Zhang *et al*., Signal response model in pandax-4t. *arXiv preprint arXiv: 2403.04239*, 2024.