Peer Review File

Manuscript Title: Electron Beam Energy Reconstruction for Neutrino Oscillation Measurements

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

Or Hen e4nu Nature Submission Review
==

This is a fine analysis with an important result. Accelerator neutrino beams are far from monochromatic. Therefore it is of prime importance to understand the extent to which one can accurately reconstruct the neutrino energy from the reaction products. The authors do a thorough job by mining an old CLAS data set, comparing the results to a well understood and popular neutrino event simulation, and then extrapolating to the specific effect on determining neutrino oscillation parameters as expected in DUNE. The case is made clearly that such effects need to be taken into account, and that more work needs to be done in order to avoid dominant systematic error in the eventual DUNE results.

Of course, this is a not a surprising conclusion. Indeed, one of the co-authors (KM) has devoted much attention, and has produced important results, bearing on this point for a number of years. Nevertheless, the idea to use CLAS data to study the issue directly (with a lepton beam of known, fixed energy) is an important step forward.

The paper appears to be technically correct and the research well documented. The wide arrow indicating "Ebeam" in Fig.2. appears in fact to be to the right of 1.159 GeV. I would have liked to see more discussion regarding effects due to the non-existent axial vector interaction in electron scattering. (Presumably they could have shown what happens if you turn off this interaction in GENIE?) Figures 3 and 4 make it clear that although GENIE does a very good job predicting electron scattering results, a discussion regarding how it might be improved would be useful, including implications for how it might effect conclusions drawn from something like Fig.1.

The text and figures in the "Methods" and "Extended Data" sections are both concise and thorough. They provide a very good summary of how the work was carried out for people interested in the details. I think that, in Extended Data Fig.2., it should be more clearly pointed out that the Ar(e,e') data was taken by a different collaboration and already published elsewhere.

This is a peripheral comment, but why is the right side of Fig.1 plotted with both scales in degrees? In addition to using instead -pi to +pi for delta, isn't it more common to plot sin^2(2theta13) instead of the angle itself? It would also be useful to include the current range of accepted values for theta13 from Daya Bay and other reactor experiments.

Overall, although this is sound and important work, I have reservations about recommending publication in Nature. There is no particular discovery presented here, but rather a solid verification and quantification of something about which people have been aware for quite some time. I would have guessed that, instead, a detailed manuscript on Physical Review C or D would be more appropriate, documenting the details of the analysis and outlining the path forward so that the accelerator neutrino community can be suitably prepared by the time DUNE (or HyperK) takes data.

Referee #2 (Remarks to the Author):

A. Summary of the key results
In this manuscript, the authors use the electron scattering data with known beam energies from the CLAS experiment at the Jefferson Lab to test incident electron energy reconstruction methods. They demonstrated that there could be large uncertainties in the modeling of reconstructed electron energy spectrum even for the simple event topology. The same uncertainties are likely to also cause a bias in the reconstructed neutrino energy due to the similarity of electron- and neutrino-nucleus interactions. The bias in the neutrino energy reconstruction can lead to completely wrong answers in the measurement of CP violating phase, which is the primary physics goal of the future neutrino experiments such as DUNE and Hyper-K. Therefore, the authors conclude it is important for the neutrino generators to describe reality as accurately as possible.

The manuscript is well written, and the conclusion is important for the planning of the future neutrino experiments DUNE and Hype-K. The results are of immediate interest to the neutrino community, especially theorists working on the neutrino generators and experimentalists working on the event reconstruction and oscillation analysis. However, several improvements are needed to strength the conclusion of this manuscript. In particular, the discussion of the systematic uncertainties needs to be expanded to help the readers understand sources of the data/simulation discrepancies in the energy reconstruction. There could be detector effects as well as physics effects. The test beam measurements from detectors such as ProtoDUNE can greatly help understand the detector response. More work is needed to improve the neutrino event generators, which is the main point of this manuscript. The authors should also comment on whether the planned DUNE-PRISM near detector would address issues brought up in this manuscript.

B. Originality and significance: if not novel, please include reference
This manuscript has three important points. First, it is crucial to use the electron scattering data to validate and improve the neutrino generator. This approach has been adopted by most of the leading neutrino generators, such as GENIE (arXiv:2006.11944), NuWro (arXiv:1510.03268), and GiBUU (a list of references in arXiv:1904.11506). Secondly, the bias in the neutrino energy reconstruction can significantly affect the measurement of CP violating phase. A similar conclusion was reached in arXiv:1507.08561 (Ref [11]). The third point is the main originality of this paper, which is to use electron scattering data to demonstrate the large uncertainty in the neutrino energy reconstruction, which has a detrimental effect on the physics measurements in the future flagship neutrino experiments. This is an important conclusion, which emphasizes the urgency to improve the neutrino generators and lends support to the importance of the near detector measurements.

C. Data & methodology: validity of approach, quality of data, quality of presentation
The approach is solid. The authors selected electron scattering events with simple topology: only one electron and zero pions or photons above threshold. They also examined a subset of events with exactly one detected proton. They considered two common methods to reconstruct the incident electron energy, one using the outgoing electron momentum and angle with the assumption of quasi-elastic scattering and the other using the calorimetry information. The former is relevant to Hyper-K and the latter is relevant to DUNE. Using the selected events, the authors clearly demonstrated two issues. First, less than 50% of the events have the correctly reconstructed energy. Secondly and more importantly, simulated reconstructed energy using the GENIE generator does not match data. Using the data- and simulation-based energy smearing matrix in the oscillation fit will change the best fit value of CP phase significantly.

The quality of data looks good. However, a quite noticeable feature in this manuscript is the data/simulation discrepancy. Of course, the purpose of this manuscript is to demonstrate the consequence of such mismodelings, but the reader may still wonder: 1) What are the causes of the discrepancy? There are very few discussions on this point in the manuscript. 2) Can the GENIE generator be improved? For example, Extended Data Fig. 2 shows the data/GENIE comparison of inclusive electron scattering cross section at different energies. The agreement in general looks OK, but at 0.961 GeV for the C target, both the QE and RES peaks deviate from data. At 2.222 GeV, there is a large discrepancy at high energy transfer, which is not discussed. Is this caused by the tension in tuning against electron scattering vs. neutrino scattering data? Another example is Fig 3, where according to the text, GENIE predicts the peak position to be about 20 MeV too low at all three energies. Is this bias due to detector calibration or physics effects in GENIE? The former can be easily improved and is less relevant to the message this manuscript is trying to convey.

The presentation is good in general. Fig. 1 is a very nice illustration of the challenge in reconstructing neutrino energy and the impact on the neutrino oscillation parameter measurements. The discussion of incident energy reconstruction and its impact on the neutrino oscillation analysis is clear. A few issues:
1. "the unoscillated flux Phi(E,0) is well known from hadronic calculations." I think in general there is a 10-20% uncertainty on the flux prediction so this is not well known especially since DUNE requires 1% systematic uncertainty.
2. "… experiments also typically select simpler charge-current event "topologies", such as events with one detected muon, one proton and zero pions": This is only true for experiments with low energy beams, such as T2K. Other experiments with higher energy beams need to include events with more complicated topologies such as MINOS, NOvA and DUNE.
3. In section V, the authors should state explicitly which dataset was used, C or Fe.
4. I don't understand the statement "While DUNE is overall similar to CLAS". They are very different detectors.
5. The quality of Extended Data Fig.9. seems lower compared with other figures.

D. Appropriate use of statistics and treatment of uncertainties
The main statistical analysis in this work is to perform the neutrino oscillation fit to derive the oscillation parameters. The authors used the standard GLoBES package for this analysis and the details are properly discussed in Methods. Regarding the data and simulation comparisons, there is no quantification of chi^2 or p-values, even though by-eye most of data/simulation agreement is quite poor.

The statistical uncertainty is included in all the data plots. The major systematic effects are discussed in Methods, which are quite small. However, there is no discussion of the detector calibration, event reconstruction and related uncertainties. It is hard to judge whether the discrepancies seen in many distributions are caused by instrumentation effects or real physics mismodeling in the generator. For example, simulation misses the peak of all the 1p0pi Ecal distributions by 20 MeV. This may indicate an issue in the proton momentum reconstruction.

The uncertainties are not treated in a consistent way in the figures. Some figures include both statistical and systematic errors, some only include statistical errors, and some do not specify what errors are included. Also, the end of section I says the total systematic error is 5% at 4.453 GeV while the caption of Fig.3. says the systematic error is 4% at 4.453. Those numbers need to be checked and made consistent. Some of the simulation curves have wiggles, which are probably due to the low statistics. It would be good to either increase the simulation statistics or add statistical uncertainty on the curves for the total prediction.

E. Conclusions: robustness, validity, reliability
The conclusions are sound and can be strengthened by expanding the discussion on systematic uncertainty on the detector effects and physics effects. The authors should mention the DUNE-PRISM near detector which was proposed to address the exact issues raised in this manuscript by providing a data-driven determination of the relationship between true and reconstructed energy that is significantly less sensitive to the neutrino interaction models.

F. Suggested improvements: experiments, data for possible revision
More information should be given on the data/simulation discrepancies as mentioned above focusing on separating detector effects from physics effects.

G. References: appropriate credit to previous work?
Other papers on using electron scattering data to constrain and improve the neutrino generators should be cited (e.g. arXiv:2006.11944, arXiv:1510.03268, arXiv:1904.11506).

MINERvA’s flux paper arXiv:1607.00704 should be added together with Ref. 15 and 16.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions
The abstract, introduction and conclusions are clear and well written.

Referee #3 (Remarks to the Author):

This paper describes a possible impact of a usage of imperfect neutrino-nucleus models for neutrino oscillation parameter measurements, especially the CP-violating phase.
Neutrino energy reconstruction is a long standing problem in the accelerator-based and atomspheric neutrino oscillation measurement. This paper, for the first time, demonstrates the energy reconstruction using electron scattering data.
The quality of electron-nucleus scattering data and their analysis, statistics and treatment of uncertainties look very solid. The obtained results for the energy reconstruction and transverse missing momenta are quite useful for the development of the neutrino-nucleus interaction model. However, evaluation of the impact to the CP-violating phase has been done for too simple case. The authors examine the case of the future experiment, DUNE, by assuming electron neutrino appearance data and concluded that an improper model could significantly bias the result. Actually, accelerator-based experiments are aiming to measure CP-violating by using not only electron appearance but also electron antineutrino appearance, muon neutrino disappearance and muon antineutrino disappearance. This would make measurements more robust. So the conclusion of this paper could be wrong.
Apart from these major comments, I list up parts which would need modification in the following.

Abstract has "We find that even in simple interactions where no pions are detected, only a small fraction of events reconstruct to the correct incident energy." and Methods "However, GENIE should describe these QE-like events more accurately than other, more complicated channels.". However, the bias of the energy reconstruction could be smaller for deep inelastic scattering at higher energy region. So are these sentences adequate?

- Ref[1][2] is one of review articles. Original articles should be cited. Or it should be noted that those are review articles.

- Ref[5] is not proper as a reference to two flavor oscillation formalism.

- In the paragraph below Eq.1, it should be noted that three flavor oscillation is necessary to introduce CP-violating phase.

- The last sentence of p.1, it should be added that Erec is also reconstructed from the observed energy, not by angle and momenta.

- Caption of Fig.1, there written "the model-derived smearing matrices". Actually, the model used is one of many models. So it should be "smearing matrices derived using a model"

- Ref.[9] is a detector paper and not adequate there.

- p.2 More correct wording is necessary for "flux Φ is well known from hadronic calculations" .

- Extended Data Fig.2 bottom shows good agreement between e-nucleus scattering weighted by 1/Q⁴ and nu-nucleus scattering. Dose this mean the contribution from the axial-vector current is negligible?

- p4 "This energy reconstruction discrepancy is observed despite the overall good reproduction of the differential inclusive cross-sections (Extended Data Fig. 2)". But there seems some discrepancy in Extended Data Fig. 2. How can you say that the energy reconstruction discrepancy is more significant than the inclusive cross section?

- Fig3 and Extended Data Fig.6. What is the meaning of 1/3 in the middle plots? Extend Data Fug.6 is made with additional cut. Then, why are the number of events larger than Fig.3?
- Sec.V Which sample was used, (e,e')_{0π} or (e,e'p)_{1p0π}?

- Sec.V 2nd paragraph. It should be noted which oscillation parameters were fit. I guess it is δ_{CP} and sin²θ₂₃, but it should be clarified in the text.

- reference. Some references have arXive, but some not. It should be consistent.

- Method. The first sentence "Eq.7" should be "Eq.1"

- Method. Experimental setup and particle identification. Figure of the setup is desired.

- Extended Data Fig.2 "ω" in the caption should be "Energy Transfer".

- Extended Data Fig.3 In caption, I guess blue should be red and black should be blue.

Referee #4 (Remarks to the Author):

The present manuscript is a notable piece of work on neutrino energy reconstruction methods and interaction simulations. Electron scattering data on a set of nuclei are collected for a fixed and known value of the beam energy. Using these data, the accuracy of energy reconstruction techniques routinely employed in neutrino experiments is tested.
The results presented in this manuscript show that the true incident energy is not reproduced at the quantitative level, implying that interaction models have to be improved to fully capitalize on planned neutrino oscillation experiments. This work proposes a novel and powerful strategy yielding to important findings. I believe that it deserves to be published in Nature, provided that the Authors address the following points.

1) At the end of the paragraph below Eq.(3) the text says "nuclear effects are irreducible and must be accounted for using theoretical models, typically implemented in neutrino event generators". I believe the Authors should include in this point some References to the models included in event generators (Nieves model, Local Fermi gas, Berger-Sehgal…).

2) In the first paragraph of Sec. I, the Authors explain why the carbon and iron data are relevant for future and present neutrino experiments. However, helium is only mentioned at the beginning of the Section and in Table I of the Extended data. The Authors should mention the reason why they decide to study this nucleus, too.

3) In Sec. III the Authors say that "the e-Genie peak is too narrow because of an inexact modeling of the nuclear ground-state momentum distribution." It would be useful for the reader to know which nuclear model is used to obtain these predictions, but I could not find this information in the text. Since these details are reported in Methods in the GENIE Simulations discussion, the Authors should either refer the reader to that part or include some of the information on the interaction models they adopted in the main text. In the same paragraph the text says: " This energy reconstruction discrepancy is observed despite the overall good reproduction of the differential inclusive cross-sections". I do not fully agree with this comment. Looking at Fig. 2 of the Extended Data, the agreement is not exactly 'overall good': there are some visible discrepancies between predictions and data in the high energy transfer region where the main contributions are RES and DIS. The modeling of these reaction mechanisms also fails to accurately reproduce the data in the low reconstructed-energy region in Fig. 2. The contribution

of MEC in Fig.2 is very small and might also contribute to the lack of strength at around 0.9 GeV. If there is a simple argument to explain why the contribution of MEC is almost negligible in Fig. 2 while it is sizable for the kinematic presented in Extended Data Fig.2, this should be included.

The Authors should acknowledge that models able to provide accurate predictions for inclusive observables are not necessarily that accurate in predicting exclusive ones. This point is discussed, for example, in Phys. Rev. C 100 (2019) 4, 044620.

4) In the right panel of Fig 4 it would be useful to separate the contributions given by QE, MEC, RES and DIS. My impression is that for P_T > 400 MeV—where the discrepancies with data are more relevant—the dominant reaction might be the RES. For the RES and DIS contribution, additional reaction mechanisms besides the interaction model, such as pion absorption entering the GENIE nuclear cascade, might play a relevant role. I would appreciate if the Authors could comment on this. Could the excess in the GENIE pion multiplicities with respect to data —Extended Data Fig. 4—indicate the presence of some inaccuracies in the way pion absorption is modeled?

5) In the GENIE Simulations discussion: the Authors say that a Local Fermi gas is used in the e-Genie simulation. However, from Fig. Extended Data Fig. 2 in the quasi-elastic region the theoretical curve appears to exhibit a tail that extends to intermediate values of the energy transfer. If that is the case, could you please explain how this tail arises?
It is not clear from the text which model has been used for the treatment of DIS in both e-Genie and nu-GENIE. As a remark, for future analyses, I would recommend a more realistic treatment of MEC than the one presented in this work. For example, using the SuSav2-MEC approach — computed for both electron and neutrino interactions within the global Fermi gas model — would provide a more consistent treatment with the Local Fermi gas adopted to describe the quasi-elastic region.

6) The analysis of the energy reconstruction performances of the GENIE event generator carried out in this work is very valuable for present and future neutrino-oscillation experiments. However, more dedicated measurements would be beneficial to accurately address DUNE and Hyper-Kamiokande. In particular, iron contains 16 additional nucleons and, as opposed to argon is a symmetric nucleus. This could introduce differences both on the accuracy of the theoretical models adopted for the interaction vertex and the nuclear cascade. Final state interactions are likely to be more relevant in iron than in argon. None of the explored energy regions seem to exactly match the one of interest for Hyper-Kamiokande. I would appreciate if the authors could address these points more in detail when stating their conclusions. The text says: "Future experiments with the improved CLAS12 spectrometer will extend these measurements to more nuclei, much smaller scattering angles, and to a wider range of momentum transfers." Will these future measurements include argon and oxygen? Will you take data for lower energies to better understand the kinematics relevant for Hyper-Kamiokande?

Author Rebuttals to Initial Comments:

Author response to referee comments on "Electron Beam Energy Reconstruction for Neutrino Oscillation Measurements"

We thank the referees for their careful review of our manuscript and constructive suggestions. We revised the manuscript accordingly. The main changes we introduced in response to the referee comments are:

1. We now report on absolute cross-sections, as oppose to event shape distributions (including additional single transverse variables).
2. A separate technical paper was published presenting in detail the electron GENIE event generator (eGENIE) developed for this work and discussing the commonalities between electron and neutrino scattering simulations [Papadopoulou et al., PRD 2021. arXiv:2009.07228].
3. The SuSAv2 MEC and QE models are now implemented in eGENIE and used in our study.

4. The paper is more general by removing the specific comparisons to DUNE θ_{CP} extraction.
5. The Methods section was extended to provide more details on the data analysis and detector calibration.

In addition, we addressed the individual points raised by the referees as detailed below.

We hope the referees will find the revised manuscript suitable for publication in Nature,

The e4ν and CLAS collaborations.

Referee #1:
This is a fine analysis with an important result. Accelerator neutrino beams are far from monochromatic. Therefore it is of prime importance to understand the extent to which one can accurately reconstruct the neutrino energy from the reaction products. The authors do a thorough job by mining an old CLAS data set, comparing the results to a well understood and popular neutrino event simulation, and then extrapolating to the specific effect on determining neutrino oscillation parameters as expected in DUNE. The case is made clearly that such effects need to be taken into account, and that more work needs to be done in orderto avoid dominant systematic error in the eventual DUNE results.

Of course, this is a not a surprising conclusion. Indeed, one of the co-authors (KM) has devoted much attention, and has produced important results, bearing on this point for a number of years. Nevertheless, the idea to use CLAS data to study the issue directly (with a lepton beam of known, fixed energy) is an important step forward.

We thank the referee for the kind words and for noting the importance of the new CLAS data.

The paper appears to be technically correct and the research well documented. The wide arrow indicating "Ebeam" in Fig.2. appears in fact to be to the right of 1.159 GeV. I would have liked to see more discussion regarding effects due to the non-existent axial vector interaction in electron scattering. (Presumably they could have shown what happens if you turn off this interaction in GENIE?) Figures 3 and 4 make it clear that although GENIE does a very good job predicting electron scattering results, a discussion regarding how it might be improved would be useful, including implications for how it might effect conclusions drawn from something like Fig.1.

We now discuss the similarities and differences of ν-A and e-A cross sections in the methodssection and the importance of constraining models of the vector part of the cross section. We also now compare the GENIE “tune” (i.e., set of models) previously used in this paper with the more- recent, theory-inspired, SuSAv2 superscaling QE and MEC models, and show how our semi- exclusive data can better discriminate between the models than the inclusive data. We also show that the GENIE resonance and DIS models (which are common to both electron and neutrino versions) overpredict the cross section by a factor of two for the 4.453 GeV data.

The text and figures in the "Methods" and "Extended Data" sections are both concise and thorough. They provide a very good summary of how the work was carried out for people interested in the details. I think that, in Extended Data Fig.2., it should be more clearly pointed out that the Ar(e,e') data was taken by a different collaboration and already published elsewhere.

The inclusive data previously shown in Extended Data Fig 2 is now part of a separate publication focusing on the electron version of GENIE and inclusive scattering, see arXiv:2009.07228 [Papadopoulou et al., PRD 2021]. We also added more references to previous works andcomplementary efforts.

This is a peripheral comment, but why is the right side of Fig.1 plotted with both scales in degrees? In addition to using instead -pi to +pi for delta, isn't it more common to plot sin^2(2theta13) instead of the angle itself? It would also be useful to include the current range of accepted values for theta13 from Daya Bay and other reactor experiments.
This figure has been replaced.

Overall, although this is sound and important work, I have reservations about recommending publication in Nature. There is no particular discovery presented here, but rather a solid verification and quantificationof something about which people have been aware for quite some time. I would have guessed that, instead,a detailed manuscript on Physical Review C or D would be more appropriate, documenting the details of the analysis and outlining the path forward so that the accelerator neutrino community can be suitably prepared by the time DUNE (or HyperK) takes data.

We agree with the referee that the reliability of neutrino event generators is a known issue. The current manuscript describes the first experiment to directly test and quantify the most crucial predictions of modern generators with high-precision mono-energetic electron beams on several nuclei and beam energies.

The use of mono-energetic electron beams with a large acceptance detector and a ‘neutrino equivalent’ events selection is done here for the first time and allows testing how well event generators describe event-by-event energy reconstruction, which is critical for reconstructing the incident neutrino flux and hence for reconstructing the oscillation parameters. Previous experiments compared event generators to data cross sections; this is the first to directly test energyreconstruction with high precision.

This is also the first experiment to test neutrino event generators with semi-exclusive electron scattering, thus testing both their inclusive cross section predictions and their previously untested (and poorly constrained) hadronization routines. Neutrino event generators have been extensivelytested against inclusive electron scattering; testing with semi-exclusive electron scattering is necessary for improving the models in the generators. For example, one significant source of uncertainty is our knowledge of the average nuclear binding energy [T2K PRL 121 (2018) 171802].We measured the average nucleon removal energy for the 1p0pi channel.

We have also added the SuSAv2 model, to show that semi-exclusive electron scattering data can also help model builders by discriminating among different models. Thus, our data can help the neutrino community better prepare for next generation experiments.

Thus, we think that the current manuscript is more than just an incremental progress over previousstudies. It is in fact ushering in a new era of an international joint electron and neutrino interactionstudies program to facilitate the highest accuracy for precision neutrino oscillation studies.

Referee #2:

A. Summary of the key results

In this manuscript, the authors use the electron scattering data with known beam energies from the CLAS experiment at the Jefferson Lab to test incident electron energy reconstruction methods. They demonstrated that there could be large uncertainties in the modeling of reconstructed electron energy spectrum even for the simple event topology. The same uncertainties are likely to also cause a bias in the reconstructed neutrino energy due to the similarity of electron- and neutrino-nucleus interactions. The bias in the neutrinoenergy reconstruction can lead to completely wrong answers in the measurement of CP violating phase, which is the primary physics goal of the future neutrino experiments such as DUNE and Hyper-K. Therefore, the authors conclude it is important for the neutrino generators to describe reality as accurately as possible.

The manuscript is well written, and the conclusion is important for the planning of the future neutrino experiments DUNE and Hype-K. The results are of immediate interest to the neutrino community, especially theorists working on the neutrino generators and experimentalists working on the event reconstruction and oscillation analysis. However, several improvements are needed to strength the conclusion of this manuscript. In particular, the discussion of the systematic uncertainties needs to be expanded to help the readers understand sources of the data/simulation discrepancies in the energy reconstruction. There could be detector effects as well as physics effects. The test beam measurements fromdetectors such as ProtoDUNE can greatly help understand the detector response. More work is needed to improve the neutrino event generators, which is the main point of this manuscript. The authors should alsocomment on whether the planned DUNE-PRISM near detector would address issues brought up in this manuscript.

We now present the data as cross sections (rather than area-normalized event yields) for more quantitative comparisons with GENIE predictions. As a result, we now discuss the systematic uncertainties in more detail. This discussion includes using the six-fold sector symmetry of CLASto calculate the systematic uncertainty due to detector effects. We use the variance of the data to GENIE ratio for the different CLAS sectors to calculate our point-to-point systematic uncertainty due to detector effects. This is explained in the Methods section.

DUNE Prism is an excellent and challenging idea. However, it is quite a few years away. Nonetheless, we now mention it in the summary and explain that it will nicely complement the electron effort, both in terms of the sensitivity to Axial currents and the very different systematics.Since the quasi-monochromatic beams of DUNE Prism will be produced by subtracting the resultsfrom two wider beams, they will be constraining small numbers by subtracting larger ones. This will have systematic challenges, like those seen by the electromagnetic community when experimentalists made quasi-monochromatic photon beams by subtracting the bremsstrahlung spectra from two narrowly separated electron beams (e.g., 4.1 and 4.0 GeV beams). Therefore, wethink it will be very valuable to combine information from both the prism and the mono-energeticapproaches.

B. Originality and significance: if not novel, please include reference

This manuscript has three important points. First, it is crucial to use the electron scattering data to validate and improve the neutrino generator. This approach has been adopted by most of the leading neutrino generators, such as GENIE (arXiv:2006.11944), NuWro (arXiv:1510.03268), and GiBUU (a list of references in arXiv:1904.11506). Secondly, the bias in the neutrino energy reconstruction can significantly affect the measurement of CP violating phase. A similar conclusion was reached in arXiv:1507.08561 (Ref [11]). The third point is the main originality of this paper, which is to use electron scattering data to demonstrate the large uncertainty in the neutrino energy reconstruction, which has a detrimental effect on the physics measurements in the future flagship neutrino experiments. This is an important conclusion, which emphasizes the urgency to improve the neutrino generators and lends support to the importance of the near detector measurements.

While electron data has been used to validate and improve event generators, to date mostly inclusivedata has been used and no wide phase-space data with similar selection to that of neutrino experiments was available. Therefore, existing data is not sufficient to adequately constrain neutrino event generators.

In addition to cross-sections differential in different variables (e.g., energy transfer), event generators must be able to accurately describe the energy reconstruction (i.e., the energy feed-downmatrices). We emphasize the need to use semi-exclusive data and we show, for the first time, the effect on neutrino energy reconstruction.

C. Data & methodology: validity of approach, quality of data, quality of presentation

The approach is solid. The authors selected electron scattering events with simple topology: only one electron and zero pions or photons above threshold. They also examined a subset of events with exactly one detected proton. They considered two common methods to reconstruct the incident electron energy, one using the outgoing electron momentum and angle with the assumption of quasi-elastic scattering and the other using the calorimetry information. The former is relevant to Hyper-K and the latter is relevant toDUNE. Using the selected events, the authors clearly demonstrated two issues. First, less than 50% of the events have the correctly reconstructed energy. Secondly and more importantly, simulated reconstructed energy using the GENIE generator does not match data. Using the data- and simulation-based energy smearing matrix in the oscillation fit will change the best fit value of CP phase significantly.

The quality of data looks good. However, a quite noticeable feature in this manuscript is the data/simulation discrepancy. Of course, the purpose of this manuscript is to demonstrate the consequence of such mismodelings, but the reader may still wonder: 1) What are the causes of the discrepancy? There are very few discussions on this point in the manuscript. 2) Can the GENIE generator be improved? For example, Extended Data Fig. 2 shows the data/GENIE comparison of inclusive electron scattering cross section at different energies. The agreement in general looks OK, but at 0.961 GeV for the C target, both the QE and RES peaks deviate from data. At 2.222 GeV, there is a large discrepancy at high energy transfer, which is not discussed. Is this caused by the tension in tuning against electron scattering vs. neutrino scattering data? Another example is Fig 3, where according to the text, GENIE predicts the peak position to be about 20 MeV too low at all three energies. Is this bias due to detector calibration or physics effects in GENIE? Theformer can be easily improved and is less relevant to the message this manuscript is trying to convey.

1) There is no single description of electron-nucleus scattering for all nuclei, incident energies, and outgoing lepton energies and angles. GENIE is a phenomenological semiclassical model, that adds the contributions of different reaction mechanisms at the cross-section level, rather than at the amplitude level. By implementing a second set of GENIE models (SuSAv2 QE and MEC), and by comparing cross sections rather than event yields, we show that the discrepancies are due to the

GENIE models. We now show specific places where specific aspects of the GENIE models are incorrect and can be improved.

2) GENIE can be improved by implementing new models of the different reaction mechanisms. For example, we helped implement the SuSAv2 QE and MEC models. We now show that our semi-exclusive data prefer the newer SuSAv2 models.
    a. The 20-MeV beam energy reconstruction error in the Rosenbluth QE model used in the G2018 tune is inherent to the model. We used the correct incident energy as an input parameter and reconstructed the incorrect energy. This does not happen in SuSAv2.
    b. We also now show that the DIS and RES models progressively overestimate the semi-exclusive data as the beam energy increases. This discrepancy between our semi-exclusive data and GENIE is of the same magnitude as the discrepancy between the inclusive 2.22 GeV data and GENIE (previously shown in Extended data Fig 2 and now shown in a separate publication). These all point to specific problems in GENIE that can now be fixed.
3) We now quantify our detector calibration more thoroughly by comparing to the known measured H(e,e'p) cross section and by more systematic uncertainty checks. These are discussed in the Methods section.
4) The quality of the CLAS energy calibration is examined using exclusive reactions such as H(e,e'pi+) and ^{3}He(e,e'pp)n etc. We now discuss it in the methods section and show relevant distributions in Extended Data Fig. 9.

The presentation is good in general. Fig. 1 is a very nice illustration of the challenge in reconstructing neutrino energy and the impact on the neutrino oscillation parameter measurements. The discussion of incident energy reconstruction and its impact on the neutrino oscillation analysis is clear.

A few issues:

1. "the unoscillated flux Phi(E,0) is well known from hadronic calculations." I think in general there is a 10-20% uncertainty on the flux prediction so this is not well known especially since DUNE requires 1% systematic uncertainty.

We now write "Phi(E,0) is calculated from hadronic reaction rates", which makes no claim about the accuracy.

2. "… experiments also typically select simpler charge-current event "topologies", such as events with one detected muon, one proton and zero pions": This is only true for experiments with low energy beams, such as T2K. Other experiments with higher energy beams need to include events with more complicated topologies such as MINOS, NOvA and DUNE.

We now write "experiments can select".

3. In section V, the authors should state explicitly which dataset was used, C or Fe.

We explicitly state that we used the C data set. (This is now section IV.)

4. I don't understand the statement "While DUNE is overall similar to CLAS". They are very different detectors.

The statement is now removed since we no longer focus on the DUNE parameter reconstruction. Its original intent was to say that while based on very different technology, both large acceptance detectors with similar particle detection thresholds with an ability to measure many particles in thefinal state.

5. The quality of Extended Data Fig. 9. seems lower compared with other figures.

This figure has been remade.

D. Appropriate use of statistics and treatment of uncertainties

The main statistical analysis in this work is to perform the neutrino oscillation fit to derive the oscillation parameters. The authors used the standard GLoBES package for this analysis and the details are properly discussed in Methods. Regarding the data and simulation comparisons, there is no quantification of chi^2 or p-values, even though by-eye most of data/simulation agreement is quite poor. The statistical uncertaintyis included in all the data plots. The major systematic effects are discussed in Methods, which are quite small. However, there is no discussion of the detector calibration, event reconstruction and related uncertainties. It is hard to judge whether the discrepancies seen in many distributions are caused by instrumentation effects or real physics mismodeling in the generator. For example, simulation misses the peak of all the 1p0pi Ecal distributions by 20 MeV. This may indicate an issue in the proton momentum reconstruction.

We now discuss detector calibration more thoroughly. By including a second QE model in GENIEwe show that the 20 MeV peak error is due to mismodeling rather than detector effects. Because the reconstructed energy data and simulation differ so significantly, quantifying this huge disagreement (e.g., via including the chi2/dof) would add little information.

The uncertainties are not treated in a consistent way in the figures. Some figures include both statistical and systematic errors, some only include statistical errors, and some do not specify what errors are included. Also, the end of section I says the total systematic error is 5% at 4.453 GeV while the caption of Fig.3. says the systematic error is 4% at 4.453. Those numbers need to be checked and made consistent. Some of the simulation curves have wiggles, which are probably due to the low statistics. It would be good to either increase the simulation statistics or add statistical uncertainty on the curves for the total prediction.

All figures are now properly normalized, and the point-to-point systematic uncertainty is includedin all primary figures.

E. Conclusions: robustness, validity, reliability

The conclusions are sound and can be strengthened by expanding the discussion on systematic uncertainty on the detector effects and physics effects. The authors should mention the DUNE-PRISM near detector which was proposed to address the exact issues raised in this manuscript by providing a data-driven determination of the relationship between true and reconstructed energy that is significantly less sensitive to the neutrino interaction models.

We expanded the discussion of systematic uncertainties and physics effects. We now mention DUNE PRISM.

F. Suggested improvements: experiments, data for possible revision

More information should be given on the data/simulation discrepancies as mentioned above focusing on separating detector effects from physics effects.

This is now discussed at greater length in the paper and in the Methods section. Since we now compare two sets of physics models (for the QE and MEC mechanisms) and we now use the six relatively-independent sector results of CLAS, we can identify and separate the detector and physics effects far more robustly.

G. References: appropriate credit to previous work?

Other papers on using electron scattering data to constrain and improve the neutrino generators should be cited (e.g. arXiv:2006.11944, arXiv:1510.03268, arXiv:1904.11506). MINERvA's flux paper arXiv:1607.00704 should be added together with Ref. 15 and 16.

We added the reference to GENIE (e,e') comparisons arXiv:2006.11944 and the MINERvA flux paper. We did not add the others due to lack of space for references (we are limited to 30 in the main paper).

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

The abstract, introduction and conclusions are clear and well written.

Thank you.

Referee #3:

This paper describes a possible impact of a usage of imperfect neutrino-nucleus models for neutrino oscillation parameter measurements, especially the CP-violating phase. Neutrino energy reconstruction is a long standing problem in the accelerator-based and atomspheric neutrino oscillation measurement. This paper, for the first time, demonstrates the energy reconstruction using electron scattering data.

The quality of electron-nucleus scattering data and their analysis, statistics and treatment of uncertainties look very solid. The obtained results for the energy reconstruction and transverse missing momenta are quite useful for the development of the neutrino-nucleus interaction model. However, evaluation of the impact to the CP-violating phase has been done for too simple case. The authors examine the case of the future experiment, DUNE, by assuming electron neutrino appearance data and concluded that an impropermodel could significantly bias the result. Actually, accelerator-based experiments are aiming to measure CP-violating by using not only electron appearance but also electron antineutrino appearance, muon neutrino disappearance and muon antineutrino disappearance. This would make measurements more robust.So the conclusion of this paper could be wrong.

The focus of this paper is on comparing energy reconstruction from electron scattering data and from GENIE, and its general impact on accelerator-based neutrino oscillation experiments. Therefore, we have removed the partial DUNE parameter reconstruction study because, as the referee points out, it is a far more complicated issue.

In addition, we have improved the paper by adding SuSAv2, a new, theoretically-motivated, set ofQE and MEC GENIE models, and by extracting absolute cross sections and not just comparing distribution shapes. This allows us to discriminate among models and to show which pieces of themodels need improvement. We show that the SuSAv2 models are much better than the older GENIE models, but that they fail at the highest incident electron energy. We also show that while the DIS and RES models agree with the data at 1.1 GeV, they dramatically overestimate it at the higher incident energies.

Apart from these major comments, I list up parts which would need modification in the following.

Abstract has "We find that even in simple interactions where no pions are detected, only a small fraction of events reconstruct to the correct incident energy." and Methods "However, GENIE should describe these QE-like events more accurately than other, more complicated channels.". However, the bias of the energy reconstruction could be smaller for deep inelastic scattering at higher energy region. So are these sentences adequate?

The statement in the Methods section has been deleted, along with the rest of the DUNE-related discussion.

- Ref[1][2] is one of review articles. Original articles should be cited. Or it should be noted that those are review articles.

These are general statements, so we think it is appropriate to cite review articles. We now state thatthese are reviews.

- Ref[5] is not proper as a reference to two flavor oscillation formalism.

Thank you. Replaced with RMP 75, 345 (2003)

- In the paragraph below Eq.1, it should be noted that three flavor oscillation is necessary to introduce CP-violating phase.

Done. We moved the CP violation comment to the three-flavor oscillation discussion.

- The last sentence of p.1, it should be added that Erec is also reconstructed from the observed energy, notby angle and momenta.

We now write "Erec is reconstructed from the measured angles, momenta, and/or energies …"since lepton-only methods use angles and momenta and calorimetric methods use energies.

- Caption of Fig.1, there written "the model-derived smearing matrices". Actually, the model used is one of many models. So it should be "smearing matrices derived using a model"

This is implied by the term "model"

- Ref.[9] is a detector paper and not adequate there.

Thank you. We replaced it.

- p.2 More correct wording is necessary for "flux Φ is well known from hadronic calculations" .

We changed this to "relatively well known"

- Extended Data Fig.2 bottom shows good agreement between e-nucleus scattering weighted by 1/Q4 and nu-nucleus scattering. Dose this mean the contribution from the axial-vector current is negligible?

No, it means that the axial-vector current does not significantly change the shape of these inclusive distributions. The plots are area normalized, so they only compare shapes.

- p4 "This energy reconstruction discrepancy is observed despite the overall good reproduction of the differential inclusive cross-sections (Extended Data Fig. 2)". But there seems some discrepancy in Extended Data Fig. 2. How can you say that the energy reconstruction discrepancy is more significant than the inclusive cross section?

The agreement in the inclusive cross sections between data and GENIE (Extended Data Fig 2) is within about 10% (in the QE, dip and first resonance region). However, Data and GENIE disagreeby up to a factor of two in our semi-exclusive cross section (Fig 3). Thus, the semi-exclusive data is much more sensitive than the inclusive data to specific problems with the event generators.

- Fig3 and Extended Data Fig.6. What is the meaning of 1/3 in the middle plots? Extend Data Fug.6 is made with additional cut. Then, why are the number of events larger than Fig.3?

The factor of 1/3 in the middle plots (which referred to scaling the peak) has been removed.

- Sec.V Which sample was used, (e,e')0π or (e,e'p)1p0π?
- Sec.V 2nd paragraph. It should be noted which oscillation parameters were fit. I guess it is δCP and sin2θ23, but it should be clarified in the text.

Section V has been removed.

- reference. Some references have arXive, but some not. It should be consistent.

Fixed.

- Method. The first sentence "Eq.7" should be "Eq.1"

Fixed.

- Method. Experimental setup and particle identification. Figure of the setup is desired.

Done. See Extended Data Fig. 9

- Extended Data Fig.2 "ω" in the caption should be "Energy Transfer".

Done.

- Extended Data Fig.3 In caption, I guess blue should be red and black should be blue.

Fixed.

Referee #4:
The present manuscript is a notable piece of work on neutrino energy reconstruction methods and interaction simulations. Electron scattering data on a set of nuclei are collected for a fixed and known valueof the beam energy. Using these data, the accuracy of energy reconstruction techniques routinely employedin neutrino experiments is tested. The results presented in this manuscript show that the true incident energy is not reproduced at the quantitative level, implying that interaction models have to be improved to fully capitalize on planned neutrino oscillation experiments. This work proposes a novel and powerful strategy yielding to important findings. I believe that it deserves to be published in Nature, provided that the Authorsaddress the following points.

1) At the end of the paragraph below Eq.(3) the text says “nuclear effects are irreducible and must be accounted for using theoretical models, typically implemented in neutrino event generators”. I believe the

Authors should include in this point some References to the models included in event generators (Nieves model, Local Fermi gas, Berger-Sehgal…).

The models are discussed and referenced in the Methods section, since we are quite limited in references in the main text.

2) In the first paragraph of Sec. I, the Authors explain why the carbon and iron data are relevant for future and present neutrino experiments. However, helium is only mentioned at the beginning of the Section and in Table I of the Extended data. The Authors should mention the reason why they decide to study this nucleus, too.

The target nuclei were chosen for the original (1999) experiment. It took little more work to analyze4He in addition to C and Fe.

3) In Sec. III the Authors say that "the e-Genie peak is too narrow because of an inexact modeling of the nuclear ground-state momentum distribution." It would be useful for the reader to know which nuclear model is used to obtain these predictions, but I could not find this information in the text. Since these details are reported in Methods in the GENIE Simulations discussion, the Authors should either refer the reader to that part or include some of the information on the interaction models they adopted in the main text. In the same paragraph the text says: "This energy reconstruction discrepancy is observed despite the overall good reproduction of the differential inclusive cross-sections". I do not fully agree with this comment. Looking at Fig. 2 of the Extended Data, the agreement is not exactly 'overall good': there are some visible discrepancies between predictions and data in the high energy transfer region where the main contributions are RES and DIS.

The modeling of these reaction mechanisms also fails to accurately reproduce the data in the low reconstructed-energy region in Fig. 2. The contribution of MEC in Fig.2 is very small and might also contribute to the lack of strength at around 0.9 GeV. If there is a simple argument to explain why the contribution of MEC is almost negligible in Fig. 2 while it is sizable for the kinematic presented in Extended Data Fig.2, this should be included.

We agree that there are discrepancies between GENIE and data for both inclusive (e,e') data and our semi-exclusive (e,e'p) data. We now compare data and model cross sections (rather than areanormalizing) and we now include the SuSAv2 QE and MEC models for comparison with the previously used Local Fermi Gas/Rosenbluth QE model and the Empirical MEC model. The SuSAv2 MEC results are much closer to expected than the (negligible) Empirical MEC. The SuSAv2 QE width in Figure 2 is also closer to that of the data than the LFG/Rosenbluth QE model. Our discussion of the data/GENIE comparisons is now more detailed.

The Authors should acknowledge that models able to provide accurate predictions for inclusive observables are not necessarily that accurate in predicting exclusive ones. This point is discussed, for example, in Phys. Rev. C 100 (2019) 4, 044620.

We thank the referee for emphasizing this. This is exactly why we are making the first comparisonof semi-exclusive electron scattering data to the models. While the SuSAv2 and G2018 GENIE tunes both describe the inclusive data reasonably well, our semi-exclusive data clearly much prefersthe SuSAv2 models (except at 4.4 GeV).

4) In the right panel of Fig 4 it would be useful to separate the contributions given by QE, MEC, RES and DIS. My impression is that for P_T > 400 MeV—where the discrepancies with data are more relevant—the dominant reaction might be the RES. For the RES and DIS contribution, additional reaction mechanisms

besides the interaction model, such as pion absorption entering the GENIE nuclear cascade, might play a relevant role. I would appreciate if the Authors could comment on this. Could the excess in the GENIE pion multiplicities with respect to data - Extended Data Fig. 4 - indicate the presence of some inaccuracies in the way pion absorption is modeled?

We now show those separate contributions. For P_T > 400 MeV/c, the dominant contributions areQE (presumably due to FSI) and MEC at larger reconstructed energies and RES at smaller reconstructed energies. Some of the model overestimate could be due to pion absorption and nucleon rescattering in the GENIE nuclear cascade. We hope to measure pion absorption in futureexperimental analyses.

5) In the GENIE Simulations discussion: the Authors say that a Local Fermi gas is used in the e-Genie simulation. However, from Fig. Extended Data Fig. 2 in the quasi-elastic region the theoretical curve appears to exhibit a tail that extends to intermediate values of the energy transfer. If that is the case, could you please explain how this tail arises? It is not clear from the text which model has been used for the treatment of DIS in both e-Genie and nu-GENIE.

There is a very small tail on the Rosenbluth/LFG QE due to the variation of k_F with density in theLFG. The SuSAv2 results have a much larger tail (as expected for their momentum distribution). This figure is now part of a forthcoming Phys Rev D paper on eGENIE and inclusive scattering, see arxiv:2009.07228.

We now explain the DIS model in the Methods section in more detail.

As a remark, for future analyses, I would recommend a more realistic treatment of MEC than the one presented in this work. For example, using the SuSav2-MEC approach — computed for both electron and neutrino interactions within the global Fermi gas model — would provide a more consistent treatment with the Local Fermi gas adopted to describe the quasi-elastic region.

We thank the referee for their suggestion and completely agree that SuSav2 is more realistic than the Rosenbluth/LFG QE and Empirical MEC models available previously. We now include the SuSav2 QE and MEC models. The comparisons among data, the G2018 tune, and SuSAv2 are quite instructive.

6) The analysis of the energy reconstruction performances of the GENIE event generator carried out in this work is very valuable for present and future neutrino-oscillation experiments. However, more dedicated measurements would be beneficial to accurately address DUNE and Hyper-Kamiokande. In particular, iron contains 16 additional nucleons and, as opposed to argon is a symmetric nucleus. This could introduce differences both on the accuracy of the theoretical models adopted for the interaction vertex and the nuclear cascade. Final state interactions are likely to be more relevant in iron than in argon. None of the explored energy regions seem to exactly match the one of interest for Hyper-Kamiokande. I would appreciate if the authors could address these points more in detail when stating their conclusions. The text says: "Future experiments with the improved CLAS12 spectrometer will extend these measurements to more nuclei, much smaller scattering angles, and to a wider range of momentum transfers." Will these future measurements include argon and oxygen? Will you take data for lower energies to better understand the kinematics relevant for Hyper-Kamiokande?

As we now write in the Methods section:
Jefferson Lab Experiment E12-17-006, ``Electrons for Neutrinos: Addressing Critical Neutrino- Nucleus Issues" (scientific rating: A) will take more data on more targets with a greater kinematicalrange using the upgraded CLAS12 detector. The approved experiment includes measurements on

^{4}He, C, Ar and Sn with 1, 2, 4 and 6 GeV electron beams, as well as measurements on O with1 and 2 GeV electron beams. The 1 and 2 GeV measurements will be performed with a minimum electron scattering angle of 5°, compared to a minimum CLAS angle of about 20°. This will extend the measurements down to the much lower momentum transfers typical of some neutrino experiments. It will therefore allow comparison with the lower beam-energy data of T2Kand HyperK. The first phase of the experiment is scheduled to run in the second half of 2021.”

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I would like to thank the authors for their careful consideration of my (and the other reviewers') comments. My comments are sufficiently addressed to proceed with publication.

This result is a clear improvement over previous comparisons of electron scattering data and neutrino even generators. I suppose the question of whether or not it is "more than just an incremental progress" can be left to the eye of the beholder.

Referee #2 (Remarks to the Author):

I would like to thank the authors for addressing my comments and making changes to the paper accordingly, especially adding details on detector calibration and systematic analysis and discussing DUNE prism, which definitely make the paper stronger. In the new version, the authors removed the discussion of impact on neutrino oscillation parameters, added a new model prediction based on SuSa and expanded the discussion on transverse variables to include two more variables delta alpha_T and delta phi_T. All these changes make the paper more focused on discussing how to use the electron scattering data to improve modeling of neutrino interactions, with a focus on the issues in the energy reconstruction. I think this paper deserves to be published in Nature, provided that the authors address the following additional comments.

1. I noticed that some of the GENIE G2018 predictions changed in the new draft. For example, in the old draft figure 4 (P_T > 400 MeV/c), there are two peaks in the GENIE prediction (at 1 GeV and 1.7 GeV). In the new draft figure 4, there is only one peak at 1.7 GeV. What have changed?

2. Fig. Extended Data Fig. 2 is interesting, but the He point at 2.257 GeV on the right plot seems to be an outlier. Is there a reason why that point is so high?

3. There are discrepancies between data and predictions in many distributions. I still think it would be good to provide more information on the mismodeling to emphasize the advantage of using electron scattering data to improve the neutrino simulation.
a) e-GENIE using the older G2018 model predicts the peak position to be 10, 25 and 36 MeV too low. The authors provided an explanation in the response to my comments. It would be good to clarify it in the paper.
b) In figure 3, SuSAv2 either underestimates or overestimates the QE peak, is there an explanation for that?
c) In places where RES and DIS dominate, the theories always overestimate the cross section. Previous neutrino cross section measurements seem to suggest GENIE overpredicts resonance production in the neutrino-nucleus interactions (e.g. T2K arXiv:1909.03936, figures 13-16). This seems to be confirmed by the electron scattering study reported in this paper. What could be the source of this mismodeling?

4. I found the section “Subtraction of undetected pions and photons” hard to follow. The concept is clear but the details are confusing. I don’t understand what this sentence means: We used that as a weight to

subtract for the undetected pion events. Another example is the sentence: this subtraction from the subtracted events increased the yield, which sounds very cryptic. Maybe this is a standard method used in electron scattering experiments, but the description of the method is not very clear for a general reader.

Referee #3 (Remarks to the Author):

The weak points of the previous version have been well refined in the updated manuscript: Estimation of the impact on DUNE under too simple assumptions etc.
As I reported in the previous review, this paper, for the first time, demonstrates the energy reconstruction used in the accelerator-based neutrino oscillation measurement by using electron scattering data, shows interesting nuclear effects and will be very important for the progresses of the neutrino oscillation measurements. The data and methodology, treatment of uncertainties are well described and look solid.
I consider that this result is appropriate to be published in high-ranked journals dedicated to physics because it would have an important impact on the field of the elementary particle physics. On the other hand, I would reserve to recommend for publication in Nature because its content is about a technical methodology and does not contain facts that are common interest to people outside the field.

A few minor comments follow below:

Equation (2) is not exact because the flux is not same even without oscillation due to acceptance difference at near and far.

p.2 right column "There is no guarantee that they are suitable for analyzing far-detector data" is too strong statement.
The constraint by near-detector is stronger for narrower-energy beam. So it is a matter of degree.

p.3 Left column, "the model dramatically overestimated the amount of mis-reconstructed events"
This is true only for \`the' model introduced later and used in this paper.

p.3 Right column, isn't "interactions of the lepton with a meson exchanged between two nucleons" wrong?
This sounds like it is interactions with a meson, but actually isn't it interactions with 'two nucleons exchanging a meson'?

p.9 Right column, I cannot understand "Inclusive lepton-nucleus scattering has the same number of independent structure functions." Actually it is different (two for electron- and three for nuetrino-scattering) as written in the paper.

Referee #4 (Remarks to the Author):

I carefully went through the updated version of both the paper and the "Response to the Referee" that the authors provided. This version of the manuscript presents significant changes with respect to the one originally submitted. I am overall happy with the way in which the authors addressed the different points I raised in the Referee report. I believe that this second version of the paper is written in a clearer way and the results of the analysis are relevant for the neutrino community. Therefore, I recommend the publication of the manuscript in Nature.

Author Rebuttals to First Revision:

Author response to referee comments on “Electron Beam Energy Reconstruction for Neutrino Oscillation Measurements”

We thank the referees for their careful review of our manuscript and constructive suggestions. We revised the manuscript accordingly and present below our point-by-point responses to the questions raised by referees #2 and #3.

We hope the referees will find the revised manuscript suitable for publication in Nature,

The e4v and CLAS collaborations.

Referee #2:

I would like to thank the authors for addressing my comments and making changes to the paper accordingly, especially adding details on detector calibration and systematic analysis and discussing DUNE prism, which definitely make the paper stronger. In the new version, the authors removed the discussion of impact on neutrino oscillation parameters, added a new model prediction based on SuSa and expanded the discussion on transverse variables to include two more variables delta alpha_T and delta phi_T. All these changes make the paper more focused on discussing how to use the electron scattering data to improve modeling of neutrino interactions, with a focus on the issues in the energy reconstruction. I think this paper deserves to be published in Nature, provided that the authors address the following additional comments.

1. I noticed that some of the GENIE G2018 predictions changed in the new draft. For example, in the old draft figure 4 (P_T > 400 MeV/c), there are two peaks in the GENIE prediction (at 1 GeV and 1.7 GeV). In the new draft figure 4, there is only one peak at 1.7 GeV. What have changed?

> The updated manuscript shows cross sections, which are corrected for CLAS acceptance effects, rather than weighted counts as shown in the previous revision. Among other changes, in order to calculate cross sections we refined our particle fiducial cuts to better define the CLAS acceptance, as detailed in the Methods section under “Experimental setup and particle identification”. We verified that these changes had the largest effect for the highest perpendicular momenta and are the cause of the difference.

2. Fig. Extended Data Fig. 2 is interesting, but the He point at 2.257 GeV on the right plot seems to be an outlier. Is there a reason why that point is so high?

> SuSAv2 predictions for few-body nuclei such as ^{4}He are very challenging due to their use of relativistic mean field calculation of the ground state, which was never intended to describe nuclei lighter than ^{12}C. While the GENIE fermi gas model describes ^{4}He better than SuSAv2 does, both models are much better approximations for medium and heavy nuclei than for light ones. We have added a note about this limitation to the Methods section and to the captions of Extended Data Table 1 and Figure 2.

3. There are discrepancies between data and predictions in many distributions. I still think it would be good to provide more information on the mismodeling to emphasize the advantage of using electron scattering data to improve the neutrino simulation.

a) e-GENIE using the older G2018 model predicts the peak position to be 10, 25 and 36 MeV too low. The authors provided an explanation in the response to my comments. It would be good to clarify it in the paper.

Added to the text.

b) In figure 3, SuSAv2 either underestimates or overestimates the QE peak, is there an explanation for that?

We show this discrepancy directly in Extended Data Table 1 and Fig 2. This is an important part of the empirical findings of our paper and we do not have specific explanations for it. This showcases the importance of our data to help theorists further test and refine their models for implementation in neutrino event generators.

c) In places where RES and DIS dominate, the theories always overestimate the cross section. Previous neutrino cross section measurements seem to suggest GENIE overpredicts resonance production in the neutrino-nucleus interactions (e.g. T2K arXiv:1909.03936, figures 13-16). This seems to be confirmed by the electron scattering study reported in this paper. What could be the source of this mismodeling?

This data-GENIE discrepancy was also seen in inclusive electron scattering from the proton and the deuteron in Ref 19 [PRD 102 113003] (see Figs 11 and 12). It appears to be due to the fundamental lepton-nucleon interaction model rather than to the nuclear modeling. Specifically, we are aware of two possible causes: (1) GENIE was tuned to ν-D data only and therefore the sum of the axial+vector response is better constrained than just the vector part. (2) the split between "RES" and "DIS" appears to double count some resonance contributions. We added a comment in the text with more explanation in the Methods section, including references to both T2K and the inclusive electron scattering study.

4. I found the section "Subtraction of undetected pions and photons" hard to follow. The concept is clear but the details are confusing. I don't understand what this sentence means: We used that as a weight to subtract for the undetected pion events. Another example is the sentence: this subtraction from the subtracted events increased the yield, which sounds very cryptic. Maybe this is a standard method used in electron scattering experiments, but the description of the method is not very clear for a general reader.

We added an example to clarify the first confusing sentence and we rewrote the paragraphs describing the higher multiplicity event subtraction.

Referee #3:

Equation (2) is not exact because the flux is not same even without oscillation due to acceptance difference at near and far.

We added the word "approximately" and explain that acceptance effects will also cause some difference.

p.2 right column "There is no guarantee that they are suitable for analyzing far-detector data" is too strong statement. The constraint by near-detector is stronger for narrower-energy beam. So it is a matter of degree.

We softened the wording and clarified its applicability: "Thus, for precision measurements using a broad-energy neutrino beam, the degree to which the near-detector data alone can constrain models is unclear, since the neutrino flux can be very different at the far detector due to oscillations."

p.3 Left column, "the model dramatically overestimated the amount of mis-reconstructed events" This is true only for `the' model introduced later and used in this paper.

We now refer to "the models used" to be more specific.

p.3 Right column, isn't "interactions of the lepton with a meson exchanged between two nucleons" wrong? This sounds like it is interactions with a meson, but actually isn't it interactions with 'two nucleons exchanging a meson'?

Changed.

p.9 Right column, I cannot understand "Inclusive lepton-nucleus scattering has the same number of independent structure functions." Actually it is different (two for electron- and three for nuetrino-scattering)as written in the paper.

We thank the referee for catching this error. We removed the sentence.

Reviewer Reports on the Second Revision:

Referees' comments:

Referee #2 (Remarks to the Author):

I would like to thank the authors for addressing my additional comments and questions. I am satisfied with the responses and recommend the paper for publication in Nature.