An Alternative LowBG Solar Neutrino Analysis by Using MultiPath Water Fitter

Jie Hu

Department of Physics, University of Alberta

1. Introduction

The SNO+ water phase data were taken from May 2017 to September 2018. The period from May 2017 to October 2018 is the first stage of the water phase. During this stage, several calibration runs were taken, including the ^{16}N calibration scans and the laserball scans. During the period from October 2018 to July 2019, over 20 tonnes of LAB (without PPO) was filled into the detector and the LAB mostly occupied the neck volume, slightly below the neck bottom. With the nitrogen cover gas on the top of the AV, the dataset taken during this period is called "low background dataset". In this study, 4838 runs of data were used, which summed up a total live time of 190.31 days after the data cleaning process.

In this chapter, I applied the MPW reconstruction algorithm described in Chapter 4 as a position and direction fitter to the raw dataset as well as the run-by-run MC simulations. The fitted event vertex was used by the energy fitter developed by the SNO+ collaboration.

First, the open dataset taken in 2017 was used to test the MPW results, compared to the official RAT results. A quantity called "KullbackLeibler Divergence" was developed to evaluate the Cherenkov signals. After that, I mainly analyzed the low background dataset. I used sub datasets of the run-by-run MC simulations to evaluate the ability of separating the solar ν_e signals from the backgrounds. The Toolkit for Multivariate Data Analysis with ROOT (TMVA) package [**? ?**] was used to train and test on the MC simulations to obtain optimized discriminants. These optimized discriminants were applied on the whole dataset to remove the backgrounds.

The outputs from the data were fitted to obtain the number of signal events and the background events. Ensemble tests were performed on fake datasets to check the fit pull and bias. The systematics obtained from the ^{16}N calibration in Chapter 5 were applied on the results. Finally, the solar

ν_e interaction rates and the ^{8}B solar neutrino flux were evaluated.

2. KullbackLeibler (KL) Divergence for High Level Cuts

The KullbackLeibler (KL) divergence (also called "relative entropy") is used to measure the dissimilarity of two probability distributions[**?**]. I used this quantity to compare the reconstructed angular distribution of an event, $\vec{u}_{fit} \cdot (\vec{X}_{PMT} - \vec{X}_{fit})/|\vec{X}_{PMT} - \vec{X}_{fit}|$, with the angular distribution of solar ν_e events extracted from the MC (expected as a Cherenkov distribution) to check the dissimilarity of the event compared with the solar ν_e event. The quantity $D_{KL}(p||q)$ is calculated as:

$$klDiv(p||q) \equiv \sum_i^N p(x_i) \log \frac{p(x_i)}{q(x_i)}, \tag{2.1}$$

where $p(x_i)$ is the angular distribution after a time residual window cut: $-5 < t_{Res} < 1\ ns$, to extract prompt Cherenkov lights. Both of the event and the MC distributions were filled into a histogram with 40 bins ranging from [-1,1] and the *klDiv* values were calculated bin by bin except the empty bins (zero count). A small *klDiv* value indicates a small dissimilarity.

These values were used for distinguishing the signal from backgrounds, which will be discussed in the section 3.2. Fig. 1 shows an example of the *klDiv* calculation. Two events are compared here. One is a randomly selected event from the solar ν_e run-by-run MC ($E = 4.78\ MeV$), the other is from the ^{214}Bi MC ($E = 2.18\ MeV$), with the same event GTID. It can be seen that the background event with lower energy is more dispersive while the signal event has a peak around the Cherenkov angle (~ 0.75) and thus its shape is more close to the pdf. The calculation of 2.1 gives $klDiv(solar\ \nu_e) = 11.78$ and $klDiv(^{214}Bi) = 22.69$, which verifies the observation.

A symmetrical form of *klDiv* can be taken as:

$$klDiv(p,q) \equiv \frac{1}{2} \sum_i^N (p \log \frac{p}{q} + q \log \frac{q}{p}), \tag{2.2}$$

Since $klDiv(p,q) = klDiv(q,p)$, it has a meaning of distance. The quantity is not used in the thesis, but can be considered in future.

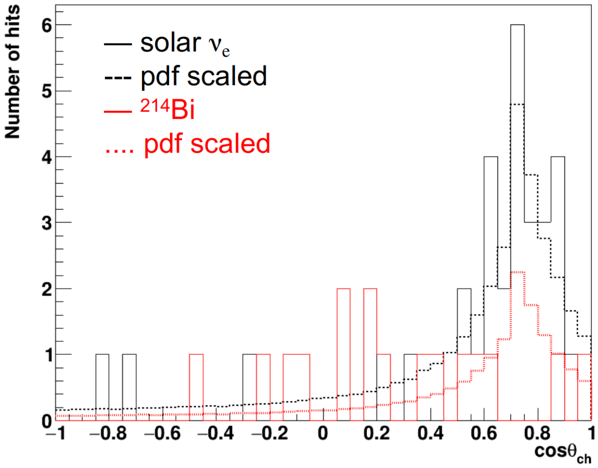

Figure 1: Angular distributions of the MC events in run-by-run simulation 206391, with the same event GTID =7. The black line is the MC solar ν_e distribution, while the red line is the MC ^{214}Bi distribution. The pdf is scaled to the number of the hits in solar ν_e event (black dashed line) and the ^{214}Bi event (red dotted line) respectively.

3. Solar ν_e Analysis and Background Separation in Water Phase

3.1. Open Dataset Analysis

This open dataset was used to compare the MPW and the RAT fitted results.

In SNO+ water phase, solar ν_es are basically measured via elastic scattering $\nu_e + e^- \rightarrow \nu_e + e^-$. The maximum kinetic energy of the recoil electron is $T_{max} = \frac{2E_\nu^2}{2E_\nu + m_e c^2}$ the cross section is $\sigma(\nu_e + e^- \rightarrow \nu_e + e^-) = 9.52 \times 10^{-44}(E_\nu/10\ MeV)\ cm^2$ the expected solar neutrino rate is $R = A\int_{T_{thresh}}^{T_{max}} \frac{d\sigma}{dE}\frac{dN}{dE_\nu}dE_\nu$. A "solar angle", θ_{sun} is the direction of the event relative to the Sun's location,

$\nu - e^-$ elastic scattering:

$$\cos\theta_{sun} = \sqrt{\frac{T_e(m_e + E_\nu)^2}{2m_e E_\nu^2 + T_e E_\nu^2}} \tag{3.1}$$

For the data, the solar angle is defined as:

$$\cos\theta_{sun} \equiv \vec{u}_{event} \cdot \frac{\vec{X}_{event} - \vec{X}_{sun}}{|\vec{X}_{event} - \vec{X}_{sun}|}, \tag{3.2}$$

where $\vec{X}_{sun}$ is taken as the Sun's location relative to the SNOLAB location since the whole lab can be treated as a point regarding the long distance to the Sun.

High level cuts mentioned in **??** were applied.

Table 1: Candidate events in the open dataset. Compared the fitted results of the candidate events with different fitters.

Fitter	Run	GTID	$z - 0.108$(m)	R(m)	$(R/R_{av})^3$	$\cos\theta_{sun}$	SNO+ Day
Rat	100093	11108354	3.49	3.57	0.21	-0.954	2683.92
MPW	–	–	3.43	3.52	0.20	-0.906	–
Rat	100207	5079885	-2.61	4.60	0.45	0.816	2687.04
MPW	–	–	-3.63	**7.61**	2.03	**0.656**	–
Rat	100632	7882360	1.77	3.19	0.15	0.937	2696.93
MPW	–	–	1.67	3.11	0.14	0.911	–
Rat	100663	15767175	-4.33	4.96	0.56	0.978	2698.18
MPW	–	–	-4.45	5.07	0.60	0.980	–
Rat	100915	169700	-1.00	5.10	0.61	0.341	2701.23
MPW	–	–	-1.08	5.08	0.61	0.337	–

Table 2: Candidate events in the open dataset, searched by the MPW fitter.

Run	GTID	energy	$z - 0.108$	R	$(R/R_{av})^3$	$\cos\theta_{sun}$
100093	11108354	5.827	3.43	3.52	0.20	-0.907005
100632	7882360	6.183	1.67	3.11	0.14	0.9146124
100663	15767175	6.182	-4.45	5.07	0.60	0.9807349
100915	169700	5.684	-1.07	5.08	0.61	0.3385341
100984	8621621	5.701	0.76	4.75	0.502	-0.647735
101075	11673714	5.667	4.43	5.18	0.64	0.5873025

solar neutrino candidate events in the open dataset.

3.2. TMVA Analysis

The MC simulations of the runs 200004 to 203602 were used. These run-by-run simulations simulated the full detector conditions for every run. This is a sub-dataset to the whole "low background dataset", with a live time of 92.54 days for testing and training the TMVA methods.

Two types of background isotopes, ^{208}Tl and ^{214}Bi were simulated in different detector regions. In this study, the background events simulated in the inner AV (internal backgrounds) , in the AV and in the external water region were checked. The solar ν_e events simulated in the inner AV were used as signals. Table. 3 summarizes the types of simulations used in this study.

Table 3: Datasets of MC simulations.

Simulations	Simulated positions in the detector
^{208}Tl	inner AV (internal ^{208}Tl)
–	AV
–	external water (external ^{208}Tl)
^{214}Bi	inner AV (internal ^{214}Bi)
–	AV
–	external water (external ^{214}Bi)
Solar ν_e	inner AV (internal ν_e)
–	AV
–	external water (external ν_e)

Different types of the simulations were merged into a mixed dataset. The simulated solar ν_e events are tagged as signals and mixed with ^{214}Bi and ^{208}Tl background events. The total dataset was divided into training and testing sets.

Fig. 2 shows the energy spectrum of simulated internal events with their fitted positions inside the 5.5-m fiducial volume, i.e., with a radial cut of $R'_{fit} < 5.5\ m$, where the R'_{fit} is the magnitude of the reconstructed event position $\vec{X}_{fit}$ after the AV coordinate correction: $R'_{fit} \equiv \sqrt{x_{fit}^2 + y_{fit}^2 + (z_{fit} - 108)^2}$. The 108 mm offset in z was discussed in Chapter 3 and Chapter 4.

In this sub-dataset, runs from 201700 to 202516 were taken as the training set (69.5% of the total sub-dataset), and the rest 30.5% were taken as testing set. Once the weights of the variables were tuned, they were put into the actual data.

Before the analysis, a few "beforehand cuts"were applied: $NHits > 20$, $R'_{fit} < 5500\ mm$, $ITR > 0.55$, $-0.12 < \beta_{14} < 0.95$. Here $NHits > 20$ was applied as a reconstruction threshold. Only the events with $NHits > 20$ were reconstructed by the MPW fitter for the solar neutrino analysis; a default fiducial volume of 5.5 m was set; the ITR and β_{14} cuts were suggested by

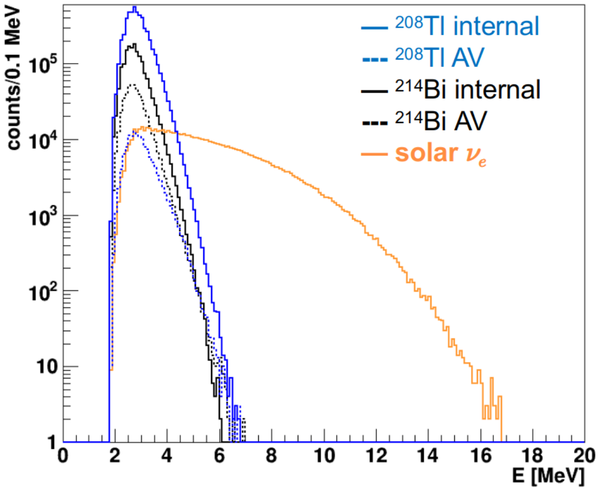

Figure 2: Energy spectrum of the simulated events for ^{214}Bi (black), ^{208}Tl (blue) and solar ν_e (orange). Solid lines show the internal events and dotted lines show the AV events.

the collaboration, which were mostly based on the experiences for removing the instrumental backgrounds[**?**].

Distributions of input variables: signal vs combined backgrounds

After applying these beforehand cuts, for $4 < E_{fit} < 15$ MeV, the training dataset had 416780 events and the testing dataset had 184330 events.

Two other ranges of E_{fit} were also tested: $4 < E_{fit} < 5$ (low energy region) and $5 < E_{fit} < 15$ ($E > 5$ region).

After the training and testing datasets were ready, three classification methods implemented in the TMVA package were used: the Fisher discriminants/linear discriminant analysis (Fisher/LD), the Boosted Decision Tree (BDT), and the Artificial Neural Networks Multilayer Perceptron (ANN-MLP, or MLP in short)[**?**].

The Fisher discriminant $y_{F_i}(i)$ for classifying event i is defined by [**?**]:

$$y_{F_i}(i) = F_0 + \sum_{k=1}^{n_{params}} F_k x_k(i), \tag{3.3}$$

where n_{params} is the number of input variables; the Fisher coefficients, F_k is given by:

$$F_k = \frac{\sqrt{N_S N_B}}{N_S + N_B} \sum_{l=1}^{n_{params}} 1/W_{kl}(\bar{x}_{S,l} - \bar{x}_{B,l}), \tag{3.4}$$

where $N_{S(B)}$ are the number of signal (background) events in the training sample; $x_{S(\bar{B}),l}$ are the means of input variables for signal (background); W_{kl} is the covariance matrix[**?**].

For the BDT method, the adaptive boosting (AdaBoost) algorithm was used; 400 trees were trained with a maximum depth of 3; gini index was used for the decision tree.

For the MLP method, sigmoid function was set as the activate function; 4 hidden layers, 200 training cycles were used.

I used 9 variables as inputs: ITR, β_{14}, E_{fit}, G_{test}, U_{test}, $scaleLogL$, Z_{factor}, $\vec{u} \cdot \vec{R}$ and $klDiv$. Among them, the input values of ITR and β_{14} were after the beforehand cuts mentioned above. The $NHits$ and θ_{ij} were not used, since the $NHits$ is correlated to the energy while the θ_{ij} is anticorrelated to the β_{14}.

The MLP method gave the best results, while it was the most CPU-consuming method.

As one of the essential TMVA output, the background rejection versus signal efficiency curve is also called a receiver operating characteristic (ROC) curve, which is usually used to test the performance of machine learning classifier. A quantity taking the integrals of the ROC curve: called the "area under the curve" (AUC) is often used to summarize the quality of a ROC curve[**?**]. Fig. 3 shows the ROC curves for different methods: where the Fisher/LD is the worst case; the BDT and MLP outputs are close to each other while the MLP gives the largest AUC values.

A typical CPU time for a certain method to train the dataset is listed in Table. 4, 5 and 6.

Table 4: Testing results of $4 < E_{fit} < 15\ MeV$ from different TMVA methods.

Method	AUC	CPU time (second/10^6 events)
Fisher/LD	0.915	0.81
BDT	0.940	249.53
MLP	0.944	1370.02

It shows that, when the energy goes lower, it is more difficult to separate the signals from the backgrounds.

The distributions of the "solar angle", $\cos\theta_{sun}$ were used to show the performance of the solar ν_e event selection and background event discrimination.

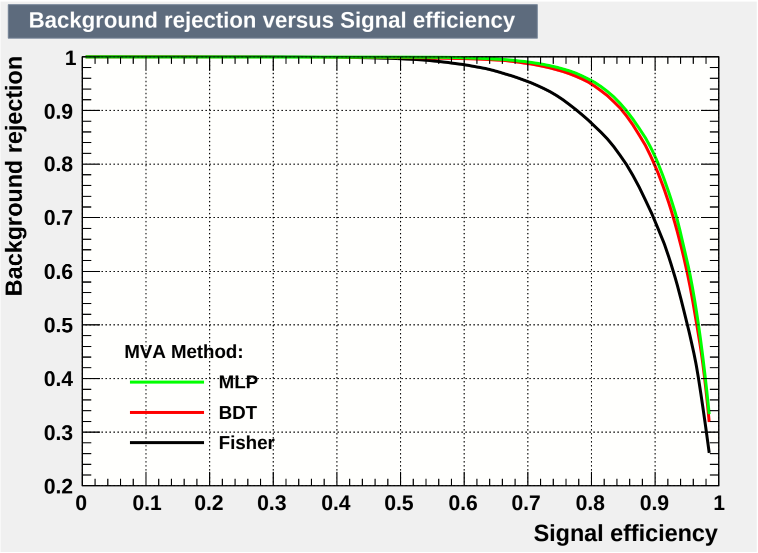

Figure 3: ROC curves from TMVA output, for event with $4 < E_{fit} < 15\ MeV$.

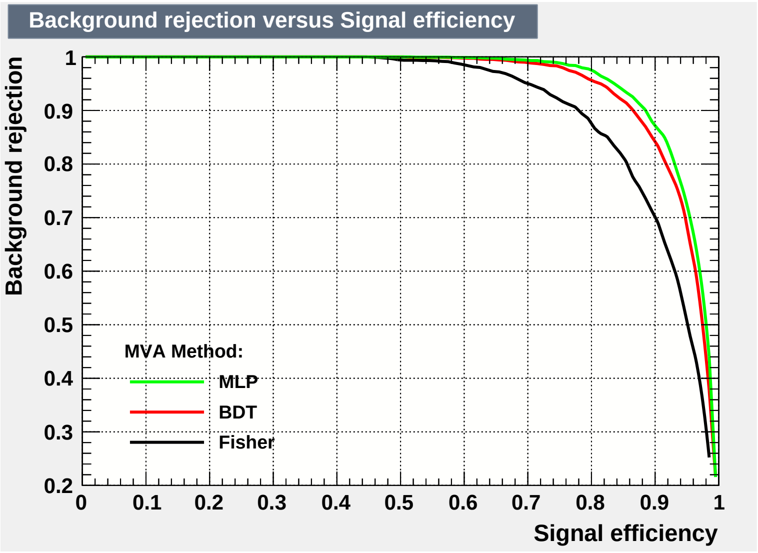

Figure 4: ROC curves from TMVA output, for event with $5 < E_{fit} < 15\ MeV$ (energy above 5 MeV).

It is also used to extract the number of signal and background events, which will be discussed in the section 3.4. Here I applied the BDT and the MPL method on the test sub dataset. For the real dataset from run-200004 to

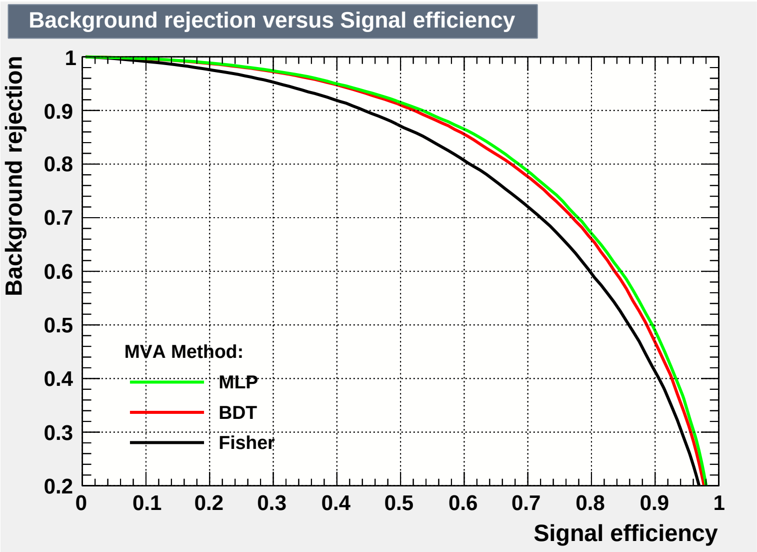

Figure 5: ROC curves from TMVA output, for event with $4 < E_{fit} < 5\ MeV$ (low energy region).

Table 5: Testing results of $5 < E_{fit} < 15\ MeV$ (above 5 MeV) from different TMVA methods.

Method	AUC	CPU time (second/10^6 events)
Fisher/LD	0.915	0.93
BDT	0.950	269.71
MLP	0.958	1450.90

207718, the trained weights and variables from the BDT and the MLP methods were applied event by event and the discriminator responses, D_{BDT} and D_{MLP} were calculated respectively. Cuts of $D_{BDT} > 0.0$ and $D_{MLP} > 0.5$ were applied to extract the solar ν_e signals from backgrounds.

3.2.1. TMVA Outputs for Data

Fig. 6 and Fig. 8 show the BDT selection outputs from the 190.33-day dataset. Fig. 7 and Fig. 9 show the MLP outputs. Table. **??** shows the number of the output events for different energy regions and from different methods.

Cuts on the position and energy FOMs suggested by the collaboration[**?**

Table 6: Testing results of $4 < E_{fit} < 5\ MeV$ (low energy region) from different TMVA methods.

Method	AUC	CPU time (second/10^6 events)
Fisher/LD	0.782	0.84
BDT	0.816	280.1
MLP	0.823	1337.9

] are [1]: $-11 < Z_{factor} < 1$, $scaleLogL > 10.85$, $0 < G_{test} < 1.9$, $U_{test} < 0.95$, $ITR > 0.55$, $-0.12 < \beta_{14} < 0.95$. Combined with the "beforehand cuts", the whole set of cuts is considered as "default cuts" here and is compared with the TMVA outputs.

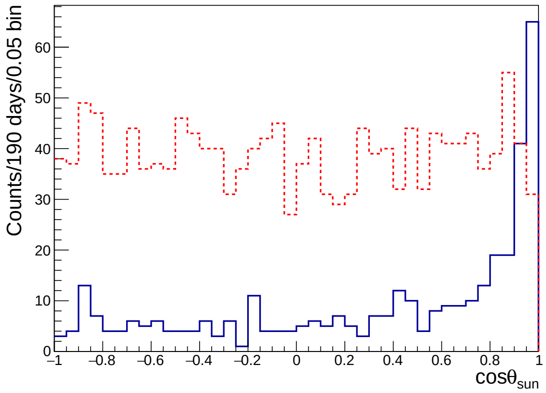

Figure 6: BDT output for $\cos\theta_{sun}$, with $4 < E_{fit} < 15\ MeV$. The solid blue line shows the selected candidate solar ν_e events while the dotted red line shows the selected background events.

The main analysis is focused on the [5,15] MeV energy region. A comparison of the outputs of $5 < E_{fit} < 15\ MeV$ from the BDT, MLP and the default cuts is shown in Fig. 10.

[1] There is also a suggested cut on the quantity of $position_error$ ($position_error < 525\ mm$). However, since this quantity was not calculated by the MPW fitter, it was not included here.

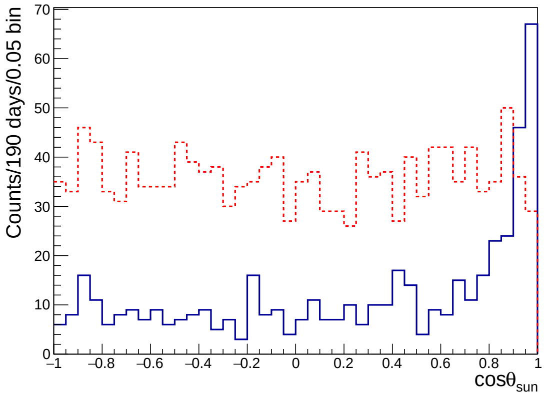

Figure 7: MLP output for $\cos\theta_{sun}$, with $4 < E_{fit} < 15\ MeV$. The solid blue line shows the selected candidate solar ν_e events while the dotted red line shows the selected background events.

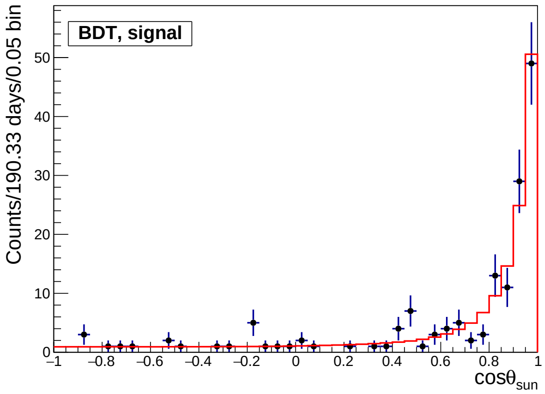

Figure 8: BDT output for $\cos\theta_{sun}$, with $4 < E_{fit} < 15\ MeV$. The solid blue line shows the selected candidate solar ν_e events while the dotted red line shows the selected background events.

3.3. Discussions on the TMVA Results

A more stringent radial cut (or tighter FV) can be applied on lower energy region $4 < E_{fit} < 5\ MeV$ to further remove the backgrounds which are dominant in lower energy region. However, this tighter cut can also reduce the number of signals.

Other packages developed for high energy particle physics, such as `StatPatternRecognition`

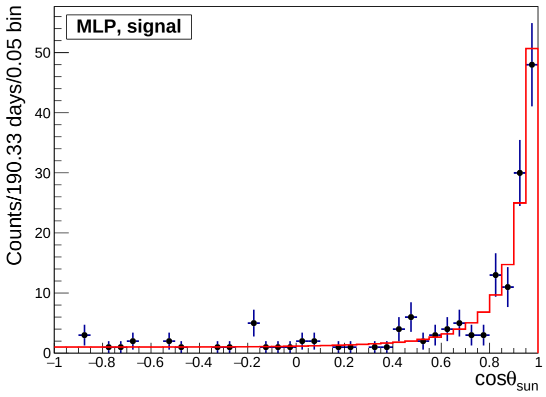

Figure 9: MLP output for $\cos\theta_{sun}$, with $5 < E_{fit} < 15\ MeV$. The solid blue line shows the selected candidate solar ν_e events while the dotted red line shows the selected background events.

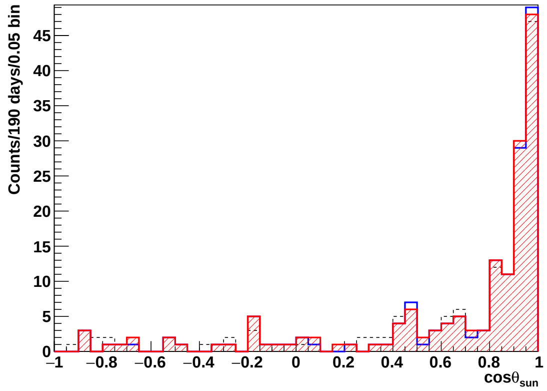

Figure 10: For the $5 < E_{fit} < 15\ MeV$, comparing the outputs of the BDT, MLP and the default cuts. The solid blue line shows the BDT results; the red slashes show the MLP results and the dotted black line shows the default cut results.

(SPR)[?], or more general tools for deep learning, such as `Keras`, `PyTorch` and `TensorFlow`, can also be considered as an alternative tool or as a reference for results comparisons.

3.4. Likelihood Fits for Solar Neutrino Candidate Events

In the previous section, the optimized cuts obtained from the TMVA analysis were applied on the dataset.

After the event selections, a distribution of $\cos\theta_{sun}$ extracted from the solar ν_e candidate events was obtained.

3.4.1. Maximum Likelihood Fit

A maximum likelihood method was applied on the distribution to extract the number of the solar ν_e interaction events (N_{sig}) as well as the number of the background events (N_{bkg}).

The values of $\cos\theta_{sun}$ from the selected events were filled into a histogram divided into bins. For each bin, the observed event number (n_{obs}) was considered as a sum of solar ν_e and background events. The n_{obs} in each bin was assumed to follow a Poisson distribution: $Poisson(n_{obs}, N_{bkg} \cdot P_{bkg} + N_{sig} \cdot P_{ES}(E))$, where P_{bkg} and $P_{ES}(E)$ are the assumed distribution of backgrounds and solar ν_e events respectively.

For the background events, a uniform distribution of $\cos\theta_{sun}$ was assumed. On the other hand, the $\cos\theta_{sun}$ distributions of solar ν_e were extracted from the realistic run simulations after applying the optimized cuts, as shown in Fig. 11.

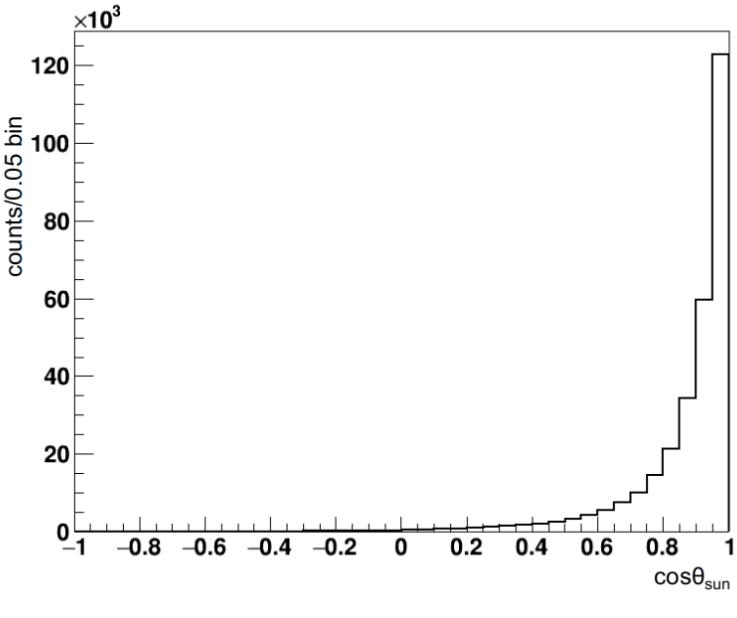

Figure 11: The $\cos\theta_{sun}$ distribution of solar ν_e extracted from the simulations, which was used as a pdf function.

Adding up each bin i and taking N_{bkg} and N_{sig} as the free parameters for fitting, the maximum likelihood function was built as[**?**]:

$$-2\ln\lambda(N_{sig},N_{bkg}) = 2\sum_{i=0}^{N_{bins}}[\mu_i(N_{sig},N_{bkg}) - n_i + n_i \ln\frac{n_i}{\mu_i(N_{sig},N_{bkg})}], \quad (3.5)$$

where $\mu_i(N_{sig},N_{bkg})$ is the expected number of events in each bin: $\mu_i(N_{sig},N_{bkg}) = N_{sig}\cdot P^i_{ES}(E^i) + N_{bkg}\cdot 1/N_{bins}$; N_{bins} is the total number of the bins, usually taken as 40 (per 0.05 bins). This quantity also includes the cases when the bin contains zero ($n_i = 0$).

Fitting the data with (N_{bkg},N_{sig}) by maximizing the quantity 3.5, N_{bkg} and N_{sig} were obtained. In the next section, an ensemble test based on fake datasets was used for testing the fit results.

3.4.2. Ensemble Test

To check the uncertainty of the Poisson fit, 5000 fake datasets were generated. Here I used the method similar to the [**?**]. The fake data were taken from the MC simulation dataset of run-200004 to 203602 after the default cuts (the same to the one used by the TMVA).

The number of backgrounds in a fake dataset, N^f_{bkg}, was assumed to be two times of the event number in the $-1 < \cos\theta_{sun} < 0$ region while the number of signals $N^f_{sig} = N^f_{total} - N^f_{bkg}$. Reading from the sub dataset of run-200004 to 203602 (see Fig. 12), it found $N^f_{bkg} = 38$ and then $N^f_{sig} = 109 - N^f_{bkg} = 71$. To do the ensemble test, for each fake dataset, two random numbers: N^r_{sig} and N^r_{bkg} were generated by the `ROOT TRandom3` random number generator class. Each of the two random numbers followed the random Poisson distribution: $e^{-\mu}\mu^{N^r}/N^r!$, where $\mu = 71$ or 38, and thus they fluctuated around N^f_{sig} or N^f_{bkg}.

To create the fake datasets, N^r_{sig} (N^r_{bkg}) events after the cuts were randomly and uniformly selected from the solar ν_e (merged backgrounds) MC simulations. For each randomly selected event, the values of E_{fit} and $\cos\theta_{sun}$ were recorded. Each dataset was fitted with the maximum likelihood function described in 3.4.1. Fig. 13 shows an example of the fitted results.

The fit pull and the fit bias were defined by [**?**]:

$$bias = \frac{N_{sig} - N^r_{sig}}{N_{sig}}, \quad (3.6)$$

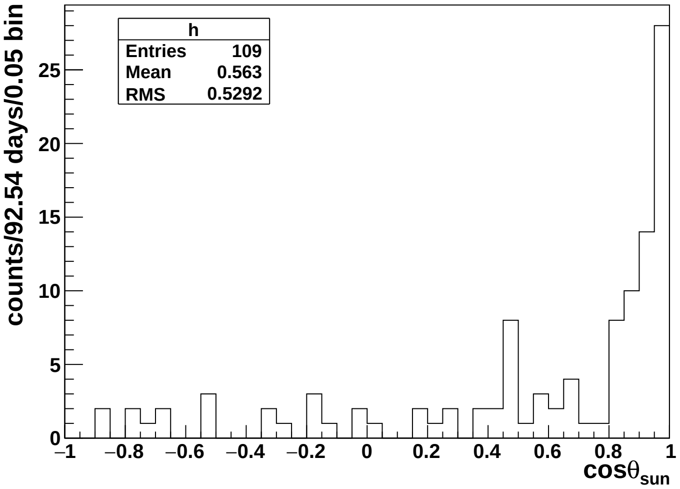

Figure 12: Real data from run-200004 to 203602 (half dataset), after the default cuts. The number of counts in $-1 < \cos\theta_{sun} < 0$ region is 19.

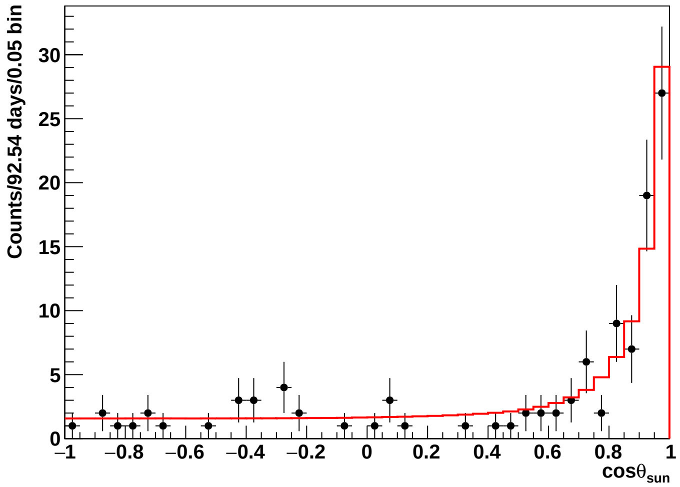

Figure 13: An example of the randomly generated $\cos\theta_{sun}$ fitted with (N_{sig}, N_{bkg}). The black dots are data points and the red line is the fitted results. For $N^r_{sig} = 73$ and $N^r_{bkg} = 44$, the fitted results give $N_{sig} = 73.42 \pm 9.42$ and $N_{bkg} = 43.58 \pm 7.73$, with a $\chi^2/ndf = 60.19/40 = 1.50$.

$$pull = \frac{N_{sig} - N^{r}_{sig}}{\sigma_{sig}}, \tag{3.7}$$

where N_{sig} is the fitted number of signal events, σ_{sig} is the statistical uncertainty of N_{sig}; N^{r}_{sig} is used as the true number of signal events in the fake dataset.

Fig. 15 and Fig. 14 show the fit pull and biases respectively. The histograms were fitted with Gaussians. For the fitted number of signal events, the Gaussian mean of the fit biases is -0.0044 ± 0.0008 for 5000 fake datasets while the Gaussian mean of the fit pulls is -0.026 ± 0.006. These pulls and biases will be applied on the data. Fig. 16 shows the distributions of the $-2\ln L$ returned by the best fitted results ($-2\ln L_{best}$). The distribution, $f(-2\ln L_{best})$, follows the asymptotic χ^2 pdf with a degree of 40 and is used to compute the p-values[**?**]. For a best-fit set $(N^{i}_{sig}, N^{i}_{bkg})$ with a value of $-2\ln L^{i}_{best}$, the p-value is calculated as $p = \int_{-2\ln L^{i}_{best}}^{-2\ln L^{max}_{best}} f(-2\ln L_{best}) d(-2\ln L_{best})$.

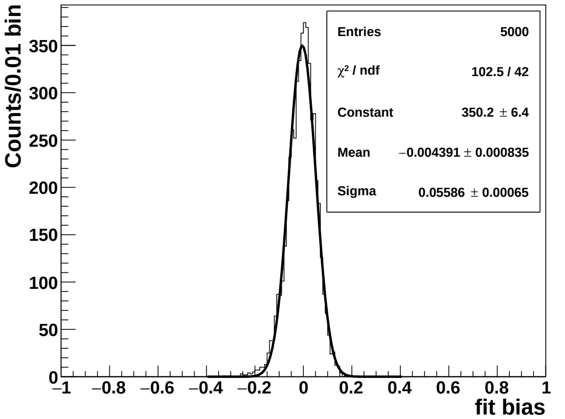

Figure 14: N_{sig} fit biases for 5000 fake datasets.

3.4.3. Fitting on the Whole Dataset (run-200004 to 207718)

The whole dataset started from run-200004 (on 24 Oct, 2018) to run-207718 (on 10 July, 2019). This dataset has a live time of 190.33 days. The BDT and MLP were applied on this dataset.

In the region of $5 < E_{fit} < 15$ MeV, the outputs from the BDT and MLP were fitted to obtain the N_{sig} and N_{bkg}. Fig. 17 and Fig. 18 show their

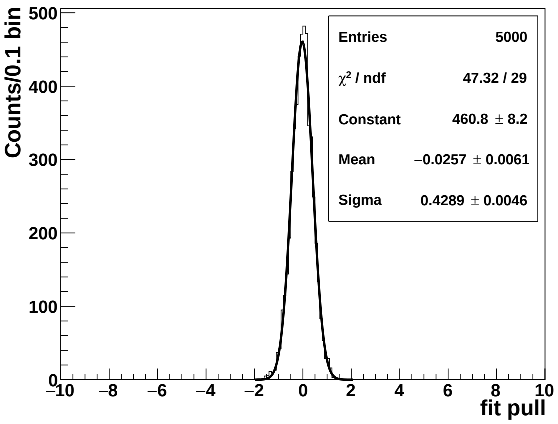

Figure 15: N_{sig} fit pulls for 5000 fake datasets.

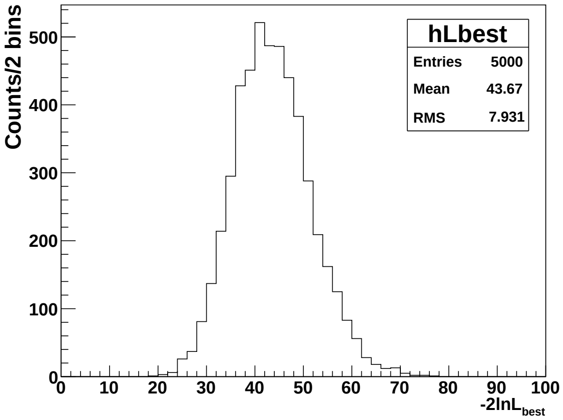

Figure 16: The $-2\ln L$ of the best fitted results for 5000 fake datasets.

results respectively.

These results are also summarized in Table. 7.

It can be concluded here that in the $[5, 15]$ MeV energy region, the BDT results are consistent with the MLP results. The estimated background rate is lower than the signal rate, which indicates that an extremely low background is achieved for the data.

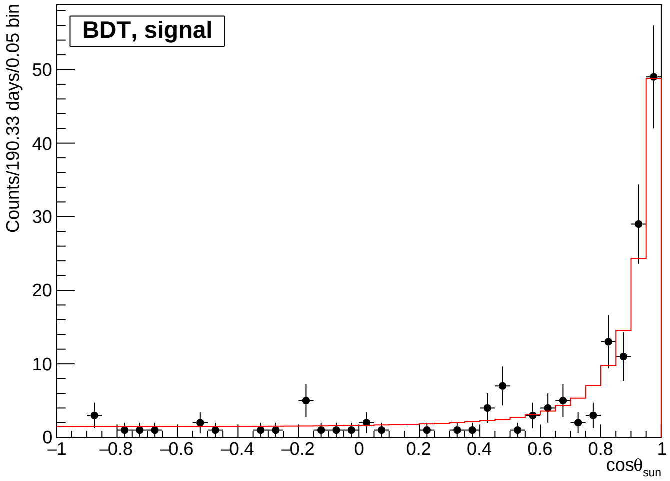

Figure 17: Fitted results for the $5 < E_{fit} < 15\ MeV$, from BDT outputs.

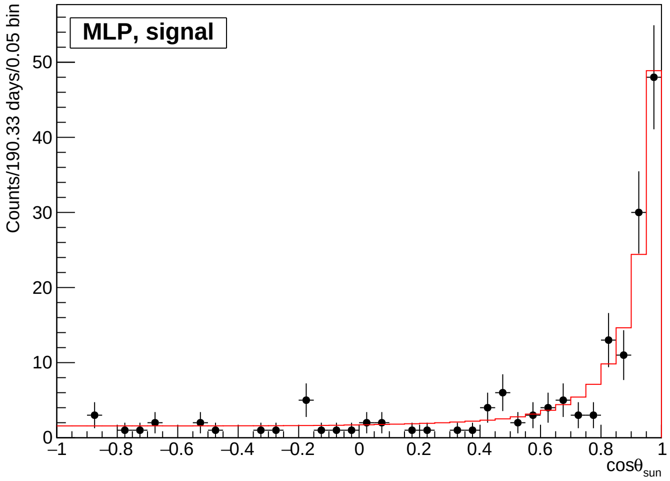

Figure 18: Fitted results for the $5 < E_{fit} < 15\ MeV$, from MLP outputs.

3.4.4. Systematics Evaluation

The systematics of position, direction and energy reconstruction were obtained from the Chapter 5. The quantities of position scale (XYZ_{scale}), position resolution (XYZ_{resol}), direction resolution (Dir_{resol}), energy scale

Table 7: Fitted results for the whole dataset ($5 < E < 15\ MeV$).

Methods	N_{sig}	N_{bkg}	R_{sig}	R_{bkg}	p-value
BDT	119.40 ± 11.84	36.55 ± 7.57	0.90 ± 0.09	0.28 ± 0.06	0.07
MLP	119.49 ± 11.89	40.51 ± 7.91	0.90 ± 0.09	0.31 ± 0.06	0.20

(E_{scale}) and energy resolution (E_{resol}) were used. Table. 8 summarizes these quantities used for this analysis.

Table 8: Systematics for the solar ν_e analysis in the water phase.

Systematics	values (positive/negative)
x shift	+6.48/-5.98 mm
y shift	+6.22/-4.06 mm
z shift	+6.62/-4.8 mm
x scale	+0.04%/-0.06%
y scale	+0.03%/-0.09%
z scale	+0.03%/-0.01%
δ_θ	+0.097/-0.008
E_{scale}	1.12%
E_{resol}	0.02%
β_{14} shifts	+0.010/-0.036

For the energy scale, the reconstructed MC energies were scaled up and down the expected number of signal events, the uncertainty on the expected number of events was obtained to be $^{+0.05}_{-0.06}$.

A remapping of $\cos\theta_{sun}$ by using the equation **??** was applied on the spectrum for evaluating the systematics. Its impact is $^{+0.3}_{-0.2}$ events.

3.4.5. Extracting the Solar Neutrino Flux

To evaluate the ^8B solar neutrino flux, the fitted number of $\nu - e^-$ elastic scattering (ES) is divided by the expected number of ES and is then multiplied by the flux in the MC:

$$\Phi_{^8B} = \Phi_{^8B,MC}\frac{N_{fit}}{N_{expected}}, \tag{3.8}$$

In the MC, the ν_e is generated by 1700 times the nominal; while the ν_μ is generated by 9600 times the nominal.

As mentioned in the last section, since the BDT selection gives a smaller p-value for the fitting, here the BDT outputs are used for the flux estimation.

In addition to the fitting of the whole energy region of $[5, 15]MeV$, different energy bins with 1 MeV step were also fitted. Fig. 19 to Fig. 24 show the results.

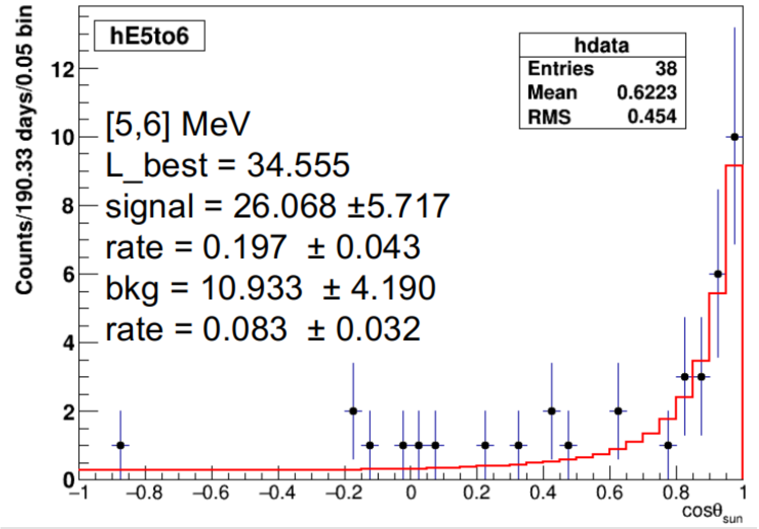

Figure 19: Fit with $5 < E < 6$ MeV.

Table. 9 shows the numbers of the MC generated ν_e and ν_μ after scaling by 1700 or 9600 as well as the BDT selections.

Table 9: Expected number of solar ν_e, ν_μ and the fitted number of signals in each energy bins.

E (MeV)	$N_{expected}(\nu_e)$	$N_{expected}(\nu_\mu)$	$N_{fit,sig}$
$[5, 15]$	336	53	119.40±11.84
$[5, 6]$	68	11	26.07±5.72
$[6, 7]$	91	14	36.52±6.70
$[7, 8]$	68	11	20.87±4.95
$[8, 9]$	47	7	16.24±4.30
$[9, 10]$	30	5	11.50±3.56
$[10, 15]$	32	5	9.09±3.06

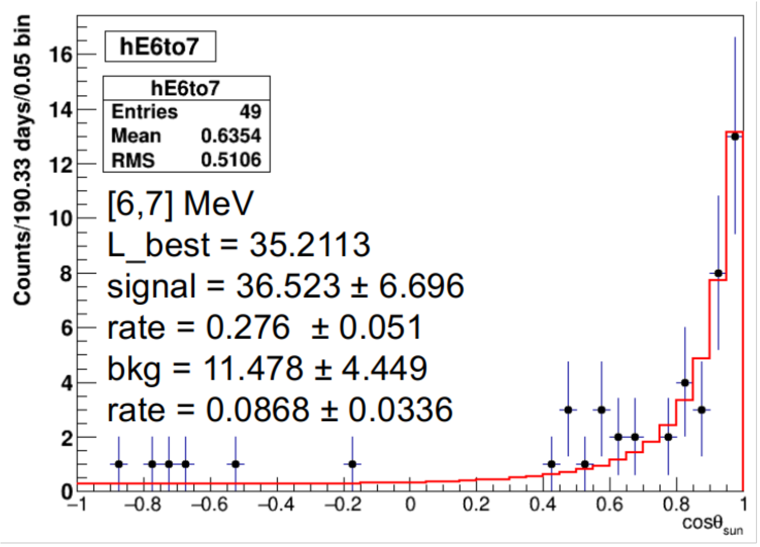

Figure 20: Fit with $6 < E < 7$ MeV.

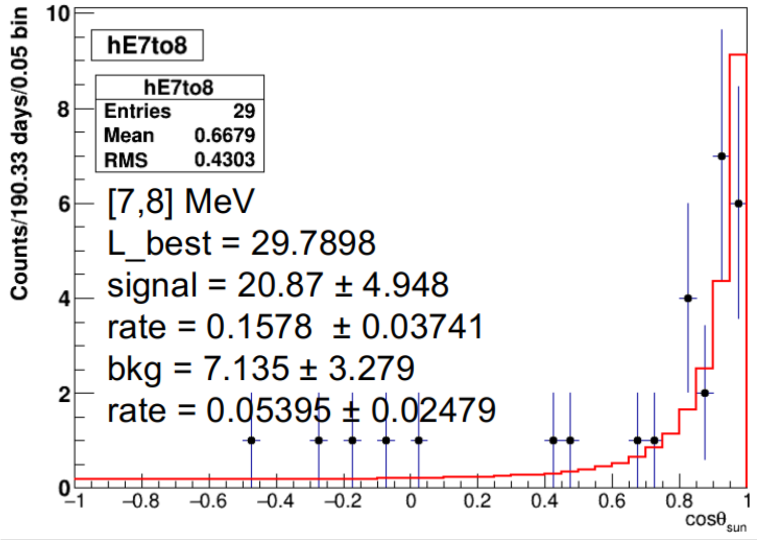

Figure 21: Fit with $7 < E < 8$ MeV.

For a nominal ^8B solar neutrino flux $\Phi_{MC} = 5.46 \times 10^6\ cm^{-2}s^{-1}$ (was 5.69), the flux is estimated to be $\Phi_{ES} = \Phi_{MC} \cdot N_{fit,sig}/336 = 1.94 \pm 0.19(stat.) \times 10^6\ cm^{-2}s^{-1}$, compared to the Super-K measurement:

$$\Phi_{ES} = (2.345 \pm 0.039) \times 10^6\ cm^{-2}s^{-1}, \tag{3.9}$$

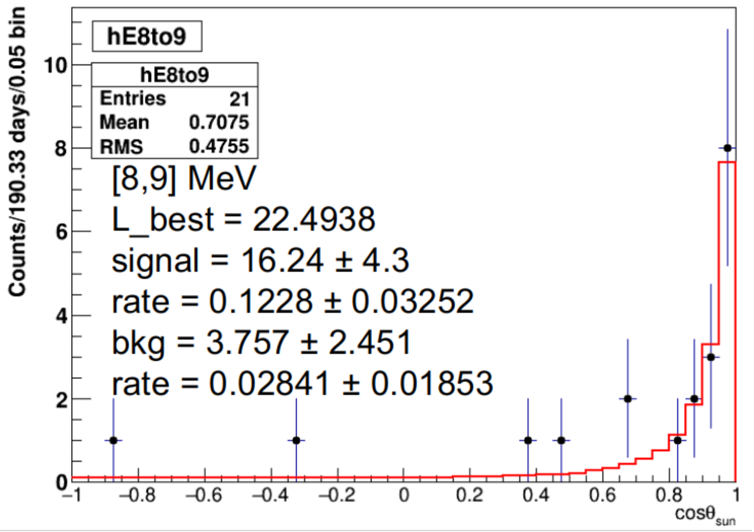

Figure 22: Fit with $8 < E < 9$ MeV.

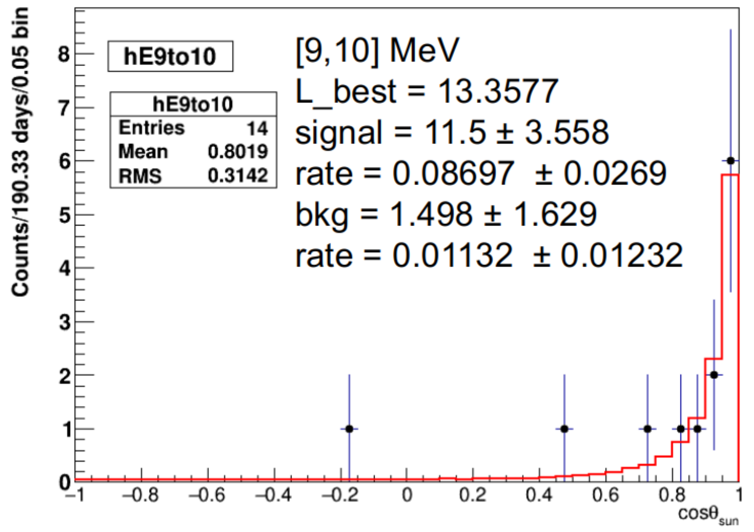

Figure 23: Fit with $9 < E < 10$ MeV.

and the SNO+ 2018 published results:

$$\Phi_{ES} = 2.53^{+0.31}_{-0.28}(stat.)^{+0.13}_{-0.10}(syst.) \times 10^{6}\ cm^{-2}s^{-1}. \qquad (3.10)$$

Fig. 25 shows the flux as a function of energies. A P_{ee} curve obtained from the `RAT` was overlay with the spectrum.

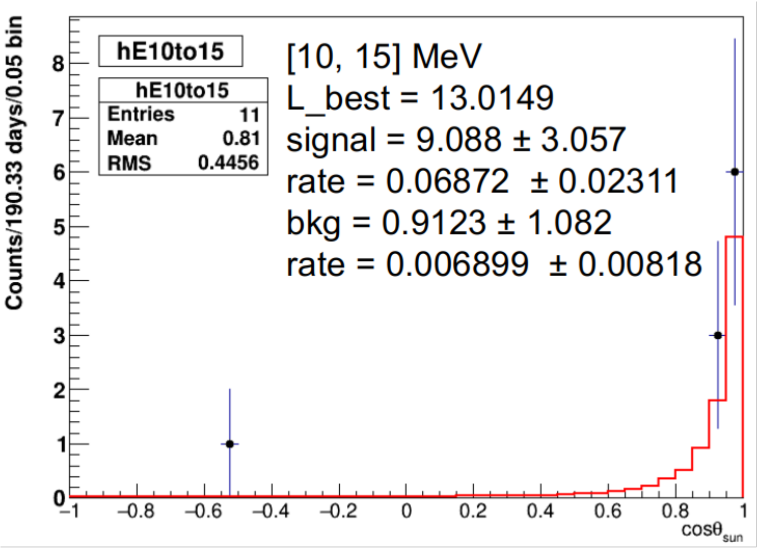

Figure 24: Fit with $10 < E < 15$ MeV.

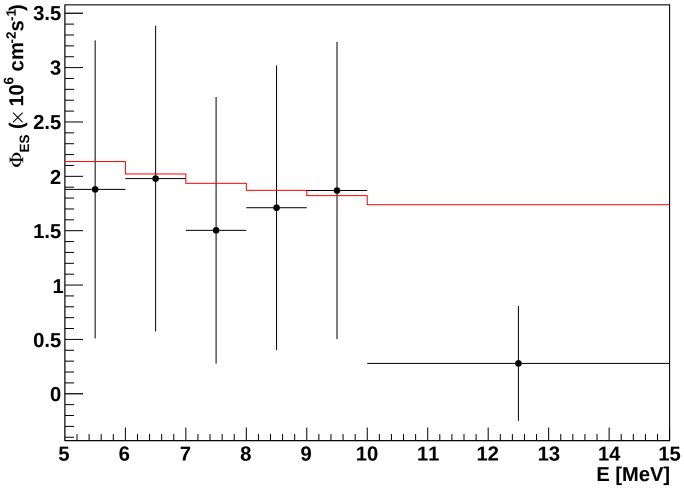

Figure 25: ^8B solar neutrino flux as a function of energies. The P_{ee} curve obtained from the `RAT` is in red.

Taking the systematics into accounts,

$$\Phi_{^8B,sys} = \Phi_{^8B,MC}\frac{N_{fit}}{N_{expected}}\sqrt{(\frac{N_{expected,sys}}{N_{expected}})^2 + (\frac{N_{fit,sys}}{N_{fit}})^2}, \tag{3.11}$$

3.4.6. Limits of this Study

Here I used the background types descried in the Table. 3. However, there are a few other backgrounds, such as the backgrounds from the AV ropes, the PMTs and the cosmogenic ones mainly caused by the cosmic muons. A more comprehensive study requires to include all possible backgrounds.

For the background events, I assumed a flat distribution of $\cos\theta_{sun}$. A more realistic shape of the distribution can be investigated to describe the backgrounds more properly.