

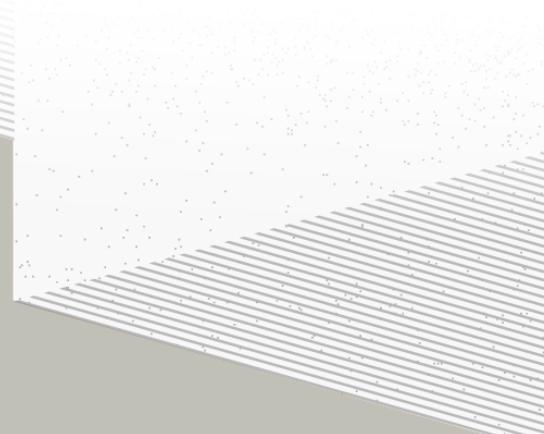
1

2

# THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 596  
*Matière, Molécules, Matériaux*  
Spécialité : *Physique Subatomique et Instrumentation Nucléaire*



Par

**Léonard Imbert**

**Deep learning methods and Dual Calorimetric analysis for high precision neutrino oscillation measurements at JUNO**

Thèse présentée et soutenue à Nantes, le 2 Decembre 2024  
Unité de recherche : Laboratoire SUBATECH, UMR 6457

## Rapporteurs avant soutenance :

Christine Marquet Directrice de recherche au CNRS, LP2I Bordeaux  
David Rousseau Directeur de recherche au CNRS, IJCLab

## Composition du Jury :

Président :	Barbara Erazmus	Directrice de recherche au CNRS, Subatech
Examinateurs :	Juan Pedro Ochoa-Ricoux	Full Professor, University of California, Irvine
	Yasmine Amhis	Directrice de recherche au CNRS, IJCLab
	Christine Marquet	Directrice de recherche au CNRS, LP2I Bordeaux
	David Rousseau	Directeur de recherche au CNRS, IJCLab
Dir. de thèse :	Frédéric Yermia	Professeur des universités, Nantes Université
Co-dir. de thèse :	Benoit Viaud	Chargé de recherche au CNRS, Subatech



<sup>3</sup> A mon Père  $\hat{\text{D}}$

<sup>4</sup> A ma Mère  $\hat{\text{A}}$



# 5 Contents

6	<b>Contents</b>	1
7	<b>Remerciements</b>	5
8	<b>Introduction</b>	7
9	<b>1 Neutrino physics</b>	9
10	1.1 Introduction to the Standard model . . . . .	9
11	1.1.1 Limits of the standard model . . . . .	11
12	1.2 The Neutrinos . . . . .	11
13	1.2.1 Oscillation . . . . .	11
14	1.2.2 Phenomenology . . . . .	11
15	1.2.3 Open questions . . . . .	11
16	<b>2 The JUNO experiment</b>	13
17	2.1 Reactor Neutrinos physics in JUNO . . . . .	14
18	2.1.1 Antineutrino spectrum measured in JUNO . . . . .	14
19	2.1.2 Background spectra . . . . .	16
20	2.2 Other physics . . . . .	17
21	2.3 The JUNO detector . . . . .	18
22	2.3.1 Detection principle . . . . .	19
23	2.3.2 Central Detector (CD) . . . . .	20
24	2.3.3 Veto detector . . . . .	24
25	2.4 Calibration strategy . . . . .	25
26	2.4.1 Energy scale calibration . . . . .	25
27	2.4.2 Calibration system . . . . .	26
28	2.4.3 Instrumental non-linearity calibration . . . . .	27
29	2.5 Satellite detectors . . . . .	28
30	2.5.1 TAO . . . . .	28
31	2.5.2 OSIRIS . . . . .	29
32	2.6 Software . . . . .	29
33	2.7 Reactor anti-neutrino oscillation analysis . . . . .	30
34	2.7.1 IBD samples selection . . . . .	30
35	2.7.2 Synthetic overview of fit procedures developed at JUNO . . . . .	31
36	2.7.3 The spectrum model and sources of systematic uncertainties . . . . .	33

37	2.7.4 Versions of the fit used in this thesis . . . . .	35
38	2.7.5 Physics results . . . . .	36
39	2.8 Summary . . . . .	36
40	<b>3 Introduction to the reconstruction methods and algorithms used in this thesis</b> . . . . .	37
41	3.1 Core concepts in machine learning and neural networks . . . . .	38
42	3.1.1 Boosted Decision Tree (BDT) . . . . .	38
43	3.1.2 Artificial Neural Network (NN) . . . . .	39
44	3.1.3 Training procedure . . . . .	40
45	3.1.4 Potential pitfalls . . . . .	43
46	3.2 Neural networks architectures . . . . .	46
47	3.2.1 Fully Connected Deep Neural Network (FCDNN) . . . . .	46
48	3.2.2 Convolutional Neural Network (CNN) . . . . .	47
49	3.2.3 Graph Neural Network (GNN) . . . . .	49
50	3.2.4 Adversarial Neural Network (ANN) . . . . .	51
51	3.3 State of the art of the Offline IBD reconstruction in JUNO . . . . .	51
52	3.3.1 Interaction vertex reconstruction . . . . .	52
53	3.3.2 Energy reconstruction . . . . .	55
54	3.3.3 Machine learning for reconstruction . . . . .	58
55	3.4 Conclusion . . . . .	61
56	<b>4 Image recognition for IBD reconstruction with the SPMT system</b> . . . . .	63
57	4.1 Method and model . . . . .	64
58	4.1.1 Model . . . . .	65
59	4.1.2 Data representation . . . . .	66
60	4.1.3 Dataset . . . . .	68
61	4.1.4 Data characteristics . . . . .	69
62	4.2 Training . . . . .	71
63	4.3 Results . . . . .	71
64	4.3.1 J21 results . . . . .	72
65	4.3.2 J21 Combination of classic and ML estimator . . . . .	74
66	4.3.3 J23 results . . . . .	77
67	4.4 Conclusion and prospect . . . . .	78
68	<b>5 Graph representation of JUNO for IBD reconstruction</b> . . . . .	81
69	5.1 Data representation . . . . .	82
70	5.2 Message passing algorithm . . . . .	85
71	5.3 Data . . . . .	87
72	5.4 Model . . . . .	88
73	5.5 Training . . . . .	89
74	5.6 Optimization . . . . .	90
75	5.6.1 Software optimization . . . . .	90
76	5.6.2 Hyperparameters optimization . . . . .	91
77	5.7 performance of the final version . . . . .	92

78	5.8 Conclusion . . . . .	95
79	<b>6 Reliability of machine learning methods</b>	99
80	6.1 BDT for energy reconstruction (BDTE) . . . . .	100
81	6.2 Adversarial method . . . . .	102
82	6.3 ANN Architecture . . . . .	104
83	6.3.1 Back-propagation problematic . . . . .	105
84	6.3.2 Reconstruction Network (FFNN) . . . . .	106
85	6.3.3 Adversarial Neural Network (ANN) . . . . .	107
86	6.4 Training of the ANN . . . . .	110
87	6.4.1 First training phase: back to physics . . . . .	110
88	6.4.2 Second training phase: Breaking of the reconstruction . . . . .	111
89	6.5 Conclusion and prospect . . . . .	114
90	<b>7 Dual calorimetric analysis with neutrino oscillation for Precision Measurement</b>	119
91	7.1 Motivations . . . . .	122
92	7.1.1 Discrepancies between the SPMT and LPMT results . . . . .	122
93	7.1.2 Charge Non-Linearity (QNL) . . . . .	122
94	7.2 Our approach to Dual Calorimetry with neutrino oscillation . . . . .	124
95	7.2.1 Toy experiments . . . . .	126
96	7.2.2 Comparing the solar parameters from individual analyses : LPMT vs SPMT . . . . .	127
97	7.2.3 Direct comparison between the SPMT and LPMT spectra . . . . .	129
98	7.2.4 Joint fit of the SPMT and LPMT spectra : $\chi^2_{H_0} - \chi^2_{H_1}$ . . . . .	131
99	7.2.5 Joint fit of the SPMT and LPMT spectra : distribution of $\delta \sin^2(2\theta_{12})$ and $\delta \Delta m^2_{21}$ . . . . .	132
100	7.2.6 Limitations . . . . .	132
101	7.3 Fit software . . . . .	133
102	7.3.1 AveNu <sub>e</sub> Standalone Generators . . . . .	134
103	7.3.2 AveNu <sub>e</sub> Fitting Package . . . . .	134
104	7.3.3 Details of the IBD generator . . . . .	135
105	7.4 Technical challenges and development . . . . .	136
106	7.5 Covariance matrix . . . . .	137
107	7.5.1 Analytical method . . . . .	137
108	7.5.2 Empirical method . . . . .	139
109	7.6 Technical Validation . . . . .	140
110	7.7 Results . . . . .	143
111	7.7.1 Effect of supplementary QNL on the LPMT spectrum . . . . .	143
112	7.7.2 Comparison and statistical tests results . . . . .	145
113	7.8 Conclusion and perspectives . . . . .	148
114	7.8.1 Empirical correlation matrix from fully simulated event . . . . .	149
115	<b>Conclusion</b>	155
116	<b>A Calculation of optimal <math>\alpha</math> for estimator combination</b>	159
117	A.1 Unbiased estimator . . . . .	159

---

118	<b>A.2 Optimal variance estimator</b> . . . . .	159
119	<b>B Charge spherical harmonics analysis</b>	161
120	<b>C Correction of <math>E_{vis}</math> bias</b>	169
121	<b>List of Tables</b>	172
122	<b>List of Figures</b>	181
123	<b>List of Abbreviations</b>	183
124	<b>Bibliography</b>	185

<sup>125</sup> **Remerciements**



# <sup>126</sup> Introduction

<sup>127</sup> The Standard Model of particle physics (SM) has been remarkably successful at accounting for,  
<sup>128</sup> or predicting experimental observations in the laboratory. However, it is the subject of several  
<sup>129</sup> limitations. For instance, it provides a mechanism to explain the existence of mass but can't predict  
<sup>130</sup> the peculiar pattern followed by fermion masses. The same applies to CP violation. The SM predicts  
<sup>131</sup> its existence but not the amplitude necessary to explain the baryonic asymmetry of the Universe. For  
<sup>132</sup> such reasons, one can assume the SM is the manifestation of a more fundamental physics, Beyond  
<sup>133</sup> the Standard Model (BSM).

<sup>134</sup> Neutrino physics is a window on BSM. Indeed, the mass of known neutrinos is at least 5 order of  
<sup>135</sup> magnitudes below that of the lightest fermion, which further deepens the issue of fermion mass  
<sup>136</sup> generation. Some solutions have implication on the nature of neutrinos – dirac or majorana fermions  
<sup>137</sup> ? – which one of the big unknowns in this domain. Additional neutrinos beyond the three presently  
<sup>138</sup> known shall also be considered. The way neutrinos mix flavor to make neutrino oscillation possible  
<sup>139</sup> is also unexplained. This is one of the tasks of BSM models to answer such questions. Before that, a  
<sup>140</sup> good part of the World experimental program in the 10 coming years is to complete the exploration  
<sup>141</sup> of 3-neutrino physics by answering mainly two questions : does CP violation exist the lepton system  
<sup>142</sup> ? What is the Neutrino Mass ordering (NMO) ? An introduction to neutrino physics will be given in  
<sup>143</sup> Chapter 1.

<sup>144</sup>

<sup>145</sup> The Jiangmen Underground Neutrino Observatory (JUNO), currently under construction in China,  
<sup>146</sup> aims to address these questions, particularly the determination of the NMO. JUNO's approach is  
<sup>147</sup> to study reactor antineutrinos emitted from nearby nuclear power plants. By precisely measuring  
<sup>148</sup> the energy spectrum of these antineutrinos after oscillation, JUNO seeks to detect the subtle inter-  
<sup>149</sup> ference patterns in the spectrum that are sensitive to the NMO. The ability to achieve this requires  
<sup>150</sup> unprecedented precision in both the energy resolution and the calibration of the detector's response  
<sup>151</sup> to neutrino events. JUNO is expected to start data collection in 2025, with the goal of determining the  
<sup>152</sup> NMO at a significance level of  $3-4\sigma$  after six years of data taking. At the heart of JUNO's experimen-  
<sup>153</sup> tal design is its dual calorimetry system, comprising two separate sets of photomultipliers—large  
<sup>154</sup> (LPMT) and small (SPMT) PMTs—that allow for independent energy measurements of the same  
<sup>155</sup> events. This dual system is not only essential for improving energy resolution but also for providing  
<sup>156</sup> cross-checks that ensure systematic uncertainties are well-understood and minimized. Achieving  
<sup>157</sup> JUNO's goals depends on this dual calorimetry system, as it will enable precise reconstruction of the  
<sup>158</sup> energy spectrum and the identification of potential discrepancies between the two systems.

<sup>159</sup>

<sup>160</sup> Another emerging area of importance in particle physics experiments is the application of machine  
<sup>161</sup> learning (ML) techniques. Over the past decade, ML methods, particularly deep learning, have been  
<sup>162</sup> increasingly used to tackle complex problems in event classification, reconstruction, and even data  
<sup>163</sup> generation like the High luminosity LHC Upgraded experiments. Performant online reconstruction,  
<sup>164</sup> critical for the trigger systems of such experiments, is another example. The complexity of the data  
<sup>165</sup> and the required precision in experiments such as JUNO make ML an attractive tool. In particular,  
<sup>166</sup> Neural Networks (NNs) and other advanced ML models have shown potential for improving the  
<sup>167</sup> accuracy of energy reconstruction and other key analysis tasks. However, for the results obtained

168 using ML methods to be trusted by the scientific community, the reliability of these methods must be  
169 rigorously demonstrated. An introduction to ML, and in particular Neural Network (NN) is given  
170 in Chapter 3.

171  
172 This thesis was performed in the framework of the Neutrino group at Subatech, since October 2021.  
173 The exploratory works reported in this manuscript addresses the subjects mentioned above, in the  
174 particular context of the measurement by JUNO of the reactor antineutrino oscillation to determine  
175 the NMO. Before the start of this thesis, several ML energy reconstruction algorithms – Boosted  
176 Decision Trees (BDT), Fully Connected Neural Networks (FCNN), Convolutional Neural Networks  
177 (CNNs) and Graph Neural Networks (GNNs) – had already been developed within the collaboration.  
178 Their performance seems to match that of the classical algorithm but not to do convincingly better.  
179 We have explored a possibility to do better by developing a GNN with an innovative architecture  
180 tailored to the JUNO experiment. Before that, we developed a CNN for the reconstruction of the  
181 anti-neutrino energy using only JUNO’s small PMTs system. This CNN is useful in particular in  
182 Chapter 7 as there is official SPMT only reconstruction in the collaboration yet. These algorithms are  
183 described in Chapters 4 and 5.

184 We have been the first in JUNO to address the issue of ML reliability. We have followed two paths  
185 for that. First, a simple approach is to compare event per event the results obtained by various algo-  
186 rithms, to find discrepancies, and more generally differences or common points in the way detector’s  
187 information is used. This requires to implement in JUNO’s official software algorithms traditionally  
188 developed standalone, as well as the necessary software tools. This was our contribution there. The  
189 second path was to explore the feasibility of an Adversarial Neural Network (ANN) to generate  
190 (and therefore identify) scenarios of discrepancies between raw data in the real detector and in the  
191 detector’s simulation. The focus here is on discrepancies that could alter JUNO’s results on NMO, but  
192 are too subtle to be detected via usual data/MC comparisons in control samples. This is presented  
193 in Chapter 6.

194  
195 We have already mentioned earlier it is crucial for JUNO to understand its energy scale with a  
196 good precision. This is the raison d’être of the existence of two calorimetric readout systems : the  
197 large (LPMT) and small (SPMT) photomultipliers systems. It allows Dual Calorimetry techniques  
198 to constrain our understanding of the reconstruction. The last subject of this thesis explores for the  
199 first time one of them : the Dual Calorimetry with neutrino oscillation, which leverages potential  
200 discrepancies between the oscillation analyses performed with each system. Our work on this is  
201 described in Chapter 7. It was also the occasion of technical developments on the analysis framework  
202 used at Subatech. These improvements will be very useful for future analyses of the group, beyond  
203 Dual calorimetry.

<sup>204</sup> **Chapter 1**

<sup>205</sup> **Neutrino physics**

*I have done a terrible thing, I have postulated a particle that cannot be detected.*

Wolfgang Pauli – "Foreword" by Frederick Reines to "Spaceship Neutrino" by Christine Sutton, (p. xi), 1992.

<sup>207</sup> **Contents**

---

<sup>208</sup> <b>1.1</b>	<b>Introduction to the Standard model</b>	<b>9</b>
<sup>210</sup> 1.1.1	Limits of the standard model	11
<sup>211</sup> <b>1.2</b>	<b>The Neutrinos</b>	<b>11</b>
<sup>212</sup> 1.2.1	Oscillation	11
<sup>213</sup> 1.2.2	Phenomenology	11
<sup>214</sup> 1.2.3	Open questions	11

---

<sup>215</sup> Our understanding of the universe describe it as composed of elementary components called elementary particles, the study of these particles is therefore particles physics. The theoretical model describing these particles and their interactions is the Standard Model (SM), with the exception of the gravitation. It has proven its robustness over the last decades, accounting for most of observed phenomena with a few exception. This exception are phenomena described as Beyond Standard Model (BSM).

<sup>224</sup> In this chapter in describe briefly the Standard model and its limitations in section 1.1, then delve a bit further in the specificity of neutrinos physics in Section 1.2.

<sup>226</sup> **1.1 Introduction to the Standard model**

<sup>227</sup> The SM categorize the elementary particles into two categories: the *fermions* constituting the matter and the *bosons* that mediate their interaction. The fermions are themselves divided in two categories, the *quark* and the *leptons*. Figure 1.1 shows the elementary particle and their classification. Each one these particle is characterized by the value of their quantum number the main one being their mass  $m$ , spin  $J$  and electric charge  $Q$ . The leptons also possess leptonic quantum number  $L = 1$  a flavor quantum number  $L_{e,\mu,\tau}$  corresponding to their family, electronic, muonic or tauic. The leptons are thus split in three family: the electronic  $L_e = 1 \rightarrow (e, \nu_e)$ , muonic  $L_\mu = 1 \rightarrow (\mu, \nu_\mu)$  and tauic  $L_\tau = 1 \rightarrow (\tau, \nu_\tau)$  families, each composed of a charged  $Q = 1$  and a neutral particle  $Q = 0$ . The neutral leptons are named the *neutrinos* represented by the character  $\nu$ .

<sup>236</sup> Each fermion also possess an antiparticle of opposite electric charge and opposite leptonic and flavour quantum number. Thus the antiparticle of the electron  $e(Q = 1, L = 1, L_e = 1)$ , the positron is defined  $e^+(Q = -1, L = -1, L_e = -1)$ .

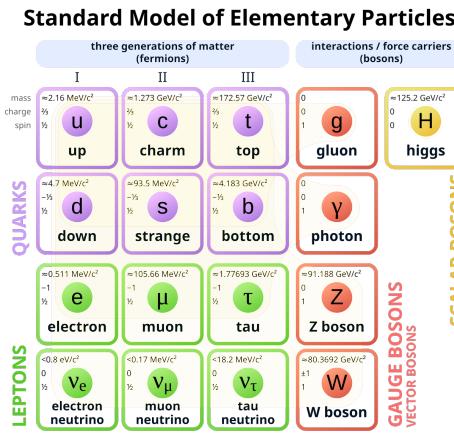


FIGURE 1.1 – List of the elementary particles in the Standard Model. The antiparticles are not displayed.

239 The particles of the SM interact with each other via four interactions or forces. Three are described  
 240 by the SM by the exchange of a boson:

- 241 — The strong force, described by the exchange of a gluon. Only the quark are sensitive to it.  
 242 This force is very small range  $\sim 10^{-15}$  m, the size of a nucleus. It's the strong force that allow  
 243 the cohesion of nucleus inside atoms. As its name indicates, it this the strongest of the four  
 244 interaction.
- 245 — The electromagnetic force, described by the exchange of a photon. This force has unlimited  
 246 range, all the charged particles – quark and charged leptons – are sensitive to it. It is responsible  
 247 for every magnetic effect like the bonding of electrons and nucleus. Its relative strength with  
 248 the strong force is 1/137.
- 249 — The weak force, carried by the  $Z^0$  and  $W^\pm$  bosons. Every fermions is sensitive to it. Its range  
 250 is  $\sim 10^{-18}$  m,  $\sim 0.1\%$  the size of a proton. Its relative strength to the strong force is  $10^{-6}$ ,  
 251 explaining its name. It is responsible for nuclear beta decay and other similar process.

252 The final force, not described by the standard model, is the gravitational force. It's range is infinite,  
 253 and concerns every massive ( $m \neq 0$ ) particles. Its relative strength to the strong force is of  $6 \times$   
 254  $10^{-39}$ . Extension to the SM propose a supplementary boson, the graviton, that be the carrier of the  
 255 gravitational force but it has yet to be detected [1, 2].

## 256 Interactions and symmetries

257 Symmetries are fundamental components of modern particle physics. As described in Noether's  
 258 theorem [3], the invariance or non-invariance of the physics law under transformations (translation,  
 259 rotation, ...), represented by the formal invariance of the SM Lagrangian  $\mathcal{L}$  under those transforma-  
 260 tion, express the conservation of a quantity.

261 The invariance of  $\mathcal{L}$  by translation in space characterize the conservation of the momentum, the  
 262 rotational invariance in space the conservation of the angular momentum, the invariance by trans-  
 263 lation in time the conservation of energy, etc. If the transformation is continuous, the sum of the  
 264 quantum numbers is conserved in a interaction. Invariances under discrete transformation also  
 265 provide conservation of quantum number. Three discrete transformations are important for the SM:

- 266 — The parity  $P$  symmetry transform  $(\vec{x}, t) \rightarrow (-\vec{x}, t)$ , reversing the handedness of space. The  
 267 momentum thus become  $\vec{p} \rightarrow -\vec{p}$  and the helicity  $\frac{\vec{p} \cdot \vec{s}}{|\vec{p}|}$ , where  $\vec{s}$  is the spin, change sign.

268 — Reversal of time  $T$  where  $(\vec{x}, t) \rightarrow (\vec{x}, -t)$ , inverting the initial and final state of an interaction  
269  $A + B \rightarrow C$  become  $C \rightarrow A + B$ . The momentum and the spin both change sign, leaving the  
270 helicity unchanged.

271 — Charge conjugation  $C$ , replacing the particles by their antiparticles counterpart and vice-versa,  
272 leaving untouched the momentum, spin and helicity.

273 The C, P and their combination CP symmetry was believed to be conserved until 1956, the discovery  
274 of their violation [4–6] in weak interaction revealed the non-triviality of its nature.

275 The fundamental symmetry CPT, the combination of C, P and T symmetry, is an exact symmetry. It  
276 mean that any process where the particles are switched with their anti-particles, spin-projection are  
277 of opposite sign and initial and final state are swapped must go with the same probability than the  
278 initial process. This implies that the mass, life times, absolute values of electric charge and magnetic  
279 moment of particles and antiparticles must be the same.

280 **1.1.1 Limits of the standard model**

281 **1.2 The Neutrinos**

282 **1.2.1 Oscillation**

283 **1.2.2 Phenomenology**

284 **1.2.3 Open questions**



<sup>285</sup> **Chapter 2**

<sup>286</sup> **The JUNO experiment**

<sup>287</sup> “*Ave Juno, rosae rosam, et spiritus rex*”. It means nothing but I found  
it in tone.

<sup>288</sup> **Contents**

---

<sup>289</sup> <b>2.1 Reactor Neutrinos physics in JUNO</b>	<sup>290</sup> . . . . .	<sup>14</sup>
<sup>291</sup> <b>2.1.1 Antineutrino spectrum measured in JUNO</b>	<sup>291</sup> . . . . .	<sup>14</sup>
<sup>292</sup> <b>2.1.2 Background spectra</b>	<sup>292</sup> . . . . .	<sup>16</sup>
<sup>293</sup> <b>2.2 Other physics</b>	<sup>293</sup> . . . . .	<sup>17</sup>
<sup>294</sup> <b>2.3 The JUNO detector</b>	<sup>294</sup> . . . . .	<sup>18</sup>
<sup>295</sup> <b>2.3.1 Detection principle</b>	<sup>295</sup> . . . . .	<sup>19</sup>
<sup>296</sup> <b>2.3.2 Central Detector (CD)</b>	<sup>296</sup> . . . . .	<sup>20</sup>
<sup>297</sup> <b>2.3.3 Veto detector</b>	<sup>297</sup> . . . . .	<sup>24</sup>
<sup>298</sup> <b>2.4 Calibration strategy</b>	<sup>298</sup> . . . . .	<sup>25</sup>
<sup>299</sup> <b>2.4.1 Energy scale calibration</b>	<sup>299</sup> . . . . .	<sup>25</sup>
<sup>300</sup> <b>2.4.2 Calibration system</b>	<sup>300</sup> . . . . .	<sup>26</sup>
<sup>301</sup> <b>2.4.3 Instrumental non-linearity calibration</b>	<sup>301</sup> . . . . .	<sup>27</sup>
<sup>302</sup> <b>2.5 Satellite detectors</b>	<sup>302</sup> . . . . .	<sup>28</sup>
<sup>303</sup> <b>2.5.1 TAO</b>	<sup>303</sup> . . . . .	<sup>28</sup>
<sup>304</sup> <b>2.5.2 OSIRIS</b>	<sup>304</sup> . . . . .	<sup>29</sup>
<sup>305</sup> <b>2.6 Software</b>	<sup>305</sup> . . . . .	<sup>29</sup>
<sup>306</sup> <b>2.7 Reactor anti-neutrino oscillation analysis</b>	<sup>306</sup> . . . . .	<sup>30</sup>
<sup>307</sup> <b>2.7.1 IBD samples selection</b>	<sup>307</sup> . . . . .	<sup>30</sup>
<sup>308</sup> <b>2.7.2 Synthetic overview of fit procedures developed at JUNO</b>	<sup>308</sup> . . . . .	<sup>31</sup>
<sup>309</sup> <b>2.7.3 The spectrum model and sources of systematic uncertainties</b>	<sup>309</sup> . . . . .	<sup>33</sup>
<sup>310</sup> <b>2.7.4 Versions of the fit used in this thesis</b>	<sup>310</sup> . . . . .	<sup>35</sup>
<sup>311</sup> <b>2.7.5 Physics results</b>	<sup>311</sup> . . . . .	<sup>36</sup>
<sup>312</sup> <b>2.8 Summary</b>	<sup>312</sup> . . . . .	<sup>36</sup>

---

<sup>313</sup> The first idea of a medium baseline ( $\sim 52$  km) experiment, was explored in 2008 [7] where it was  
<sup>314</sup> demonstrated that the Neutrino Mass Ordering (NMO) could be determined by a medium baseline  
<sup>315</sup> experiment if  $\sin^2(2\theta_{13}) > 0.005$  without the requirements of accurate knowledge of the reactor  
<sup>316</sup> antineutrino spectra and the value of  $\Delta m_{32}^2$ . From this idea is born the Jiangmen Underground  
<sup>317</sup> Neutrino Observatory (JUNO) experiment.

<sup>318</sup> JUNO is a neutrino detection experiment under construction located in China, in Guangdong prov-  
<sup>319</sup> ing, near the city of Kaiping. Its main objectives are the determination of the mass ordering at the  
<sup>320</sup> 3-4 $\sigma$  level in 6 years of data taking and the measurement at the sub-percent precision of the oscillation  
<sup>321</sup> parameters  $\Delta m_{21}^2$ ,  $\sin^2 \theta_{12}$ ,  $\Delta m_{32}^2$  and with less precision  $\sin^2 \theta_{13}$ [8].



FIGURE 2.1 – **On the left:** Location of the JUNO experiment and its reactor sources in southern china. **On the right:** Aerial view of the experimental site

For this JUNO will measure the electronic anti-neutrinos ( $\bar{\nu}_e$ ) flux coming from the nuclear reactors of Taishan, Yangjiang, for a total power of 26.6 GW<sub>th</sub>, and the Daya Bay power plant to a lesser extent. All of those cores are the second-generation pressurized water reactors CPR1000, which is a derivative of Framatome M310. Details about the power plants characteristics and their expected flux of  $\bar{\nu}_e$  can be found in the table 2.1. The distance of 53 km has been specifically chosen to maximize the disappearance probability of the  $\bar{\nu}_e$ . The data taking is scheduled to start early 2025.

## 2.1 Reactor Neutrinos physics in JUNO

JUNO will try to determine the NMO and to bring at the few per mille level our knowledge of  $\Delta m_{31}^2$ ,  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$  via the precision analysis of the spectrum of the visible energy left by reactor antineutrinos in its detector.

### 2.1.1 Antineutrino spectrum measured in JUNO

To some extent, this analysis is equivalent to extracting from this spectrum the oscillation probability [8] :

$$P(\bar{\nu}_e \rightarrow \bar{\nu}_e) = 1 - \sin^2 2\theta_{12} c_{13}^4 \sin^2 \frac{\Delta m_{21}^2 L}{4E} - \sin^2 2\theta_{13} \left[ c_{12}^2 \sin^2 \frac{\Delta m_{31}^2 L}{4E} + s_{12}^2 \sin^2 \frac{\Delta m_{32}^2 L}{4E} \right]$$

Where  $s_{ij} = \sin \theta_{ij}$ ,  $c_{ij} = \cos \theta_{ij}$ ,  $E$  is the  $\bar{\nu}_e$  energy and  $L$  is the baseline. We can see the sensitivity to the NMO in the dependency to  $\Delta m_{32}^2$  and  $\Delta m_{31}^2$  causing a phase shift of the spectrum as we can see in the Figure 2.2.

In practice, a fit to the grey distribution of Figure 2.3 will be performed. It is the sum of two components :signal (black) and bacgrounds (colored). Reactor antineutrinos are detected by JUNO via Inverse Beta Decays (IBD) :  $\bar{n}\bar{\nu}_e + p^- > e^+ + n$ . The energy spectrum under investigation is therefore that of the reconstructed  $e^+$  visible energy. The black signal spectrum is therefore the sum of the antineutrino differential fluxes from all reactors and reaching the detecteur, weighted by the oscillation probability of Eq 2.1.1 and the IBD differential cross section and convoluted with detection effects. These various ingredients are theoretically modelled in order to provide the probability density function (PDF) to be used in the fit.

To reach JUNO's goals, it takes that this experimental spectrum still bears sizeable traces of the very small phase shift mentioned above. Most notably, the following requirements must be fulfilled :

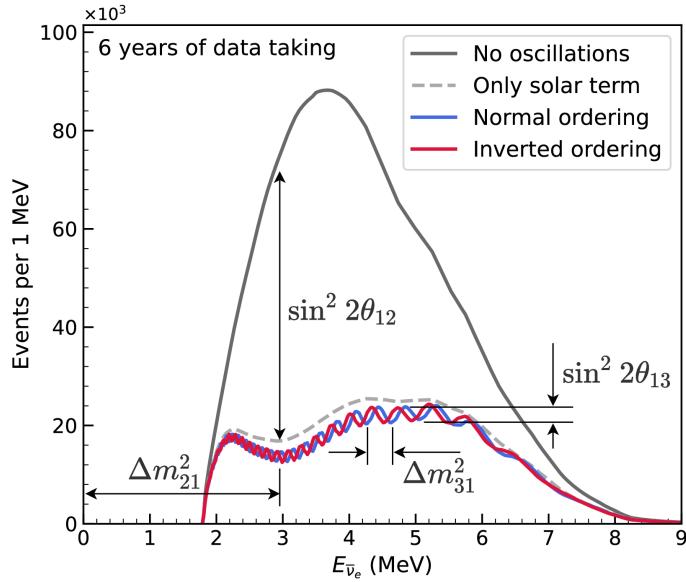


FIGURE 2.2 – Expected number of neutrinos event per MeV in JUNO after 6 years of data taking. The black curve shows the flux if there was no oscillation. The light gray curve shows the oscillation if only the solar terms are taken in account ( $\theta_{12}$ ,  $\Delta m_{21}^2$ ). The blue and red curve shows the spectrum in the case of, respectively, NO and IO. The dependency of the oscillation to the different parameters are schematized by the double sided arrows. We can see the NMO sensitivity by looking at the fine phase shift between the red and the blue curve.

- 349 1. An energy resolution of  $3\%/\sqrt{E(\text{MeV})}$  to be able to distinguish the fine structure of the fast  
350 oscillation.
- 351 2. An energy scale known at the better than the 1% level.
- 352 3. A baseline between 40 and 65 km to maximise the  $\bar{\nu}_e$  oscillation probability. The optimal  
353 baseline would be 58 km and JUNO baseline is 53 km.
- 354 4. At least  $\approx 100,000$  events. This is the necessary statistics to reach JUNO's canonical sensitivity  
355 after 6 years of data taking.

### 356 $\bar{\nu}_e$ flux coming from nuclear power plants

357 To get such high measurements precision, it is necessary to have a very good understanding of the  
358 sources characteristics. For its NMO and precise measurement studies, JUNO will observe the energy  
359 spectrum of neutrinos coming from the nuclear power plants Taishan and Yangjiang's cores, located  
360 at 53 km of the detector to maximise the disappearance probability of the  $\bar{\nu}_e$ .

361 The  $\bar{\nu}_e$  coming from reactors are emitted from  $\beta$ -decay of unstable fission fragments. The Taishan  
362 and Yangjiang reactors are Pressurised Water Reactor (PWR), the same type as Daya Bay. In those  
363 type of reactor more the 99.7 % and  $\bar{\nu}_e$  are produced by the fissions of four fuel isotopes  $^{235}\text{U}$ ,  $^{238}\text{U}$ ,  
364  $^{239}\text{Pu}$  and  $^{241}\text{Pu}$ . The neutrino flux per fission of each isotope is determined by the inversion of the  
365 measured  $\beta$  spectra of fission product [10–14] or by calculation using the nuclear databases [15, 16].

366 The neutrino flux coming from a reactor at a time  $t$  can be predicted using

$$\phi(E_\nu, t)_r = \frac{W_{th}(t)}{\sum_i f_i(t) e_i} \sum_i f_i(t) S_i(E_\nu) \quad (2.1)$$

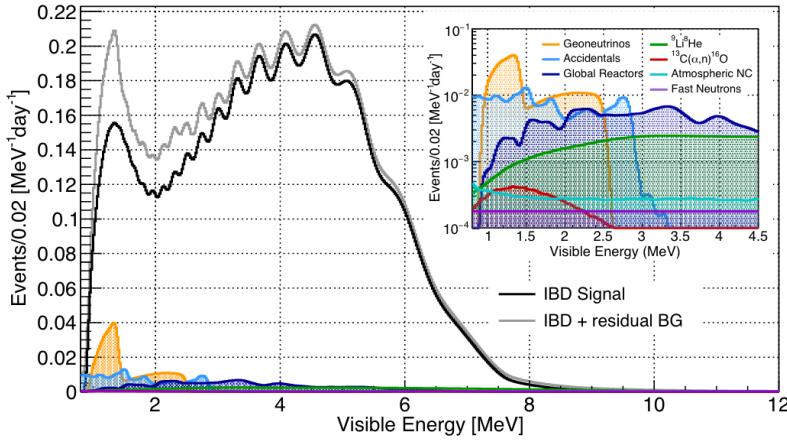


FIGURE 2.3 – Expected visible energy spectrum measured with the LPMT system with (grey) and without (black) backgrounds. The background amount for about 7% of the IBD candidate and are mostly localized below 3 MeV [9]

Reactor	Power (GW <sub>th</sub> )	Baseline (km)
Taishan	9.2	52.71
Core 1	4.6	52.77
Core 2	4.6	52.64
Yangjiang	17.4	52.46
Core 1	2.9	52.74
Core 2	2.9	52.82
Core 3	2.9	52.41
Core 4	2.9	52.49
Core 5	2.9	52.11
Core 6	2.9	52.19
Daya Bay	17.4	215
Huizhou	17.4	265

TABLE 2.1 – Characteristics of the nuclear power plants observed by JUNO.

where  $W_{th}(t)$  is the thermal power of the reactor,  $f_i(t)$  is the fraction fission of the  $i$ th isotope,  $e_i$  its thermal energy released in each fission and  $S_i(e_\nu)$  the neutrino flux per fission for this isotope.

The latter flux is difficult to predict. To evaluate JUNO's sensitivity and to serve as a starting point in the spectrum PDF, the Huber-Mueller model is used [11], corrected using Daya Bays data [17] to account for a ~5% deficit with respect to models, referred to as the reactor antineutrino anomaly [18], and for a discrepancy between models and data in the spectral shape (the so call 5 MeV bump).

In addition to those prediction, a satellite experiment named TAO[19] will be setup near the reactor core Taishan-1 to measure with an energy resolution of 2% at 1 MeV the neutrino flux coming from the core, more details can be found in Section 2.5.1. It will help identifying unknown fine structure and give more insight on the  $\bar{\nu}_e$  flux coming from this reactor.

### 2.1.2 Background spectra

Considering the close reactor neutrinos flux as the main signal, the signals that are considered as background are:

- The geoneutrinos producing background in the 0.511 ~ 2.7 MeV region.

- The neutrinos coming from the other nuclear reactors around Earth.
- In addition to all those physics signal, non-neutrinos signal that would mimic an IBD will also be present. It is composed of:
- The signal coming from radioactive decay ( $\alpha$ ,  $\gamma$ ,  $\beta$ ) from natural radioactive isotopes in the material of the detector.
  - Cosmogenic event such as fast neutrons and activated isotopes induced by muons passing through the detector, most notably the spallation on  $^{12}\text{C}$ .
- All those events represent a non-negligable part of the spectrum as shown in Figure 2.3.

## 2.2 Other physics

While the design of JUNO is tailored to measure  $\bar{\nu}_e$  coming from nuclear reactor, JUNO will be able to detect neutrinos coming from other sources thus allowing for a wide range of physics studies as detailed in the table 2.2 and in the following sub-sections.

Research	Expected signal	Energy region	Major backgrounds
Reactor antineutrino	60 IBDs/day	0–12 MeV	Radioactivity, cosmic muon
Supernova burst	5000 IBDs at 10 kpc	0–80 MeV	Negligible
DSNB (w/o PSD)	2300 elastic scattering		
Solar neutrino	2–4 IBDs/year	10–40 MeV	Atmospheric $\nu$
Atmospheric neutrino	hundreds per year for $^8\text{B}$	0–16 MeV	Radioactivity
Geoneutrino	hundreds per year	0.1–100 GeV	Negligible
	$\approx 400$ per year	0–3 MeV	Reactor $\nu$

TABLE 2.2 – Detectable neutrino signal in JUNO and the expected signal rates and major background sources

### Geoneutrinos

Geoneutrinos designate the antineutrinos coming from the decay of long-lived radioactive elements inside the Earth. The 1.8 MeV threshold necessary for the IBD makes it possible to measure geoneutrinos from  $^{238}\text{U}$  and  $^{232}\text{Th}$  decay chains. The studies of geoneutrinos can help refine the Earth crust models but is also necessary to characterise their signal, as they are a background to the mass ordering and oscillations parameters studies.

### Atmospheric neutrinos

Atmospheric neutrinos are neutrinos originating from the decay of  $\pi$  and  $K$  particles that are produced in extensive air showers initiated by the interactions of cosmic rays with the Earth atmosphere. Earth is mostly transparent to neutrinos below the PeV energy, thus JUNO will be able to see neutrinos coming from all directions. Their baseline range is large (15km  $\sim$  13000km), they can have energy between 0.1 GeV and 10 TeV and will contain all neutrino and antineutrinos flavour. Their studies is complementary to the reactor antineutrinos and can help refine the constraints on the NMO [8].

407 **Supernovae burst neutrinos**

408 Neutrinos are crucial component during all stages of stellar collapse and explosion. Detection of  
409 neutrinos coming from core collapse supernovae will provide us important informations on the mech-  
410 anisms at play in those events. Thanks to its 20 kt sensible volume, JUNO has excellent capabilities  
411 to detect all flavour of the  $\mathcal{O}(10 \text{ MeV})$  postshock neutrinos, and using neutrinos of the  $\mathcal{O}(1 \text{ MeV})$   
412 will give informations about the pre-supernovae neutrinos. All those informations will allow to  
413 disentangle between the multiple hydro-dynamic models that are currently used to describe the  
414 different stage of core-collapse supernovae.

415 **Diffuse supernovae neutrinos background**

416 Core-collapse supernovae in our galaxy are rare events, but they frequently occur throughout the  
417 visible Universe sending burst of neutrinos in direction of the Earth. All those events contributes to  
418 a low background flux of low-energy neutrinos called the Diffuse Supernovae Neutrino Background  
419 (DSNB). Its flux and spectrum contains informations about the red-shift dependent supernovae rate,  
420 the average supernovae neutrino energy and the fraction of black-hole formation in core-collapse su-  
421 pernovae. Depending of the DSNB model, we can expect 2-4 IBD events per year in the energy range  
422 above the reactor  $\bar{\nu}_e$  signal, which is competitive with the current Super-Kamiokande+Gadolinium  
423 phase [20].

424 **Beyond standard model neutrinos interactions**

425 JUNO will also be able to probe for beyond standard model neutrinos interactions. After the main  
426 physics topics have been accomplished, JUNO could be upgraded to probe for neutrinoless beta  
427 decay ( $0\nu\beta\beta$ ). The detection of such event would give critical informations about the nature of  
428 neutrinos, is it a majorana or a dirac particle. JUNO will also be able to probe for neutrinos that  
429 would come for the decay or annihilation of Dark Matter inside the sun and neutrinos from putative  
430 primordial black hole. Through the unitary test of the mixing matrix, JUNO will be able to search for  
431 light sterile neutrinos. Thanks to JUNO sensitivity, multiple other exotic research can be performed  
432 on neutrino related beyond standard model interactions.

433 **Proton decay**

434 Proton decay is a potential unobserved event where the proton decay by violating the baryon num-  
435 ber. This violation is necessary to explain the baryon asymmetry in the universe and is predicted  
436 by multiple Grand Unified Theories which unify the strong, weak and electromagnetic interactions.  
437 Thanks to its large active volume, JUNO will be able to take measurement of the potential proton  
438 decay channel  $p \rightarrow \bar{\nu}K^+$  [21] thanks to the timing resolution of the SPMT system. Studies show  
439 that JUNO should be competitive with the current best limit at  $5.9 \times 10^{33}$  years from Super-K. This  
440 studies show that JUNO, considering no proton decay events observed, would be able to rules a  
441 limit of  $9.6 \times 10^{33}$  years at 90 % C.L.

442 **2.3 The JUNO detector**

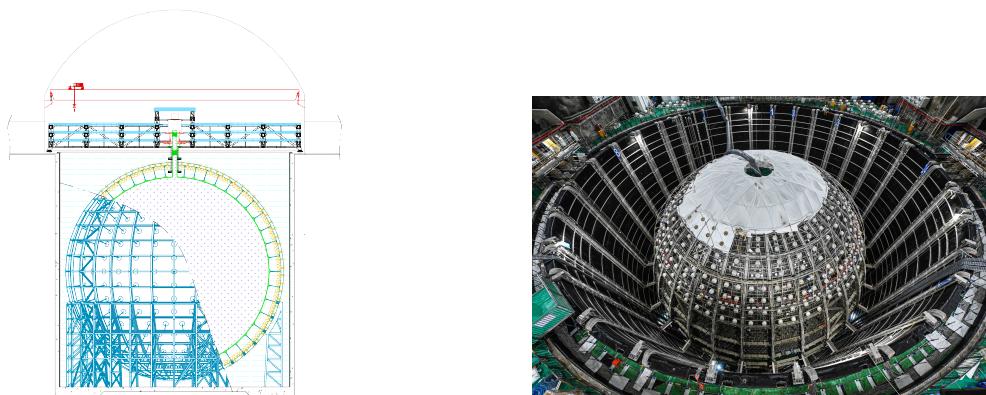
443 The JUNO detector is a scintillator detector buried 693.35 meters under the ground (1800 meters  
444 water equivalent). It consist of Central Detector (CD), a water pool and a Top Tracker (TT) as showed  
445 in Figure 2.4a. The CD is an acrylic vessel containing the 20 ktons of Liquid Scintillator (LS). It is  
446 supported by a stainless steel structure and is immersed in that water pool that is used as shielding

447 from external radiation and as a cherenkov detector for the background. The top of the experiment  
 448 is partially covered by the Top Tracker (TT), a plastic scintillator detector which is use to detect the  
 449 atmospheric muons background and is acting as a veto detector.

450 The top of the experiment also host the LS purification system, a water purification system, a venti-  
 451 lation system to get rid of the potential radon in the air. The CD is observed by two system of  
 452 Photo-Multipliers Tubes (PMT). They are attached to the steel structure and their electronic readout  
 453 is submersed near them. A third system of PMT is also installed on the structure but are facing  
 454 outward of the CD, instrumenting the water to be cherenkov detector. The CD and the cherenkov  
 455 detector are optically separated by Tyvek sheet. A chimney for LS filling and purification and for  
 456 calibration operations connects the CD to the experimental hall from the top.

457 The CD has been dimensioned to meet the requirements presented in Section 2.1.1:

- 458 — Its 20 ktons monolithic LS provide a volume sizeable enough, in combination with the expected  
 459  $\bar{\nu}_e$  flux, to reach the desired statistic in 6 years. Its monolithic nature also allow for a full  
 460 containment of most of the events, preventing the energy loss in non-instrumented parts that  
 461 would arise from a segmented detector.
- 462 — Its large overburden shield it from most of the atmospheric background that would pollute the  
 463 signal.
- 464 — The localization of the experiment, chosen to maximize the disappearance with a 53km baseline  
 465 and in a region that allow two nuclear power plant to be used as sources.



(A) Schematics view of the JUNO detector.

(B) Top down view of the JUNO detector under construction

FIGURE 2.4

466 This section cover in details the different components of the detector and the detection systems.

### 467 2.3.1 Detection principle

The CD will detect the neutrino and measure their energy mainly via an Inverse Beta Decay (IBD) interaction with proton mainly from the  $^{12}\text{C}$  and H nucleus in the LS:

$$\bar{\nu}_e + p \rightarrow n + e^+$$

468 Kinematics calculation shows that this interaction has an energy threshold for the  $\bar{\nu}_e$  of  $(m_n + m_e -$   
 469  $m_p) \approx 1.806$  MeV [22]. This threshold make the experiment blind to very low energy neutrinos.  
 470 The residual energy  $E_\nu - 1.806$  MeV is be distributed as kinetic energy between the positron and the

<sup>471</sup> neutron. The energy of the emitted positron  $E_e$  is given by [22]

$$E_e = \frac{(E_\nu - \delta)(1 + \epsilon_\nu) + \epsilon_\nu \cos \theta \sqrt{(E_\nu - \delta)^2 + \kappa m_e^2}}{\kappa} \quad (2.2)$$

<sup>472</sup> where  $\kappa = (1 + \epsilon_\nu)^2 - \epsilon_\nu^2 \cos^2 \theta \approx 1$ ,  $\epsilon_\nu = \frac{E_\nu}{m_p} \ll 1$  and  $\delta = \frac{m_n^2 - m_p^2 - m_e^2}{2m_p} \ll 1$ . We can see from this  
<sup>473</sup> equation that the positron energy is strongly correlated to the neutrino energy.

<sup>474</sup> The positron and the neutron will then propagate in the detection medium, the Liquid Scintillator  
<sup>475</sup> (LS), loosing their kinetic energy by exciting the molecule of the LS (more details in Section 2.3.2).  
<sup>476</sup> Once stopped, the positron will annihilate with an electron from the medium producing two 511  
<sup>477</sup> KeV gamma. Those gamma will themselves interact with the LS, exciting it before being absorbed  
<sup>478</sup> by photoelectrical effect. The neutron will be captured by an hydrogen, emitting a 2.2 MeV gamma  
<sup>479</sup> in the process. This gamma will also deposit its energy before being absorbed by the LS.

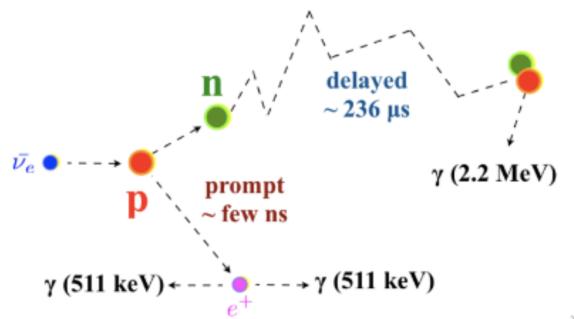


FIGURE 2.5 – Schematics of an IBD interaction in the central detector of JUNO

<sup>480</sup> The scintillation photons have frequency in the UV and will propagate in the LS, being re-absorbed  
<sup>481</sup> and re-emitted by compton effect before finally be captured by PMTs instrumenting the acrylic  
<sup>482</sup> sphere. The analog signal of the PMTs digitized by the electronic is the signal of our experiment.  
<sup>483</sup> The signal produced by the positron is subsequently called the prompt signal, and the signal coming  
<sup>484</sup> from the neutron the delayed signal. This naming convention come from the fact that the positron  
<sup>485</sup> will deposit its energy rather quickly (few ns) where the neutron will take a bit more time ( $\sim 236 \mu\text{s}$ ).

### <sup>486</sup> 2.3.2 Central Detector (CD)

<sup>487</sup> The central detector, composed of 20 ktons of Liquid Scintillator (LS), is the main part of JUNO. The  
<sup>488</sup> LS is contained in a spherical acrylic vessel supported by a stainless steel structure. The CD and  
<sup>489</sup> its structural support are submerged in a cylindrical water pool of 43.5m diameter and 44m height.  
<sup>490</sup> We're confident that the water pool provide sufficient buffer protection in every direction against the  
<sup>491</sup> rock radioactivity.

#### <sup>492</sup> Acrylic vessel

<sup>493</sup> The acrylic vessel is a spherical vessel of inner diameter of 35.4 m and a thickness of 120 mm. It is  
<sup>494</sup> assembled from 265 acrylic panels, thermo bonded together. The acrylic recipes has been carefully  
<sup>495</sup> tuned with extensive R&D to ensure it does not include plasticizer and anti-UV material that would  
<sup>496</sup> stop the scintillation photons. Those panels requires to be pure of radioactive materials to not  
<sup>497</sup> cause background. Current setup where the acrylic panels are molded in cleanrooms of class 10000,  
<sup>498</sup> let us reach a uranium and thorium contamination of <0.5 ppt. The molding and thermoforming

processes is optimized to increase the assemblage transparency in water to >96%. The acrylic vessel is supported by a stainless steel structure via supporting node (fig 2.6). The structure and the nodes are designed to be resilient to natural catastrophic events such as earthquake and can support many times the effective load of the acrylic vessel.

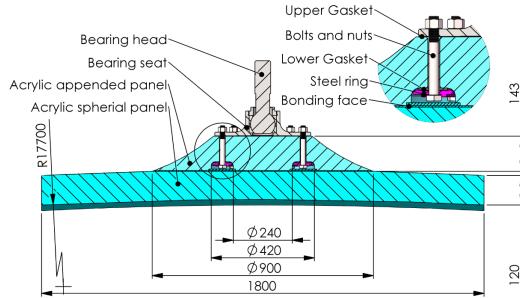


FIGURE 2.6 – Schematics of the supporting node for the acrylic vessel

### 503 Liquid scintillator

504 The Liquid Scintillator (LS) has a similar recipe as the one used in Daya Bay [23] but without gadolinium  
505 doping. It is made of three components, necessary to shift the wavelength of emitted photons to  
506 prevent their reabsorption and to shift their wavelength to the PMT sensitivity region as illustrated  
507 in Figure 2.7:

- 508 1. The detection medium, the *linear alkylbenzene* (LAB). Selected because of its excellent trans-  
509 parency, high flash point, low chemical reactivity and good light yield. Accounting for ~ 98%  
510 of the LS, it is the main component with which ionizing particles and gamma interact. Charged  
511 particles will collide with its electronic cloud transferring energy to the molecules, gamma will  
512 interact via compton effect with the electronic cloud before finally be absorbed via photoelectric  
513 effect.
- 514 2. The second component of the LS is the *2,5-diphenyloxazole* (PPO). A fraction of the excitation  
515 energy of the LAB is transferred to the PPO, mainly via non radiative process [24]. The PPO  
516 molecules de-excites in the same way, transferring their energy to the bis-MSB. The PPO makes  
517 for 1.5 % of the LS.
- 518 3. The last component is the *p-bis(o-methylstyryl)-benzene* (bis-MSB). Once excited by the PPO, it  
519 will emit photon with an average wavelength of ~ 430 nm (full spectrum in Figure 2.7) that  
520 can thus be detected by our photo-multipliers systems. It amount for ~ 0.5% of the LS.

521 This formula has been optimized using dedicated studies with a Daya Bay detector [23, 26] to reach  
522 the requirements for the JUNO experiment:

- 523 — A light yield / MeV of the amount of  $10^4$  photons to maximize the statistic in the energy  
524 measurement.
- 525 — An attenuation length comparable to the size of the detector to prevent losing photons during  
526 their propagation in the LS. The final attenuation length is 25.8m [27] to compare with the CD  
527 diameter of 35.4m.
- 528 — Uranium/Thorium radiopurity to prevent background signal. The reactor neutrino program  
529 require a contamination fraction  $F < 10^{-15}$  while the solar neutrino program require  $F < 10^{-17}$ .

530 The LS will frequently be purified and tested in the Online Scintillator Internal Radioactivity In-  
531 vestigation System (OSIRIS) [28] to ensure that the requirements are kept during the lifetime of the  
532 experiment, more details to be found in Section 2.5.2.

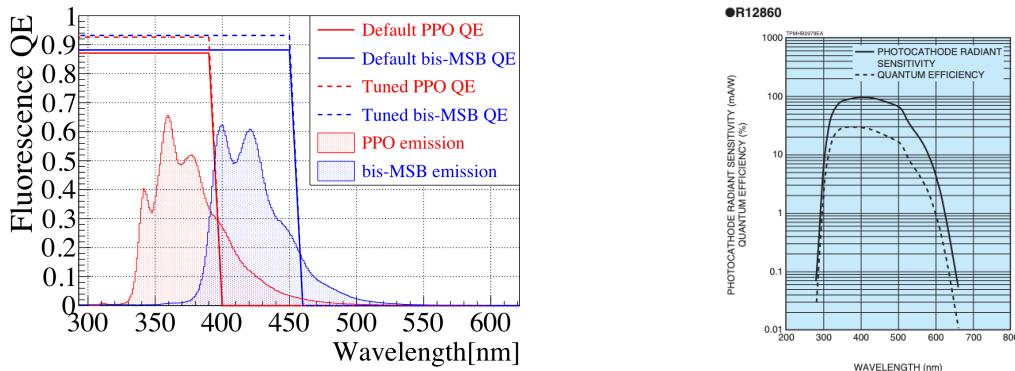


FIGURE 2.7 – On the left: Quantum efficiency (QE) and emission spectrum of the LAB and the bis-MSB [23]. On the right: Sensitivity of the Hamamatsu LPMT depending on the wavelength of the incident photons [25].

### 533 Large Photo-Multipliers Tubes (LPMTs)

534 The scintillation light produced by the LS is then collected by Photo-Multipliers Tubes (PMT) that  
 535 transform the incoming photon into an electric signal. As described in Figure 2.8, the incident  
 536 photons interact with the photocathode via photoelectric effect producing an electron called a Photo-  
 537 Electron (PE). This PE is then focused on the dynodes where the high voltage will allow it to be  
 538 multiplied. After multiple amplification the resulting charge - in coulomb [C] - is collected by the  
 539 anode and the resulting electric signal can be digitalized by the readout electronics from which the  
 540 charge and timing can be extracted.

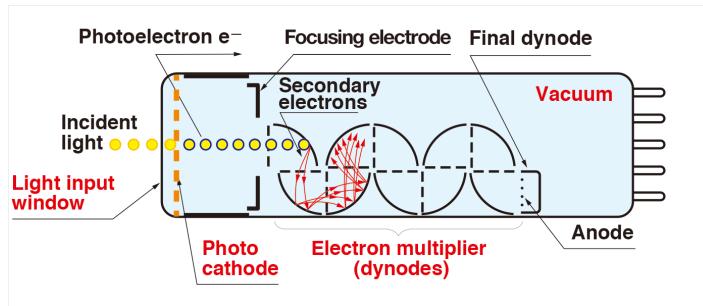


FIGURE 2.8 – Schematic of a PMT

541 The Large Photo-Multipliers Tubes (LPMT), used in the central detector and in the water pool, are  
 542 20-inch (50.8 cm) radius PMTs.  $\sim 5000$  dynode-PMTs [25] were produced by the Hamamatsu<sup>©</sup>  
 543 company and  $\sim 15000$  Micro-Channel Plate (MCP) [29] by the NNVT<sup>©</sup> company. This system is  
 544 the one responsible for the energy measurement with a energy resolution of  $3\%/\sqrt{E}$ , resolution  
 545 necessary for the mass ordering measurement. To reach this precision, the system is composed of  
 546 17612 PMTs quasi uniformly distributed over the detector for a coverage of 75.2% reaching  $\sim 1800$   
 547 PE/MeV or  $\sim 2.3\%$  resolution due to statistic, leaving  $\sim 0.7\%$  for the systematic uncertainties. They  
 548 are located outside the acrylic sphere in the water pool facing the center of the detector. To maintain  
 549 the resolution over the lifetime of the experiment, JUNO require a failure rate  $< 1\%$  over 6 years.

550 The LPMTs electronic are divided in two parts. One "near", located underwater, in proximity of the  
 551 LPMT to reduce the cable length between the PMT and early electronic. A second one, outside of the  
 552 detector that is responsible for higher level analysis before sending the data to the DAQ.

553 The light yield per MeV induce that a LPMT can collect between 1 and 1000 PE per event, a wide

dynamic range, causing non linearity in the PMT response that need to be understood and calibrated, see Section 2.4 for more details.

Before performing analysis, the analog readout of the LPMT need to be amplified, digitised and packaged by the readout electronics schematized in Figure 2.9. This electronic is splitted in two parts: *wet* electronic that are located near the LPMTs, protected in an Underwater Box (UWB) and the *dry* electronics located in deicated rooms outside of the water pool.

The LPMTs are connected to the UWB by groups of three. Each UWB contains:

- Three high voltage units, each one powering a PMT.
- A global control unit, responsible for the digitization of the waveform, composed of six analog-digital units that produce digitized waveform and a Field Programmable Gate Array (FPGA) that complete the waveform with metadatas such as the local timestamp trigger, etc... The FPGA also act as a data buffer when needed by the DAQ and trigger system.
- Additional memory in order to temporally store the data in case of sudden burst of the input rate (such as in the case of nearby supernovae).

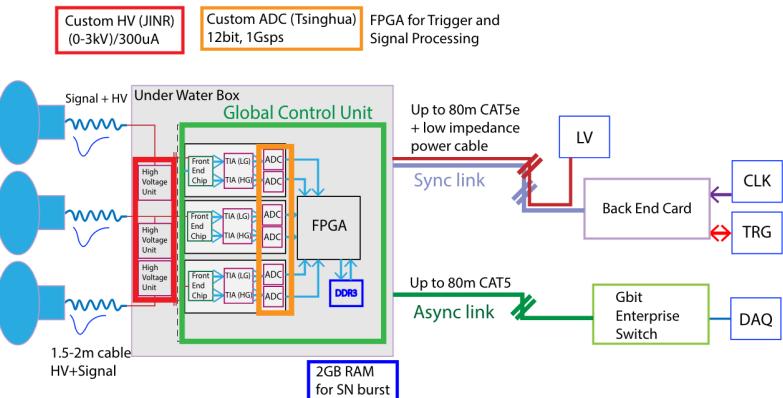


FIGURE 2.9 – The LPMT electronics scheme. It is composed of two part, the *wet* electronics on the left, located underwater and the *dry* electronics on the right. They are connected by Ethernet cable for data transmission and a dedicated low impedance cable for power distribution

The *dry* electronic synchronize the signals from the UWBs abd centralise the information of the CD LPMTs. It act as the Global Trigger by sending the UWB data to DAQ in the case if the LPMT multiplicity condition is fulfilled.

### 571 Small Photo-Multipliers Tubes (SPMTs)

The Small PMT (SPMTs) system is made of 3-inch (7.62 cm) PMTs. They will be used in the CD as a secondary detection system. Those 25600 SPMTs will observe the same events as the LPMTs, thus sharing the physics and detector systematics up until the photon conversion. With a detector coverage of 2.7%, this system will collect  $\sim 43$  PE/MeV for a final energy resolution of  $\sim 17\%$ . This resolution is not enough to measure the NMO,  $\theta_{13}$ ,  $\Delta m_{31}^2$  but will be sufficient to independently measure  $\theta_{12}$  and  $\Delta m_{21}^2$ .

The benefit of this second system is to be able to perform another, independent measure of the same events as the LPMTs, constituting the Dual Calorimetry useful for calibrationa and, as it we will explore in this thesis, for physics analysis. Due to the low PE rate, SPMTs will be running in photo-counting mode in the reactor range and thus will be insensitive to LPMT intrinsic effect (see

582 Section 2.4). Using this property, the intrinsic charge non linearity of the LPMTs can be measured by  
 583 comparing the PE count in the SPMTs and LPMTs [30]. Also, due to their smaller size and electronics,  
 584 SPMTs have a better timing resolutions than the LPMTs. At higher energy range, like supernovae  
 585 events, LPMTs will saturate where SPMTs due to their lower PE collection will to produce a reliable  
 586 measure of the energy spectrum.

587 The SPMTs will be grouped by pack of 128 to an UWB hosting their electronics as illustrated in Figure  
 588 2.10. This underwater box host two high voltage splitter boards, each one supplying 64 SPMTs, an  
 589 ASIC Battery Card (ABC) and a global control unit.

590 The ABC board will readout and digitize the charge and time of the 128 SPMTs signals and a FPGA  
 591 will joint the different metadata. The global control unit will handle the powering and control of the  
 592 board and will be in charge of the transmission of the data to the DAQ.

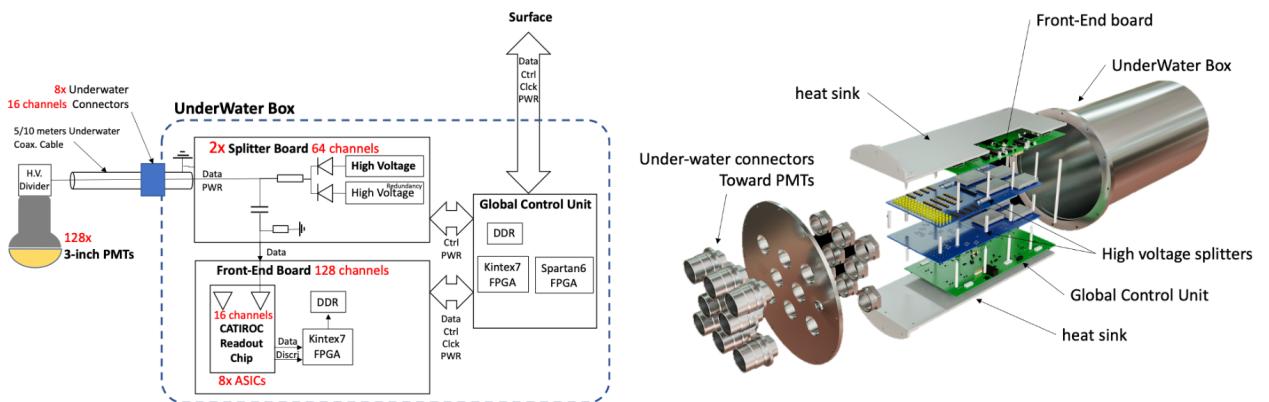


FIGURE 2.10 – Schematic of the JUNO SPMT electronic system (left), and exploded view of the main component of the UWB (right)

### 593 2.3.3 Veto detector

594 The CD will be bathed in constant background noise coming from numerous sources : the radioac-  
 595 tivity from surrounding rock and its own components or from the flux of cosmic muons. This  
 596 background needs to be rejected to ensure the purity of the IBD spectrum. To prevent a big part  
 597 of them, JUNO use two veto detector that will tag events as background before CD analysis.

### 598 Cherenkov in water pool

599 The Water Cherenkov Detector (WCD) is the instrumentation of the water buffer around the CD.  
 600 When high speed charged particles will pass through the water, they will produced cherenkov  
 601 photons. The light will be collected by 2400 MCP LPMTs installed on the outer surface of the CD  
 602 structure. The muons veto strategy is based on a PMT multiplicity condition. WCD PMTs are  
 603 grouped in ten zones: 5 in the top, 5 in the bottom. A veto is raised either when more than 19  
 604 PMTs are triggered in one zone or when two adjacent zones simultaneously trigger more than 13  
 605 PMTs. Using this trigger, we expect to reach a muon detection efficiency of 99.5% while keeping the  
 606 noise at reasonable level.

607 **Top tracker**

608 The JUNO Top Tracker (TT) is a plastic scintillator detector located on the top of the experiment (see  
 609 Figure 2.11). Made from plastic scintillator from OPERA [31] layered horizontally in 3 layers on the  
 610 top of the detector, the TT will be able to detect incoming atmospheric muons. With its coverage,  
 611 about 1/3 of the of all atmospheric muons that passing through the CD will also pass through the 3  
 612 layer of the detector. While it does not cover the majority of the CD, the TT is particularly effective  
 613 to detect muons coming through the filling chimney region which might present difficulties from the  
 other subsystems in some classes of events.

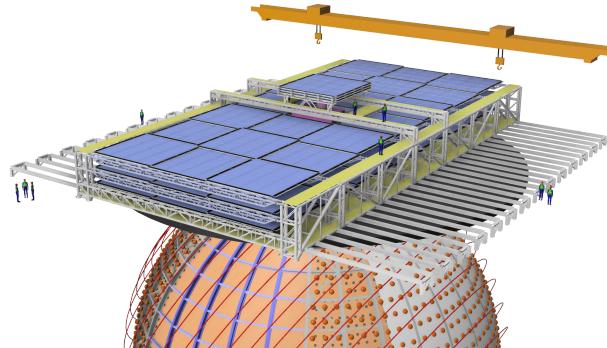


FIGURE 2.11 – The JUNO top tracker

614

615 **2.4 Calibration strategy**

616 The calibration is a crucial part of the JUNO experiment. The detector will continuously bath in  
 617 neutrinos coming from the close nuclear power plant, from other sources such as geo neutrinos,  
 618 the sun and will be exposed to background noise coming from atmospheric muons and natural  
 619 radioactivity. Because of this continuous rate, low frequency signal event, we need high frequency,  
 620 recognisable sources in the energy range of interest : [0-12] MeV for the positron signal and 2.2 MeV  
 621 for the neutron capture. It is expected that the CD response will be different depending on the type  
 622 of particle, due to the interaction with LS, the position on the event and the optical response of the  
 623 acrylic sphere (see Section 3.3). We also expect a non-linear energy response of the CD due to the LS  
 624 properties [23] but also due to the reponse of the LPMTs system when collecting a large amount of  
 625 PE [30].

626 **2.4.1 Energy scale calibration**

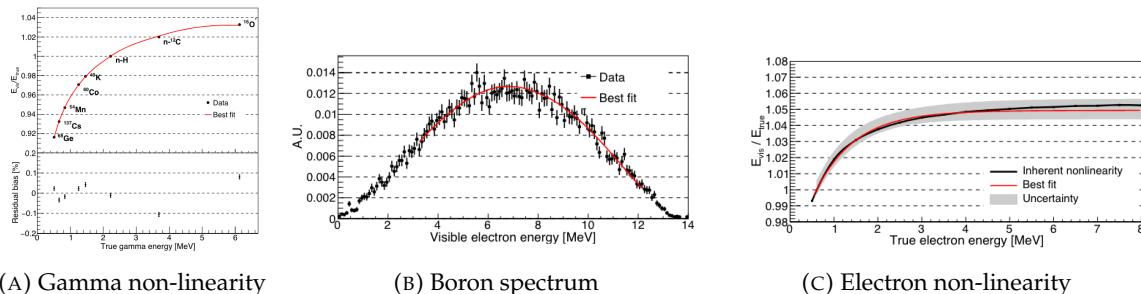
627 While electrons and positrons sources would be ideal, for a large LS detector thin-walled electrons  
 628 or positrons sources could lead to leakage of radionucleides causing radioactive contamination.  
 629 Instead, we consider gamma sources in the range of the prompt energy of IBDs. The sources are  
 630 reported in table 2.3.

631 For the  $^{68}\text{Ge}$  source, it will decay in  $^{68}\text{Ga}$  via electron capture, which will itself  $\beta^+$  decay into  $^{68}\text{Zn}$ .  
 632 The positrons will be absorbed by the enclosure so only the annihilation gamma will be released. In  
 633 addition,  $(\alpha, n)$  sources like  $^{241}\text{Am-Be}$  and  $^{241}\text{Am-}^{13}\text{C}$  are used to provide both high energy gamma  
 634 and neutrons, which will later be captured in the LS producing the 2.2 MeV gamma.

Sources / Processes	Type	Radiation
$^{137}\text{Cs}$	$\gamma$	0.0662 MeV
$^{54}\text{Mn}$	$\gamma$	0.835 MeV
$^{60}\text{Co}$	$\gamma$	$1.173 + 1.333$ MeV
$^{40}\text{K}$	$\gamma$	1.461 MeV
$^{68}\text{Ge}$	$e^+$	annihilation $0.511 + 0.511$ MeV
$^{241}\text{Am-Be}$	$n, \gamma$	neutron + 4.43 MeV ( $^{12}\text{C}^*$ )
$^{241}\text{Am-}^{13}\text{C}$	$n, \gamma$	neutron + 6.13 MeV ( $^{16}\text{O}^*$ )
$(n, \gamma)p$	$\gamma$	2.22 MeV
$(n, \gamma)^{12}\text{C}$	$\gamma$	4.94 MeV or $3.68 + 1.26$ MeV

TABLE 2.3 – List of sources and their process considered for the energy scale calibration

From this calibration we call  $E_{\text{vis}}$  the "visible energy" that is reconstructed by our current algorithms and we compare it to the true energy deposited by the calibration source. The results shown in Figure 2.12 show the expected response of the detector from calibration sources. The non-linearity is clearly visible from the  $E_{\text{vis}} / E_{\text{true}}$  shape. See [32] for more details.

FIGURE 2.12 – Fitted and simulated non linearity of gamma, electron sources and from the  $^{12}\text{B}$  spectrum. Black points are simulated data. Red curves are the best fits. Figures taken from [32].

## 2.4.2 Calibration system

The non-uniformity due to the event position in the detector (more details in Section 3.3) will be studied using multiples systems that are schematized in Figure 2.13. They allow to position sources at different location in the CD.

- For a one-dimension vertical calibration, the Automatic Calibration Unit (ACU) will be able to deploy multiple radioactive sources or a pulse laser diffuser ball along the central axis of the CD through the top chimney. The source position precision is less than 1cm.
- For off-axis calibration, a calibration source attached to a Cable Loop System (CLS) can be moved on a vertical half-plane by adjusting the length of two connection cable. Two set of CSL will be deployed to provide a 79% effective coverage of a vertical plane.
- A Guiding Tube (GT) will surround the CD to calibrate the non-uniformity of the response at the edge of the detector
- A Remotely Operated under-LS Vehicle (ROV) can be deployed to desired location inside LS for a more precise and comprehensive calibration. The ROV will also be equipped with a camera for inspection of the CD.

The preliminary calibration program is depicted in table 2.4.

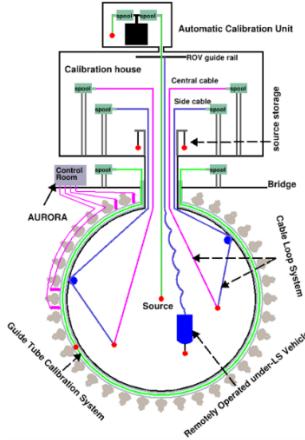


FIGURE 2.13 – Overview of the calibration system

Program	Purpose	System	Duration [min]
Weekly calibration	Neutron (Am-C)	ACU	63
	Laser	ACU	78
Monthly calibration	Neutron (Am-C)	ACU	120
	Laser	ACU	147
	Neutron (Am-C)	CLS	333
	Neutron (Am-C)	GT	73
Comprehensive calibration	Neutron (Am-C)	ACU, CLS and GT	1942
	Neutron (Am-Be)	ACU	75
	Laser	ACU	391
	$^{68}\text{Ge}$	ACU	75
	$^{137}\text{Cs}$	ACU	75
	$^{54}\text{Mn}$	ACU	75
	$^{60}\text{Co}$	ACU	75
	$^{40}\text{K}$	ACU	158

TABLE 2.4 – Calibration program of the JUNO experiment

### 655 2.4.3 Instrumental non-linearity calibration

656 One of the main interests of Dual Calorimetry is to calibrate away an instrumental effect called charge  
 657 non linearity (QNL), which will be described in more detail in Chapter 7.

658 In short, during a typical IBD event, between 0 and 100 PEs can be produced in a given LPMT  
 659 (depending on the position of the interaction and the positron energy). This is a large dynamic range.  
 660 When the number of PEs is high, the reconstruction of the LPMT charge can become inaccurate,  
 661 underestimating the actual number of PEs as illustrated in Figure 2.14. This QNL is difficult to  
 662 separate from other non linearities (like the non linearity in the LS photon yield as a function of the  
 663 deposit energy). In chapter 5 and 6 of this thesis [30], a calibration method that constitutes the core of  
 664 dual calorimetry are described. They are based on the comparisons between signals seen in LPMTs  
 665 and signals seen in SPMTs. In the latter system, due to its small angular coverage, individual SPMT  
 666 rarely see more than 1 PE per event, and therefore are essentially immune against QNL. The method  
 667 described in [30] uses a tunable light source covering the range of 0 to 100 PE per LPMT channel

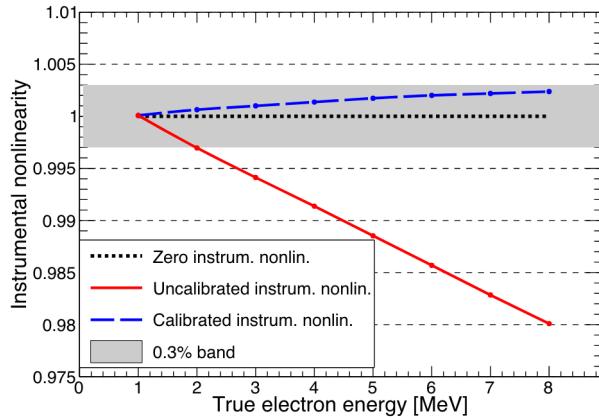


FIGURE 2.14 – Event-level instrumental non-linearity, defined as the ratio of the total measured LPMT charge to the true charge for events at the center of the detector. The solid red line represents event-level non-linearity without the channel-level correction in an extreme hypothetical scenario of 50% non-linearity over 100 PEs for the LPMTs. The dashed blue line represents that after the channel-level correction. The gray band shows the residual uncertainty of 0.3%, after the channel-level correction. Figure taken from [32].

## 668 2.5 Satellite detectors

669 As introduced in Section 2.1.1 and section 2.3.2, the precise knowledge and understanding of the  
 670 detector condition is crucial for the measurements of the NMO and oscillation parameters. Thus two  
 671 satellite detectors will be setup to monitor the experiment condition. TAO to monitor and understand  
 672 the  $\bar{\nu}_e$  flux and spectrum coming from the nuclear reactor and OSIRIS to monitor the LS response.

### 673 2.5.1 TAO

674 The Taishan Antineutrino Observatory (TAO) [19, 33] is a ton-level gadolinium doped liquid scin-  
 675 tillator detector that will be located near the Taishan-1 reactor. It aim to measure the  $\bar{\nu}_e$  spectrum at  
 676 very low distance (44m) from the reactor to measure a quasi-unoscillated spectrum. TAO also aim to  
 677 provide a major contribution to the so-called reactor anomaly [18]. Its requirement are to the level of  
 678 2 % energy resolution at 1 MeV.

#### 679 Detector

680 The TAO detector is close, in concept, to the CD of JUNO. It is composed of an acrylic vessel  
 681 containing 2.8 tons of gadolinium-loaded LS instrumented by an array of silicon photomultipliers  
 682 (SiPM) reaching a 95% coverage. To efficiently reduce the dark count of those sensors, the detector  
 683 is cooled to -50 °C. The  $\bar{\nu}_e$  will interact with the LS via IBD, producing scintillation light, that will  
 684 be detected by the SiPMs. From this signal the  $\bar{\nu}_e$  energy and the full spectrum reconstructed. This  
 685 spectrum will then be used by JUNO to calibrate the unoscillated spectrum, most notably the fission  
 686 product fraction that impact the rate and shape of the spectrum. A schema of the detector is presented  
 687 in Figure 2.15a.

### 688 2.5.2 OSIRIS

689 The Online Scintillator Internal Radioactivity Investigation System (OSIRIS) [28] is an ultralow back-  
 690 ground,  $20 \text{ m}^3$  LS detector that will be located in JUNO cavern. It aim to monitor the radioactive  
 691 contamination, purity and overall response of the LS before it is injected in JUNO. OSIRIS will  
 692 be located at the end of the purification chain of JUNO, monitoring that the purified LS meet the  
 693 JUNO requirements. The setup is optimized to detect the fast coincidences decay of  $^{214}\text{Bi} - ^{214}\text{Po}$   
 694 and  $^{212}\text{Bi} - ^{212}\text{Po}$ , indicators of the decay chains of U and Th respectively.

695 **Detector**

696 OSIRIS is composed of an acrylic vessel that will contains 17t of LS. The LS is instrumented by  
 697 a PMT array of 64 20 inch PMTs on the top and the side of the vessel. To reach the necessary  
 698 background level required by the LS purity measurements, in addition to being 700m underground  
 699 in the experiment cavern, the acrylic vessel is immersed in a tank of ultra pure water. The water is  
 700 itself instrumented by another array of 20 inch PMTs, acting as muon veto. A schema of the detector  
 701 is presented in Figure 2.15b.

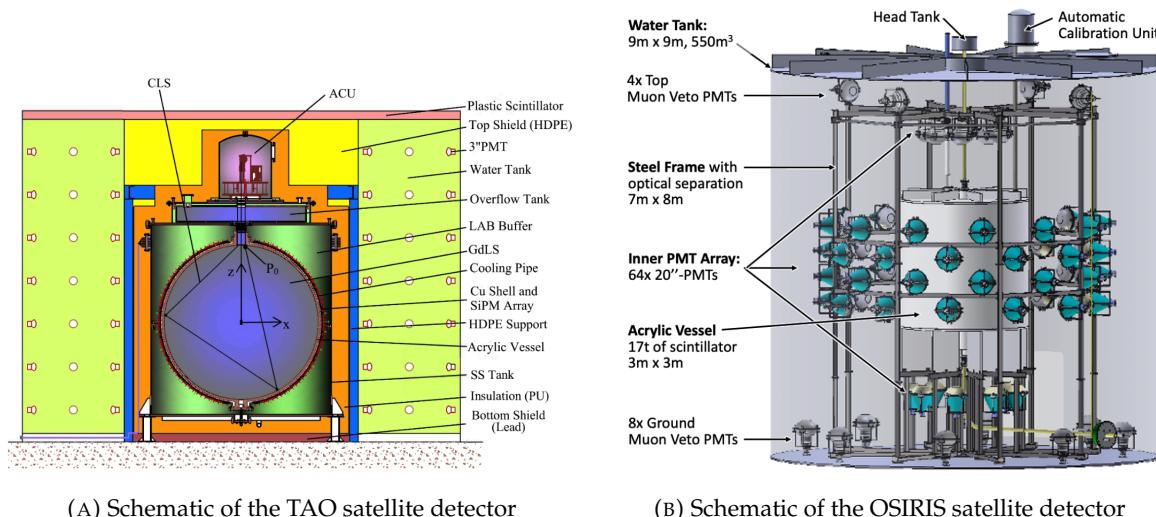


FIGURE 2.15

702 **2.6 Software**

703 The simulation, reconstruction and analysis algorithms are all packaged in the JUNO software,  
 704 subsequently called the software. It is composed of multiple components integrated in the SNiPER  
 705 [34] framework:

- 706 — Various primary particles simulators for the different kind of events, background and calibra-  
 707 tion sources.
- 708 — A Geant4 [35–37] Monte Carlo (MC) simulation containing the detectors geometries, a custom  
 709 optical model for the LS and the supporting structures of the detectors. The Geant4 simulation  
 710 integrate all relevant physics process for JUNO, validated by the collaboration. This step of the  
 711 simulation is commonly called *Detsim* and compute up to the production of photo-electrons in

the PMTs. The optics properties of the different materials and detector components have been measured beforehand to be used to define the material and surfaces in the simulation.

- An electronic simulation, simulating the response waveform of the PMTs, tracking it through the digitization process, accounting for effects such as non-linearity, dark noise, Time Transit Spread (TTS), pre-pulsing, after-pulsing and ringing of the waveform. It's also the step handling the event triggers and mixing. This step is commonly referenced as *ElecSim*.
- A waveform reconstruction where the digitized waveform are filtered to remove high-frequency white noise and then deconvoluted to yield time and charge informations of the photons hits on the PMTs. This step is commonly referenced as *Calib*.
- The charge and time informations are used by reconstruction algorithms to reconstruct the interaction vertex and the deposited energy. This step is commonly reported as *Reco*. See Section 3.3 for more details on the reconstruction.
- Once the singular events are reconstructed, they go through event pairing and classification to select IBD events. This step is named Event Classification.
- The purified signal is then analysed by the analysis framework which depend of the physics topic of interest. An introduction to the reactor  $\bar{n}u_e$  is presented in Section 2.7.

The steps Reco and Event Classification are divided into two category of algorithm. Fast but less accurate algorithms that are running during the data taking designated as the *Online* algorithms. Those algorithm are used to take the decision to save the event on tape or to throw it away. More accurate algorithms that run on batch of events designated *Offline* algorithms. They are used for the physics analysis. The Offline Reco will be one of the main topic of interest for this thesis.

## 2.7 Reactor anti-neutrino oscillation analysis

### 2.7.1 IBD samples selection

The  $\bar{\nu}_e$  coming from nuclear reactor will, for the most part, interact with proton, hydrogen nucleus, via Inverse Beta Decay (IBD). The first step of the oscillation analysis is to constitute a sample of IBD candidates, dominated by actual IBDs. The IBD interaction, schematised in Figure 2.5, will produce two particle, with differentiable signals.

The first signal comes from the positron slowdown and its annihilation with an electron of the LS. This is the *prompt* signal, happening a few ns after the IBD. The positron takes most of the  $\bar{\nu}_e$  kinetic energy, as detailed in Section 2.3.1.

The leftover kinetic energy is taken by the neutron that, after thermalisation in the LS, will be captured by an hydrogen and produce a 2.2 MeV gamma, or by a carbon emitting a 4.9 MeV gamma. This is the *delayed* signal, happening  $\sim 236 \mu\text{s}$  after the IBD. This second mono-energetic event serve as a marker for the IBD.

The IBD selection is thus based on the selection of a prompt event, with an energy between 0.8 and 12 MeV, and a delayed event with an energy in the ranges [1.9, 2.5] MeV or [4.4, 5.5] MeV. Those two signal needs to be in a 1 ms time window and within 1.5 m from each other. Additionally the two signal needs to be in a radius of 17.2m from the detector center (0.5 m from the edge) to protect from accidental background formed by two uncorrelated signals [38]. Those values will be further refined after once JUNO data-taking starts.

In addition, specials veto are setup to protect from cosmic muons and their aftermath. The details of those veto and selection can be found in [38].

754 The expected rate and selection efficiency on IBD can be found in table 2.5. After these selection, the  
 755 residual background, including  $\bar{\nu}_e$  coming from other sources than the reactor can be found in table  
 756 2.6.

Selection Criterion	Efficiency [%]	IBD Rate [day <sup>-1</sup> ]
All IBDs	100.0	57.4
Fiducial Volume	91.5	52.5
IBD Selection	98.1	51.5
Energy Range	99.8	-
Time Correlation ( $\Delta T_{p-d}$ )	99.0	-
Spatial Correlation ( $\Delta R_{p-d}$ )	99.2	-
Muon Veto (Temporal + Spatial)	91.6	47.1
Combined Selection	82.2	47.1

TABLE 2.5 – Summary of cumulative reactor antineutrino selection efficiencies. The reported IBD rates (with baselines <300 km) refer to the expected events per day after the selection criteria are progressively applied. Table taken from [38]

Backgrounds	Rate [day <sup>-1</sup> ]	B/S [%]
Geoneutrinos	1.2	2.5
World reactors	1.0	2.1
Accidentals	0.8	1.7
<sup>9</sup> Li/ <sup>8</sup> He	0.8	1.7
Atmospheric neutrinos	0.16	0.34
Fast neutrons	0.1	0.21
<sup>13</sup> C( $\alpha$ ,n) <sup>16</sup> O	0.05	0.01
Total backgrounds	4.11	8.7

TABLE 2.6 – Expected background rates, background to signal ratio (B/S), and rate and shape uncertainties. The B/S ratio is calculated by using the IBD signal rate of 47.1/day. Table taken from [38]

757 Once a sample is obtained, the oscillation analysis will consist essentially on the fit of a spectrum  
 758 model to the spectrum observed in the selected sample. More specifically, the spectrum under  
 759 analysis is the spectrum of the reconstructed visible energy of the positron :  $E_{vis}^{vis}$ . The reconstruction  
 760 is presented in detail in Section 3.3. For 6 years of data taking, it will resemble that on Figure 2.3.  
 761 In the next sections, I describe the fit procedures developed in JUNO. This will be the occasion to  
 762 introduce notions useful for Chapter 7. Besides, I'll also describe the versions of the fit used in this  
 763 Chapter 7.

### 764 2.7.2 Synthetic overview of fit procedures developed at JUNO

765 Several fit procedures are being developed by JUNO collaborators (half a dozen of groups work  
 766 in parallel within the collaboration). We do not have the ambition of a thorough description here.  
 767 Instead, we try to introduce the main elements useful to the reader to understand JUNO's future  
 768 results, and the fit procedures used Chapter 7.

769 In most cases, the fit is a binned fit to the histogrammed spectrum of  $E_{vis}^{vis}$ , like the one in Figure 2.3.  
 770 It is based on the minimization of a  $\chi^2$  test statistics. Generically, it can be written this way :

$$\chi^2 = (\mathbf{T}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathbf{D})^T V^{-1} (\mathbf{T}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathbf{D}) + \chi^2_{nuis}(\boldsymbol{\eta}) \quad (2.3)$$

771 where the components of data vector  $\mathbf{D}$  are the number of events found in individual bins of the

fitted histogram,  $T(\theta, \eta)$  is the vector of the predicted number of entries in each bins. This prediction is the integration over the width of the bins of the spectrum model for a given NMO (described latter in this section).

This model depends on the oscillation parameters  $\theta = (\Delta m_{21}^2, \sin^2(2\theta_{12}), \Delta m_{31}^2, \sin^2(2\theta_{13}))$ , and on nuisance parameters  $\eta$  involved in the fit model and associated with systematic uncertainties. Uncertainties are treated in two ways : statistical and some of the systematic uncertainties are accounted for via the covariance matrix  $V = V_{stat} + V_{syst}$ ; remaining systematic uncertainties are treated via the penalty term  $\chi^2_{nuis}$ , which is written this way :

$$\chi^2_{nuis}(\eta) = (\eta - \bar{\eta})^T \cdot V_\eta^{-1}(\eta) \cdot (\eta - \bar{\eta}) \quad (2.4)$$

where  $\bar{\eta}$  is the vector containing the most probable values of the nuisance parameters according to our knowledge prior to the fit, and where  $V_\eta$  is the covariance matrix accounting of the uncertainty on these values, and the potential correlations between them. In principles, a likelihood could be used instead of a  $\chi^2$ . However, some of the systematic uncertainties are not trivial to parameterize, therefore treating them as nuisance parameters is not trivial.

An example of nuisance parameters are the  $A$ ,  $B$  and  $C$  parameters of equation 7.19, which can be used to describe the resolution on the reconstructed energy. The fit model leading to  $T(\theta, \eta)$  indeed incorporates this resolution.

## 788 Treatment of uncertainties

Differences between various fit procedures developed within JUNO often lies in the choice of the systematic uncertainties that are treated via  $V$  or  $\chi^2_{nuis}(\eta)$ . Among the reasons behind these differences is the necessity to compare several approaches to ensure the robustness JUNO's oscillation analysis results. This approach was already adopted in the recent evaluations of JUNO's potential [9, 38]. Studies carried out so far at Subatech assumes a treatment entirely via  $V$ .

Other differences lies in the choice of the way to evaluate  $V_{stat}$ . Two common approaches used in  $\chi^2$  fit are the Neyman and the Pearson approaches. If the size of the fitted sample is high enough, the variation of  $D_i$ , the number of entries in bin  $i$ , around its true expectation value  $\bar{D}_i$  is  $\sqrt{\bar{D}_i}$ . To evaluate this number, the Neyman approach uses simply the number of entries observed in the sample under analysis :  $\sqrt{D_i}$ . The Pearson approach uses the prediction by the fit model :  $\sqrt{T(\theta, \eta)_i}$ .

Both cases are approximations which lead to biases that are not tolerable given the precision JUNO must aim at for a successful oscillation analysis. To reduce this bias, most of JUNO groups employ the "Combined Neyman Pearson" approach introduced in [39]. Schematically, it consists on combining both approaches :  $(V_{stat})_{ii} = 3 / \left( \frac{1}{T(\theta, \eta)_i} + \frac{2}{D_i} \right)$ . Weights in this relation are chosen in order to cancel typical biases. The validity of this method is not guaranteed universally. In particular, limitations appear when a complex systematic matrix  $V_{syst}$  is added to  $V_{stat}$ .

This is the case in the approach followed at Subatech, were all sources of systematic uncertainties are treated via this matrix. Dedicated studies run at Subatech observed biases in the fitted oscillation parameters using CNP in this case. Subatech's group therefore adopted another approach (verified to be unbiased).

Originally, fitting the  $E_{vis}^{e+}$  spectrum should mean maximising a likelihood, equal to the product over all bins of the probabilities to find  $D_i$  in bin  $i$ . With a large enough samples, this product tends to a multidimensional gaussian (one dimension per bin) :

$$\mathcal{L} = 2\pi^{-\frac{N}{2}} |V|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{D} - \mathbf{T}(\theta, \eta))^T V^{-1} (\mathbf{D} - \mathbf{T}(\theta, \eta))} \quad (2.5)$$

812 Replacing  $\mathcal{L}$  by  $-2 \ln \mathcal{L}$  one obtains :

$$\chi_{PV}^2 = (\mathbf{T}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathbf{D})^T V^{-1} (\mathbf{T}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \mathbf{D}) + \ln(|V|) \quad (2.6)$$

813 where  $V$  is the total covariance matrix with its statistical component evaluated according to the  
 814 Pearson approach. The  $\ln |V|$  term, often neglected in  $\chi^2$  fits, ensures that biases, essentially related  
 815 to the normalisation of the fitted distribution, are avoided. This "PearsonV"  $\chi^2$  is the one that we  
 816 minimize in the fits used in Chapter 7.

817 Another difference between the various procedures developed at JUNO is the choice of the spectrum  
 818 range and binning. So far, at Subatech, we use an histogram defined between 0.8 and 9 MeV, and a  
 819 regular binning involving 20 keV wide bins.

820 **Joint fit of JUNO and TAO spectra**

821 Another difference between the various fit procedures developed in the collaboration is the inclusion  
 822 of the data collected by TAO (see Section 2.5.1). The spectrum prediction  $\mathbf{T}(\boldsymbol{\theta}, \boldsymbol{\eta})$  involves predictions  
 823 on the differential flux of  $\bar{\nu}_e$  as a function of  $E_{\bar{\nu}_e}$  produced in reactors. This is one of the main  
 824 systematic uncertainties affecting the oscillation analysis. This can be constrained using the data  
 825 of TAO. An efficient way to use them is via a simultaneous fit, which will constrain the part of the  $\boldsymbol{\eta}$   
 826 parameters related to the reactor predictions. In this case, equation 2.3 becomes :

$$\chi^2 = \sum_d \left( \mathbf{T}^d(\boldsymbol{\theta}^d, \boldsymbol{\eta}) - \mathbf{D}^d \right)^T V^{-1} \left( \mathbf{T}^d(\boldsymbol{\theta}^d, \boldsymbol{\eta}) - \mathbf{D}^d \right) + \chi_{nuis}^2(\boldsymbol{\eta}) \quad (2.7)$$

827 where the  $d$  superscript stands for the spectrum measured in JUNO or TAO.

828 Finally, it must be noted that JUNO's sensitivity to  $\sin^2(2\theta_{13})$  is too weak for a competitive measure-  
 829 ment. In most versions of the oscillation analyses carried out within JUNO, it will be considered as a  
 830 nuisance parameter. In practice, the various  $\chi^2$ 's presented earlier will receive an additional term :

$$\chi_{\sin^2(2\theta_{13})}^2 = \frac{(\sin^2(2\theta_{13}) - \overline{\sin^2(2\theta_{13})})^2}{\sigma_{\sin^2(2\theta_{13})}^2} \quad (2.8)$$

831 where  $\overline{\sin^2(2\theta_{13})}$  and the denominators can be provided, for instance, by the world average on this  
 832 parameter.

833 **2.7.3 The spectrum model and sources of systematic uncertainties**

834 The  $E_{vis}^{e+}$  spectrum observed in data (Fig 2.3) is the sum of the IBD spectrum and of the various  
 835 backgrounds spectra (see table 2.6). The spectrum prediction  $\mathbf{T}(\boldsymbol{\theta}, \boldsymbol{\eta})$  is therefore the sum of IBD and  
 836 backgrounds predictions. The latter are provided by MC simulations. The former results from the  
 837 theoretical description of the series of phenomena that lead to the observed IBD spectrum. In a given  
 838 bin  $i$ , it can be expressed this way :

$$T^i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_j C_{ij}^{E_{rec}} \int_{E_j^{vis}}^{E_{j+1}^{vis}} dE^{vis} \int_{-1}^1 d\cos\theta \Phi(E^\nu) \frac{d\sigma}{d\cos\theta}(E^\nu, \cos\theta) \frac{dE^\nu}{dE^{dep}} \frac{dE^{dep}}{dE^{vis}} \quad (2.9)$$

839 In the above equation, 4 kinds of energies appears: following the IBD, the antineutrino energy  $E^\nu$  is  
 840 quasi entirely transferred to the positron, of energy  $E_e$ . It eventually annihilates, so the actual energy  
 841 released in the LS is  $E_{dep}$ , which includes the mass of the annihilated electron. The production optical

photons is not linear in  $E_{dep}$  (see Section 2.4), so that the visible energy (that will be reconstructed) is  $E_{vis}$ . This reconstruction comes with resolution effects, leading to  $E_{rec}$ .

Equation 2.9 describe the passage from the original differential flux (as a function of  $E^\nu$ ) of antineutrinos reaching the detector to the reconstructed spectrum:

- $\Phi(E^\nu)$  is the differential antineutrino flux reaching JUNO.
- $\frac{d\sigma}{d\cos\theta}(E^\nu, \cos\theta)$  account for the IBD cross section, which depends on the antineutrino energy and on the incidence angle.
- The last two terms of the integrand are the differential relations linking  $E^\nu$ ,  $E^{dep}$  and  $E^{vis}$ .
- Reconstruction effects are described via  $C_{ij}^{rec}$ 's, that make the link between the true and reconstructed visible energy. In a simple case, it is equivalent to a convolution product. The matrix formalism here prepares the fact that a realistic analysis might employ a more empirical way, based on MC.

854

The differential flux is expressed this way:

$$\Phi(E^\nu) = \sum_r \left( \frac{\mathcal{P}_{\bar{\nu}_e \rightarrow \bar{\nu}_e}(E^\nu, L_r)}{4\pi L_r^2} \frac{W_r}{\sum_i f_{i,r} e_i} \sum_i f_{i,r} s_i(E^\nu) \right) \quad (2.10)$$

856 where:

- $\mathcal{P}_{\bar{\nu}_e \rightarrow \bar{\nu}_e}(E^\nu, L_r)$  is the antineutrino survival probability at distance  $L_r$  from the production point in reactor  $r$ , dictated by the oscillation probability.
- $e_i$  stands for the mean energy released per fission for isotope  $i$ .
- $W_r$  is the thermal power of reactor  $r$ .
- $f_{i,r}$  is the fission fraction in reactor  $r$  of isotope  $i$  among the four.
- $s_i(E^\nu)$  is the  $\bar{\nu}_e$  energy spectrum - at emission point - per fission for each isotope, as emitted by the reactor.

864

## 865 Sources of systematic uncertainties

The numerous quantities appearing in the spectrum model embody a good part of the systematic uncertainties. Among the leading contributions are those related to the knowledge of the reactor related quantities. Of importance are also the uncertainties related to the modelling of the non linearity of the photon emission (passage from  $E^{dep}$  to  $E^{vis}$ ) and of the reconstruction resolution. The shape and rate of the backgrounds are also a leading source of systematic uncertainties. The uncertainty on IBD selection efficiency also has a notable role.

## 872 Sensitivities to NMO and oscillation parameters

JUNO will start taking data in 2025. During the months and years to come, oscillation analyses will naturally be optimized regularly. What we described here represent the state of the art mid 2024, and was used for the sensitivity studies published in [9, 38] and are presented in table 2.7

	Central Value	PDG 2020	100 days	6 years	20 years
$\Delta m_{31}^2 (\times 10^{-3} \text{eV}^2)$	2.5283	$\pm 0.034$ (1.3%)	$\pm 0.021$ (0.8%)	$\pm 0.0047$ (0.2%)	$\pm 0.0029$ (0.1%)
$\Delta m_{21}^2 (\times 10^{-3} \text{eV}^2)$	7.53	$\pm 0.18$ (2.4%)	$\pm 0.074$ (1.0%)	$\pm 0.024$ (0.3%)	$\pm 0.017$ (0.2%)
$\sin^2 \theta_{12}$	0.307	$\pm 0.013$ (4.2%)	$\pm 0.0058$ (1.9%)	$\pm 0.0016$ (0.5%)	$\pm 0.0010$ (0.3%)
$\sin^2 \theta_{13}$	0.0218	$\pm 0.0007$ (3.2%)	$\pm 0.010$ (47.9%)	$\pm 0.0026$ (12.1%)	$\pm 0.0016$ (7.3%)

TABLE 2.7 – A summary of precision levels for the oscillation parameters. The reference value (PDG 2020 [40]) is compared with 100 days, 6 years and 20 years of JUNO data taking.

### 876 Asimov studies

877 To study the behavior and performance of fit procedures with enough realism, one should perform  
 878 fits to a large number of toy spectra, generated with a number events equal to what one expects in  
 879 real data, for the given exposure under consideration. This allows to study the impact of realistic  
 880 statistical fluctuations. This is, however, time consuming, since thousands of spectra have to be  
 881 generated and fitted.

882 When subtle details are not crucial, another approach is possible to estimate sensitivities to the NMO  
 883 and oscillation parameters, as well as (for instance) to verify the technical implementation of fitter  
 884 (as we will do in Chapter 7 for the implementation of the joint fit). It consists on generating only 1  
 885 pseudo-data sample, where the content of each bin  $D^i$  is set to the predicted value  $T^i$ , computed with  
 886 a reasonable choice for the values of the model parameters (for instance, with the recent PDG values  
 887 for the oscillation parameters). This is equivalent to a spectrum with fluctuations. It provides valid  
 888 sensitivities if the expected statistics in the real data sample is high enough in each bin to assume a  
 889 gaussian behavior.

### 890 2.7.4 Versions of the fit used in this thesis

891 In Chapter 7, we'll study the potential of a particular application of Dual Calorimetry, call "Dual  
 892 Calorimetry with neutrino oscillation." This approach require to perform fits to the  $E^{vis}$  spectrum  
 893 reconstructed with the LPMT system, with the SPMT system, and a joint fit to both spectra.

894 In the two former cases, the PearsonV  $\chi^2$  introduced above will be used. In the latter case, it will  
 895 be extended in the following way : The  $D$  data vector now possess 820 elements. Indeed, the fit is  
 896 performed to a joint spectrum, where the LPMT spectrum is juxtaposed with the SPMT spectrum  
 897 (see Figure 2.16).

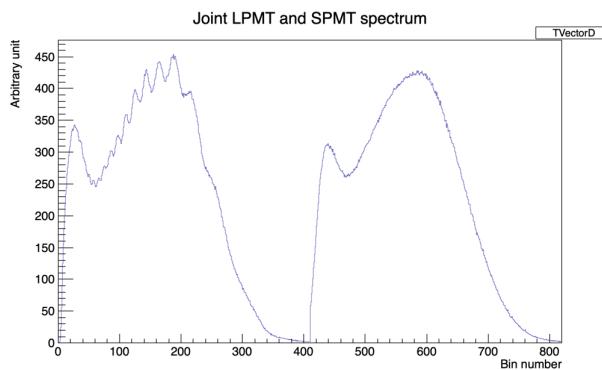


FIGURE 2.16 – Illustration of the spectrum considered when joint fitting

898 The prediction vector  $T(\theta^d, \eta)$  is naturally extended in the same way. Its components 1 to 410  
 899 predict the number of entries in the LPMT part of the LPMT+SPMT joint spectrum, while its com-  
 900 ponents from 411 to 820 predict the contents of the SPMT part. Note that the list of oscillation  
 901 parameters in  $T_{411}(\theta^d, \eta)$  to  $T_{820}(\theta^d, \eta)$  is the same as usual. However,  $T_1(\theta^d, \eta)$  to  $T_{410}(\theta^d, \eta)$  2  
 902 additional parameters,  $\delta \sin^2(2\theta_{12})$  and  $\delta \Delta m_{21}^2$ , are added to the corresponding oscillation parameters  
 903 to account for a potential unexpected problem in the LPMT reconstruction or calibration.

904 In the case of this joint fit, the covariance matrix  $V$  is extended to a  $(820 \times 820)$  matrix. It is a central  
 905 element of this study, as will be explained in Chapter 7, since the LPMT and SPMT data spectrum  
 906 are correlated, even at the statistical level. The determination of this matrix will be an important and  
 907 original point.

908 Fits will be performed to an histogram spectrum defined over the 0.8-9 MeV range, with a flat binning  
 909 (20 keV wide bins), often restricted to the 335 lowest  $E^{vis}$  bins.

910 In this Section 2.7, we have provided a theoretical description of the fit procedures developed at  
 911 JUNO. Software frameworks are necessary to use them in practice. The framework developed at  
 912 Subatech will be described in Chapter 7.

### 913 2.7.5 Physics results

914 The oscillation parameters are directly extracted from the minimization procedure and the error can  
 915 be estimated directly from the procedure. For the NMO, the data are fitted under the two assumption  
 916 of NO and IO. The difference in  $\chi^2$  give us the preferred ordering and the significance of our test.  
 917 Latest studies show that the precision on oscillation parameters after six year of data taking will be  
 918 of 0.2%, 0.3%, 0.5% and 12.1% for  $\Delta m_{31}^2$ ,  $\Delta m_{21}^2$ ,  $\sin^2 \theta_{12}$  and  $\sin^2 \theta_{13}$  respectively [9]. The expected  
 919 sensitivity to mass ordering is  $3\sigma$  after 6.5 years [41].

## 920 2.8 Summary

921 JUNO is one the biggest new generation neutrino experiment. Its goal, the measurements of oscil-  
 922 lation parameters with unprecedented precision and an NMO preference at the 3 sigma confidence  
 923 level, needs an in depth knowledge and understanding of the detector and the physics at hand. The  
 924 characterisation and calibration of the detector are of the utmost importance and the understanding  
 925 of the detector response in its resolution and bias is capital to be able to correctly carry the high  
 926 precision physics analysis of the neutrino oscillation.

927 In this thesis, I explore the usage of data-driven reconstruction methods to validate and optimize the  
 928 reconstruction of IBD events in JUNO in the chapters 4, 5 and 6 and the usage of the dual calorimetry  
 929 in the detection of possible mis-modelisation in the theoretical spectrum 7.

930 **Chapter 3**

931 **Introduction to the reconstruction  
methods and algorithms used in this  
thesis**

932

933

934 “I have the shape of a human being and organs equivalent to those of a  
935 human being. My organs, in fact, are identical to some of those in a  
936 prosthethized human being. I have contributed artistically, literally, and  
937 scientifically to human culture as much as any human being now  
938 alive. What more can one ask?”

Isaac Asimov, *The Complete Robot*

935 **Contents**

---

936 <b>3.1 Core concepts in machine learning and neural networks</b> . . . . .	38
937        3.1.1 Boosted Decision Tree (BDT) . . . . .	38
938        3.1.2 Artificial Neural Network (NN) . . . . .	39
939        3.1.3 Training procedure . . . . .	40
940        3.1.4 Potential pitfalls . . . . .	43
941 <b>3.2 Neural networks architectures</b> . . . . .	46
942        3.2.1 Fully Connected Deep Neural Network (FCDNN) . . . . .	46
943        3.2.2 Convolutional Neural Network (CNN) . . . . .	47
944        3.2.3 Graph Neural Network (GNN) . . . . .	49
945        3.2.4 Adversarial Neural Network (ANN) . . . . .	51
946 <b>3.3 State of the art of the Offline IBD reconstruction in JUNO</b> . . . . .	51
947        3.3.1 Interaction vertex reconstruction . . . . .	52
948        3.3.2 Energy reconstruction . . . . .	55
949        3.3.3 Machine learning for reconstruction . . . . .	58
950 <b>3.4 Conclusion</b> . . . . .	61

---

951 Machine Learning (ML) and more specifically Neural Network (NN) are families of data-driven  
952 algorithms. They are used in a wide variety of domains including natural language processing,  
953 computer vision, speech recognition and, the subject of this thesis, scientific studies.

954 Machine learning models aim to learn underlying patterns from finite datasets in order to make  
955 general predictions or classifications. For example, in our case, it could be an algorithm that would  
956 differentiate the nature of a particle interacting in the liquid scintillator, between a positron and an  
957 electron, based on the readout charge and time ( $Q, t$ ) of the 17612 LPMT of the JUNO experiment.  
958 During a first training phase, it would learn the discriminative features between the two in the 35224-  
959 dimensional charge and time distribution, built from samples of  $e^+$  and  $e^-$  events.

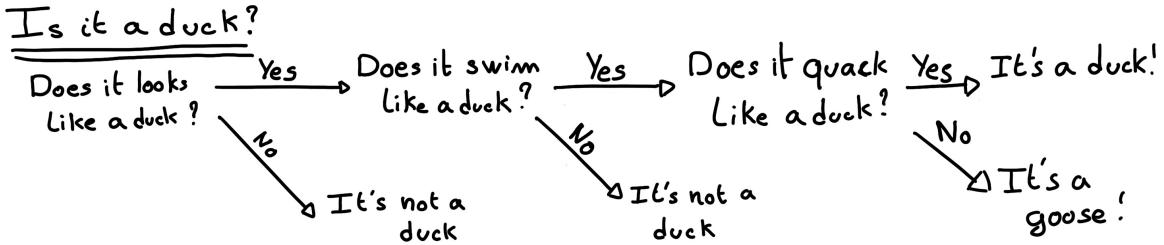


FIGURE 3.1 – Example of a BDT that determine if the given object is a duck

964 It extracts essential features from a highly complex and multi-dimensional dataset that describe the  
 965 physical interactions: a three body energy deposition (the positron and two annihilation gammas)  
 966 and the single deposit from an electron.

967 Ideally, the algorithm would learn to recognize those informations on its own, regardless of the input  
 968 size and complexity. In practice, however, these algorithms are guided by human design through  
 969 their architectures and training conditions. We can still hope that they can use more thoroughly the  
 970 detector informations while traditional methods are often subject to assumptions or simplifications  
 971 to make the task easier (see for instance the algorithm in Section 3.3).

972 The role of machine learning algorithms has expanded rapidly in the past decade, either as the  
 973 main or secondary algorithm for a wide variety of tasks: event reconstruction, event classification,  
 974 waveform reconstruction and so on. In particular in domains where the underlying physic and  
 975 detector processes are complex and highly dimensional, and when large amount of data must be  
 976 processed quickly.

977 This chapter present an overview of the different kind of machine learning methods and neural  
 978 networks that will be discussed in this thesis, and the state of the art of the reconstructions methods  
 979 in JUNO our ML algorithms will be compared to.

## 980 3.1 Core concepts in machine learning and neural networks

981 In this section, we discuss the core concepts in machine learning that will be used thorough this the-  
 982 sis. We place particular emphasis on Neural Networks, as it's the family of the algorithms described  
 983 in chapters 4, 5 and 6.

### 984 3.1.1 Boosted Decision Tree (BDT)

985 One of the most classic machine learning algorithm used in particle physics is Boosted Decision Tree  
 986 (BDT) [42] (or more recently Gradient Boosting Machine [43]).

987 BDTs operate by making a series of decisions based on a set of input features, with each decision  
 988 represented as a node in the tree. Each decision point, or node, takes its decision based on a set of  
 989 trainable parameters leading to a subtree of decisions. The process is repeated until it reach the final  
 990 node, yielding the prediction. A simplistic example is given in Figure 3.1.

991 The training procedure follows a reward-based approach where the algorithm predictions are com-  
 992 pared to the true outcomes. During the training phase the prediction of the BDT is compared to a  
 993 known truth about the data. The score is then used to backpropagate corrections to the parameters  
 994 of the tree. Modern BDT use gradient boosting where the gradient of the loss is calculated for each  
 995 of the BDT parameters. Following the gradient descent, we can reach the, hopefully, global minima  
 996 of the loss for our set of parameters.

### 3.1.2 Artificial Neural Network (NN)

One of the modern ML family is the Neural Network, historical name as their design was inspired by the behaviour of biological neurons in the brain. As schematized in Figure 3.2, the input, output

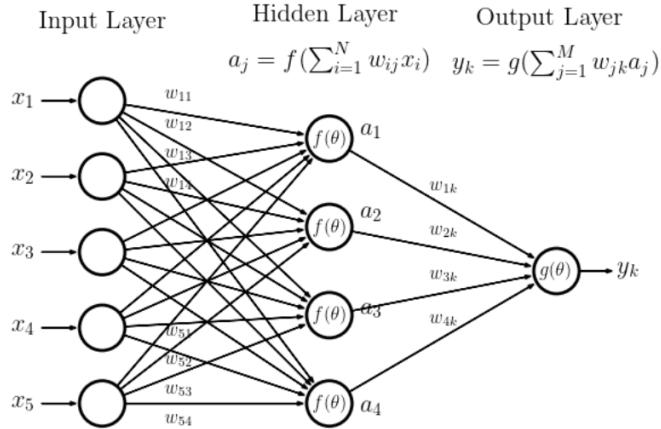


FIGURE 3.2 – Schema of a simple neural network

and steps inside the NN is described as neuron *layers*. The neurons of the layers take as input a set of values from the preceding layer, here the  $a_i$  takes every informations of the  $x_i$  input layer, and aggregate those values following learnable *parameters*  $w_{ij}$ . In the example in Figure 3.2, fully connected layers are used, meaning that each neuron in one layer is connected to every neuron in the previous layer.

The aggregation procedure is core of defining the architecture of the NN. The different architectures used in this thesis will be discussed in Section 3.2. The process is repeated until reaching the output layer.

For example, let's take the network in Figure 3.2 and say that  $a_1$ ,  $a_2$  and  $a_3$  are the neurons of the output layer. We try to produce a vertex reconstruction algorithm that will approach the charge barycentre. Let's limit the input  $x_i$  to the charge of the  $i$ th PMT, one of the solution is to aggregate on  $a_1$  the  $x$  coordinate of the barycenter. The network would thus adapt the  $w_{i1}$  parameters so they correspond to the  $x$  coordinates of the  $i$ th PMT. Same for the  $y$  and  $z$  coordinate on  $a_2$  and  $a_3$  respectively.

The layers used in the example above are designated as *Fully connected* layers, where every neurons of the layer is connected to the every neurons of the preceding layer. The layer can be expressed using the Einstein summation and in bold the learnable parameters

$$O_j = I_i + \mathbf{W}_j^i \quad (3.1)$$

where  $O_j$  is the output neurons vector (the  $a_i$ ),  $I_i$  is the preceding layer neurons vector (the  $x_i$ ) and  $\mathbf{W}$  is the parameters, or weights, matrix (composed of the  $w_{ij}$ ). In practice, this fully connected layer is often adjoined a bias  $B$  and an *activation function*  $F$ .

$$I_j = F(I_i \mathbf{W}_j^i + \mathbf{B}_j) \quad (3.2)$$

This is the fundamental component of the Fully Connected Deep NN (FCDNN) family presented in Section 3.2.1.

This description of neural networks as layers introduce the principles of *depth* and *width*, the number

of layers in the NN and the number of neurons in each layer respectively. Those quantities that not directly used for the computation of the results but describes the NN or its training are designated as *hyperparameters*.

Now we just need to adapt the parameters so that this network learn that  $w_{ij}$  are the PMT coordinate. We describe the space produced by the parameters of the network as the *parameter phase space* or *latent space*. The optimization of the network and exploration of this phase space is done through training over a *training dataset* as described in next section.

### 3.1.3 Training procedure

To adapt the parameters we need an object that describe how well the network perform. This is the *loss* of our neural networks  $\mathcal{L}$ . In our barycenter example, it could be the distance between the reconstructed and real barycenter. Using this metric we can adjust the parameters of our network.

Depending if we try to minimize or maximize it, it need to posses a minima or a maxima. For example when doing *regression*, i.e. produce a scalar result like the coordinates of a barycenter, a common loss is the Mean Square Error (MSE). Let  $i$  be our dataset, the  $N$  events considered for training,  $y_i$  be the target scalar, the barycenter positions of each events,  $x_i$  the input data, the charge vector, and  $f(x_i, \theta)$  the result of the network. The network here is modelled by  $f$ , and its parameter  $\theta$

$$\mathcal{L} \equiv MSE = \frac{1}{N} \sum_i^N (y_i - f(x_i, \theta))^2 \quad (3.3)$$

Another common loss function is the Mean Absolute Error (MAE)

$$\mathcal{L} \equiv MAE = \frac{1}{N} \sum_i^N |y_i - f(x_i, \theta)| \quad (3.4)$$

We see that those loss function possess a minima when  $f(x_i, \theta) = y_i$ .

Modern neural networks typically use gradient descent to optimize their parameters by minimizing the loss function. The gradient of the parameter  $w$ , designated in literature as  $\theta$ , with respect of the loss function  $\mathcal{L}$  is subtracted each optimisation step  $t$

$$\theta_{t+1} = \theta_t - \frac{\partial \mathcal{L}}{\partial \theta} \quad (3.5)$$

This induce  $\mathcal{L}$  needs to be differentiable with respect to  $\theta$ , thus the layers and their activation functions also need to be differentiable. This simple gradient descent, designated as Stochastic Gradient Descent (SGD), can be extended with first and second order momentums like in the Adam optimizer [44]. More details about the optimizers can be found in Section 3.1.3.

### Training lifecycle

The training process of neural networks can vary depending on the application and dataset, but in this thesis, we follow a standard approach. As shown in Fig. 3.3, training is organized into *epochs*, each of which consists of several *steps*. During each step, the neural network optimizes its parameters using a *batch*, a subset of the entire training dataset.

The ideal batch size, meaning the number of events in each batch, would encompass the entire dataset to avoid bias introduced by sub-sample specificity. However, in large-scale experiments like JUNO, the batch size is often constrained by memory limitations due to the massive volume of data

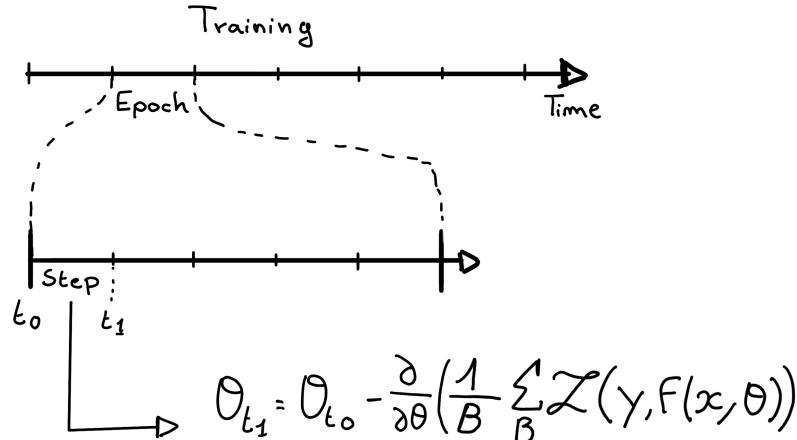


FIGURE 3.3 – Illustration of the training lifecycle

1056 generated by the photomultiplier tubes (PMTs). Balancing batch size with memory capacity is crucial  
1057 to ensure efficient and accurate training.

1058 At the end of each epoch, the neural network is evaluated on a validation dataset, which is not  
1059 used during training. This dataset serves as a reference to assess the network's performance and to  
1060 monitor for signs of overfitting. In JUNO, this is critical because the model needs to generalize well  
1061 to unseen experimental data and avoid overfitting to noise in the training set (see Section 3.1.4).

1062 Hyperparameters that can be optimized during the training can be optimized at each epoch, for  
1063 example the learning rate, or each step, the optimizer momentum for example.

1064 There is not really a typical number of epochs or steps for the training. The number steps can be  
1065 defined such as in one epoch, the NN see the entirety of the dataset but the number of steps and  
1066 epochs are hyperparameters that are optimized over the each subsequent training. We adjust them  
1067 by looking at the loss evolution profile over time.

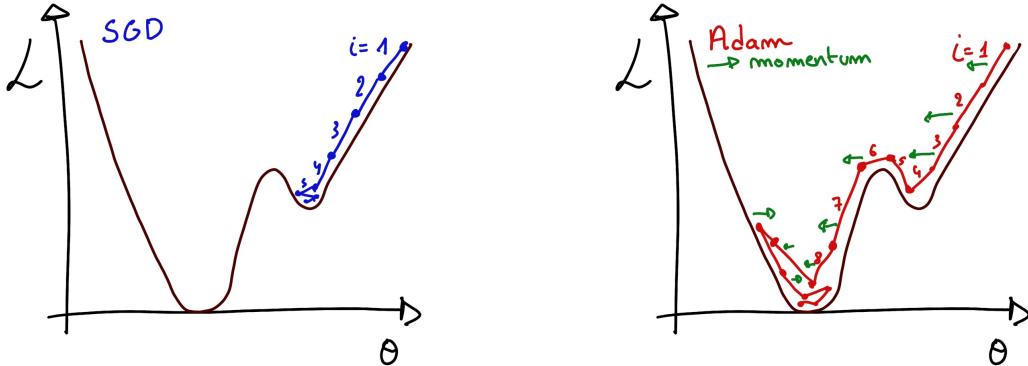
1068 Most training are started with a fixed number of epochs, i.e. from what we've seen from precedent  
1069 training, the network stop learning, the loss is constant, after  $N$  epoch so we run the training for  
1070  $N + \delta$  epochs to see if the modification brings improvements to the loss profile. We can implement  
1071 *early stopping policies* to halt training if certain conditions are met, such as a sudden increase in loss  
1072 or when the loss plateaus. However, for the JUNO experiment, where training time is not a strict  
1073 limitation, early stopping is less critical, though it may still be useful to prevent overfitting in some  
1074 cases

### 1075 The optimizer

1076 As briefly introduced at the beginning of this section, the parameters of the neural network are  
1077 optimized using the gradient descent method. We compute the gradient of the mean loss over the  
1078 batch with respect of each parameters and we update the parameters in accord to minimize the loss.  
1079 The gradient is computed backward from the loss up to the first layer parameters using the chain  
1080 rule, in this case with only one parameter at each step for simplicity:

$$\frac{\partial \mathcal{L}}{\partial \theta_1} = \frac{\partial \theta_2}{\partial \theta_1} \frac{\partial \mathcal{L}}{\partial \theta_2} = \frac{\partial \theta_2}{\partial \theta_1} \frac{\partial \theta_3}{\partial \theta_2} \frac{\partial \mathcal{L}}{\partial \theta_3} = \frac{\partial \theta_2}{\partial \theta_1} \prod_{i=2}^{N-1} \frac{\partial \theta_{i+1}}{\partial \theta_i} \frac{\partial \mathcal{L}}{\partial \theta_N} \quad (3.6)$$

1081 where  $\theta$  is a parameter,  $i$  is the layer index. We see here that the gradient of the first layer is  
1082 dependent of the gradient of all the following layers. Because the only value known at the start



(A) Illustration of SGD falling into a local minima

(B) Illustration of the Adam momentum allowing it to overcome local minima

FIGURE 3.4

of the optimization procedure is  $\mathcal{L}$  we compute  $\frac{\partial \mathcal{L}}{\partial \theta_N}$  then,  $\frac{\partial \theta_N}{\partial \theta_{N-1}}$ , etc... This is called the *backward propagation*.

This update of the parameters is done following an optimizer policy. Those optimizers depends on hyperparameters. The ones used in this thesis are:

1. Stochastic Gradient Descent (SGD). A simple but widely used optimizer that relies on one key hyperparameter, the learning rate (LR) /  $\lambda$ . It update each step the parameters  $\theta$  following

$$\theta_{t+1} = \theta_t - \lambda \frac{\partial \mathcal{L}}{\partial \theta} \Big|_{\theta_t} \quad (3.7)$$

where  $t$  is the step index. It is a powerful optimizer but is very sensible to local minima of the loss in the parameters phase space as illustrated in Figure 3.4a.

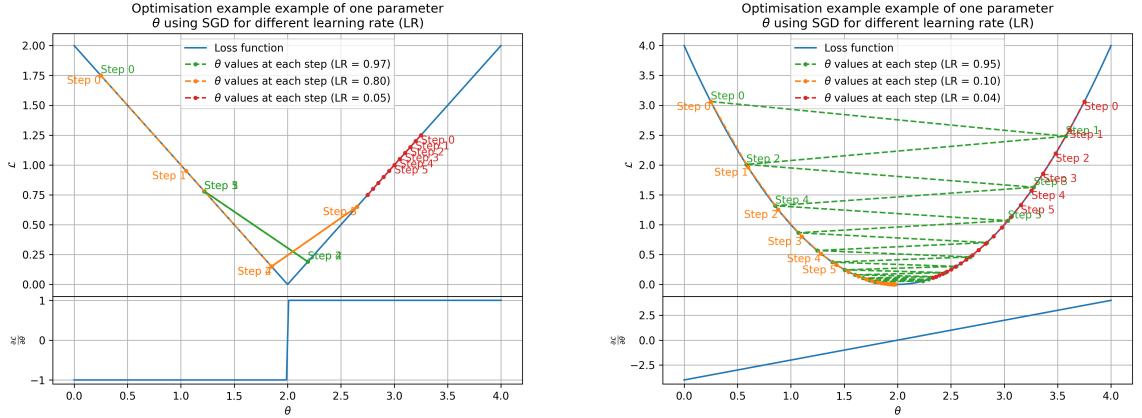
2. Adam Optimizer [44]. The concept is, in short, to have and SGD but with momentum. Adam possess two momentum  $m(\beta_1)$  and  $v(\beta_2)$  which are respectively proportional to  $\frac{\partial \mathcal{L}}{\partial \theta}$  and  $(\frac{\partial \mathcal{L}}{\partial \theta})^2$ .  $\beta_1$  and  $\beta_2$  are hyperparameters that dictate the moment update at each optimization step. The parameters are then upgraded following

$$m_{t+1} = \beta_1 m_t + (1 - \beta_1) \frac{\partial \mathcal{L}}{\partial \theta} \quad (3.8)$$

$$v_{t+1} = \beta_2 v_t + (1 - \beta_2) \left( \frac{\partial \mathcal{L}}{\partial \theta} \right)^2 \quad (3.9)$$

$$\theta_{t+1} = \theta_t - \lambda \frac{m_{t+1}}{\sqrt{v_{t+1}} + \epsilon} \quad (3.10)$$

where  $\epsilon$  is a small number to prevent divergence when  $v$  is close to 0. These momentums allow to overcome small local minima in the parameters phase. Imagine ball going down a slope as illustrated in 3.4a, if you ignore the stored momentum you get SGD and get stuck as on the left plot. Now if you consider the momentum you get over the hill and end up in the global minima.



(A) Illustration of the SGD optimizer on one parameter  $\theta$  on the MAE Loss. We see here that it has trouble reaching the minima due to the gradient being constant.

(B) Illustration of the SGD optimizer on one parameter  $\theta$  on the MSE Loss. We see two different behavior: A smooth one (orange and red) when the LR is small enough and a more chaotic one when the LR is too high.

FIGURE 3.5 – Illustration of the SGD optimizer. In blue is the value of the loss function, orange, green and red are the path taken by the optimized parameter during the training for different LR.

### 1096 Learning Rate (LR) Schedules

1097 The learning rate plays a crucial role in determining how fast or slow the model converges. If the  
 1098 learning rate is too high (Fig. 3.5a), the model may skip over the optimal solution, whereas a low  
 1099 learning rate (Fig. 3.5b) can slow down the convergence process, leading to inefficient training. To  
 1100 address this, learning rate schedulers are employed.

1101 Using a learning rate scheduler allows the optimizer to take larger steps in the early stages of training,  
 1102 where rapid learning is beneficial, and progressively smaller steps as the model approaches convergence.  
 1103 This strategy is especially useful in JUNO, where early learning from noisy data may require  
 1104 coarse adjustments, but fine-tuning is needed later to accurately capture subtle event characteristics.

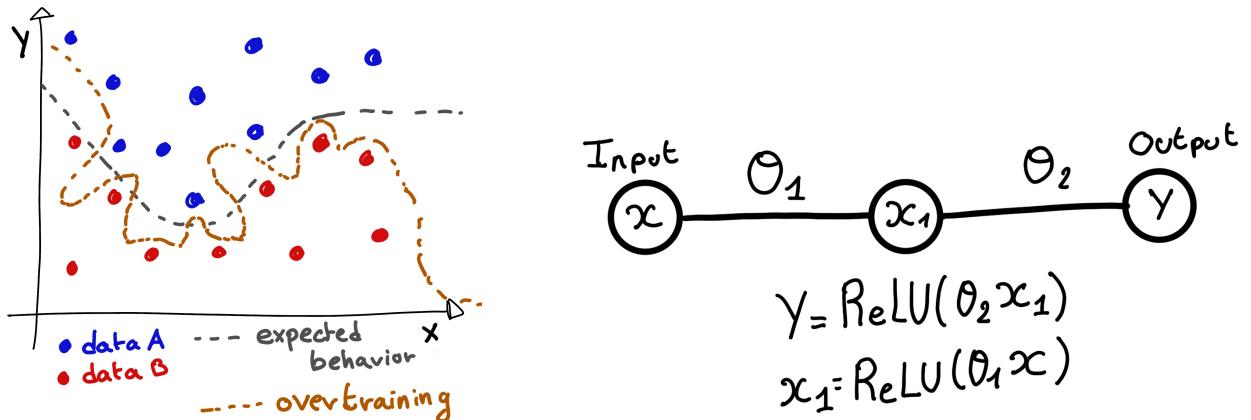
1105 Another policy that is often used is the save of the best model. In some situations, the loss value after  
 1106 each epoch will strongly oscillate or even worsen. This policy allows us to keep the best version  
 1107 of the model attained during the training phase.

### 1108 3.1.4 Potential pitfalls

1109 Apart from being stuck in local minima, there are also other behaviors and effects we want to prevent  
 1110 during training.

#### 1111 Overtraining

1112 Overfitting occurs when a neural network memorizes specific details or noise from the training  
 1113 dataset rather than learning a general representation of the underlying data. This is common when  
 1114 the training dataset is small relative to the number of parameters in the network or when the dataset  
 1115 contains specific features that do not generalize well to unseen data. Additionally, training the



(A) Illustration of overtraining. The task at hand is to determine depending on two input variable  $x$  and  $y$  if the data belong to the dataset  $A$  or the dataset  $B$ . The expected boundary between the two dataset is represented in grey. A possible boundary learnt by overtraining is represented in brown.

(B) Illustration of a very simple NN

FIGURE 3.6

network for too many epochs can exacerbate this issue. Figure 3.6a illustrates the impact of overfitting, where the model fits the training data too closely, compromising its ability to generalize. To detect overfitting, techniques like monitoring the validation loss, early stopping, or employing cross-validation can be employed. In JUNO's context, managing overfitting is critical due to the large volume of data generated by the photomultiplier tubes (PMTs), which may include noise or other artifacts.

Overtraining can be fought in multiple ways, for example:

- **More data.** By having more data in the training dataset, the network will not be able to learn the specificities of every data.
- **Less parameters.** By reducing the number of parameters, we reduce the computing and learning capacities of the network. This will force it to fallback to generalist behaviours.
- **Dropout.** This technique implies to randomly set some neurons to 0, i.e. cutting the relation between two neurons in a layer. By doing this, we force the network to allocate more of its parameter to the features learning, preventing those parameters to be used for overtraining.
- **Early stopping.** During the training we monitor the network performance over a validation dataset. The network does not train on this dataset and thus cannot learn its specificities. If the loss on the training dataset diverge too much from the loss on the validation dataset, we can stop the training earlier to prevent it from overtraining.

### Gradient vanishing

Gradient vanishing is the effect of the gradient being so small for the early layers that the parameters are barely updated after each step. This cause the network to be unable to converge to the minima.

This comes from the way the gradient descent is calculated. Imagine a simple network composed of three fully connected layers: the input layer, a intermediate layer and the output layer. Let  $L$  be the loss,  $\theta_1$  the parameter between the input and the intermediate layer and  $\theta_2$  the parameter between the intermediate and output layer. This network is schematized in Figure 3.6b.

1141 The gradient for  $\theta_1$  will be computed using the chain rule presented in equation 3.6. Because  $\theta_1$   
 1142 depends on  $\theta_2$ , if the gradient of  $\theta_2$  is small, so will be the gradient of  $\theta_1$ . Now if we would have  
 1143 much more layer, we can see how the subsequent multiplication of small gradients would lead to  
 1144 very small update of the parameters thus “*vanishing gradient*”.

1145 Multiple actions can be taken to prevent this effect such as:

- 1146 — **Batch normalization:** In this case we apply a normalization layer that will normalize the data.  
 1147 It means that we transform the input variable  $X$  into a variable  $D$  which distribution follow  
 1148  $\langle D \rangle = 0$  and  $\sigma_D = 1$ . This helps the parameters of the network to maintain an appropriate  
 1149 scale.
- 1150 — **Residual Network (ResNet) [45]:** Residual network is a technique for neural network in which,  
 1151 instead of just sequentially feeding the results of each layer to the next one, you compute a  
 1152 residual over the input data. This technique is illustrated in Figure 3.7. The reference [45] show  
 1153 empirical evidence of its relevance.

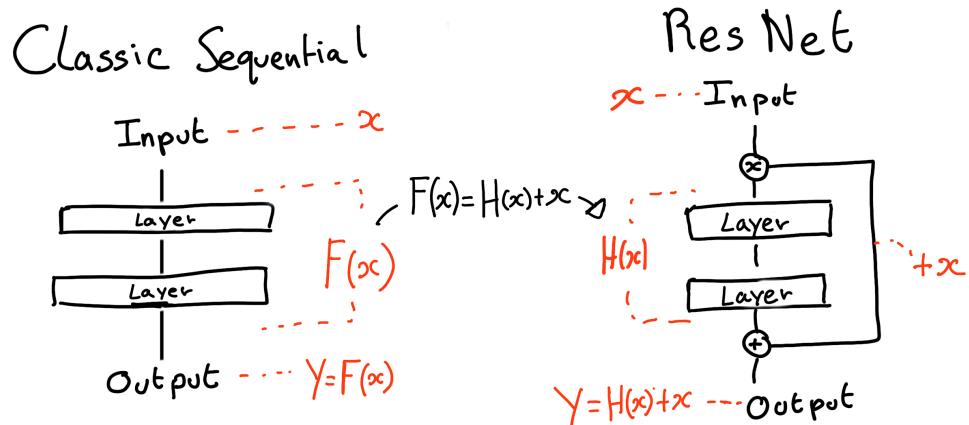


FIGURE 3.7 – Illustration of the ResNet framework

#### 1154 Gradient explosion

Gradient explosion occurs when gradients grow exponentially during backpropagation, causing parameter values to increase dramatically. This is particularly problematic in deep networks where the product of large gradients across layers can lead to unstable updates. In practice, gradient explosion is often caused by large learning rates, poor weight initialization, or nonlinearities in the network. For illustration, consider that the loss dependency in  $\theta$  follow

$$\begin{aligned}\mathcal{L}(\theta) &= \frac{\theta^2}{2} + e^{4\theta} \\ \frac{\partial \mathcal{L}}{\partial \theta} &= \theta + 4e^{4\theta}\end{aligned}$$

1155 The explosion is illustrated in Figure 3.8 where we can see that the loss degrades with each step of  
 1156 optimization. In this illustration it is clear that reducing the learning rate suffice but this behaviour  
 1157 can happens in the middle of the training where the learning rate schedule does not permit reactivity.

1158 There exist solutions to prevent this explosions:

- 1159 — **Gradient clipping:** In this case we work on the gradient so that the norm of gradient vector  
 1160 does not exceed a certain threshold. In our illustration in Figure 3.8 the gradient for  $\theta > 0$   
 1161 could be clipped at 3 for example.

- 1162 — **Batch normalization:** For the same reasons as for gradient vanishing, normalizing the input  
 1163 data help reduce erratic behaviour.

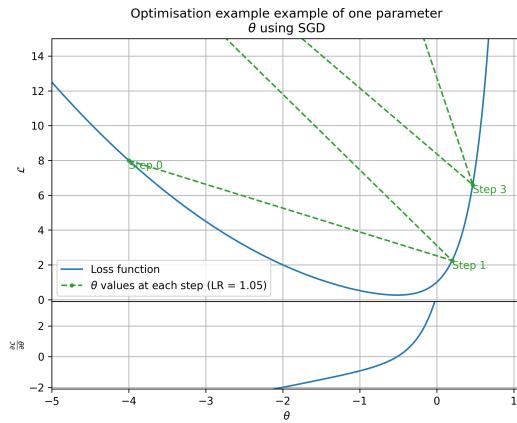


FIGURE 3.8 – Illustration of the gradient explosion. Here it can be solved with a lower learning rate but its not always the case.

## 1164 3.2 Neural networks architectures

### 1165 3.2.1 Fully Connected Deep Neural Network (FCDNN)

1166 In this thesis, FCDNN serves as a baseline architecture for comparison with more specialized models  
 1167 like CNNs (see Section 3.2.2) and GNNs (see section 3.2.3), which are better suited to structured or  
 1168 graph-based data. However, FCDNNs are still useful when modeling highly abstract relationships,  
 1169 such as aggregating features from the JUNO PMTs. While they are powerful, their main drawback  
 1170 lies in their inefficiency when dealing with high-dimensional or spatially structured data, which  
 1171 will be addressed with convolutional architectures. This architecture is the stack of multiple fully  
 1172 connected layers as presented in the Figure 3.9a. Most of the time, the classic ReLU function

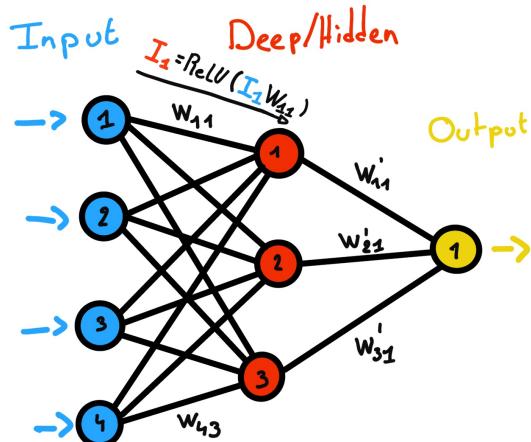
$$\text{ReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (3.11)$$

1173 is used as activation function. Prelu and Sigmoid are also popular choices:

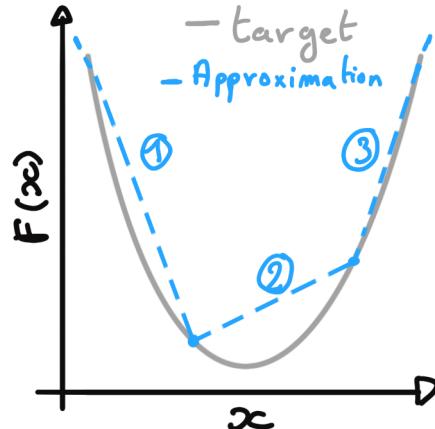
$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (3.12) \quad \text{PReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{otherwise} \end{cases} \quad (3.13)$$

1175 The reasoning behind ReLU and PReLU is that with enough of them, you can mimic any continuous  
 1176 function as illustrated in Figure 3.9b. Sigmoid is more used in case of classification, its behavior  
 1177 going hand in hand with the Cross Entropy loss function used in classification problems.

1178 Due to its simplicity, FCDNN are also used as basic pieces for more complex architectures such as  
 1179 the CNN and GNN that will be presented in the next sections.



(A) Schema of a FCDNN



(B) Illustration of a composition of ReLU "approximating" a function. (1) No ReLU is taking effect (2) One ReLU is activating (3) Another ReLU is activating

FIGURE 3.9

### 1180 3.2.2 Convolutional Neural Network (CNN)

1181 It's not trivial to describe in text the principles of Convolutional Neural Network (CNN) and how  
 1182 they works. We try a general description below followed by a step by step description of a concrete  
 1183 example.

1184 Convolutional Neural Networks are a family of neural networks that use discrete convolution filters,  
 1185 as illustrated in an example in Figure 3.10, to process the input data, often images. They are com-  
 1186 monly used in image recognition [46] for classification or regression problematics. Concretely, you  
 1187 multiply element-wise a portion of the input data, in the case of an image, a small part of the image,  
 1188 with a kernel of same dimension. In Figure 3.10, we multiply the  $3 \times 3$  pixels sub-image with the  
 1189  $3 \times 3$  kernel.

1190 Their filters scan the input data, highlighting patterns of interest, this scanning procedure making  
 1191 them translation-invariant. In the concrete case of Figure 3.10, for each pixel of the input image, we  
 1192 group it with the 8 neighbours pixel and produce a new pixel that correspond to the output image.  
 1193 For the pixel on the edges that do not have neighbours, we either create "imaginary" pixel with the  
 1194 value 0 or we just ignore them. If we ignore them, the output image will posses fewer pixels than the  
 1195 input image. We see that the operation do not care where is the pattern of interest in the images, the  
 1196 filter output will be *invariant* whatever *translation* is applied to the image.

1197 This invariance mean that they are capable of detecting oriented features independently of their  
 1198 location on the image. These filters scan the input, highlighting important features like edges or  
 1199 textures, which in JUNO's case could represent spatial correlations in the timing and charge data  
 1200 across the detector. As the network goes deeper, it can capture more complex and abstract features,  
 1201 making it ideal for detecting nuanced particle interactions. Again taking 3.10 as an example, with  
 1202 only the 9 parameters composing the kernel, we can highlight the contour of the duck by looking at  
 1203 the "yellowness" of the pixels.

1204 The learning parameters of CNNs are the kernels components, the network thus learn the optimal  
 1205 filters to extract the desired features.

1206 The convolution layers are commonly chained [47], reducing the input dimension while increasing  
 1207 the number of filters. The idea behind is that the first layers will process local informations and  
 1208 the latest layers will process more global informations, as the latest convolution filters will process

1209 the results of the preceding that themself have processed local information. To try to preserve the  
 1210 amount of information, we tend to grow the numbers of filters for each division of the input data.  
 1211 The results of the convolution filters is commonly then flattened and feed to a smaller FCDNN which  
 1212 will process the filters results to yield the desired output.

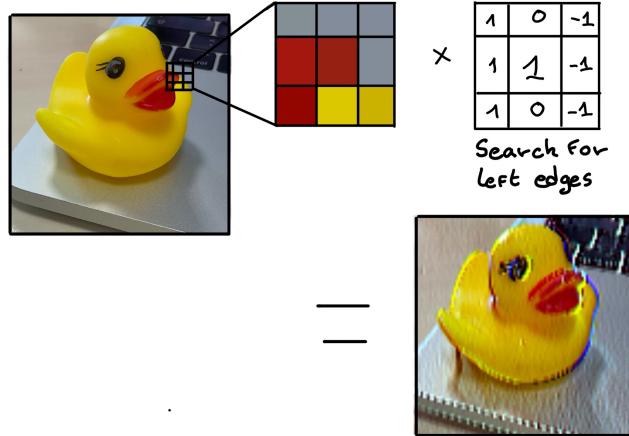
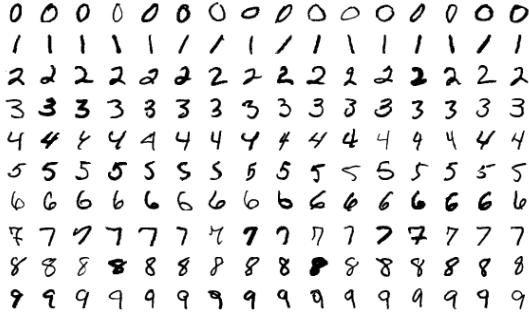


FIGURE 3.10 – Illustration of the effect of a convolution filter. Here we apply a filter with the aim do detect left edges. We see in the resulting image that the left edges of the duck are bright yellow where the right edges are dark blue indicating the contour of the object. The convolution was calculated using [48].

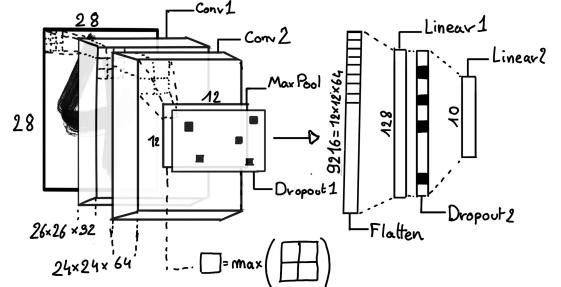
1213 As an example, let's take the Pytorch [49] example for the MNIST [50], a dataset of black and white  
 1214 images of handwritten digits. Those images are  $28 \times 28$  pixels with only one channel corresponding  
 1215 to the grey level of the pixel. Example of images from this dataset are presented in Figure 3.11a

1216 A schema of the CNN used in the Pytorch example is presented in Figure 3.11b. Using this schema  
 1217 as a reference, the trained network is made of:

- 1218 1. A convolutional layer of  $(3 \times 3)$  filters yielding 32 channels. A bias parameter is applied to each  
 1219 channel for a total of  $(32 \cdot (3 \times 3) + 32) = 320$  parameters. The resulting image is  $(26 \times 26 \times 32)$   
 1220 (26 per 26 pixels with 32 channels). The ReLU activation function is applied to each pixel.
- 1221 2. A second convolutional layer of  $(3 \times 3)$  filters yielding 64 channels. This channel also posses  
 1222 a bias parameter for a total of  $(64 \cdot (3 \times 3) + 64) = 640$  parameters. Resulting image is  $(24 \times$   
 1223  $24 \times 64)$ . This channel also apply a ReLU activation function.
- 1224 3. Then comes a  $(2 \times 2)$  max pool layer with a stride of 1 meaning that for each channel the max  
 1225 value of pixels in a  $(2 \times 2)$  block is condensed in a single resulting pixel. The resulting image  
 1226 is  $(12 \times 12 \times 64)$ .
- 1227 4. This image goes through a dropout layer which will set the pixel to 0 with a probability of 0.25.  
 1228 This help prevent overtraining the neural network (see Section 3.1.4 for more details).
- 1229 5. The data is the flattened i.e. condensed into a vector of  $(12 \times 12 \times 64) = 9216$  values.
- 1230 6. Then comes a fully connected linear layer (Eq. 3.2) with a ReLU activation that output 128  
 1231 feature. It needs  $(9216 \cdot 128) + 128 = 1'179'776$  parameters.
- 1232 7. This 128 item vector goes through another dropout layer with a probability of 0.5
- 1233 8. The vector is then transformed through a linear layer with ReLU activation. It output 10 values,  
 1234 one for each digit class (0, 1, 2, ..., 9). It need  $(128 \cdot 10) + 128 = 1408$  parameters.
- 1235 9. Finally the 10 values are normalized using a log softmax function  $\text{LogSoftmax}(x_i) = \log \left( \frac{\exp(x_i)}{\sum_j \exp(x_j)} \right)$ .  
 1236 Each of those values are the probability of the input image to be a certain digit.



(A) Example of images in the MNIST dataset



(B) Schema of the CNN used in Pytorch example to process the MNIST dataset

FIGURE 3.11

1237 The final network needs 1'182'144 parameters or, if we consider each parameters to be a double  
 1238 precision floating point, 9.45 MB of data. To gives a order of magnitude, such neural network is  
 1239 considered “simple”, train in a matter of minutes on T4 GPU [51] (14 epochs) and reach an accuracy  
 1240 in its prediction of 99%.

### 1241 3.2.3 Graph Neural Network (GNN)

1242 In GNNs, data is represented as nodes and edges in a graph, which allows us to model the JUNO  
 1243 detector as a network of PMTs, where each PMT is a node and the edges represent relationships  
 1244 such as spatial distance or timing correlations between PMTs. This flexibility enables GNNs to  
 1245 capture complex interactions across the detector geometry that would be difficult to represent with a  
 1246 CNN. Furthermore, GNNs excel at processing non-Euclidean data, making them a natural fit for the  
 1247 irregular layout of the PMTs in JUNO. In this thesis, GNNs are applied to model the spatial and tem-  
 1248 poral relationships between PMTs, enabling more precise event classification and reconstruction. By  
 1249 leveraging the message-passing framework, the GNN can aggregate information from neighboring  
 1250 PMTs, allowing it to detect subtle patterns in the detector’s data.

1251 To get deeper in details, we have seen in the previous section, the CNNs are powerful for image  
 1252 processing, and more generally any data that can be expressed as a regular, discrete space and from  
 1253 which the information reside in the dispersion in this space. For an image, the edges of an object  
 1254 and how they assemble. A red square, straight edges with a sharp angle between them, is much less  
 1255 representative of a duck than an yellow sphere, round edges without sharp angles.

1256 This “image” projection is not fitted for every problematics. The signals produced by a detector does  
 1257 not always have the properties of images. In the case of JUNO for example, we can create an image  
 1258 of two channels, one for the charge  $Q$  and one for the timing  $t$  but this image should be spheric.  
 1259 Furthermore JUNO is by nature inhomogeneous, using two different systems : The LPMT and the  
 1260 SPMT. Those two systems have different regime, and thus should be processed differently. We could  
 1261 imagine images with four channels, two for the LPMT and two for the SPMT, or even a branched  
 1262 CNN with one convolution branch for the LPMT and another one for the SPMT. Anyway, the CNN  
 1263 will need to combine the two systems.

1264 To get around the restrictions of data representation imposed by CNNs, we can use the more flexible  
 1265 *graph* representation. A graph  $G(\mathcal{N}, \mathcal{E})$  is composed of vertex or node  $n \in \mathcal{N}$  and edges  $e \in \mathcal{E}$ . The  
 1266 edges are associated to two nodes  $(u, v) \in \mathcal{N}^2$ , “connecting” them. The node and the edges can hold  
 1267 features, commonly represented as vector  $n \in \mathbb{R}^{k_n}$ ,  $e \in \mathbb{R}^{k_e}$  with  $k_n$  and  $k_e$  the number of features on  
 1268 the nodes and edges respectively. We can thus define a graph using two tensors  $A_e^{ij}$  the adjacency

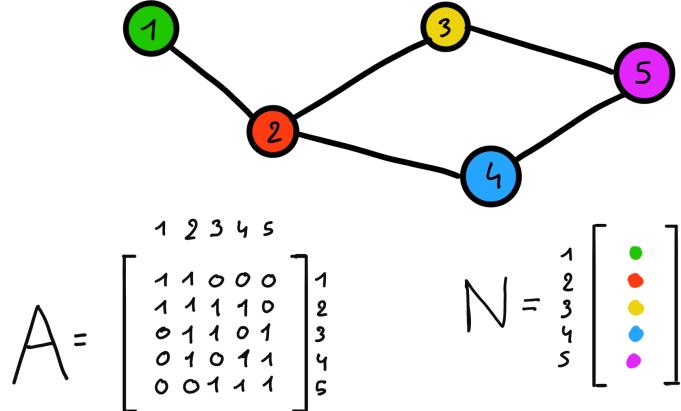
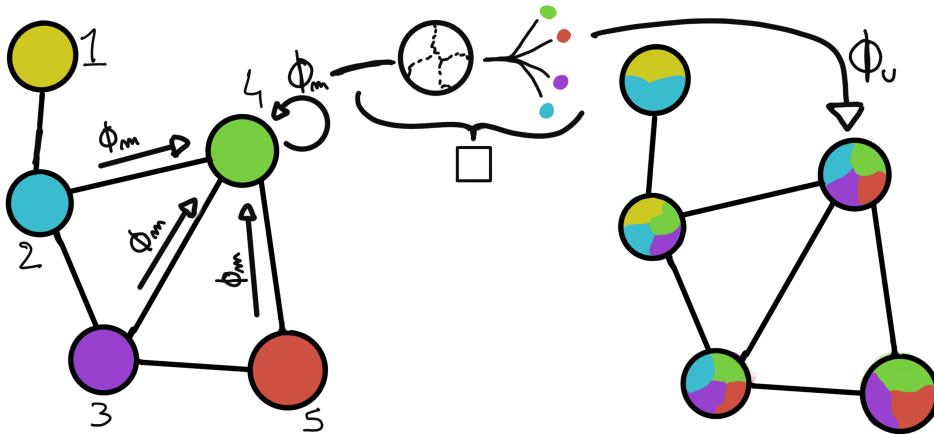


FIGURE 3.12 – Illustration of a graph and its tensor representation.

1269 tensor that hold the features  $\epsilon \in [0, k_e]$  of the edge connecting the node  $i$  and  $j$  and the tensor  $N_v^i$  that  
1270 hold the features  $v \in [0, k_n]$  of a node  $i$ .

1271 More figuratively, using the example in Figure 3.12, we have a graph of 5 nodes with a color as  
1272 feature. The edges have no features, we thus encode their existences as 0 or 1. In a realistic examples  
1273 as JUNO we could represent each PMTs as nodes and the edges between them as their relation such  
1274 as distance, timing difference, etc... There no strict rules about what is a node or how they should be  
1275 linked together. This abstraction allow us to represent virtually any type of detector of any geometry.

FIGURE 3.13 – Illustration of the message passing algorithm. The detailed explanation  
can be found in Section 3.2.3

1276 To process such object we need specific machine learning algorithms we call Graph neural network.  
1277 To efficiently manipulate graph we need to structurally encode their property in the neural network  
1278 computing architecture: each node is equivalent (as opposite to ordered data in a vector), each node  
1279 has a set of neighbours, ... One of this method is the message passing algorithm presented historically  
1280 in "Neural Message Passing for Quantum Chemistry" [52]. In this algorithm, with each layer of  
1281 message passing a new set of features is computed for each node following

$$n_i^{k+1} = \phi_u(n_i^k, \square_j \phi_m(n_i^k, n_j^k, e_{ij}^k)); n_j \in \mathcal{N}_i' \quad (3.14)$$

1282 where  $\phi_u$  is a differentiable *update* function,  $\square_j$  is a differentiable *aggregation* function and  $\phi_m$  is a

1283 differentiable *message* function.  $\mathcal{N}'_i = \{n_j \in \mathcal{N} | (n_i, n_j) \in \mathcal{E}\}$  is the set of neighbours of  $n_i$ , i.e. the  
 1284 nodes  $n_j$  from which it exist an edge  $e_{i,j} \rightarrow (n_i, n_j)$ .  $k$  is the layer on which the message passing  
 1285 algorithm is applied. The update function need also a few other property if we want to keep the  
 1286 graph property, most notably the permutational invariance of its parameters (example: mean, std,  
 1287 sum, ...). The differents message, update and aggregation functions can really be any kind of function  
 1288 if they follow the constraint presented before, even small Neural Network.

1289 The edges features can also be updated, either by directly taking the results of  $\phi_m$  or by using another  
 1290 message function  $\phi_e$ .

1291 To explain this process, let's take the situation presented in Figure 3.13. We start with an input graph  
 1292 on left, in this case the message passing algorithm is mixing the color on each nodes and produce  
 1293 nodes of mixed color. For simplicity, the  $\phi_m$  and  $\phi_u$  function are the identity, they take a color and  
 1294 output the same color.

1295 Let's look at what's happening in the node 4. It has 3 neighbours and is a neighbour of itself. The four  
 1296 resulting  $\phi_m$  extract the color of each nodes and then feed them to the  $\square$  function. The  $\square$  function  
 1297 just equally distribute the color in the node. Finally the  $\phi_u$  function just update the node with the  
 1298 output of  $\square$ .

1299 Interestingly we see that the new node 4 does not have any yellow, the color of node 1. But if we were  
 1300 to run the message passing algorithm again, it would get some as node 2 is now partially yellow. If  
 1301 color here represent information, we see that multiple step are needed so that each node is "aware"  
 1302 of the informations the other nodes possess.

1303 Message passing is a very generic way of describing the process of GNN and it can be specialized  
 1304 for convolutional filtering [53], diffusion [54] and many other specific operation. GNN are used in a  
 1305 wide variety of application such as regression problematics, node classification, edge classification,  
 1306 node and edge prediction, ...

1307 It is a very versatile but complex tool.

### 1308 3.2.4 Adversarial Neural Network (ANN)

1309 The adversarial machine learning, Adversarial Neural Networks (ANN) in the case of neural net-  
 1310 work, is a family of unsupervised machine learning algorithms where the learning algorithm (gen-  
 1311 erator) is competing against another algorithm (discriminator). Taking the example of Generative  
 1312 Adversarial Networks, concept initially developed by Goodfellow et al. [55], the discriminator goal  
 1313 is to discriminate between data coming from a reference dataset and data produced by the generator.  
 1314 The generator goal, on the other hand, is to produce data that the discriminator would not be able to  
 1315 differentiate from data from the reference dataset. The expression of duality between the two models  
 1316 is represented in the loss where, at least a part of it, is driven by the results of the discriminator.

## 1317 3.3 State of the art of the Offline IBD reconstruction in JUNO

1318 The main reconstruction method currently run in JUNO is OMILREC, a data-driven method based  
 1319 on a likelihood maximization [56, 57] using only the LPMTs. The first step is to reconstruct the  
 1320 interaction vertex from which the energy reconstruction is dependent. It is also necessary for event  
 1321 pairing and classification.

### 3.3.1 Interaction vertex reconstruction

To start the likelihood maximization, a rough estimation of the vertex and of the event timing is needed. We start by estimating the vertex position using a charge based algorithm.

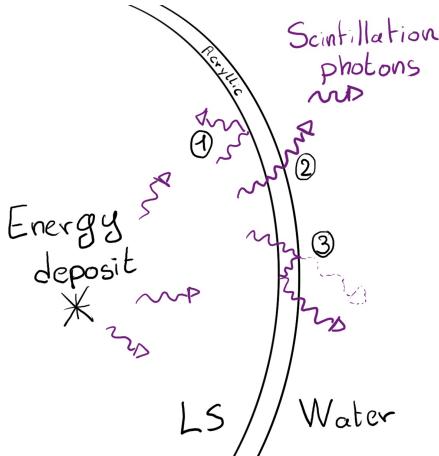
#### Charge based algorithm

The charge-based algorithm is basically base on the charge-weighted average of the PMT position.

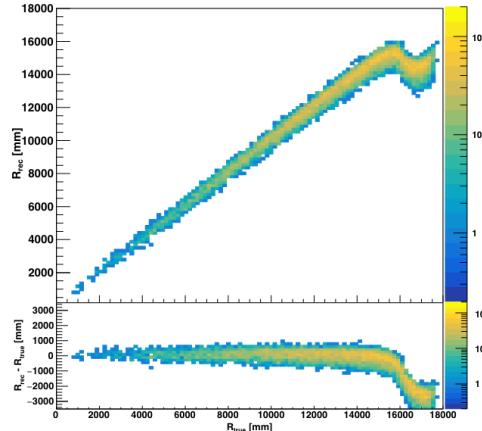
$$\vec{r}_{cb} = a \cdot \frac{\sum_i q_i \cdot \vec{r}_i}{\sum_i q_i} \quad (3.15)$$

Where  $q_i$  is the reconstructed charge of the pulse of the  $i$ th PMT and  $\vec{r}_i$  is its position.  $\vec{r}_0$  is the reconstructed interaction position.  $a$  is a scale factor introduced because a weighted average over a 3D sphere is inherently biased. Using calibration we can estimate  $a \approx 1.3$  [58]. The results in Figure 3.14b shows that the reconstruction is biased from around 15m and further. This is due to the phenomena called “total reflection area” or TR Area.

As depicted in the Figure 3.14a the optical photons, given that they have a sufficiently large incidence angle, can be deviated of their trajectories when passing through the interfaces LS-acrylic and water-acrylic due to the optical index difference. This cause photons to be lost or to be detected by PMT further than anticipated if we consider their rectilinear trajectories. This cause the charge barycenter to be located closer to the center than the event really is.



(A) Illustration of the different optical photons reflection scenarios. 1 is the reflection of the photon at the interface LS-acrylic or acrylic-water. 2 is the transmission of the photons through the interfaces. 3 is the conduction of the photon in the acrylic.



(B) Heatmap of  $R_{rec}$  and  $R_{rec} - R_{true}$  as a function of  $R_{true}$  for 4MeV prompt signals uniformly distributed in the detector calculated by the charge based algorithm

FIGURE 3.14

It is to be noted that charge based algorithm, in addition to be biased near the edge of the detector, does not provide any information about the timing of the event. Therefore, a time based algorithm needs to be introduced to provide initial values.

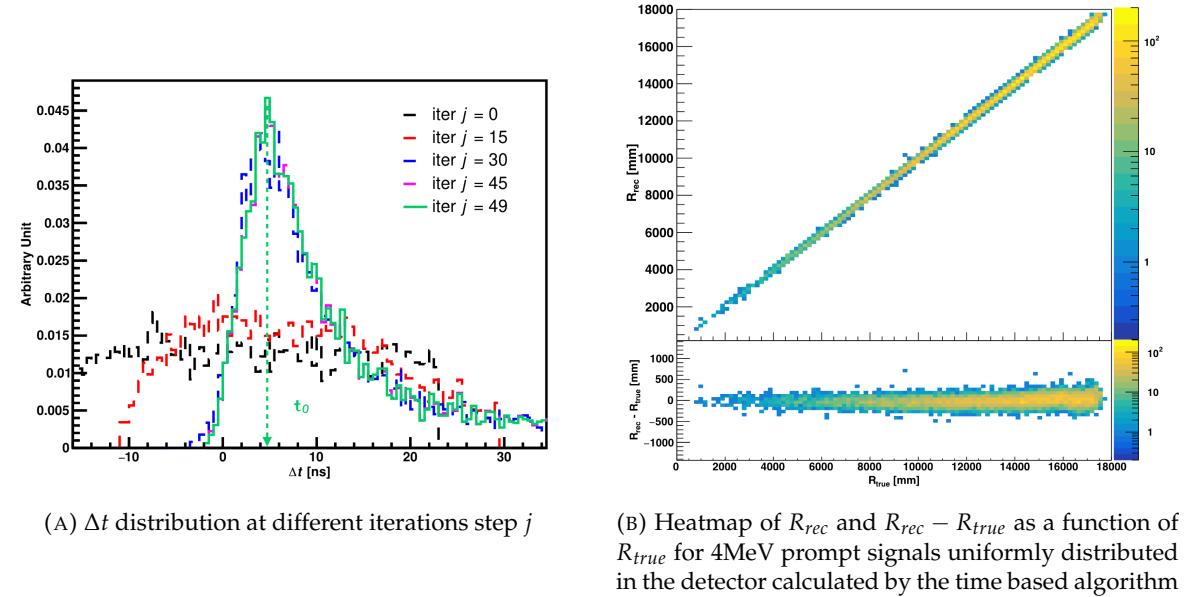


FIGURE 3.15

1340 **Time based algorithm**

1341 The time based algorithm use the distribution of the time of flight corrections  $\Delta t$  (Eq 3.16) of an event  
1342 to reconstruct its vertex and  $t_0$ . It follow the following iterations:

- 1343 1. Use the charge based algorithm to get an initial vertex to start the iteration.  
1344 2. Calculate the time of flight correction for the  $i$ th PMT using

$$\Delta t_i(j) = t_i - \text{tof}_i(j) \quad (3.16)$$

1345 where  $j$  is the iteration step,  $t_i$  is the timing of the  $i$ th PMT, and  $\text{tof}_i$  is the time-of-flight of the  
1346 photon considering an rectilinear trajectory and an effective velocity in the LS and water (see  
1347 [58] for detailed description of this effective velocity). Plot the  $\Delta t$  distribution and label the  
1348 peak position as  $\Delta t^{\text{peak}}$  (see fig 3.15a).

- 1349 3. Calculate a correction vector  $\vec{\delta}[\vec{r}(j)]$  as

$$\vec{\delta}[\vec{r}(j)] = \frac{\sum_i \left( \frac{\Delta t_i(j) - \Delta t^{\text{peak}}(j)}{\text{tof}_i(j)} \right) \cdot (\vec{r}_0(j) - \vec{r}_i)}{N^{\text{peak}}(j)} \quad (3.17)$$

1350 where  $\vec{r}_0$  is the vertex position at the beginning of this iteration,  $\vec{r}_i$  is the position of the  $i$ th PMT.  
1351 To minimize the effect of scattering, dark noise and reflection, only the pulse happening in a  
1352 time window (-10 ns, +5 ns) around  $\Delta t^{\text{peak}}$  are considered.  $N^{\text{peak}}$  is the number of PE collected  
1353 in this time-window.

- 1354 4. if  $\vec{\delta}[\vec{r}(j)] < 1\text{mm}$  or  $j \geq 100$ , stop the iteration. Otherwise  $\vec{r}_0(j+1) = \vec{r}_0(j) + \vec{\delta}[\vec{r}(j)]$  and go to  
1355 step 2.

1356 However because the earliest arrival time is used,  $t_i$  is related to the number photoelectrons  $N_i^{\text{pe}}$   
1357 detected by the PMT [59–61]. To reduce bias in the vertex reconstruction, the following equation is

1358 used to correct  $t_i$  into  $t'_i$ :

$$t'_i = t_i - p_0 / \sqrt{N_i^{\text{pe}}} - p_1 - p_2 / N_i^{\text{pe}} \quad (3.18)$$

1359 The parameters  $(p_0, p_1, p_2)$  were optimized to  $(9.42, 0.74, -4.60)$  for Hamamatsu PMTs and  $(41.31,$   
1360  $-12.04, -20.02)$  for NNVT PMTs [58]. The results presented in Figure 3.15b shows that the time based  
1361 algorithm provide a more accurate vertex and is unbiased even in the TR area. This results  $(\vec{r}_0, t_0)$  is  
1362 used as initial value for the likelihood algorithm.

### 1363 Time likelihood algorithm

1364 The time likelihood algorithm use the residual time expressed as follow

$$t_{\text{res}}^i(\vec{r}_0, t_0) = t_i - \text{tof}_i - t_0 \quad (3.19)$$

1365 In a first order approximation, the scintillator time response Probability Density Function (PDF) can  
1366 be described as the emission time profile of the scintillation photons, the Time Transit Spread (TTS)  
1367 and the dark noise of the PMTs. The emission time profile  $f(t_{\text{res}})$  is described like

$$f(t_{\text{res}}) = \sum_k \frac{\rho_k}{\tau_k} e^{-\frac{t_{\text{res}}}{\tau_k}}, \sum_k \rho_k = 1 \quad (3.20)$$

1368 as the sum of the  $k$  component that emit light in the LS each one characterised by it's decay time  $\tau_k$   
1369 and intensity fraction  $\rho_k$ . The TTS component is expressed as a gaussian convolution

$$g(t_{\text{res}}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t_{\text{res}}-\nu)^2}{2\sigma^2}} \cdot f(t_{\text{res}}) \quad (3.21)$$

1370 where  $\sigma$  is the TTS of PMTs and  $\nu$  is the average transit time. The dark noise is not correlated with any  
1371 physical events and considered as constant rate over the time window considered  $T$ . By normalizing  
1372 the dark noise probability  $\epsilon(t_{\text{res}})$  as  $\int_T \epsilon(t_{\text{res}}) dt_{\text{res}} = \epsilon_{\text{dn}}$ , it can be integrated in the PDF as

$$p(t_{\text{res}}) = (1 - \epsilon_{\text{dn}}) \cdot g(t_{\text{res}}) + \epsilon(t_{\text{res}}) \quad (3.22)$$

1373 The distribution of the residual time  $t_{\text{res}}$  of an event can then be compared to  $p(t_{\text{res}})$  and the best  
1374 fitting vertex  $\vec{r}_0$  and  $t_0$  can be chosen by minimizing

$$\mathcal{L}(\vec{r}_0, t_0) = -\ln \left( \prod_i p(t_{\text{res}}^i) \right) \quad (3.23)$$

1375 The parameter of Eq. 3.22 can be measured experimentally. The results shown in Figure 3.16 used  
1376 PDF from monte carlo simulation. The results shows that  $R_{\text{rec}} - R_{\text{true}}$  is biased depending on the  
1377 energy. While this could be corrected using calibration, another algorithm based on charge likelihood  
1378 was developed to correct this problem.

### 1379 Charge likelihood algorithm

1380 Similarly to the time likelihood algorithms that use a timing PDF, the charge likelihood algorithm  
1381 use a PE PDF for each PMT depending on the energy and position of the event. With  $\mu(\vec{r}_0, E)$  the  
1382 mean expected number of PE detected by each PMT, the probability to observe  $N_{\text{pe}}$  in a PMT follow  
1383 a Poisson distribution. Thus

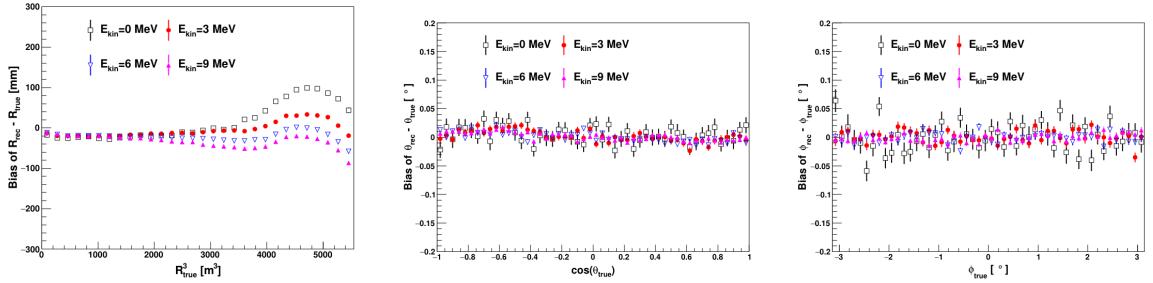


FIGURE 3.16 – Bias of the reconstructed radius  $R$  (left),  $\theta$  (middle) and  $\phi$  (right) for multiple energies by the time likelihood algorithm

- The probability to observe no hit ( $N_{pe} = 0$ ) in the  $j$ th PMT is  $P_{nohit}^j(\vec{r}_0, E) = e^{-\mu_j}$
- The probability to observe  $N_{pe} \neq 0$  in the  $i$ th PMT is  $P_{hit}^i(\vec{r}_0, E) = \frac{\mu^{N_{pe}} e^{-\mu_i}}{N_{pe}^i!}$

Therefore, the probability to observe a specific hit pattern can be expressed as

$$P(\vec{r}_0, E) = \prod_j P_{nohit}^j(\vec{r}_0, E) \cdot \prod_i P_{hit}^i(\vec{r}_0, E) \quad (3.24)$$

The best fit values of  $\vec{R}_0$  and  $E$  can then be calculated by minimizing the negative log-likelihood

$$\mathcal{L}(\vec{r}_0, E) = -\ln(P(\vec{r}_0, E)) \quad (3.25)$$

In principle,  $\mu_i(\vec{r}_0, E)$  could be expressed

$$\mu_i(\vec{r}_0, E) = Y \cdot \frac{\Omega(\vec{r}_0, r_i)}{4\pi} \cdot \epsilon_i \cdot f(\theta_i) \cdot e^{-\sum_m \frac{d_m}{\zeta_m}} \cdot E + \delta_i \quad (3.26)$$

where  $Y$  is the energy scale factor,  $\Omega(\vec{r}_0, r_i)$  is the solid angle of the  $i$ th PMT,  $\epsilon_i$  is its detection efficiency,  $f(\theta_i)$  its angular response,  $\zeta_m$  is the attenuation length in the materials and  $\delta_i$  the expected number of dark noise.

However Eq. 3.26 assume that the scintillation light yield is linear with energy and describe poorly the contribution of indirect light, shadow effect due to the supporting structure and the total reflection effects. The solution is to use data driven methods to produce the pdf by using the calibrations sources and position described in Section 2.4. In the results presented in Figures 3.17, the PDF was produced using MC simulation and 29 specific calibrations position [58] along the Z-axis of the detector. We see that the charge likelihood algorithm show little bias in the TR area and a better resolution than the time likelihood. The Figure 3.18 shows the radial resolution of the different algorithm presented for this section, we can see the refinement at each step and that the charge likelihood yield the best results.

The charge based likelihood algorithms already give some information on the energy as Eq. 3.25 is minimized but the energy can be further refined as shown in the next section.

### 3.3.2 Energy reconstruction

As explained in Section 2.1.1, energy resolution is crucial for the NMO and oscillation parameters measurements. Thus the energy reconstruction algorithm should take into consideration as much

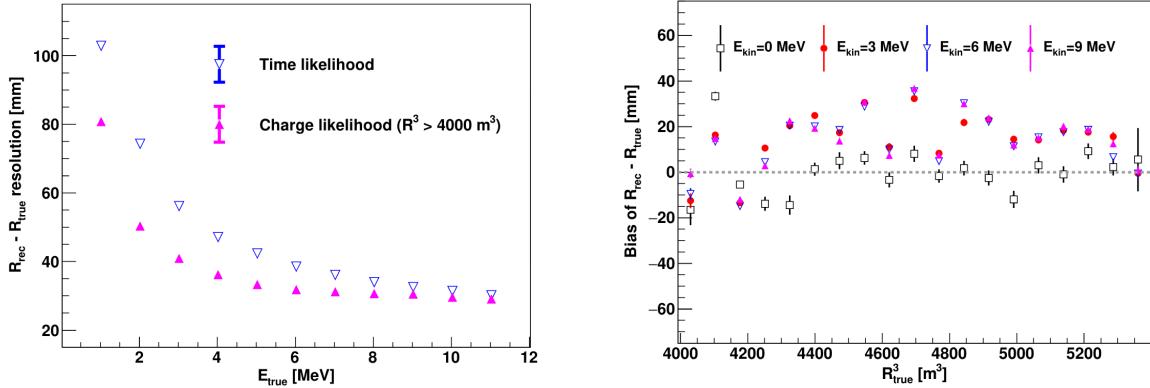


FIGURE 3.17 – On the left: Resolution of the reconstructed  $R$  as a function of the energy in the TR area ( $R^3 > 4000 \text{ m}^3 \equiv R > 16 \text{ m}$ ) by the charge and time likelihood algorithms. On the right: Bias of the reconstructed  $R$  in the TR area for different energies by the charge likelihood algorithm

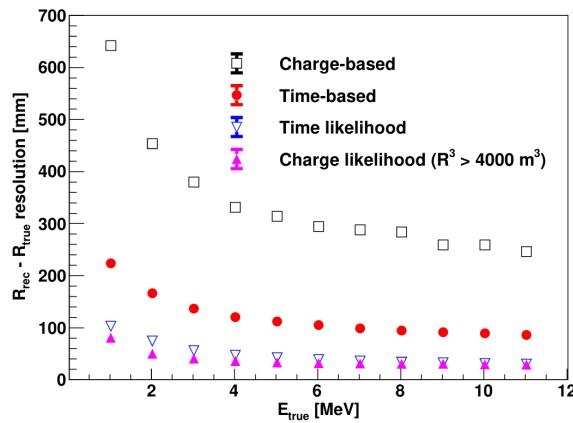


FIGURE 3.18 – Radial resolution of the different vertex reconstruction algorithms as a function of the energy

<sup>1406</sup> detector effect as possible. The following method is a data driven method based on calibration  
<sup>1407</sup> samples inspired by the charge likelihood algorithm described above [62].

#### <sup>1408</sup> Charge estimation

<sup>1409</sup> The most important element in the energy reconstruction is  $\mu_i(\vec{r}_0, E)$  described in Eq. 3.26. For  
<sup>1410</sup> realistic cases, we also need to take into account the electronics effect that were omitted in the  
<sup>1411</sup> previous section. Those effect will cause a charge smearing due to the uncertainties in the  $N_{pe}$   
<sup>1412</sup> reconstruction. Thus we define  $\hat{\mu}^L(\vec{r}_0, E)$  which is the expected  $N_{pe}/E$  in the whole detector for an  
<sup>1413</sup> event with visible energy  $E_{vis}$  and position  $\vec{r}_0$ . The position of the event and PMTs are now defined  
<sup>1414</sup> using  $(r, \theta, \theta_{pmt})$  as defined in Figure 3.19b.

$$\hat{\mu}(r, \theta, \theta_{pmt}, E_{vis}) = \frac{1}{E_{vis}} \frac{1}{M} \sum_i^M \frac{\bar{q}_i}{Q_i} - \mu_i^D, \quad \mu_i^D = \text{DNR}_i \cdot L \quad (3.27)$$

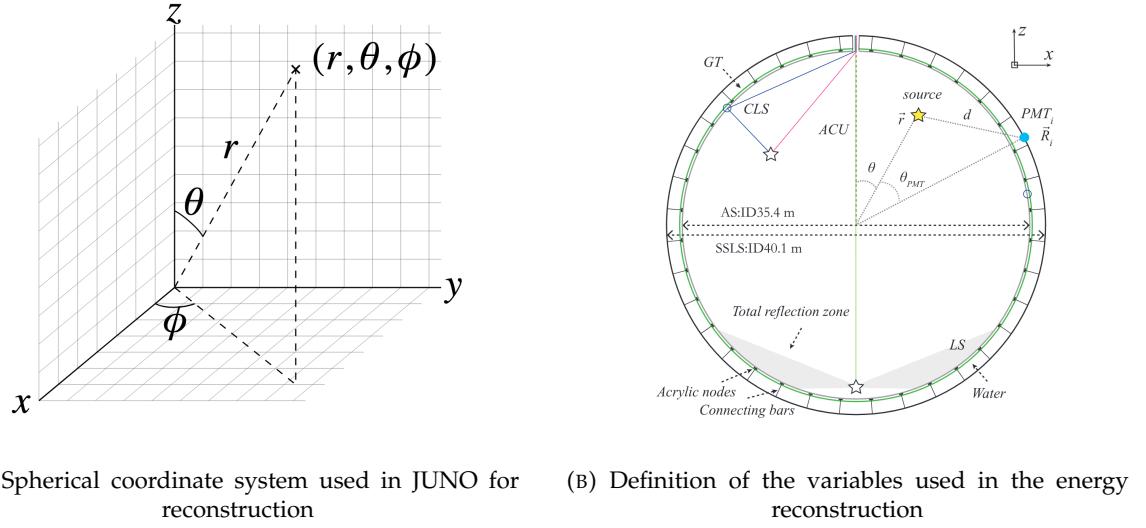


FIGURE 3.19

1415 where  $i$  runs over the PMTs with the same  $\theta_{pmt}$ ,  $DE_i$  is the detection efficiency of the  $i$ th PMT.  $\mu_i^D$   
 1416 is the expected number of dark noise photoelectrons in the time window  $L$ . The time window have  
 1417 been optimized to  $L = 280$  ns [62].  $\bar{q}_i$  is the average recorded photoelectrons in the time window  
 1418 and  $\hat{Q}_i$  is the expected average charge for 1 photoelectron. The  $N_{pe}$  map is constructed following the  
 1419 procedure described in [57].

#### 1420 Time estimation

1421 The second important observable is the hit time of photons that was previously defined in Eq. 3.19.  
 1422 It is here refined as

$$t_r = t_h - \text{tof} - t_0 = t_{LS} + t_{TT} \quad (3.28)$$

1423 where  $t_h$  is the time of hit,  $t_{LS}$  is the scintillation time and  $t_{TT}$  the transit time of PMTs that is described  
 1424 by a gaussian

$$t_{TT} = \mathcal{N}(\mu_{TT} + t_d, \sigma_{TT}) \quad (3.29)$$

1425 where  $\mu_{TT}$  is the mean transit time in PMTs,  $\sigma_{TT}$  is the Transit Time Spread (TTS) of the PMTs and  $t_d$   
 1426 is the delay time in the electronics. The effective refraction index of the LS is also corrected to take  
 1427 into account the propagation distance in the detector.

1428 The timing PDF  $P_T(t_r|r, d, \mu_l, \mu_d, k)$  can now be generated using calibration sources [62]. This PDF  
 1429 describe the probability that the residual time of the first photon hit is in  $[t_r, t_r + \delta]$  with  $r$  the radius  
 1430 of the event vertex,  $d = |\vec{r} - \vec{r}_{PMT}|$  the propagation distance,  $\mu_l$  and  $\mu_d$  the expected number of PE  
 1431 and dark noise in the electronic reading window and  $k$  is the detected number of PE.

1432 Now let denote  $f(t, r, d)$  the probability density function of "photoelectron hit a time t" for an event  
 1433 happening at  $r$  where the photons traveled the distance  $d$  in the LS

$$F(t, r, d) = \int_t^L f(t', r, d) dt' \quad (3.30)$$

1434 Based on the PDF for one photon  $k = 1$ , one can define

$$P_T^l(t|k = n) = I_n^l[f_l(t)F_l^{n-1}(t)] \quad (3.31)$$

1435 where the indicator  $l$  means that the photons comes from the LS and  $I_n^l$  a normalisation factor. To this  
 1436 pdf we add the probability to have photons coming from the dark noise indicated by the indicator  $d$   
 1437 using

$$f_d(t) = 1/L, F_d(t) = 1 - \frac{t}{L} \quad (3.32)$$

1438 and so for the case where only one photon is detected by the PMT ( $k = 1$ )

$$P_T(t|\mu_l, \mu_d, k=1) = I_1[P(1, \mu_l)P(0, \mu_d)f_l(t) + P(0, \mu_l)P(1, \mu_d)f_d(t)] \quad (3.33)$$

1439 where  $P(k_\alpha, \mu_\alpha)$  is the Poisson probability to detect  $k_\alpha$  PE from  $\alpha \in \{l, d\}$  with the condition  $k_l + k_d =$   
 1440  $k$ .

1441 Now that we have the individual timing and charge probability we can construct the charge likeli-  
 1442 hood referred as QMLE:

$$\mathcal{L}(q_1, q_2, \dots, q_N | \vec{r}, E_{vis}) = \prod_{j \in \text{unfired}} e^{-\mu_j} \prod_{i \in \text{fired}} \left( \sum_{k=1} P_Q(q_i|k) \cdot P(k, \mu_i) \right) \quad (3.34)$$

1443 where  $\mu_i = E_{vis}\hat{\mu}_i^L + \mu_i^D$  and  $P(k, \mu_i)$  is the Poisson probability of observing  $k$  PE.  $P_Q(q_i|k)$  is the  
 1444 charge pdf for  $k$  PE. And we can also construct the time likelihood referred as TMLE:

$$\mathcal{L}(t_{1,r}, t_{2,r}, \dots, t_{N,r} | \vec{r}, t_0) = \prod_{i \in \text{hit}} \frac{\sum_{k=1}^K P_T(t_{i,r}|r, d, \mu_i^l, \mu_i^d, k) \cdot P(k, \mu_i^l + \mu_i^d)}{\sum_{k=1}^K P(k, \mu_i^l + \mu_i^d)} \quad (3.35)$$

1445 where  $K$  is cut to 20 PE and hit is the set of hits satisfying  $-100 < t_{i,r} < 500$  ns.

1446 Merging those two likelihood give the charge-time likelihood QTMLE, the core algorithm of OMIL-  
 1447 REC.

$$\mathcal{L}(q_1, q_2, \dots, q_N; t_{1,r}, t_{2,r}, \dots, t_{N,r} | \vec{r}, t_0, E_{vis}) = \mathcal{L}(q_1, q_2, \dots, q_N | \vec{r}, E_{vis}) \cdot \mathcal{L}(t_{1,r}, t_{2,r}, \dots, t_{N,r} | \vec{r}, t_0) \quad (3.36)$$

1448 The radial and energy resolutions of the different likelihood are presented in Figure 3.20 (from [62]).  
 1449 We can see the improvement of adding the time information to the vertex reconstruction and that  
 1450 an increase in vertex precision can bring improvement in the energy resolution, especially at low  
 1451 energies.

1452 Data driven methods prove to be performant in the energy and vertex reconstruction given that we  
 1453 have enough calibrations sources to produce the PDF. In addition to this, member of JUNO have  
 1454 developed ML algorithms for reconstruction. The one focused on IBD reconstruction are presented  
 1455 in the next section.

### 1456 3.3.3 Machine learning for reconstruction

1457 The power of ML is the ability to model complex response to a specific problem. In JUNO the  
 1458 reconstruction problematic can be expressed as follow: knowing that each PMT, large or small,  
 1459 detected a given number of PE  $Q$  at a given time  $t$  and their position is  $x, y, z$  where did the energy  
 1460 was deposited and how much energy was it, modeling a function that naively goes:

$$\mathbb{R}^{5 \times N_{pmt}} \mapsto \mathbb{R}^4 \quad (3.37)$$

1461 It is worth pointing that while this is already a lot in informations, this is not the rawest representa-  
 1462 tion of the experiment. We could indeed replace the charge and time by the waveform in the time  
 1463 window of the event but that would lead to an input representation size that would exceed our  
 1464 computational limits. Also, due to those computational limits, most of the ML algorithm reduce this

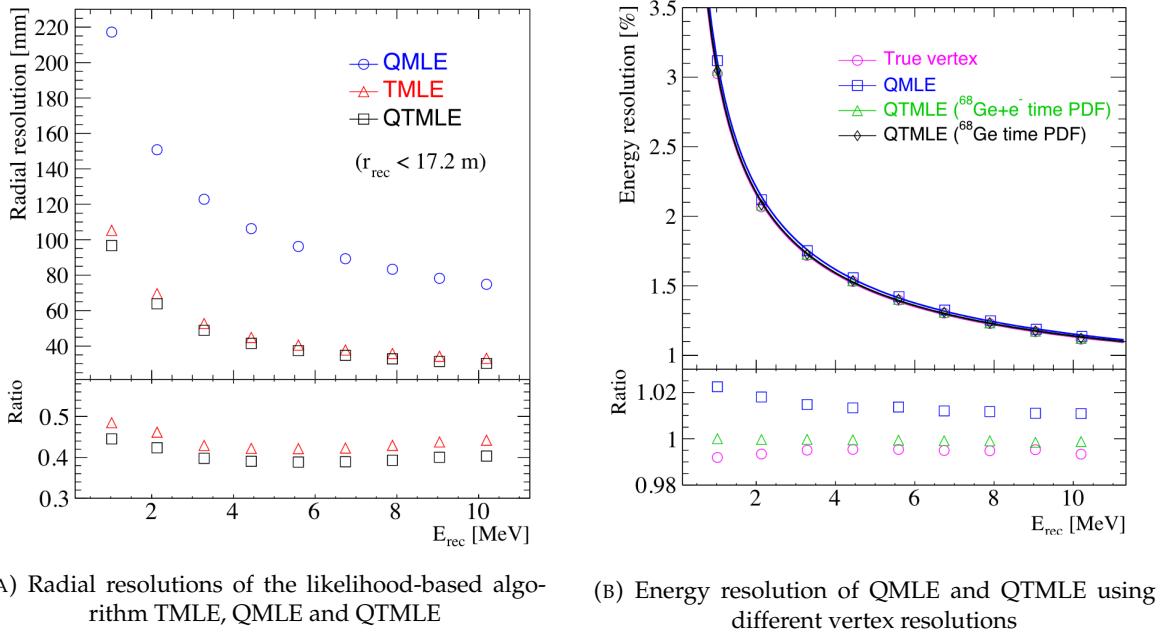


FIGURE 3.20

1465 input phase space either by structurally encoding the information (pictures, graph), by aggregating  
 1466 it (mean, variance, ...) or by exploiting invariance and equivariance of the experiment (rotational  
 1467 invariance due to the sphericity, ...).

1468 For machine learning to converge to performant algorithm, a large dataset exploring all the phase  
 1469 space of interest is needed. For the following studies, data from the monte carlo simulation presented  
 1470 in Section 2.6 are used for training. When the detector will be finished calibrations sources will be  
 1471 complementarily be used.

#### 1472 Boosted Decision Tree (BDT)

1473 On of the most classic ML method used in physics in last years is the Boosted Decision Tree. They  
 1474 have been explored for vertex reconstruction [63] et for energy reconstruction [63, 64].

1475 For vertex and energy reconstruction a BDT was developed using the aggregated informations pre-  
 1476 sented in 3.1.

Parameter	description
$n_{\text{Hits}}$	Total number of hits
$x_{cc}, y_{cc}, z_{cc}, R_{cc}$	Coordinates of the center of charge
$ht_{\text{mean}}, ht_{\text{std}}$	Hit time mean and standard deviation

TABLE 3.1 – Features used by the BDT for vertex reconstruction

1477 Its reconstruction performances are presented in Figure 3.22.

1478 A second and more advanced BDT, subsequently named BDTE, that only reconstruct energy use a  
 1479 different set of features [64]. They are presented in the table 3.2

AccumCharge	$ht_{5\%-2\%}$
$R_{cht}$	$pe_{mean}$
$z_{cc}$	$J_{cht}$
$pe_{std}$	$\phi_{cc}$
nPMTs	$ht_{35\%-30\%}$
$ht_{kurtosis}$	$ht_{20\%-15\%}$
$ht_{25\%-20\%}$	$pe_{35\%}$
$R_{cc}$	$ht_{30\%-25\%}$

TABLE 3.2 – Features used by the BDTE algorithm.  $pe$  and  $ht$  reference the charge and hit-time distribution respectively and the percentages are the quantiles of those distributions.  $cht$  and  $cc$  reference the barycenters of hit time and charge respectively

#### 1480 Neural Network (NN)

1481 Three type of neural networks have explored for event reconstruction in JUNO Deep Neural Net-  
 1482 work (DNN), Convolutional Neural Network (CNN) and Graph Network (GNN).

1483 The CNN are using 2D projection of the detector representing it as an image with two channel, one  
 1484 for the charge  $Q$  and one for the time  $t$ . The position of the PMTs is structurally encoded in the pixel  
 1485 containing the information of this PMT. In [63], the pixel is chosen based on a transformation of  $\theta$   
 1486 and  $\phi$  coordinates to the 2D plane and rounded to the nearest pixel. A sufficiently large image has  
 1487 been chosen to prevent two PMT to be located in the same pixel. An example of this projection can  
 1488 be found in Figure 3.21. The performances of the CNN can be found in Figure 3.22.

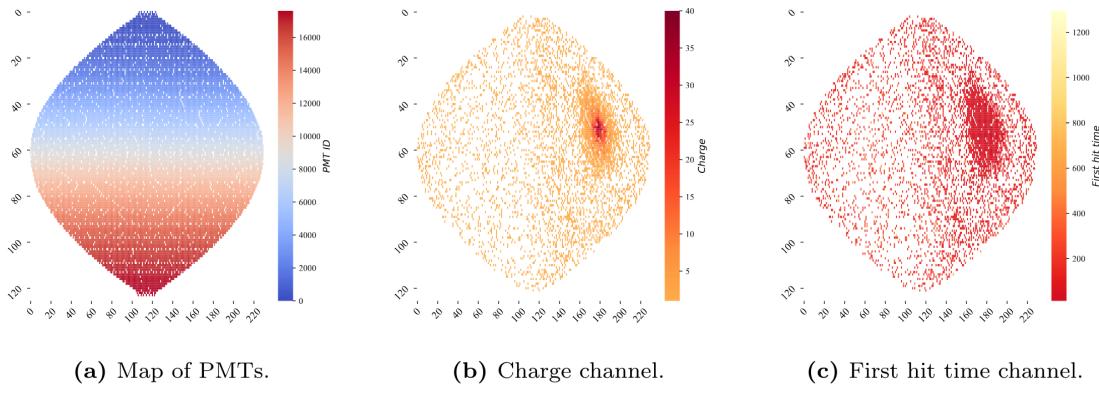


FIGURE 3.21 – Projection of the LPMTs in JUNO on a 2D plane. (a) Show the distribution of all PMTs and (b) and (c) are example of what the charge and time channel looks like respectively

1489 Using 2D have the upside of encoding a large part of the informations structurally but loose the rota-  
 1490 tional invariance of the detector. It also give undefined information to the neural network (what is a  
 1491 pixel without PMT ? What should be its charge and time ?), cause deformation in the representation  
 1492 of the detector (sides of projection) and loose topological informations.

1493 One of the way to present structurally the sphericity of JUNO to a NN is to use a graph: A collection  
 1494 of objects  $V$  called nodes and relations  $E$  called edges, each relation associated to a couple  $v_1, v_2$   
 1495 forming the graph  $G(E, V)$ . Nodes and edges can hold informations or features. In [63] the nodes,  
 1496 are geometrical region of the detector as defined by the HealPix [65]. The features of the nodes are  
 1497 aggregated informations from the PMTs it contains. The edges contains geographic informations of  
 1498 the nodes relative positions.

1499 This data representation has the advantages to keep the topology of the detector intact. It also permit  
 1500 the use of rotational invariant algorithms for the NN, thus taking advantage of the symmetries of the  
 1501 detector.

1502 The neural network then process the graph using Chebyshev Convolutions [53]. The performances  
 1503 of the GNN are presented in Figure 3.22.

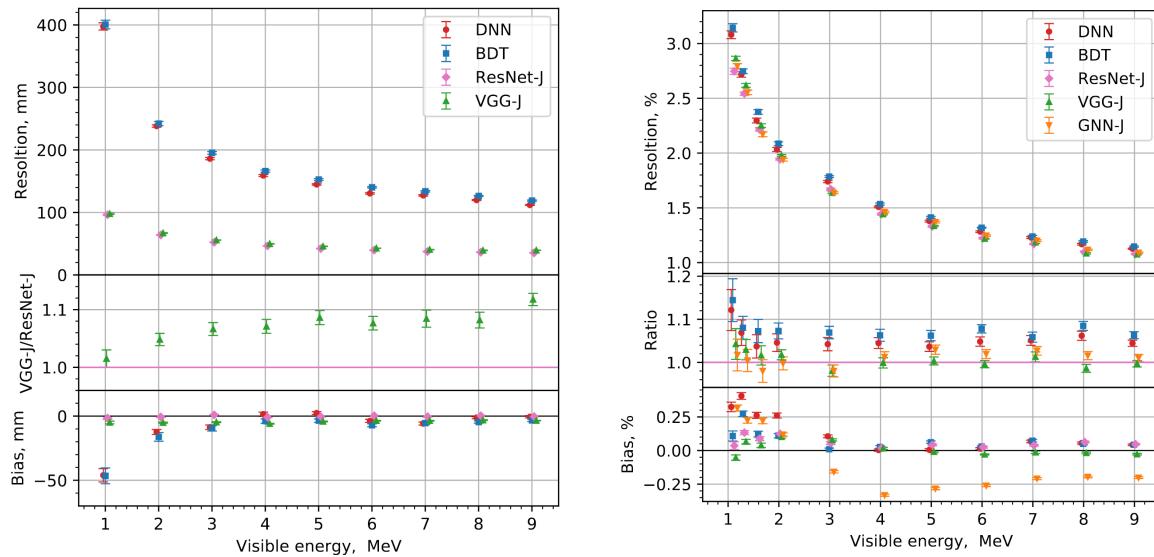


FIGURE 3.22 – Radial (left) and energy (right) resolutions of different ML algorithms.  
 The results presented here are from [63]. DNN is a deep neural network, BDT is a BDT,  
 ResNet-J and VGG-J are CNN and GNN-J is a GNN.

1504 Overall ML algorithms show similar performances as classical algorithms in term of energy recon-  
 1505 structions with the more complex structure CNN and GNN showing better performances than BDT  
 1506 and DNN. For vertex reconstruction, the BDT and DNN show poor performance while CNN are on  
 1507 the level of the classical algorithms.

## 1508 3.4 Conclusion

1509 That these first DL algorithms tried at JUNO to reconstruct IBDs do not outperform the classical  
 1510 method can be explained. They constitute a first exploration of these methods potential, as do the  
 1511 original GNN we describe in Chapter 5. Indeed, the likelihood method is also based on the full list of  
 1512 the charges ( $Q$ ) and times ( $t$ ) all PMTs, and the PDF's design accounts for an advanced knowledge of  
 1513 the detector (with a lot of human expertise). The fact that the methods presented in this chapter can  
 1514 learn enough from just the  $Q, t$  list, to reach similar performance, is already an interesting result. But  
 1515 this is not decisive yet, in my opinion.

1516 Actually, is there hope that one day DL methods reach better results at JUNO than classical's ? This  
 1517 is not a trivial question. A possibility would be to let them start from an even rawer level (involving a  
 1518 number of variables which would make a likelihood intractable). This would mean, instead of  $Q$  and  
 1519  $t$ , the full waveform in each PMT. With such a quantity of input information to analyse to identify  
 1520 patterns, even DL methods can be limited. The choice of architecture is then important, to guide the  
 1521 algorithm towards pertinent features. We doubt whether CNN's would be the best choice here. We  
 1522 bet that GNN's could be better tools, with more flexibility to hierachise information (the choice of  
 1523 which PMTs to link already helps here, as well as the possible usage of higher order quantities). The

1524 first GNN developped in JUNO (described above, [63]) does not do that. It's still only based on  $(Q, t)$   
1525 couples and link only neighbour PMTs in its first layer. It serves essentially as a way to avoid the  
1526 problems encountered by CNNs due to the planar projection of a spherical image.

1527 In chapter 5, we tried an original GNN architecture. The goal was not yet to include a rawer  
1528 information, but to see if this architecture would perform as well as the one described above when  
1529 using  $Q$ 's and  $t$ 's as the rawest information. If so, then there is hope that when rawer information  
1530 will be included, this orginal architecture will be the one able to best use it.

1531 **Chapter 4**

1532 **Image recognition for IBD  
reconstruction with the SPMT system**

1534

*Dave - Give me the position and momentum, HAL.  
 HAL - I'm afraid I can't do that Dave.  
 Dave - What's the problem ?  
 HAL - I think you know what the problem is just as well as I do.  
 Dave - What are you talking about, HAL?  
 HAL -  $\sigma_x \sigma_p \geq \frac{\hbar}{2}$*

1535

## Contents

---

<small>1536</small>	<b>4.1 Method and model</b> . . . . .	<b>64</b>
<small>1537</small>	4.1.1 Model . . . . .	65
<small>1538</small>	4.1.2 Data representation . . . . .	66
<small>1539</small>	4.1.3 Dataset . . . . .	68
<small>1540</small>	4.1.4 Data characteristics . . . . .	69
<small>1541</small>		
<small>1542</small>	<b>4.2 Training</b> . . . . .	<b>71</b>
<small>1543</small>	<b>4.3 Results</b> . . . . .	<b>71</b>
<small>1544</small>	4.3.1 J21 results . . . . .	72
<small>1545</small>	4.3.2 J21 Combination of classic and ML estimator . . . . .	74
<small>1546</small>	4.3.3 J23 results . . . . .	77
<small>1547</small>		
<small>1548</small>	<b>4.4 Conclusion and prospect</b> . . . . .	<b>78</b>
<small>1550</small>		

---

1551

As explained in Chapter 2, JUNO is an experiment composed of two systems, the Large Photomultiplier (LPMT) system and the Small Photomultiplier (SPMT) system. Both of them observe the same physics events inside of the same medium but they differ in their photo-coverage, respectively 75.2% and 2.7%, their dynamic range (see Section 2.3.2), a thousands versus a few dozen, and their front-end electronics (see section 2.3.2).

1556

The SPMT system is essential to the deployment of the Dual Calorimetry techniques, already mentioned in Section 3.3 and described in [30, 32, 66]. It is indeed less subject than the LPMTs to charge non linearity effects (QNL). This topic will be studied in more detail in Chapter 7, where the potential of one of the Dual Calorimetry techniques is explored. It consists on combined oscillation analyses based on two antineutrino energy spectra : one reconstructed with the LPMT system, the other one with the SPMT system. For that purpose, it is therefore necessary to have reconstruction tools available. Well maintained tools using the LPMT are available in the collaboration's official software. This is not the case concerning the SPMT system, where algorithms were developed more sporadically. This is one of the reasons why we developed the CNN described in this chapter.

1564

1565 Our efforts on it were limited to the early months of this thesis: it was above all a way to learn about  
 1566 ML and about JUNO's detector and software. We benchmarked its performance against a classical  
 1567 algorithm developed in Chapter 4 of [67] but not yet implemented in JUNO's software.

1568 As discussed in Chapter 3, Machine Learning (ML) algorithms shine when modeling highly dimen-  
 1569 sional data from a given dataset. In our case, we have access to complete monte-carlo simulation of  
 1570 our detector to produce large datasets that could represent multiple years of data taking. Ideally ML  
 1571 algorithms would be able to consider the entirety of the information in the detector and converge on  
 1572 the best parameters to yield optimal results.

1573 The difference between this ideal and what can be achieved in reality is an important subject. In  
 1574 particular, we wonder if an exhaustive usage of the information present in the detector could lead to  
 1575 use informations that are mismodelled in our simulated training samples (or present only in these  
 1576 samples) and therefore lead to biases when the algorithm is applied to real data. A simple way  
 1577 to start addressing this reliability issue is to try to evaluate to which extent various reconstruction  
 1578 methods use the same information. An attempt at this is presented at the end of this chapter. This is  
 1579 also the subject of Chapter 6.

## 1580 4.1 Method and model

1581 One of simplest way to look at JUNO data is to consider the detector as an array of geometrically  
 1582 distributed sensors on a sphere. Their repartition is almost homogeneous, on this sphere surface  
 1583 providing an almost equal amount of information per unit surface. It is then tempting to represent  
 1584 the detector as a spherical image with the PMTs in place of pixels. Two events with two different  
 1585 energy or position would produce two different images.

1586 The most common approach in machine learning for image processing and image recognition is the  
 1587 Convolutional Neural Network (CNN). It is widely used in research and industry [47, 68–70] due to  
 1588 its strengths (see Section 3.2.2) and has proven its relevance in image processing.

1589 Some CNN are developed to process spherical images [71] but for the sake of simplicity and as a  
 1590 first approach we decided to go with a planar projection of the detector, approach that has proven its  
 1591 efficiency using the LPMT system (see Section 3.3.3). The details about this planar projection will be  
 1592 discussed in section 4.1.2.

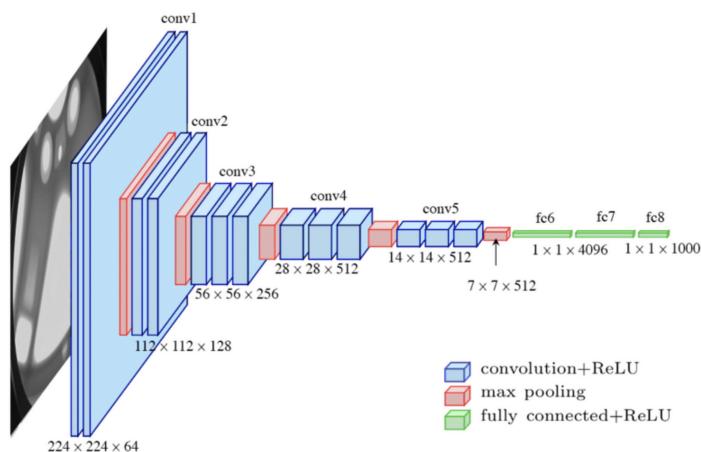


FIGURE 4.1 – Graphic representation of the VGG-16 architecture, presenting the different kind of layer composing the architecture.

### 1593 4.1.1 Model

1594 The architecture we use is derived from the VGG-16 architecture [47] illustrated in Figure 4.1. We  
 1595 define a set of hyperparameters that will define the size, complexity and computational power of the  
 1596 NN. The chose hyperparameters are detailed below and their values are presented in table 4.1.

- 1597 — **N<sub>blocks</sub>**: the number of convolution blocks, a block being composed of two convolutional layers  
 1598 with  $3 \times 3$  filters using ReLU activation function, a  $3 \times 3$  kernel max-pooling layer (except for  
 1599 the last block).
- 1600 — **N<sub>channels</sub>**: The number of channels in the first block. The number of channels in the subsequent  
 1601 blocks is computed using  $N_{channels}^i = i * N_{channels}$ ,  $i \in [1..N_{blocks}]$ .
- 1602 — **FCDNN configuration**: The result of the last convolution layer is flattened then fed to a FCDNN.  
 1603 Its configuration is expressed as the ouputs of sequenced fully connected linear layer using the  
 1604 PReLU activation function. For example  $2 * 1024 + 2 * 512$  is the sequence of 2 layers which  
 1605 output is 1024 followed by 2 other layers with an output of 512. Finally the last layer is a linear  
 1606 layer outputing 4 features wihtout activation function. Each feature of the last layer represent  
 1607 a component of the interaction vertex: Energy, X, Y, Z.
- 1608 — **Loss**: The loss function. In this work we study two different loss function ( $E + V$ ) and ( $E_r + V_r$ )  
 1609 detailed below.

$$(E + V)(E, x, y, z) = (E - E_{dep})^2 + 0.85 \sum_{\lambda \in [x, y, z]} (\lambda - \lambda_{true})^2 \quad (4.1)$$

$$(E_r + V_r)(E, x, y, z) = \frac{(E - E_{dep})^2}{E_{dep}} + \frac{10}{R} \sum_{\lambda \in [x, y, z]} (\lambda - \lambda_{true})^2 \quad (4.2)$$

1610 where  $E_{dep}$  is the deposited energy and  $R$  is the radius of JUNO's CD. With the energy in MeV and  
 1611 the distance in meters, we use the factor 0.85 and 10 to balance the two term of the loss function so  
 1612 they have the same magnitude.

1613 The loss function ( $E + V$ ) is close to a simple Mean Squared Error (MSE). MSE is one of the most  
 1614 basic loss function, the derivative is simple and continuous in every point. It is a strong starting  
 1615 point to explore the possibility of CNNs. The loss ( $E_r + V_r$ ) can be seen as a relative MSE.

1616 The idea is that: due to the inherent statistic uncertainty over the number of collected Number of  
 1617 Photo Electrons (NPE), the absolute resolution  $\sigma(E - E_{true})$  will be larger at higher energy than at  
 1618 low energy. But we expect the *relative* energy resolution  $\frac{\sigma(E - E_{true})}{E_{true}}$  to be smaller at high energy than  
 1619 lower energy as illustrated in Figure 3.20. Because of this, by using simple MSE the most important  
 1620 part in the loss come from the high energy part of the dataset whereas with a relative MSE, the  
 1621 most important part become the low energy events in the dataset. We hope that by using a relative  
 1622 MSE, the neural network will focus on low energy events where the reconstruction is considered the  
 1623 hardest.

1624 The above losses and their parameters values results from fine-tuning after multiples runs and  
 1625 adjustments of the full random search.

1626 Each combinations of those hyperparameters (for example ( $N_{blocks} = 2, N_{channels} = 32, \text{FCDNN} =$   
 1627  $(2 * 1024)$ , Loss = ( $E + V$ ))) produce models, hereinafter referred as configurations, are then tested  
 1628 and compared to each other over an analysis sample.

1629 On top those generated models, we define 4 hand tailored models:

- 1630 — Gen<sub>0</sub>:  $N_{blocks} = 4, N_{channels} = 64$ , FCDNN configuration:  $1024 * 2 + 512 * 2$ , Loss  $\equiv E + V$
- 1631 — Gen<sub>1</sub>:  $N_{blocks} = 4, N_{channels} = 64$ , FCDNN configuration:  $1024 * 2 + 512 * 2$ , Loss  $\equiv E_r + V_r$
- 1632 — Gen<sub>2</sub>:  $N_{blocks} = 5, N_{channels} = 64$ , FCDNN configuration:  $4096 * 2 + 1024 * 2$ , Loss  $\equiv E + V$

1633 — Gen<sub>3</sub>:  $N_{blocks} = 5$ ,  $N_{channels} = 64$ , FCDNN configuration:  $4096 * 2 + 1024 * 2$ , Loss  $\equiv E_r + V_r$

1634  
 1635 The resulting models possess between 2'041'034, for Gen<sub>52</sub> and Gen<sub>53</sub>, and 5'759'839'242 parameters,  
 1636 for Gen<sub>26</sub> and Gen<sub>27</sub>. The models of interest in this thesis, from which the results are discussed  
 1637 in Section 4.3, possess 86'197'196 parameters for Gen<sub>30</sub> and 332'187'530 parameters for Gen<sub>42</sub>. For  
 1638 comparison the model of CNN developed in JUNO before posses 38'352'403 parameters [63].

$N_{blocks}$	{2, 3, 4}
$N_{channels}$	{32, 64, 128}
FCDNN configurations	$2 * 1024$ $2 * 2048 + 2 * 1024$ $3 * 2048 + 3 * 512$ $2 * 4096$
Loss	{ $E + V$ , $E_r + V_r$ }

TABLE 4.1 – Sets of hyperparameters values considered in this study

1639 To rank the various configuration we cannot used directly the mean loss over the validation dataset  
 1640 as  $(E + V)$  and  $(E_r + V_r)$  are not numerically comparable. We thus use the following quantities,  
 1641 directly related to the reconstruction performances:

- 1642 — The mean absolute energy error  $\langle E \rangle = \langle |E - E_{true}| \rangle$ . It is an indicator of the energy bias of our  
 1643 reconstruction.
- 1644 — The standard deviation of the energy error  $\sigma E = \sigma(E - E_{true})$ . This the indicator on our  
 1645 precision in energy reconstruction.
- 1646 — The mean distance between the reconstructed vertex and the true vertex  $\langle V \rangle = \langle |\vec{V} - \vec{V}_{true}| \rangle$ .  
 1647 This an indicator of the bias and precision of our vertex reconstruction.
- 1648 — The standard deviation of the distance between the true and reconstructed vertex  $\sigma V = \sigma|\vec{V} - \vec{V}_{true}|$ . This is an indicator if the precision in our vertex reconstruction.

1650  
 1651 The models were developped in Python using the Pytorch framework [49] using NVIDIA A100 [72]  
 1652 and NVIDIA V100 [73] gpus. The A100 was split in two, thus the accessible gpu memory was  
 1653 the same as V100, 20 Gb, making it impossible to train some of the architectures due to memory  
 1654 consumption.

1655 The training was monitored in realtime by a custom tooling that was developed during this thesis,  
 1656 DataMo [74].

1657 The training of one model takes between 4h and 15h depending of its size, overall training the full  
 1658 72 models takes around 500 GPU hours. Even with parallel training, this random search hyper-  
 1659 optimisation was time consuming.

### 1660 4.1.2 Data representation

1661 This data is represented as  $240 \times 240$  images with a charge  $Q$  channel and a time  $t$  channel. The  
 1662 SPMTs are then projected on the plane as illustrated in Figure 4.2b using the coordinate system  
 1663 presented in 4.2a. The  $P_y$  coordinate, the row corresponding to the SPMT in the projection, is  
 1664 proportional to  $\theta$ . The  $P_x$  coordinate, the column corresponding to the SPMT in the projection, is  
 1665 defined by  $\phi \sin \theta$  in spherical coordinates.  $\theta = 0$  is defined as being the top of the detector and  $\phi = 0$   
 1666 is defined as an arbitrary direction in the detector. In practice,  $\phi = 0$  is given by the MC simulation.

$$P_y = \left\lfloor \frac{\theta \cdot H}{\pi} \right\rfloor, \theta \in [0, \pi] \quad (4.3)$$

$$P_x = \left\lfloor \frac{(\phi + \pi) \sin \theta \cdot W}{2\pi} \right\rfloor, \phi \in [-\pi, \pi], \theta \in [0, \pi] \quad (4.4)$$

1667 where  $H$  is the height of the image,  $W$  the width of the image and  $(0, 0)$  the top left corner of the  
 1668 image.

1669 This projection keep the SPMT position in the image proportional to their spherical coordinates while  
 1670 keeping the neighbouring information. This proportionality allow us to keep the specificities of the  
 1671 detector structure, the vertical bands visible in 4.2b.

1672 When two SPMTs in the same pixel are hit in the event time window, the charges are summed and  
 1673 the lowest of the hit-time is chosen. The time window depends on the datasets and are detailed in  
 1674 Section 4.1.2. The SPMTs being located close to each other, we expect the time difference between  
 1675 two successive physics signals, two photons being collected, to be small. The first hit time is chosen  
 1676 because it can be considered as the relative propagation time of the photons that went the "straightest", i.e.  
 1677 that went under the less perturbation of the two. The timing is thus more representative of  
 1678 the event location.

1679 The only potential problem in using this first time come from the Dark Noise (DN). Its time distribution  
 1680 is uniform over the signal and could come before a physics signal on the other SPMT in the pixel.  
 1681 In that case, the time information in the pixel become irrelevant and we lose the timing information  
 1682 for this part of the detector. As illustrated in Figure 4.2b the image dimension have been optimized  
 1683 so that at most two SPMTs are in the same pixel while keeping the number of empty pixels relatively  
 1684 low to prevent this kind of issue.

1685 While it could be possible to use larger images (more pixel) to prevent overlapping, keeping image  
 1686 small images gives multiple advantages:

- 1687 — As presented in Section 4.1.1, the convolution filter we use are  $3 \times 3$  convolution filter, meaning  
 1688 that if SPMTs would be separated by more than one pixel, the first filter would only see one  
 1689 SPMT per filter. This behavior would be kind of counterproductive as the first convolution  
 1690 block would basically be a transmission layer and would just induce noise in the data.
- 1691 — It keep the network relatively small, while this do not impact the convolution layers, the flatten  
 1692 operation just before the FCDNN make the number parameters in the first layer of it dependent  
 1693 on the size of the image.
- 1694 — It reduce the number of empty pixel in the image.

1695  
 1696 The question of empty pixel is an important question in this data representation. There is two kind  
 1697 of empty pixels in the data.

1698 The first kind is pixel that contain a SPMT but the SPMT did not get hit nor registered any dark noise  
 1699 during the event. In this case, the charge channel is zero, which have a physical meaning but then  
 1700 come the question of the time layer. One could argue that the correct time would be infinity (or the  
 1701 largest number our memory allows us) because the hit "never" happened, so extremely far from the  
 1702 time of the event. This cause numerical problem as large number, in the linear operation that are  
 1703 happening in the convolution layers, are more significant than smaller value. We could try to encode  
 1704 this feature in another way but no number have any significance due to our time being relative to  
 1705 the trigger of the experiment so -1 for example is out of question. Float and Double gives us access  
 1706 to special value such as NaN (Not a Number) [75] but the behavior is to propagate the NaN which  
 1707 leaves us with NaN for energy and position. We choose to keep the value 0 because it's the absorbing  
 1708 element of multiplication, absorbing the "information" of the parameter it would be multiplied by.

1709 It also can be thought as no activation in the ReLU activation function. It's important to keep in mind  
 1710 the fact that a part of the detector that has not been hit is also an information: There is no signal in  
 1711 this part of the detector. This problematic will be explored in more details in Chapter 5.

1712 The second kind of pixels are the one that do not represent parts of the detector such as the corners  
 1713 of the image. The question is basically the same, what to put in the charge and the time channel. The  
 1714 decision is to set the charge and time to 0 following the above reasoning.

1715 Another problematic that happens with this representation, and this is not dependent of the chosen  
 1716 projection, is the deformation in the edges of the image and the loss of the neighbouring information  
 1717 in the for the SPMTs at the edge of the image  $\phi \sim 180^\circ$ . This deformation and neighbouring loss  
 1718 could be partially circumvented as explained in Section 4.4

### 1719 4.1.3 Dataset

1720 In this study we will discuss two datasets of one millions prompt signal of IBD events.

#### 1721 J21

1722 The first one comes from the JUNO official MC simulation J21v1r0-Pre2 (released the 18th August  
 1723 2021). This historical version is the one on which the classical SPMT reconstruction algorithm was  
 1724 developed. This classical methods is based on the time likelihood presented Section 3.3 for the vertex  
 1725 reconstruction, and compute the energy by correcting the detector effect on the ration  $N_{pe}/E_{dep}$ . It is  
 1726 detailed in Chapter 4 of [67]. This dataset is used as a reference for comparison to classical algorithm  
 1727 performances. The data in this dataset is *detsim* level (see Section 2.6) which includes no digitization,  
 1728 no DAQ and therefore no reconstruction of PMT signals. Only the number of PEs that hit a PMT and  
 1729 the hit times are provided. A fast simulation based on gaussian drawings produces charges, with  
 1730 bias and variability, and the equivalent for times. The drawings parameters were adjusted based on  
 1731 [29, 76]. Because there is no charge reconstruction, the timing on the event is based on the Geant4  
 1732 simulation, and so  $t = 0$  is the moment the positron is created in the CD. To prevent correlation  
 1733 between the numerical value of the time of the first hit  $t_0$  and the radius of the event, we offset all  
 1734 time by this first hit time. Without simulation of the charge reconstruction, we cannot simulate the  
 1735 event trigger, we thus add an arbitrary time cut at a  $t_0 + 1000$  ns.

#### 1736 J23

1737 The second comes from the JUNO official monte-carlo simulations J23.0.1-rc8.dc1 (released the 7th  
 1738 January 2024). The data is *calib* level (see Section 2.6). Here the charge comes from the waveform  
 1739 integration, the time window resolution and trigger decision are all simulated inside the software.

1740 To put in perspective this amount of data, the expected IBD rate in JUNO is 47 / days. Taking into  
 1741 account the calibration time, and the source reactor shutdown, it amounts to  $\sim 94'000$  IBD events  
 1742 in 6 years. With this million of event, we are training the equivalent of  $\sim 10$  years of data. With  
 1743 this amount we reach a density of  $4783 \frac{\text{event}}{\text{m}^3 \cdot \text{MeV}}$ , meaning our dataset is representative of the multiple  
 1744 event scenarios that could be happening in the detector.

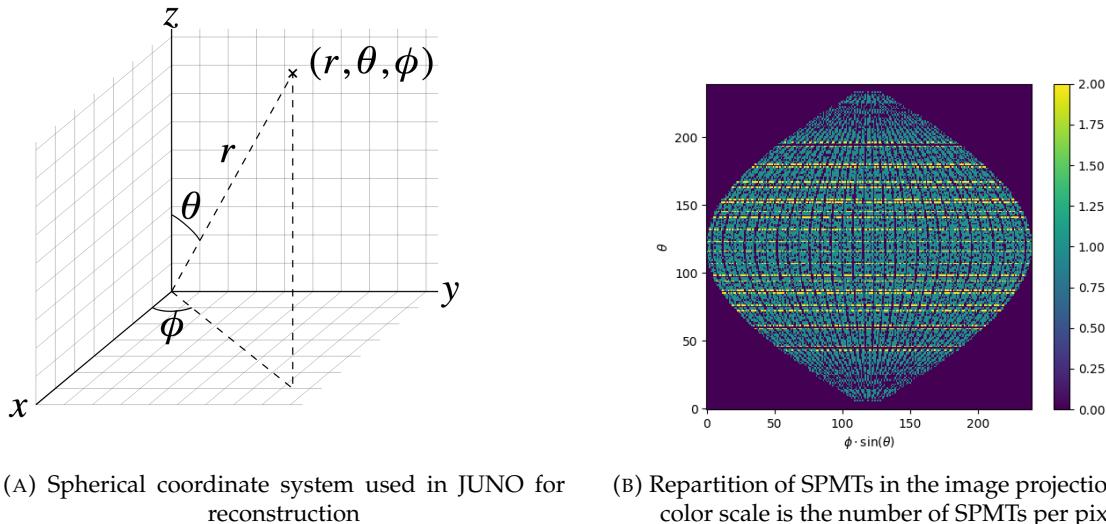
1745 While we expect and hope the MC simulation to give a realistic representation of the detector,  
 1746 there could be effects, even after the fine-tuning on calibration data, that the simulation cannot handle.  
 1747 Thus, once the calibration will be available, we will need to evaluate, and if needed retrain, the  
 1748 network on calibration data to establish definitive performances.

1749 The simulated data is composed of positron events, uniformly distributed in the CD volume and in  
 1750 kinetic energy over  $E_k \in [0; 9]$  MeV producing a deposited energy  $E_{dep} \in [1.022; 10.022]$  MeV. This is

1751 done to mimic the signal produced by the IBD prompt signal. Uniform distributions are used so that  
 1752 the CNN does not learn a potential energy distribution, favoring some part of the energy spectrum  
 1753 instead of other.

#### 1754 4.1.4 Data characteristics

1755 To delve a bit into the kind of data we will use, you can find in Figure 4.2b the repartition of the  
 1756 SPMTs in the image. The color represent the number of SPMTs per pixel.



(A) Spherical coordinate system used in JUNO for reconstruction

(B) Repartition of SPMTs in the image projection. The color scale is the number of SPMTs per pixel

FIGURE 4.2

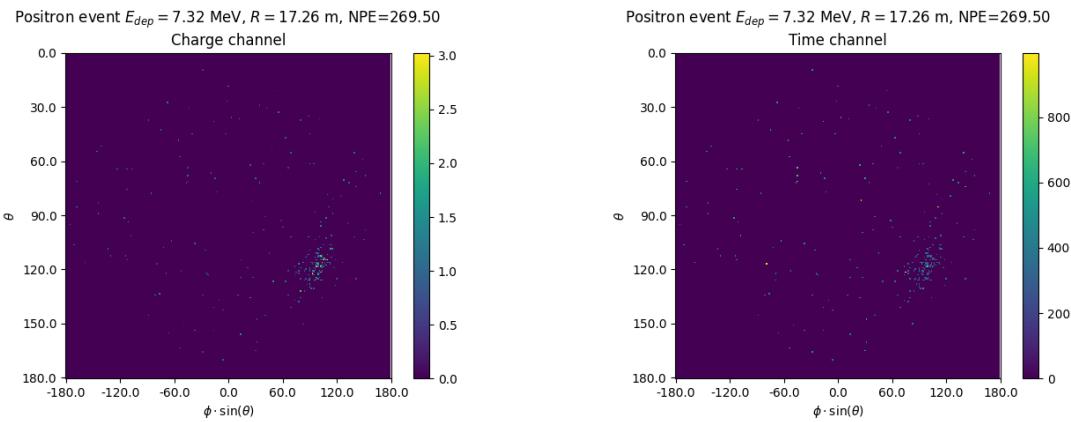


FIGURE 4.3 – Example of a high energy, radial event. We see a concentration of the charge on the bottom right of the image, clear indication of a high radius event. **On the left:** the charge channel. The color is the charge in each pixel in NPE equivalent. **On the right:** The time channel in nanoseconds.

1757 See also Figures 4.3 to 4.6 - and the explanation in their captions - which present events from J23 for  
 1758 different positions and energies. We see some characteristics and we can instinctively understand  
 1759 how the CNN could discriminate different situations.

To give an idea of the strength of the signal in comparison to the dark noise background, Figure 4.7a present the distribution of the ratio of NPE per deposited energy. Assuming a linear response of the

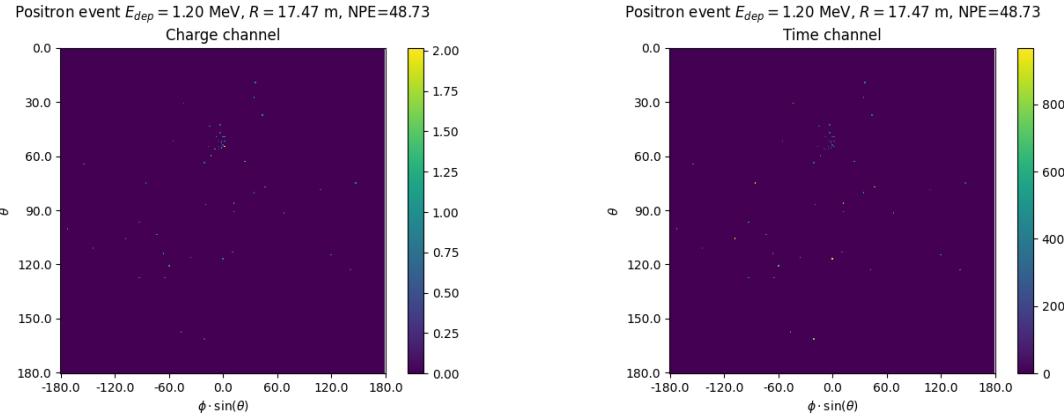


FIGURE 4.4 – Example of a low energy, radial event. The signal here is way less explicit, we can kind of guess that the event is located in the top middle of the image. **On the left:** the charge channel. The color is the charge in each pixel in NPE equivalent. **On the right:** The time channel in nanoseconds.

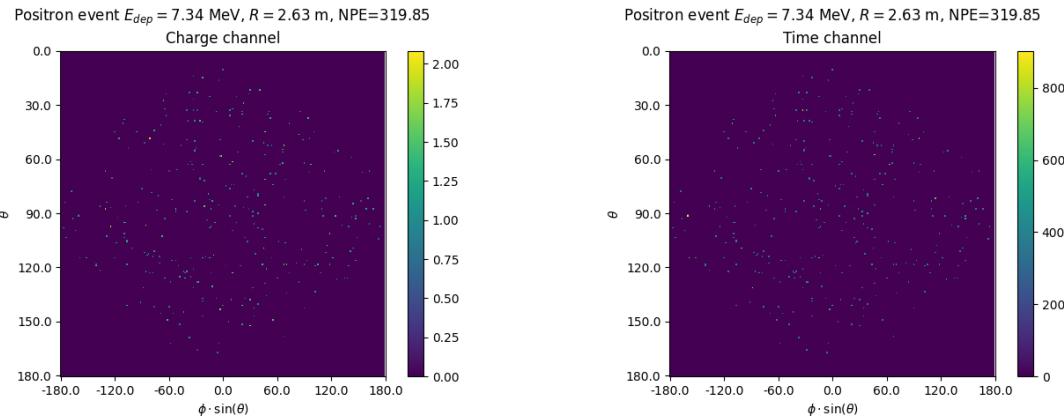


FIGURE 4.5 – Example of a high energy, central event. In this image we can see a lot of signal but uniformly spread, this is indicative of a central event. **On the left:** the charge channel. The color is the charge in each pixel in NPE equivalent. **On the right:** The time channel in nanoseconds.

LS we can model:

$$NPE_{tot} = E_{dep} \cdot P_{mev} + D_N \quad (4.5)$$

$$\frac{NPE_{tot}}{E_{dep}} = P_{mev} + \frac{D_N}{E_{dep}} \quad (4.6)$$

where  $NPE_{tot}$  is the total number of PE detected by the event,  $P_{mev}$  is the mean number of PE detected per MeV and  $D_N$  is the dark noise contribution that is considered energy independent. In the case where the readout time window is dependent of the energy the dark noise contribution become energy dependant, also the LS response is realistically energy dependant but Figure 4.7a shows that we are heavily dominated by the stochastic behavior of light emission and detection.

The fit shows a light yield of 40.78 PE/MeV and a dark noise contribution of 4.29 NPE. As shown in Figure 4.7b, the physics makes for 90% of the signal at low energy.

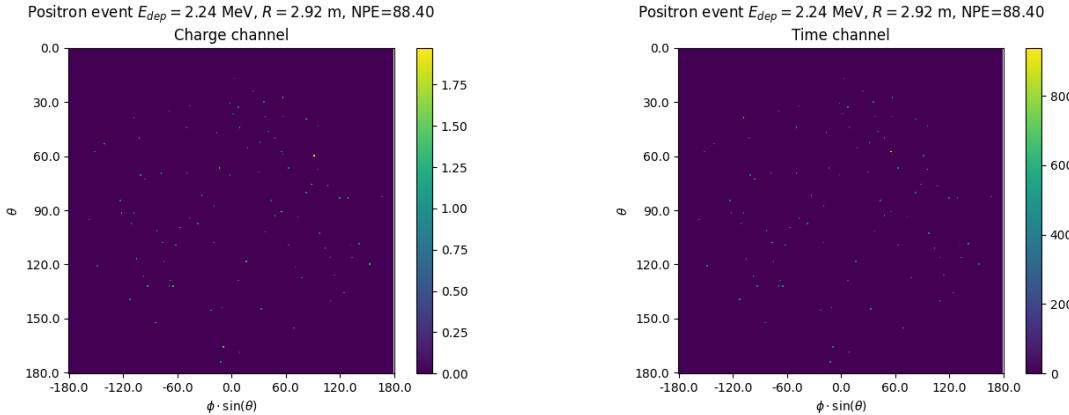


FIGURE 4.6 – Example of a low energy, central event. Here there is no clear signal, the uniformity of the distribution should make it central. **On the left:** the charge channel. The color is the charge in each pixel in NPE equivalent. **On the right:** The time channel in nanoseconds.

## 4.2 Training

The optimizer used for the training is the Adam [44] optimizer, with a learning rate  $\lambda$  of  $1e-3$ . The other hyperparameters were left to their default value ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and  $\epsilon = 1e^{-8}$ ). The learning rate was reduced exponentially during the training at a rate of  $\gamma = 0.95$ , thus  $\lambda_{i+1} = 0.95\lambda_i$  where  $i$  is the epoch.

Following the lifecycle presented in Section 3.1.3, the training used a batch size of 64 events meaning that, each step, the loss is computed on 64 events before updating the NN parameters. An epoch is composed of 10k steps, thus each epoch, the NN sees 640k events. The training last for 30 epochs, so overall the NN goes through 19.2 millions events or 19.2 times the dataset.

The number of epoch, batch size, learning rate and its decay were fine-tuned during the development of the CNN.

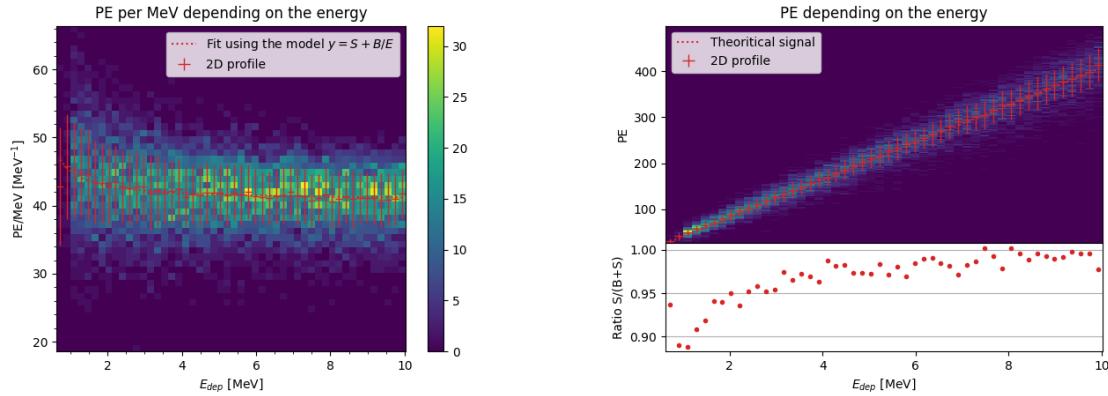
## 4.3 Results

Before presenting the results, let's discuss the different observables.

The events are considered point-like in this study. The target truth position, or vertex, is the mean position of the energy deposits of the positron and the two annihilation gammas. This approximation for point-like interaction is also used for the likelihood study presented in Section 3.3 and in previous ML studies presented in section 3.3.3 [63].

Due to the symmetries of the detector, we mainly consider and discuss the bias and precision evolution depending on the radius  $R$  but we will still monitor the performances depending on the spherical angle  $\theta$  and  $\phi$ . From the detector construction and effect we expect dependency in radius due to the TR area effect presented in Section 3.3 and the possibility for the positron or the gammas to escape from the CD for positrons interacting near the edge. We also expect dependency on  $\theta$ , the top of the experiment being non-instrumented due to the filling chimney. It is also to be noted that the events in the dataset are uniformly distributed in the CD, and so are uniformly distributed in  $R^3$  and  $\phi$ . The  $\theta$  distribution is not uniform and we will have more events for  $\theta \sim 90^\circ$  than  $\theta \sim 0^\circ$  or  $\theta \sim 180^\circ$ .

We define multiple energy in JUNO:



(A) Distribution of PE/MeV in the J23 Dataset. This distribution is profiled and fitted using equation 4.6

(B) On top: Distribution of PE vs Energy. On bottom: Using the values extracted in 4.7a, we calculate the ration signal over background + signal

FIGURE 4.7

- 1793  $E_\nu$ : The energy of the neutrino.
- 1794  $E_k$ : The kinetic energy of the resulting positron from the IBD.
- 1795  $E_{dep}$ : The deposited energy of the positron and the two annihilation gammas.
- 1796  $E_{vis}$ : The equivalent visible energy, so  $E_{dep}$  after the detector effect such as the LS response non-linearity.
- 1797  $E_{rec}$ : The reconstructed energy by the reconstruction algorithm. The expected value depend on the algorithm we discuss about. For example the algorithm presented in Section 3.3 reconstruct  $E_{vis}$  while the ones presented in section 3.3.3 reconstruct  $E_{dep}$ .

1800 In this study, we will set  $E_{dep}$  as our target for energy reconstruction. This choice is motivated by the  
1801 ease with which we can retrieve this information in the monte-carlo data while  $E_{vis}$  is less trivial to  
1802 retrieve.  
1803

### 1804 4.3.1 J21 results

- 1805 The best results comes from the Gen<sub>30</sub> model, meaning then 30th model generated using the table  
1806 4.1: Gen<sub>30</sub>:  $N_{blocks} = 3$ ,  $N_{channels} = 32$ , FCDNN configuration:  $2048 * 2 + 1024 * 2$ , Loss  $\equiv E + V$ .  
1807 The performances of its reconstruction are presented in blue in Figure 4.8. Superimposed in black is  
1808 the performances of the classical algorithm from [67].

### 1809 Energy reconstruction

- 1810 By looking at the Figure 4.8a and 4.8b, the CNN has similar performances in its energy resolution.  
1811 Important biases, however, appear at low and high energy.  
1812 This is explained by looking at the true and reconstructed energy distributions in Figure 4.10a.  
1813 We see that the distributions are similar for energies before 8 MeV but there is an excess of event  
1814 reconstructed with energies around 9 MeV while a lack of them for 10 MeV. The neural network  
1815 seems to learn the energy distribution and learn that it exist almost no event with an energy inferior  
1816 to 1.022 MeV and not event with an energy superior to 10 MeV.

1817 The first observation is a physics phenomena: for a positron, its minimum deposited energy is the  
 1818 mass energy coming from its annihilation with an electron 1.022 MeV. There is a few event with  
 1819 energies inferior to 1.022 MeV, in those case the annihilation gammas or even the positron escape the  
 1820 detector. The deposited energy in the LS is thus only a fraction of the energy of the event.

1821 The second observation is indeed true in this dataset but has no physical meaning, it is an arbitrary  
 1822 limit because the physics region of interest is mainly between 1 and 9 MeV of deposited energy  
 1823 (Figure 2.2). By learning the energy distribution, the CNN pull event from the border of it to more  
 1824 central value. That's why the energy resolution is better: the events are pulled in a small energy  
 1825 region , thus a small variance but the bias become very high (Figure 4.8a).

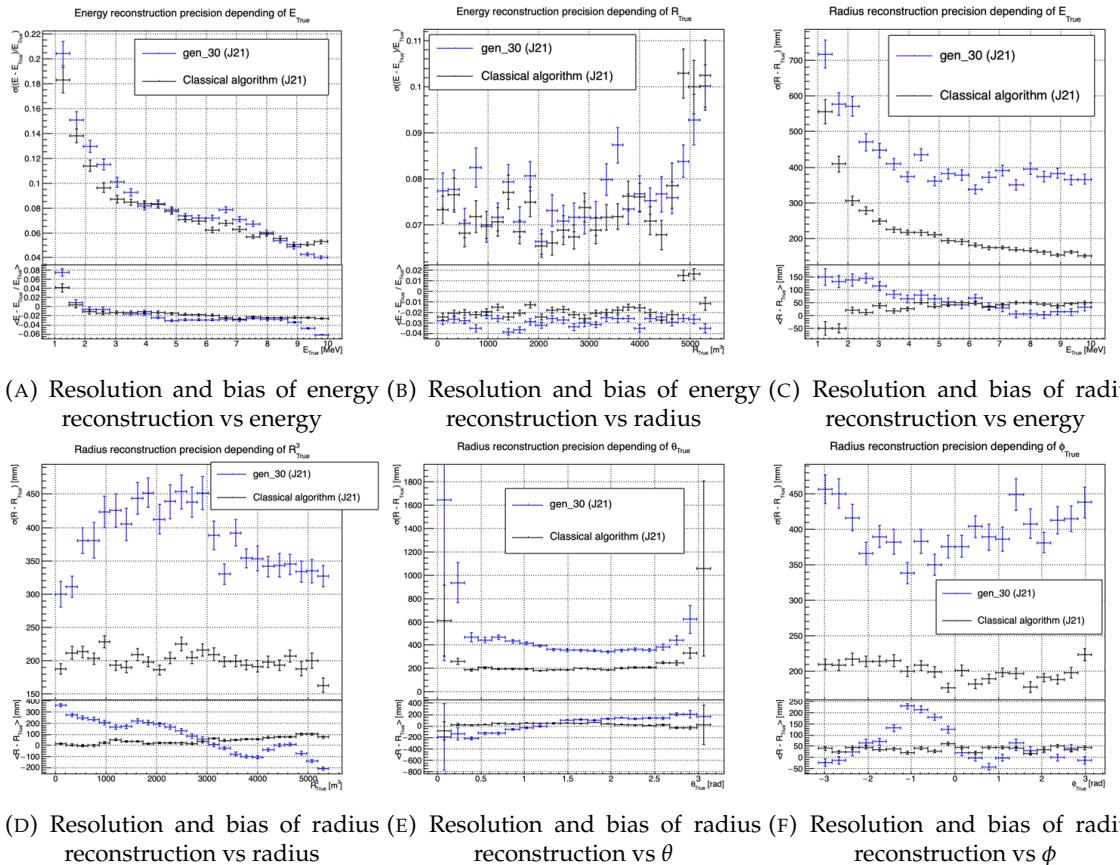


FIGURE 4.8 – Reconstruction performance of the Gen30 model on J21 data and its comparison to the performances of the classic algorithm “Classical algorithm” from [67]. The top part of each plot is the resolution and the bottom part is the bias.

1826 This behavior also explain the heavy bias at low energy in Figure 4.8a. The energy bias of the CNN  
 1827 if fairly constant over the energy range, it is interesting to note that the energy bias depending on the  
 1828 radius is a bit worse than the classical method.

### 1829 Vertex reconstruction

1830 For the vertex reconstruction we do not study  $x$ ,  $y$  and  $z$  independently but we use  $R$  as a proxy  
 1831 observable. Figure 4.9 shows the residual distribution of the different vertex coordinates. We see that  
 1832  $R$  errors and biases are slightly superior to the cartesian coordinates, thus  $R$  is a conservative proxy  
 1833 observable to discuss the subject of vertex reconstruction.

The comparison of radius reconstruction between the classical algorithm and Gen<sub>30</sub> are presented in the Figures 4.8c, 4.8d, 4.8e and 4.8f. The resolution obtained by the CNN is twice worse in average, and worse in all studied regions. In energy, Figure 4.8c, where we see a degradation of almost 20cm over the energy range. When looking over the true event radius, Figure 4.8d, we lose between 30 and 45cm of resolution. The performances are the best for central and radial event.

The precision also worsen when looking at the edge of the image  $\theta \approx 0, \theta \approx 2\pi$  respectively the top and bottom of the image, and when  $\phi \approx -\pi$  and  $\phi \approx \pi$  respectively the left and right side of the image.

The bias in radius reconstruction is about the same order of magnitude depending of the energy but is of opposite sign. As for the energy, this behavior is studied in more details in Section 4.3.2. Over radius,  $\theta$  and  $\phi$  the bias is inconsistent, sometimes event better than the classical reconstruction but can also be much worse than the classical method. This could come from the specialisation of some filters in the convolutional layers for specific part of the detector that would still work “correctly” for other parts but with much less precision.

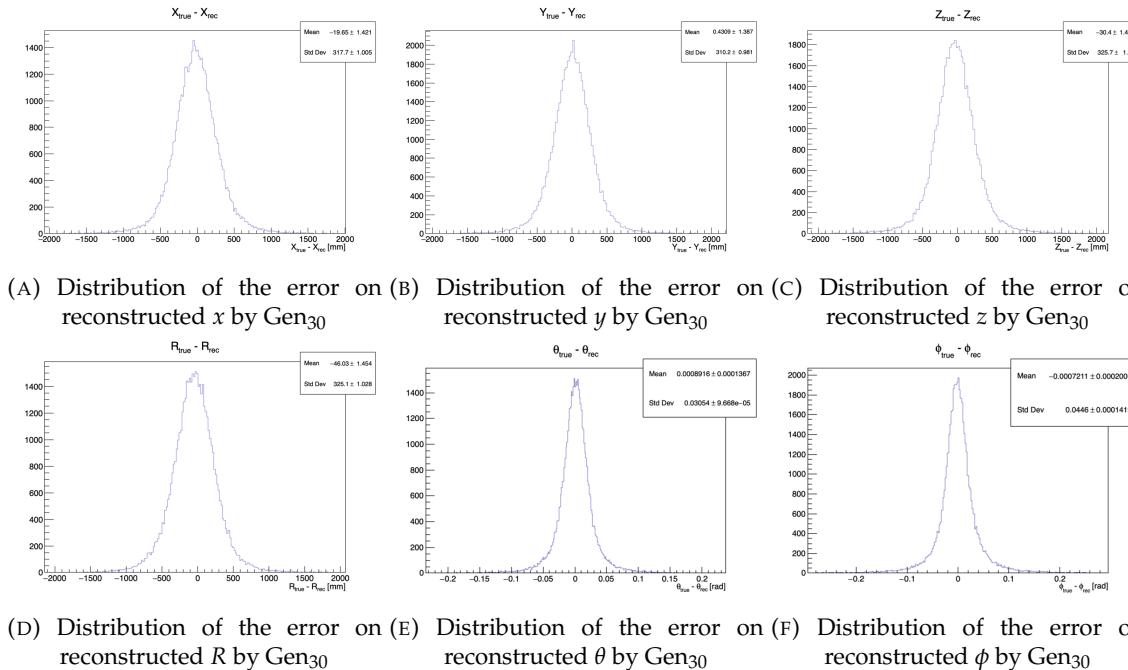


FIGURE 4.9 – Residual distribution of the different component of the vertex by Gen<sub>30</sub>. The reconstructed component are  $x$ ,  $y$  and  $z$  but we see similar behavior in the error of  $R$ ,  $\theta$  and  $\phi$ .

As mentioned in the introduction of this chapter, this CNN initially served as a tool for learning about machine learning and JUNO’s detector and software. It eventually became necessary for use as an SPMT reconstruction tool in Chapter 7, so we made some optimizations. However, we did not invest much time in fully addressing its issues.

### 4.3.2 J21 Combination of classic and ML estimator

As it has been presented in previous section, there is instances where the reconstructed energy and vertex behaves differently between the neural network and the classic algorithm. For instance, if we look at Figure 4.8c, we see that while the CNN tend to overestimate the radius at low energy while the classical algorithm seems to underestimate it. Let’s designate the two reconstruction algorithms

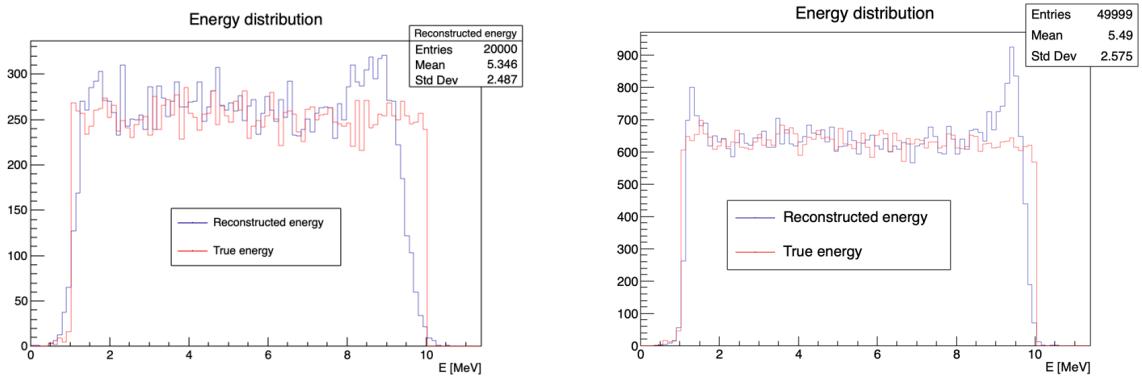
(A) Distribution of Gen<sub>30</sub> reconstructed energy and true energy of the analysis dataset (J21)(B) Distribution of Gen<sub>42</sub> reconstructed energy and true energy of the analysis dataset (J23)

FIGURE 4.10

as estimator of  $X$ , the truth about the event in the phase space  $(E, x, y, z)$ . The CNN and the classical algorithm are respectively designated as  $\theta_N(X)$  and  $\theta_C(X)$ .

$$E[\theta_N] = \mu_N + X; \text{Var}[\theta_N] = \sigma_N^2 \quad (4.7)$$

$$E[\theta_C] = \mu_C + X; \text{Var}[\theta_C] = \sigma_C^2 \quad (4.8)$$

1853 where  $\mu$  is the bias of the estimator and  $\sigma^2$  its variance.

1854 Now if we were to combine the two estimators using a simple mean

$$\hat{\theta}(X) = \frac{1}{2}(\theta_N(X) + \theta_C(X)) \quad (4.9)$$

then the variance and mean would follow

$$E[\hat{\theta}] = \frac{1}{2}E[\theta_N] + \frac{1}{2}E[\theta_C] \quad (4.10)$$

$$= \frac{1}{2}(\mu_N + X + \mu_C + X) \quad (4.11)$$

$$= \frac{1}{2}(\mu_N + \mu_C) + X \quad (4.12)$$

$$\text{Var}[\hat{\theta}] = \frac{1}{4}\sigma_N^2 + \frac{1}{4}\sigma_C^2 + 2 \cdot \frac{1}{4} \cdot \sigma_{NC} \quad (4.13)$$

$$= \frac{1}{4}\sigma_N^2 + \frac{1}{4}\sigma_C^2 + \frac{1}{2} \cdot \sigma_{NC} \quad (4.14)$$

$$= \frac{1}{4}\sigma_N^2 + \frac{1}{4}\sigma_C^2 + \frac{1}{2} \cdot \sigma_N \sigma_C \rho_{NC} \quad (4.15)$$

1855 Where  $\sigma_{NC}$  is the covariance between  $\theta_N$  and  $\theta_C$  and  $\rho_{NC}$  their correlation.

1856 We see immediately that if the two estimators are of opposite bias, the bias of the resulting estimator  
1857 is reduced. For the variance, it depends of  $\rho_{NC}$  but in this case if  $\sigma_C^2$  is close to  $\sigma_N^2$  then even for  
1858  $\rho_{NC} \lesssim 1$  then we can gain in resolution.

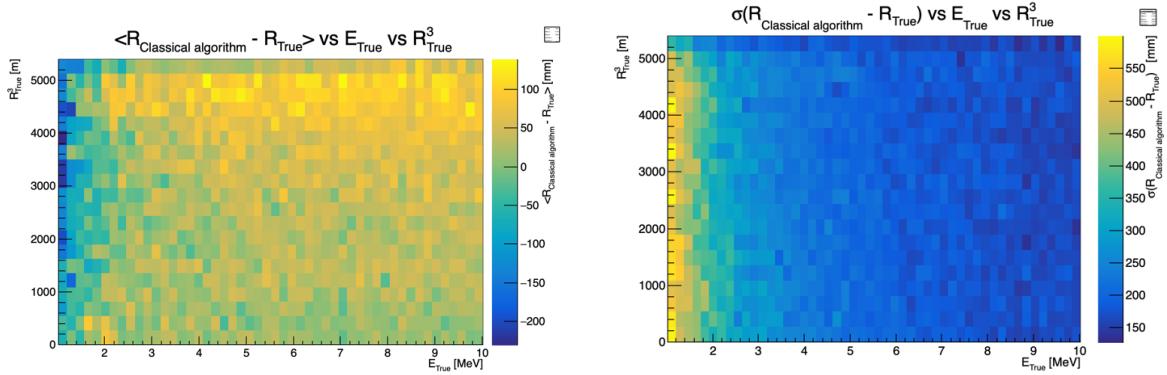


FIGURE 4.11 – Radius bias (on the left) and resolution (on the right) of the classical algorithm in a  $E, R^3$  grid

1859 By generalising the equation 4.9 to

$$\hat{\theta}(X) = \alpha\theta_N + (1 - \alpha)\theta_C; \alpha \in [0, 1] \quad (4.16)$$

1860 we can determine an optimal  $\alpha$  for two combined estimators. The estimators with the smallest  
1861 variance

$$\alpha = \frac{\sigma_C^2 - \sigma_N\sigma_C\rho_{NC}}{\sigma_N^2 + \sigma_C^2 - 2\sigma_N\sigma_C\rho_{NC}} \quad (4.17)$$

1862 and the estimator without bias

$$\alpha = \frac{\mu_C}{\mu_C - \mu_N} \quad (4.18)$$

1863 See annex A for demonstration.

1864 We present in this section the result of the estimator with the smallest variance.

1865 Its pretty clear from the results shown in Figure 4.8 that the bias, variances and correlation are not  
1866 constant across the  $(E, R^3)$  phase space. We thus compute those parameters in a grid in  $E$  and  $R^3$  for  
1867 the following results as illustrated in 4.11.

1868 The map we are using are composed of 20 bins for  $R^3$  going from 0 to 5400 m<sup>3</sup> (17.54 m) and 50 bins  
1869 in energy ranging from 1.022 to 10.022 MeV. In the case where we are outside the grid, we use the  
1870 closest cell.

1871 The performance of this weighted mean is presented in Figure 4.12. We can see that even when the  
1872 CNN resolution is much worse than the classical algorithm, it can still bring some information thus  
1873 improving the resolution. This comes from the correlation of the reconstruction error to be smaller  
1874 than 1 as presented in Figure 4.13. We even see some anticorrelation in the radius reconstruction for  
1875 High radius, high energy, event.

1876 This technique is not suited for realistic reconstruction, we rely too much on the knowledge of the  
1877 resolution, bias and correlation between the two methods. While this is possible to determine using  
1878 simulated data or calibration sources, the real data might differ from our model and we would need  
1879 to really well understand the behavior of the two system. But this is a good tool to detect that  
1880 algorithms don't all use the same information, and is a first step to identify new information that  
1881 could be brought to the best algorithms, to improve their performance.

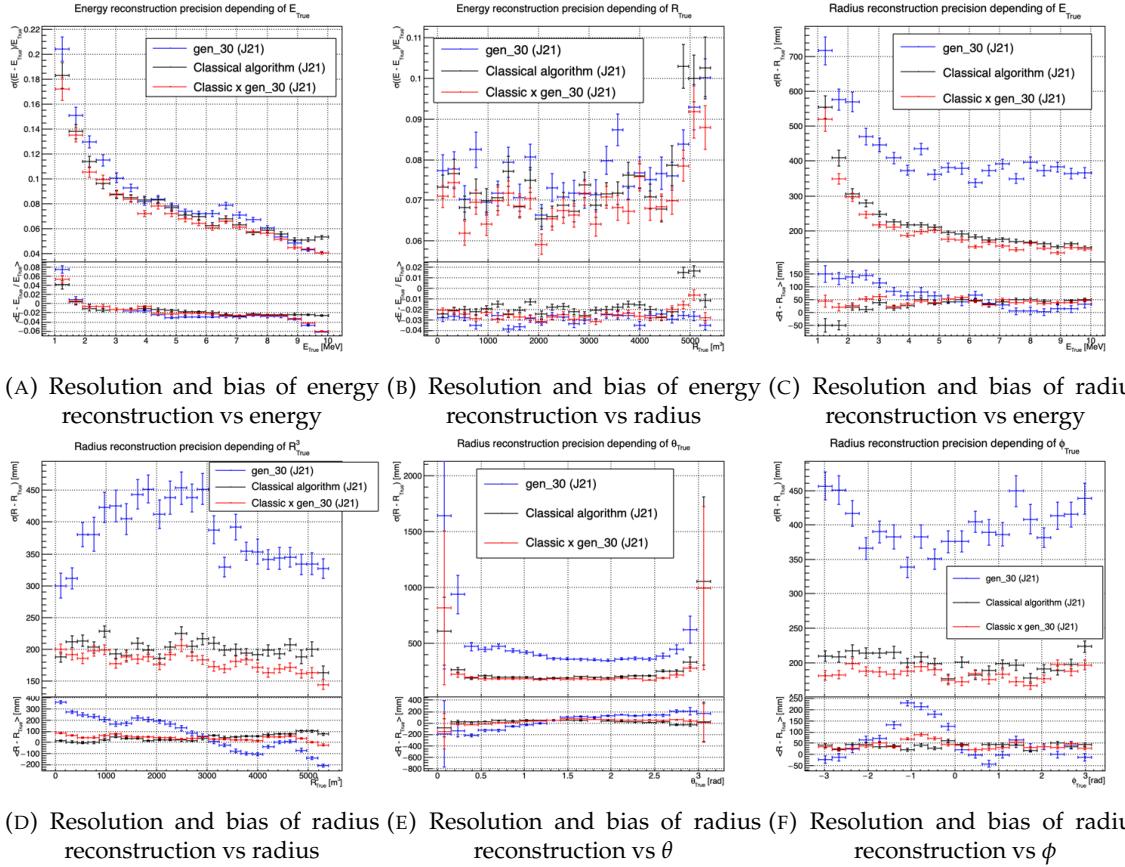


FIGURE 4.12 – Reconstruction performance of the Gen30 model on J21, the classic algorithm “Classical algorithm” from [67] and the combination of both using weighted mean. The top part of each plot is the resolution and the bottom part is the bias.

### 4.3.3 J23 results

We needed for Chapter 7 a SPMT reconstruction tool to run the comparison with LPMT. We thus retrained the SPMT CNN on newer, more realistic data.

The J21 simulation is fairly old and newer version, such as J23, include refined measurements of the light yield, reflection indices of materials of the detector, structural elements such as the connecting structure and more realistic dark noise. Additionally, the trigger, waveform integration and time window are defined using the algorithms that will ultimately be used by the collaboration to process real physics events.

We retrained the models defined in 4.1.1 on the J23 data and used the same hyperparameter optimisation procedure. The results from the best architecture, Gen<sub>42</sub>, are presented in Figure 4.14. Following the table 4.1, Gen<sub>42</sub>:  $N_{blocks} = 3$ ,  $N_{channels} = 64$ , FCDNN configuration:  $4096 * 2$ , Loss  $\equiv E + V$ .

#### Energy reconstruction

The results of the energy reconstruction are presented in Figures 4.14a and 4.14b. The resolution is close to the one of the classical algorithm with the exception of the start and end of the spectrum. This is the same effect that we saw with Gen<sub>30</sub>, events are pulled from the edge of the distribution, resulting in smaller resolution but heavy biases.

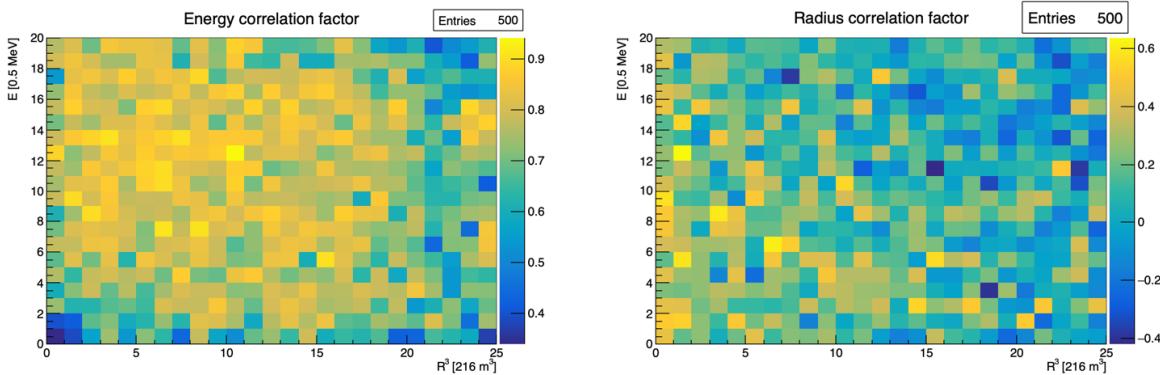


FIGURE 4.13 – Correlation between CNN and classical method reconstruction (on the left) for energy and (on the right) for radius in a  $E, R^3$  grid

### 1898 Vertex reconstruction

1899 The vertex reconstruction, presented in Figures 4.14c, 4.14d, 4.14e and 4.14f is not yet to the level of  
 1900 the classical reconstruction but the degradation is smaller than for Gen<sub>30</sub> being at most a difference  
 1901 of 15cm of resolution and closing to the performance of the classical algorithm in the most favourable  
 1902 condition. Gen<sub>42</sub> has also very little bias in comparison with the classical method with the exception  
 1903 of the transition to the TR area and at the very edge of the detector.

1904 With a more realistic description of the propagation and collection of scintillation photons, of the  
 1905 charge and time resolutions, of the DN and of the trigger, it seems new features can be identified by  
 1906 the CNN.

1907 Unfortunately could not rerun the classical algorithm over the J23 data, as the algorithm was op-  
 1908 timised for J21 and was not included and maintained over J23. The combination method need for  
 1909 the two estimators to be run on the same set of event, which was impossible without the classical  
 1910 algorithm being maintained for J23.

## 1911 4.4 Conclusion and prospect

1912 In this chapter we have developed a CNN for the reconstruction of IBD prompt signals. This work  
 1913 was the opportunity to learn about machine learning and neural networks, and familiarise ourselves  
 1914 with JUNO's detector and software.

1915 This work was revisited for the needs of Chapter 7, providing a reconstruction tools for the SPMT.

1916 The CNN we developed suffers limitations in its performance. We think one of the reasons for this  
 1917 lies in the data representation. First, a lot of training time and resources is consumed going and  
 1918 optimizing over pixels with no physical meaning, notably the time information in case of no hit.  
 1919 This problem origin from the planar projection and is also a specificity of the SPMT system, where  
 1920 a low number of PMT fire per event resulting in empty pixels. To overcome this problematic, i.e.  
 1921 what is the time of a PMT that was never hit, we could transform this channel into a dimension. This  
 1922 would results in an image with multiple charge channels, each one representing the charge sum in a  
 1923 time interval.

1924 Even the best CNN design should at some point hit another limitation : the necessity to project the  
 1925 spherical image on a sphere. It would then need to optimize itself to take into account edges cases  
 1926 such as event at the edge of the image and deformation of the charge distribution. we could imagine  
 1927 a two part CNN where the first part reconstruct the  $\theta$  and  $\phi$  spherical coordinates and then rotate the

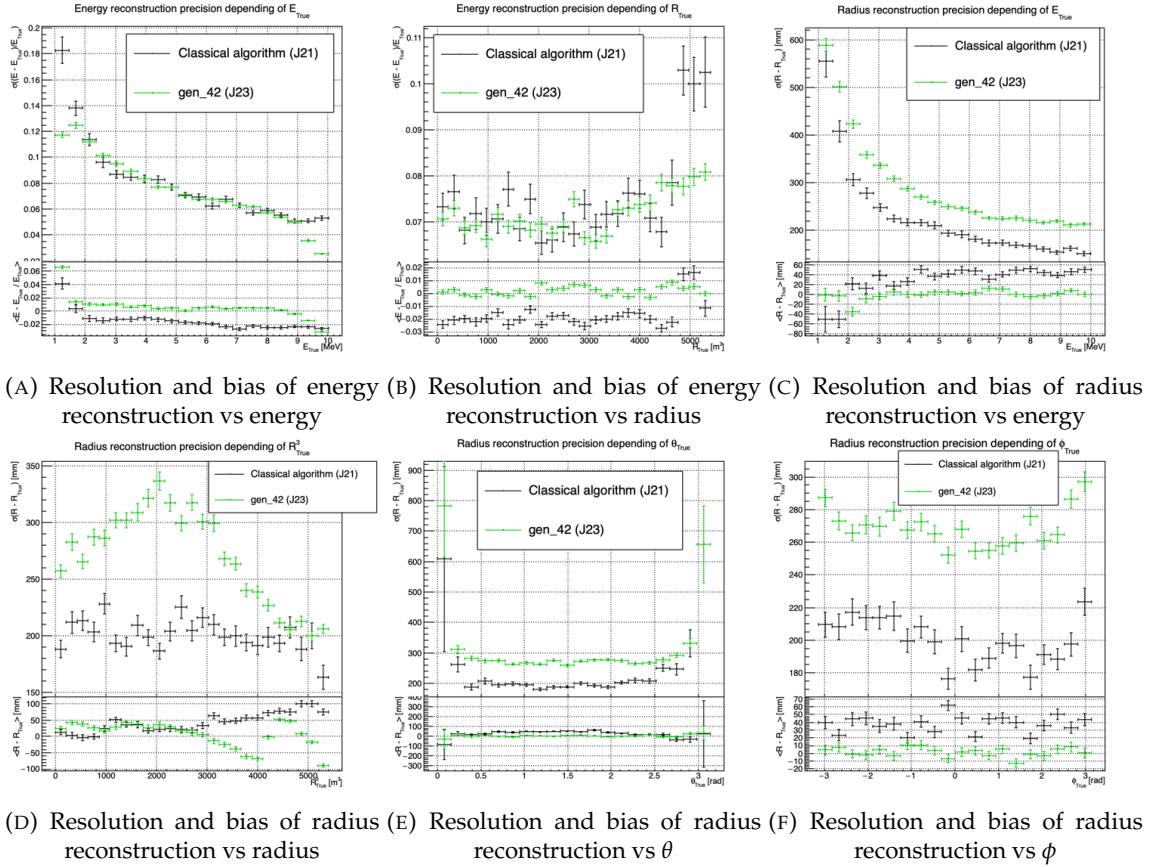


FIGURE 4.14 – Reconstruction performance of the Gen42 model on J23 data and its comparison to the performances of the classic algorithm “Classical algorithm” from [67]. The top part of each plot is the resolution and the bottom part is the bias.

image to locate the event in the center of the image. The second part, from this rotated image, would reconstruct the radius and energy of the event. Another possibility is to use a kind of algorithm that does not impose a planar projection, like a GNN. It has other advantages, as will be presented in the next chapter, where we propose a GNN to reconstruct IBD’s with the LPMT and SPMT systems.

The CNN we developed suffers limitations in its performance. We think one of the reasons for this lies in the data representation. A lot of training time and resources is consumed going and optimizing over pixel with no physical meaning, the NN needs to optimized itself to take into account edges cases such as event at the edge of the image and deformation of the charge distribution.

Those problems could be circumvented, we could imagine a two part CNN where the first part reconstruct the  $\theta$  and  $\phi$  spherical coordinates and then rotate the image to locate the event in the center of the image. The second part, from this rotated image, would reconstruct the radius and energy of the event.

To overcome the time problematic, i.e. what is the time of a PMT that was never hit, we could transform this channel into a dimension. This would results in an image with multiple charge channels, each one representing the charge sum in a time interval.

Another possibility is to use a kind of algorithm that does not impose a planar projection, like a GNN. It has other advantages, as will be presented in the next chapter, where we propose a GNN to reconstruct IBD’s with the LPMT system.



1946 **Chapter 5**

1947 **Graph representation of JUNO for  
IBD reconstruction**

1949

*"The Answer to the Great Question of Life, the Universe and  
Everything is Forty-two"*

*Douglas Adams, The Hitchhiker's Guide to the Galaxy*

1950


---

## Contents

19511952195319541955195619571958195919601961196219631964

<b>5.1</b>	<b>Data representation</b>	82
<b>5.2</b>	<b>Message passing algorithm</b>	85
<b>5.3</b>	<b>Data</b>	87
<b>5.4</b>	<b>Model</b>	88
<b>5.5</b>	<b>Training</b>	89
<b>5.6</b>	<b>Optimization</b>	90
5.6.1	Software optimization	90
5.6.2	Hyperparameters optimization	91
<b>5.7</b>	<b>performance of the final version</b>	92
<b>5.8</b>	<b>Conclusion</b>	95

---

196519661967196819691970

In Section 3.3.3, we showed that all ML methods developed before this thesis to reconstruct IBDs have similar results, and that their performance is very similar to that of the classical, likelihood-based algorithm. We think these similarities can reasonably be explained by this: the input data used by all these methods to compute  $E$  or  $\vec{X}$  is the same full list of PMT integrated signals  $\{(Q_i, t_i); i \in 1, \dots, N_{PMTs}\}$ , and by the high level of sophistication of the detector's description in the likelihood. It's probable that the likelihood method looses very little information.

197119721973197419751976

May be some was, but that the ML algorithms were not designed well enough to recover it. It's also reasonable to think that ML algorithms will make a difference when, instead of the list of  $(Q_i, t_i)$ , a rawer information will be used in input, like the full waveform. To actually be able to learn from such a complex and high dimensional input, well designed architectures (that would guide the learning toward the solution) are necessary. In any case, it seemed welcome to us to propose an additional algorithm, with an original architecture.

1977197819791980198119821983

For the fist stage of its development, the purpose of this part of my thesis, we considered it was enough to also take the  $(Q_i, t_i)$  list as the input. While achieving equivalent performance with simpler input might suggest that the architecture is not immediately advantageous, it remains crucial to explore the performance with more complex, rawer inputs such as full waveforms. This is where the true potential of the architecture could emerge, as it could better capture the intricacies that simpler inputs fail to represent. If performance does not improve with these richer inputs, it would then be appropriate to question the relevance of this approach.

The algorithm we propose is a GNN. It also has the advantage of addressing sphericity issues described in Chapter 4. From this graph representation, we can construct a neural network that will process the data while keeping some interesting properties. For example the rotational invariance, i.e. the energy and radius of the event do change by rotation our referential. For more details see Section 3.2.3. Graph representation also has the advantage to be able to encode global and higher order informations.

## 5.1 Data representation

In Section 3.3.3, we mentioned a GNN developed before the beginning of this thesis to reconstruct IBD energies in JUNO [63]. In their approach: nodes of the graph correspond to 3072 pixels representing geometric regions of the detector and the information of the  $\sim 6$  LPMTs found in a pixel are then aggregated on those nodes. This aggregation serves to simplify the data input, though at the potential cost of losing finer-grained details. The network then process the data using the equivalent of convolution but on graph [53]. In the first layer, each node is connected only with its direct neighbours.

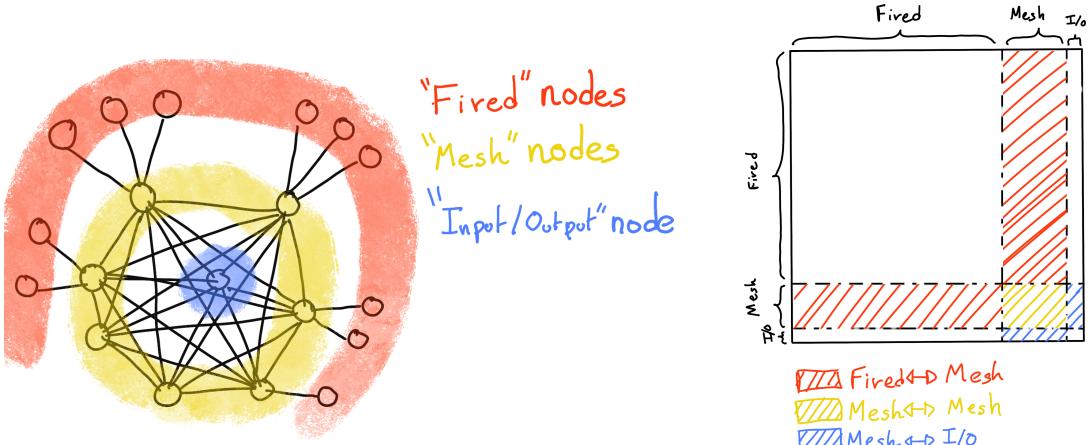
To determine the energy released by an IBD in the LS, it is helpful to determine the position of the main energy deposit. Therefore, relative Q and t's of PMTs all around the sphere is a useful information. If in the first layer only neighbour nodes are linked, several layers are necessary to access this detector-wide information. In an ideal world, we would develop a Graph NN where each PMT is a node (even if it has not been hit in the event under consideration, since this is in itself an information) and where each node is connected to all the other ones. This makes the detector-wide information available as early as the first layer. This architecture might help the network to better learn. Such an architecture can also be motivated this way: one of the strength of GNN's is their capacity to encompass the characteristics of a detector. A node can be the representation of a detector element, and the edge can represent its relationship with other elements. In the case of JUNO, any measurement is collective: an interaction is seen by all the PMTs, with no a priori hierarchy in the role of each. A fully connected GNN is particularly advantageous in JUNO's case, as the lack of a priori hierarchy among the PMTs makes it important to ensure that information is shared globally from the outset. This architecture allows the network to access detector-wide information as early as the first layer, potentially improving learning efficiency. However, this comes at a significant computational cost, which necessitates careful balancing between memory usage and model performance

Another advantage of a GNN is also that it is well adapted to inhomogenous detectors. We therefore tried to build GNNs including both LPMTs and SPMTs.

With 17612 LPMTs and 25600 SPMTs, the ideal fully connected Graph mentioned above is impossible: even excluding self relation and considering the relation to be undirected (the edge from a node A to a node B being the same from as the one from B to A) the amount of necessary edges would be  $n(n - 1)/2$  with  $n = 43212$  nodes. This amounts to 933'616'866 edges. If we encode an information with double precision (64 bits) in what we call an adjacency matrix, illustrated in Figure 3.12, each information we want to encode in the relation would consume 4 GB of data. When adding the overhead due to gradient computation during training, this would put us over the memory capacity of a single V100 gpu card (20 GB of memory). We could use parallel training to distribute the training over multiple GPU but we considered that the technical challenge to deploy this solution was too high.

We finally decided of a middle ground where we define three *families* of nodes:

- The core of the graph is composed of nodes representing geometric regions of the detector. We call those nodes **mesh** nodes. Those mesh nodes are all connected to each other. We keep their number low to gain in memory consumption.



(A) Illustration of the different nodes in our graphs and their relations.

(B) Illustration of what a dense adjacency matrix would look like and the part we are really interested in. Because Fired → Mesh and Mesh → I/O relations are undirected, we only consider in practice the top right part of the matrix for those relations.

FIGURE 5.1

- PMTs in which Photo-Electrons (PE) are found are represented by **fired** nodes. Fired nodes are connected to the mesh node they geometrically belong to.
- A final node is called the input/output node (**I/O**). It is connected to every mesh node. Its features are combinations of signals found in the whole detector.

Those nodes and their relations are illustrated in Figure 5.1a. From this representation, we end up with three distinct adjacency matrix

- A  $N_{\text{fired}} \times N_{\text{mesh}}$  adjacency matrix, representing the relations between fired and mesh. Those relations are undirected.
- A  $N_{\text{mesh}} \times N_{\text{mesh}}$  adjacency matrix, representing the relation between meshes. Those relation are directed.
- A  $N_{\text{mesh}} \times 1$  adjacency between the mesh and I/O nodes. Those relations are undirected.

The adjacency matrix representing those relation is illustrated in Figure 5.1b.

The mesh segmentation is following the Healpix segmentation [77]. This segmentation offer the advantage that almost each mesh have the same number of direct neighbours and it guarantee that each mesh represent the same extent of the detector surface. The segmentation can be infinitely subdivided to provide smaller and smaller pixels. The number of pixel follow the order  $n$  with  $N_{\text{pix}} = 12 \cdot 4^n$ . This segmentation is illustrated in Figure 5.2. To keep the number of mesh small, we use the segmentation of order 2,  $N_{\text{pix}} = 12 \cdot 4^2 = 192$ .

We decided on having the different kind of nodes **mesh (M)**, **fired (F)** and **I/O** have different set of features. The features used in the graph are presented in tables 5.1 and 5.2. Most of the features are low level informations such as the charge or time information but we include some high order features such as

1.  $P_l^h$ : Is the normalized power of the  $l$ th spherical harmonic. For more details about spherical harmonics in JUNO, see annex B.



FIGURE 5.2 – Illustration of the Healpix segmentation. **On the left:** A segmentation of order 0. **On the right:** A segmentation of order 1

2. **A** and **B** are informations that are related the likeliness of the interaction vertex to be on the segment between the center of two meshes.

$$\mathbb{A}_{ij} = (\vec{j} - \vec{i}) \cdot \frac{l_1}{D_{ij}} + \vec{i} \quad (5.1)$$

$$\mathbb{B}_{ij} = \frac{Q_i}{Q_j} \left( \frac{l_2}{l_1} \right)^2 \quad (5.2)$$

$$l_1 = \frac{1}{2} (D_{ij} - \Delta t \frac{c}{n}) \quad (5.3)$$

$$l_2 = \frac{1}{2} (D_{ij} + \Delta t \frac{c}{n}) \quad (5.4)$$

where  $\vec{i}$  is the position vector of the mesh  $i$ ,  $D_{ij}$  is the distance between the center of the meshes  $i$  and  $j$ ,  $Q_i$  the sum of charges on the mesh  $i$ ,  $\Delta t = t_i - t_j$  where  $t_i$  the earliest time on the mesh  $i$  and  $n$  the optical index of the LS. **A** is the vertex between center of meshes distance ratio between  $i$  and  $j$  based on the time information. For **B**, the charge ratio evolve with the square of the distance, so the mesh couple with the smallest **B** should be the one with the interaction vertex between its two center.

Fired	Mesh	I/O
$Q$	$\langle Q_m \rangle$	$\langle X \rangle$
$t$	$\sigma Q_m$	$\langle Y \rangle$
$x$	$\min(t_m)$	$\langle Z \rangle$
$y$	$\max(t_m)$	$\sum Q$
LPMT/SPMT: 1/-1	$\sigma t_m$ $X_m$ $Y_m$ $Z_m$	$P_l^h; l \in [0, 8]$

TABLE 5.1 – Features on the nodes of the graph. All charge are in [nPE], time in [ns] and position in [m].

$Q$  and  $t$  are the reconstructed charge and time of the hit PMTs.  $(x, y, z)$  is the position of the PMTs and the last parameter represent the type of the PMT. It's 1 for LPMT and -1 for SPMT

$Q_m$  and  $t_m$  is the set of charges and time of the PMT belonging the mesh  $m$ .  $(X_m, Y_m, Z_m)$  i the position of the center of the geometric region represented by the mesh  $m$

$(\langle X \rangle, \langle Y \rangle, \langle Z \rangle)$  is the position of the charge barycenter,  $\sum Q$  the sum of the collected charge in the detector and  $P_l^h$  is the relative power of the  $l$ th harmonic. See annex B for details.

Fired → Mesh	Mesh ( $m_1$ ) → Mesh ( $m_2$ )	Mesh → I/O
$x - X_m$	$X_{m1} - X_{m2}$	$\langle X \rangle - X_m$
$y - Y_m$	$Y_{m1} - Y_{m2}$	$\langle Y \rangle - Y_m$
$z - Z_m$	$Z_{m1} - Z_{m2}$	$\langle Z \rangle - Z_m$
$t - \min(t_m)$	$\min(t_{m1}) - \min(t_{m2})$	$\sum Q_m / \sum Q$
$Q / \sum Q_m$	$\frac{\langle Q_{m1} \rangle - \langle Q_{m2} \rangle}{\langle Q_{m1} \rangle + \langle Q_{m2} \rangle}$ $D_{m1 \rightarrow m2}^{-1}$ $\mathbb{A}$ $\mathbb{B}$	$\langle t_m \rangle$

TABLE 5.2 – Features on the edges on the graph. It use the same notation as in table 5.1.  $D_{m1 \rightarrow m2}^{-1}$  is the inverse of the distance between the mesh  $m_1$  and the mesh  $m_2$ . The features  $\mathbb{A}$  and  $\mathbb{B}$  are detailed in Section 5.1

2061 Since our different nodes do not have the same number of features, they exist in distinct spaces.  
 2062 Traditional graph neural networks only handle homogeneous graphs, where the nodes and edges  
 2063 have the same number of features at each layer. Therefore, the libraries and publicly available  
 2064 algorithms we found were not suited to our needs. As a result, we had to develop and implement a  
 2065 custom message-passing algorithm capable of handling our heterogeneous graph.

## 2066 5.2 Message passing algorithm

2067 The message passing algorithm define the way the GNN will compute and update its graph. As it is  
 2068 detailed in Section 3.2.3, the message-passing algorithm allows each node in the graph to update its  
 2069 features based on information from its neighboring nodes. This update process enables the network  
 2070 to propagate information through the graph, allowing nodes to gradually integrate knowledge about  
 2071 the entire detector. This step is crucial for ensuring that each node can take into account not only its  
 2072 local neighborhood but also the broader context of the event.

2073 As introduced in previous section and in the tables 5.1 and 5.2, our graphs nodes and edges will  
 2074 have different number of features depending on their nature, meaning that we cannot have a single  
 2075 message passing function. We thus need to define a message passing function for each transition  
 2076 inside or outside a family. Using the notation presented in Section 3.2.3:

$$n_i^{k+1} = \phi_u(n_i^k, \square_j \phi_m(n_i^k, n_j^k, e_{ij}^k)); n_j \in \mathcal{N}'_i \quad (5.5)$$

and denoting the mesh nodes  $M$ , the fired nodes  $F$  and the I/O node  $IO$ , we need to define

$$\begin{aligned} & \phi_{u;F \rightarrow M}; \phi_{m;F \rightarrow M} \\ & \phi_{u;M \rightarrow F}; \phi_{m;M \rightarrow F} \\ & \phi_{u;M \rightarrow M}; \phi_{m;M \rightarrow M} \\ & \phi_{u;M \rightarrow IO}; \phi_{m;M \rightarrow IO} \\ & \phi_{u;IO \rightarrow M}; \phi_{m;IO \rightarrow M} \end{aligned}$$

2077 to update the nodes after each layers. Following the illustration in Figure 5.3, for each transition  
 2078 between families or inside a family we need an aggregation, a message and an update function. For  
 2079 the aggregation, we use the sum. We use the same, simple, formalism for every  $\phi_u$ :

$$\phi_u \equiv I_{i'}^{n'} = I_i^n A_{i',e}^i W_n^{e,n'} + I_i^n S_n^{n'} + B^{n'} \quad (5.6)$$

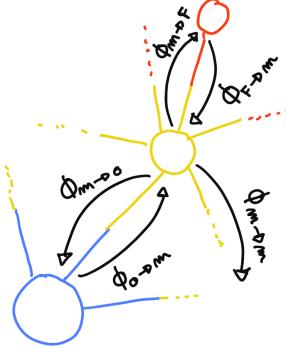


FIGURE 5.3 – Illustration of the different update function needed by our GNN

using the Einstein summation notation. The second order tensor, or matrix,  $I_i^n$  is holding the nodes informations with  $i$  the node index and  $n$  the feature index.  $n$  represent the features of the previous layer and  $n'$  the features of this layer.

$A_{i',e}^i$  is the adjacency tensor, discussed in the previous section, representing the edges between the node  $i'$  and the node  $i$ , each edges holding the features indexed by  $e$ . If the edge does not exist, the features are set to 0. This choice is justified by the linearity of the operation in equation 5.6 : whatever the weights, when multiplied by 0 the results is 0 and the sum result is unchanged.

The learnable parameters are composed of:

- The third order tensor  $W_n^{e,n'}$  which represent the passage from the previous combined feature space between the node and the edge features  $n \otimes e$ , the previous layer, to the current space  $n'$ , this layer.
- The first order tensor  $B^{n'}$  which is a learnable bias on the new features  $n'$ .
- The second order tensor  $S_n^{n'}$ , which can be viewed as a self loop relation where the node update itself based on the previous layer informations, going from the previous space  $n$  to the current space  $n'$ .

If a node have neighbours in different families, the different IAW coming from the different families are summed.

$$I' = \sum_{\mathcal{N}} [I_{\mathcal{N}} AW] + IS + B \quad (5.7)$$

where  $\mathcal{N}$  are the neighbouring family. In our case, dropping the tensor indices and indexing by family for readability, we get

$$I'_F = I_M A_{M \rightarrow F} W_{M \rightarrow F} + I_F S_F + B_F \quad (5.8)$$

$$I'_M = I_F A_{F \rightarrow M} W_{F \rightarrow M} + I_M A_{M \rightarrow M} W_{M \rightarrow M} + I_{IO} A_{IO \rightarrow M} W_{IO \rightarrow M} + I_M S_M + B_M \quad (5.9)$$

$$I'_{IO} = I_M A_{M \rightarrow IO} W_{IO \rightarrow M} + I_{IO} S_{IO} + B_{IO} \quad (5.10)$$

We thus have a  $S$ ,  $W$  and  $B$  for each of the  $\phi_u$  function we defined above. The IAW sum can be seen as the  $\phi_m$  function and  $IS + B$  as the second part of the  $\phi_u$  function. Eq 5.5 gave the generic form of message passing : to update a node  $i$ , one first combines informations from the surrounding nodes and edges and then combine the result ( $\square_j \phi_m$ ) with the current features of node  $i$ . Many practical ways to combine can be tried. In our implementation of message passing (Eq. 5.6 and 5.7) the latter combination is the simple sum of the former (IAW, the equivalent of  $\square_j \phi_m$ ) with a linear combination of the current features of node  $i$  ( $IS + B$ ).

2104 Interestingly, the number on learnable weight in those layer is independent of the number of nodes  
 2105 in each family and depends solely on the number of features on the nodes and the edges.

2106 The expression above only update the node features. We could update the edges, using the results of  
 2107  $\phi_m$  for example, but for technical simplicity we only update the nodes and keep the edges constant.  
 2108 Preserving the edges after each layers allow to share the adjacency matrix between all layers, saving  
 2109 memory and computing time.

2110 This operation of message passing is the constituent of our message passing layers, designed in this  
 2111 work as *JWGLayer*, each of them owning there own set of parameter  $W$ ,  $S$  and  $B$ . To those layers, we  
 2112 can adjoin an activation function such as *PReLU*

$$I' = \text{PReLU} \left( \sum_{\mathcal{N}} \left[ I_{\mathcal{N}} A W \right] + IS + B \right) \quad (5.11)$$

### 2113 5.3 Data

2114 The dataset consists of 1M simulated positron events from the JUNO official simulation version  
 2115 J23.0.1-rc8.dc1. This version of the simulation incorporates both the physics of the detector and its  
 2116 electronics, ensuring that the events closely reflect real detector conditions. Importantly, this version  
 2117 includes advanced digitization and trigger modeling, making it suitable for testing the reconstruction  
 2118 capabilities of our GNN model. Those events are uniformly distributed in energy with  $E_k \in [0, 9]$   
 2119 MeV and distributed in the detector.

2120 All the event are *calib* level, with simulation of the physics, electronics, digitizations and triggers.  
 2121 900k events will be used for the training, 50k for validation and loss monitoring and 50k for the  
 2122 results analysis in Section 5.7. Each events is between 2k and 12k fired PMTS, resulting in fired  
 2123 nodes being the largest family in our graphs in all circumstances as illustrated in Figure 5.4c.

2124 As expected, by comparing the scale between the Figure 5.4a and 5.4b we see that the LPMT system  
 2125 is predominant in term of informations in our data. The number of PMT hits grow with energy but  
 2126 do not reach 0 for low energy event due to the dark noise contribution which seems to be around  
 2127 1000 hits per event for the LPMT system (left limit of Figure 5.4a) and around 15 hits per event for  
 2128 the SPMT system (left limit of Figure 5.4b) which is consistent with the results show in Section 4.1.2.

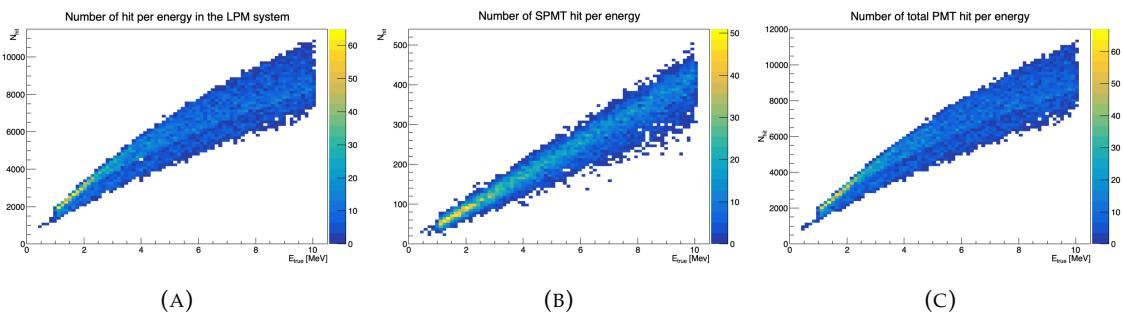


FIGURE 5.4 – Distribution of the number of hits depending on the energy. **On the right:** for the LPMT system. **In the middle :** for the SPMT system. **On the left:** For both system.

2129 The structure seen in the distribution in Figure 5.4a comes from the shape of the number of hits  
 2130 depending on the radius as shown in Figures 5.5a and 5.5b where the number of hit decrease with  
 2131 radius. It is important to understand that this is not representative of the number of PE per event  
 2132 and the decrease in hits over the radius means that the PE are just more concentrated in a smaller  
 2133 number of PMTs.

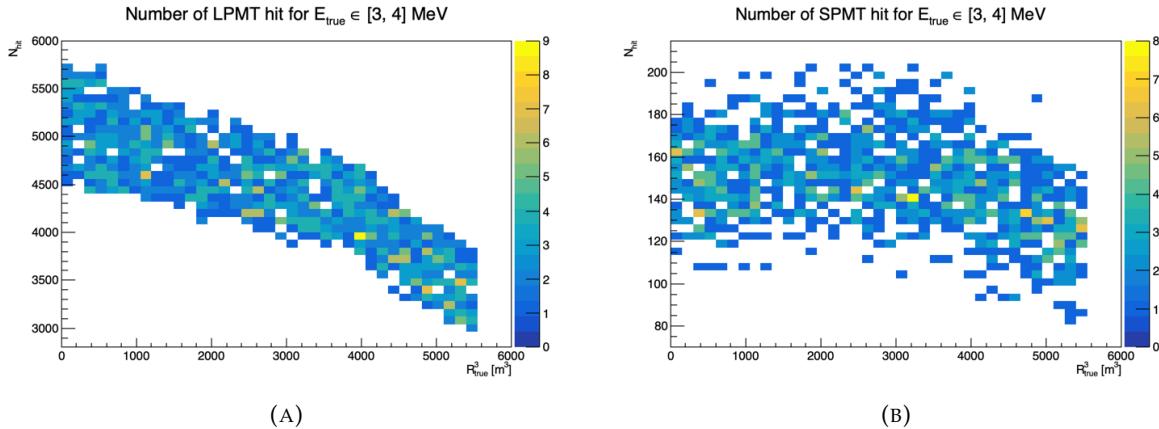


FIGURE 5.5 – Distribution of the number of hits depending on the radius. **On the right:** for the LPMT system. **On the right :** for the SPMT system. To prevent the superposition of structure of different scales we limit ourselves to the energy range  $E_{true} \in [0, 9]$ .

2134 No quality cut is applied here, we rely only on the trigger system. It means that event that would not  
 2135 trigger are not present in the dataset but for events that triggered twice, it happens rarely, the two  
 2136 trigger are considered as two separate event.

## 2137 5.4 Model

2138 In this section, we discuss the different layers that compose the final version of the model. The num-  
 2139 ber of layers, their dimensions, and their arrangement were fine-tuned through multiple iterations.  
 2140 As mentioned earlier, each JWGLayer is defined by the number of features on the nodes and edges of  
 2141 the output graph, assuming it takes as input the graph from the previous layer. For simplicity, when  
 2142 discussing a graph configuration, it will be presented as follow: {  $N_f$ ,  $N_m$ ,  $N_{IO}$ ,  $N_{f \rightarrow m}$ ,  $N_{m \rightarrow m}$ ,  $N_{m \rightarrow f}$   
 2143 } where

- 2144 —  $N_f$  is the number of feature on the fired nodes.
- 2145 —  $N_m$  is the number of features on the mesh nodes.
- 2146 —  $N_{IO}$  is the number of features on the I/O node.
- 2147 —  $N_{f \rightarrow m}$  is the number of features on the edges between the fired and mesh nodes.
- 2148 —  $N_{m \rightarrow m}$  is the number of features on the edges between two mesh nodes.
- 2149 —  $N_{m \rightarrow f}$  is the number of features on the edges between the mesh nodes and the I/O node.

2150 Because we do not change the number of features on the edges, we can simplify the notation to {  $N_f$ ,  
 2151  $N_m$ ,  $N_{IO}$  }. As an example, the input graph configuration, following the tables 5.1 and 5.2 is { 6, 8, 13,  
 2152 5, 8, 5 } or, without the edge features, { 6, 8, 13 } .

2153 The final version of the model, called JWGV8.4.0 is composed of

- 2154 — An JWGLayer, converting the input graph { 6, 8, 13 } to { 64, 512, 2048 } with a PReLU activation  
 2155 function.
- 2156 — 3 resnet layers, each of them composed of
  - 2157 1. 2 JWG layers with a PReLU activation function. They do not change the dimension of the  
 2158 graph
  - 2159 2. A sum layer that sums the features in the input graph with the one computed from the  
 2160 JWG layers

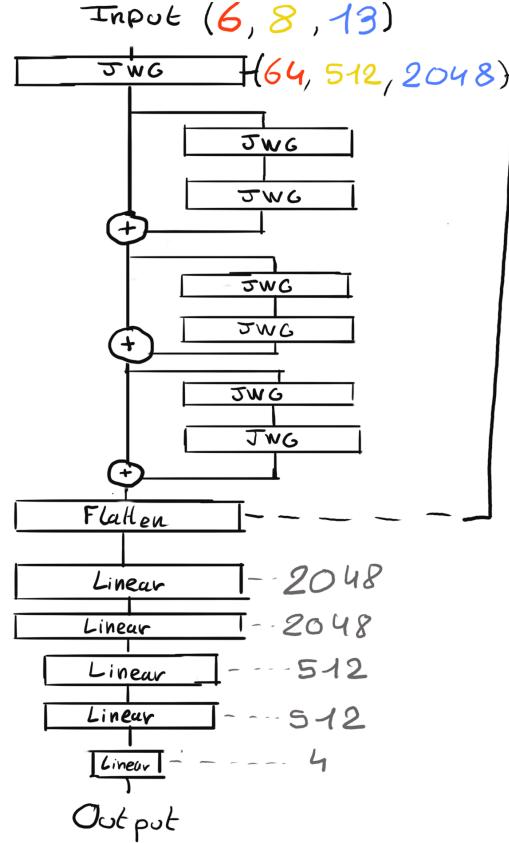


FIGURE 5.6 – Schema of the JWGv8.4.0 architecture, the colored triplet is the graph configuration after each JWG layers

- A flatten layer that flatten the features of the I/O and mesh nodes in a vector.
- 2 fully connected layers of 2048 neurons with a PReLU activation function.
- 2 fully connected layers of 512 neurons with a PReLU activation function.
- A final, fully connected layer of 4 neurons acting as the output of the network.

2165 A schematic of the model is presented in Figure 5.6.

2166 We use the Mean Square Error (MSE) for the loss

$$\mathcal{L} = (E_{rec} - E_{dep})^2 + (X_{rec} - X_{true})^2 + (Y_{rec} - Y_{true})^2 + (Z_{rec} - Z_{true})^2 \quad (5.12)$$

2167 as it was the best resulting loss in Chapter 4.

## 2168 5.5 Training

2169 The optimizer used for training is the Adam optimizer (see Section 3.1.3) and default hyperparameters ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and  $\epsilon = 1e-8$ ) with a learning rate  $\lambda = 1e-8$ . The training last 200 epochs  
2170 of 800 steps. We use a batch size of 32, the largest we can have with 40GB of GPU ram. The learning  
2171 rate is constant during the first 20 epochs then exponentially decrease with a rate of 0.99. We save

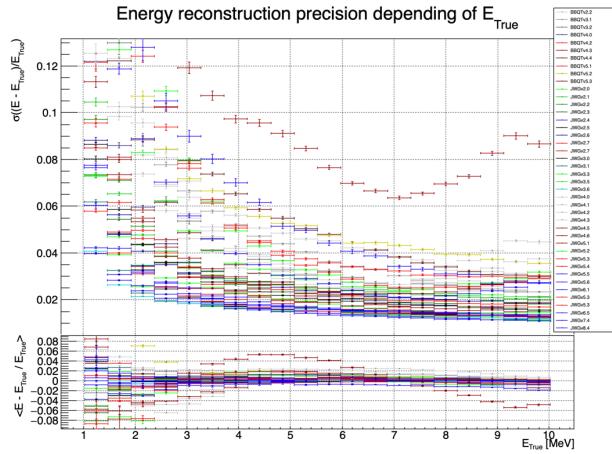


FIGURE 5.7 – Energy reconstruction depending on the true energy for samples of the different versions of the GNN

2173 two set of parameters, the set of parameters the set that yield the lowest validation loss and the set  
 2174 of parameters at the end of the training. The validation is computed over a single batch.

## 2175 5.6 Optimization

2176 The GNN model presented in previous sections is the result of a long work of optimization. Indeed,  
 2177 the innovative architecture we propose left us with an infinity of possible configurations with no  
 2178 guidance from prior works in literature nor in JUNO.

2179 In the end, more than 60 different configurations have been tested. This effort is illustrated on Figure  
 2180 5.7<sup>1</sup>, where the 40 configurations are compared in their ability to reconstruct the positron energy.  
 2181 Although all configurations share the fundamental principles we base our innovative architecture  
 2182 on (three different kinds of nodes and edges, usage of raw level features on some of them, usage of  
 2183 higher level data on others, division of JUNO's surface into regional pixels to form mesh nodes, the  
 2184 very large number of edges connected to each mesh node, etc.), performances can vary a lot between  
 2185 our first attempts (far beyond any acceptable energy resolution, and not even on this figure) and  
 2186 recent ones. Therefore: the precise way to choose hyperparameters mattered a lot, regardless of the  
 2187 relevance of the global architectural principles.

2188 The spectacular improvement between early and later configurations also explains the length of this  
 2189 process : for long we hoped we would finally reach the classical performance, and it was tempting  
 2190 to test yet another configuration.

### 2191 5.6.1 Software optimization

2192 A substantial effort was devoted to the data processing workflow. Transforming JUNO simulation  
 2193 outputs into graphs is a computationally expensive task. Furthermore, due to the ever-changing  
 2194 nature of the graph dimensions and features during optimization, preprocessing JUNO's files by  
 2195 precalculating the graphs and then reading them from files was not viable, as it would require a  
 2196 large amount of disk space to store events for each version of the graph.

1. Note that this figure was prepared on idealized data with no dark noise and perfect hit time determination.

2197 Therefore, the software does not rely on preprocessed data and instead computes the observables,  
 2198 adjacency matrix, etc., during training. This data processing is performed in parallel on the CPU.  
 2199 The raw data comes from ROOT files produced by the collaboration software, and the Event Data  
 2200 Model (EDM), used internally by the collaboration [78], had to be interfaced with our software,  
 2201 an interface that had to be maintained as the collaboration's software evolved. For the harmonic  
 2202 power calculation, we migrated from the Healpix library to Ducc0 [79] for more precise control over  
 2203 multithreading.

### 2204 5.6.2 Hyperparameters optimization

2205 The first kind of hyper-parameters that received a lot of effort concern the network's detailed architecture:

- 2207 — Message passing layers where originally not JWG layers, we started by using small FCDNN in  
 2208 place of  $\phi_u$  and  $\phi_m$ . Due to low performances and memory consumption issues, we pivoted to  
 2209 the message passing algorithm presented in Section 5.2.
- 2210 — The ResNet architecture was brought after issue with the gradient vanishing.
- 2211 — The number of layers was varied between 5 and 12.
- 2212 — The number of node features after each given message passing layer (64, 512, 2048 in the final  
 2213 version) was varied.
- 2214 — The Final FCDNN after the message passing layers is not present in all versions.
- 2215 — At some point, the PReLU activation function replaced the ReLU function.

2217 For some of them, software work was necessary. In any case, each configuration required a training  
 2218 of about 90h. Adding the analysis time necessary to the verification of its performance and the  
 2219 comparison with other versions, one understands the number of tests had to be limited.

2220 Other hyperparameters were also tested :

- 2221 — The higher level variables described in Section 5.1 (powers of various spherical harmonics,  $\mathbb{A}$ ,  
 2222  $\mathbb{A}_s$ ,  $(Q_{m1} - Q_{m2})/(Q_{m1} + Q_{m2})$ ) were added progressively. Notice that our choice to focus  
 2223 our search on this kind of variables is also due to the fact that JWGLayer involves linear  
 2224 operations. It is therefore difficult for such a network to propose variables of this kind among  
 2225 the node features learned layers after layers (i.e. it's difficult for the network to understand  
 2226 these variables are important, or only after many layers).
- 2227 — Time allocated to training, the Learning Rate, the size of batches, etc.
- 2228 — The number of pixels (ie of mesh nodes) was varied between 192 and 768.
- 2229 — Several definitions loss functions where tried. In particular, we tried some focussed only on the  
 2230 E resolution, only on the vertex resolution (R) or trying to optimize both.

2231

2232 To make a long story short, each new configuration was the result of our reflections after having  
 2233 analysed the previous configurations, or after having thought over again about JUNO's detailed  
 2234 response to energy deposits – seeking for variables that could help the GNN.

2235 Another, quite common, approach was in principle possible : a random search. However, due to the  
 2236 extensive training time, up to 90h per training, the heavy memory consumption of the models that  
 2237 would often exceed the 20GB limit of the V100, this approach was not realistic in our case, though we  
 2238 were able to extend the memory limit to 40GB thanks to a local A100 GPU card available at Subatech.

## 2239 5.7 performance of the final version

2240 The reconstruction performance of "JWGv8.4" are presented in Figures 5.8, 5.9, 5.10 and compared  
 2241 to the "Omilrec" algorithm, the official IBD reconstruction algorithm in JUNO. Omilrec is based on  
 2242 the QTMLE reconstruction method that was presented in Section 3.3.

2243 This comparison required to use a consistent definition of  $E_{true}$ . This is not trivial since at JUNO,  
 2244 ML method reconstruct the true energy deposited by the positron+annihilation gammas (that's the  
 2245 target implemented in the loss function), while Omilrec, which is based on probabilities to observe a  
 2246 given number of PE in a given PMT, reconstruct the "visible energy". It reflects the total number of  
 2247 radiated and detectable scintillation or Cherenkov photons (and is subject to non linear effects like  
 2248 quenching).

2249 The conversion we use to obtain comparable  $E_{true}$  is explained in Appendix C.

2250 On Figures 5.8 to 5.10, we notice that the best GNN does not match the performance of the OMILREC  
 2251 algorithm. Generically, Energy resolution is 50% worse, while the resolution on R is three times  
 2252 worse. Reconstruction biases are not better either with the GNN. We have tried to understand the  
 2253 origin of this limited performance.

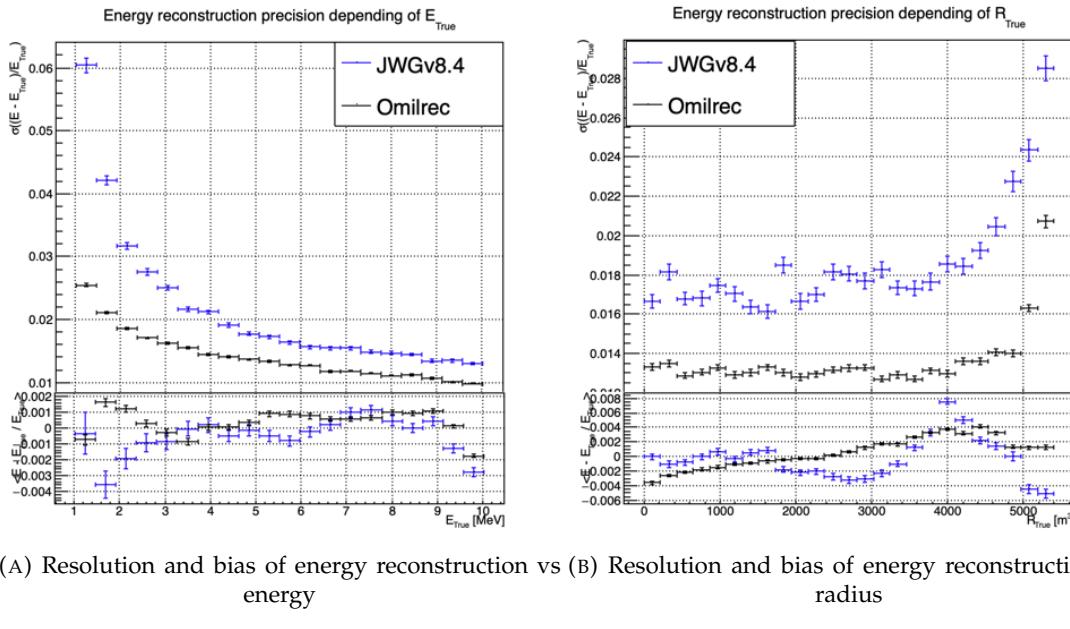


FIGURE 5.8 – Reconstruction performance of the Omilrec algorithm based on QTMLE presented in Section 3.3, JWGv8.4 presented in this chapter. The top part of each plot is the resolution and the bottom part is the bias.

2254 The first action that can be carried out in this direction was to determine if some information used  
 2255 by OMILREC was not used properly by JWGv8.4. For that purpose, we used again the approach  
 2256 presented in Chapter 4 (Sec 4.3.2 and annex A ) to combine JWGv8.4 and OMILREC. We observe on  
 2257 Figures 5.11 and 5.12 that this combination brings no sizeable improvement of the best of the two  
 2258 combined methods. The combination remains very close to OMILREC alone. This is an indication  
 2259 that JWGv8.4 does not use informations that would be overlooked by OMILREC, and that on the  
 2260 contrary, that's JWGv8.4 that fails to use properly important informations.

2261 The problem described above could be inherent to our GNN's original architecture. Discussions with  
 2262 JUNO's colleagues when these results were presented at the collaboration pointed to the role of PMT  
 2263 time information ( $t$ , in the  $(Q, t)$  pairs we use as our algorithm input features). The thousands of

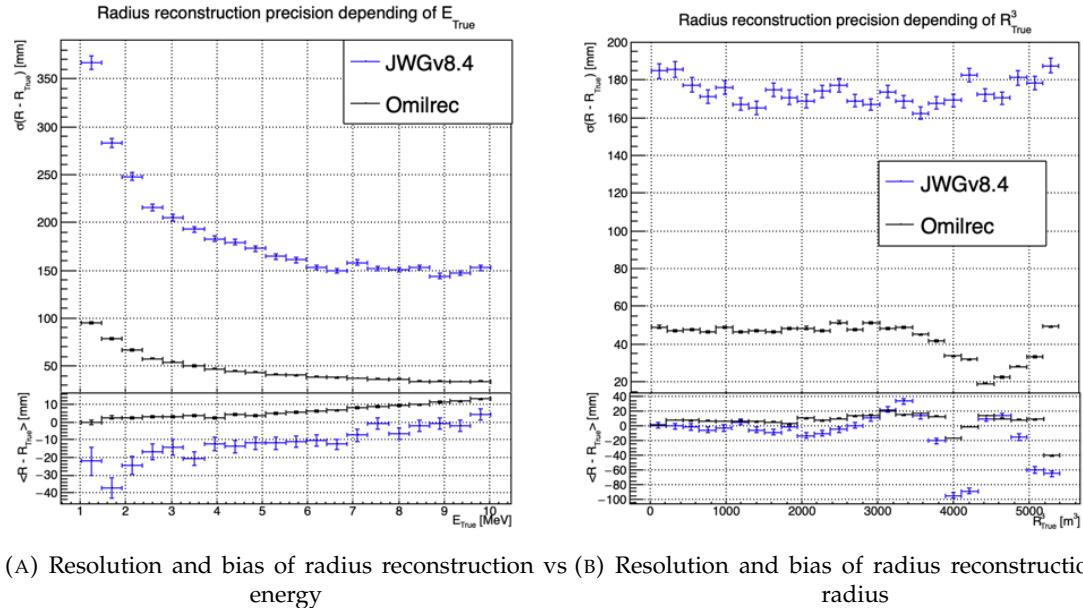


FIGURE 5.9 – Reconstruction performance of the Omilrec algorithm based on QTML presented in Section 3.3, JWGv8.4 presented in this chapter. The top part of each plot is the resolution and the bottom part is the bias.

values found in the *fired* nodes might not be aggregated well enough when transmitted to the mesh nodes, causing a loss in the redundancy of this important information.

We tested this idea in several manners, described below.

### Finer granularity

We tried to recover some redundancy by increasing the number of mesh nodes from 198 to 768. The improvement we observed was small, and did not allow to get close to OMILREC's performance.

To explore further in this direction, we would ideally try 3072 pixels (the next HEALPIX rank). However, this is not possible for our GNN due to hardware limitations, mainly the available GPU memory. Instead, we discussed the problem with Gilles Grasseau, calculus research engineer with whom we collaborate on the subject of ML reliability (see Chapter 6). In the framework of this activity, Gilles needs to develop reconstruction algorithms to be "attacked" by a prototype Adversarial NN. One of them is a pseudo-spherical CNN using oriented filters, called HCNN.

To produce its input image, this algorithms split the Sphere into 3072 pixels. Each channel of this image is an aggregation of the  $(Q, t)$  values found in all the PMTs. The charge are summed and the lowest time is kept. The performance of this algorithm can be seen on Figures 5.13 and 5.14, compared to OMILREC. With 3072 pixels, the performance of HCNN does not match that of OMILREC, but is closer to it than our GNN. The granularity of the pixels, and the way to summarize the individual PMTs information when going from 17000 LPMTs to only 3072 pixels indeed seems to play a role.

This is consistent with the results obtained by the first GNN tried at JUNO on reactor neutrinos (already described in Section 3.3.3). It used 3072 pixels, and also obtained an uncompetitive R reconstruction.

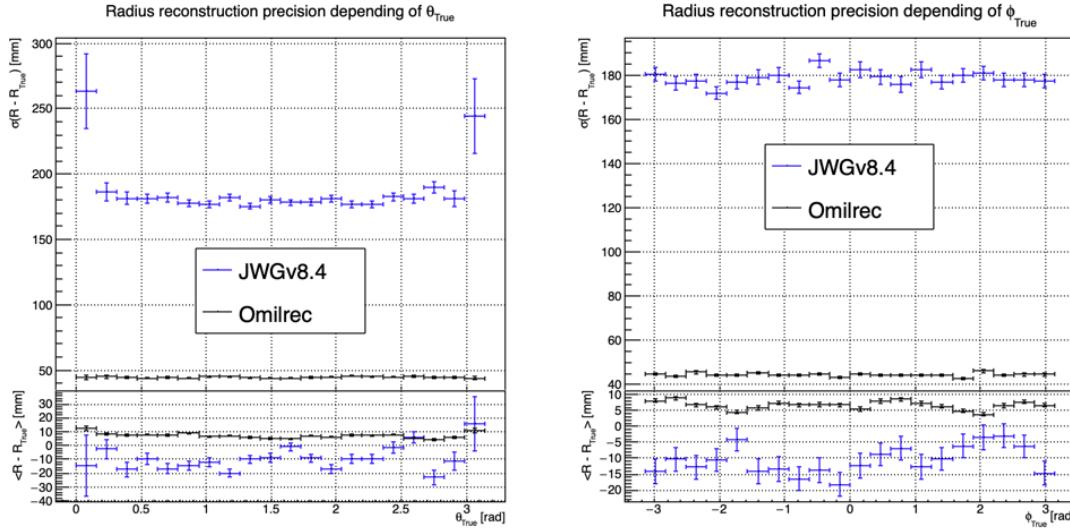
(A) Resolution and bias of radius reconstruction vs  $\theta$  (B) Resolution and bias of radius reconstruction vs  $\phi$ 

FIGURE 5.10 – Reconstruction performance of the Omilrec algorithm based on QTMLE presented in Section 3.3, JWGv8.4 presented in this chapter. The top part of each plot is the resolution and the bottom part is the bias.

## 2286 Information reduction, from fired to Meshes

2287 The problem described above is somehow classical. ML algorithms, ideally, would start from the full  
2288 information present in the detector, and learn to reduce it optimally.

2289 In cases where only 3072 pixels can be used instead of the complete information from 17000 PMTs,  
2290 one needs to understand how to combine the individual from the 5 or 6 PMT found in each pixel into  
2291 pixel-level features, without loosing important information.

2292 In the case of our GNN, we hoped that by connecting each mesh node to its corresponding 5 or 6  
2293 fired nodes, we could keep the full information. In reality, it seems that the message passing between  
2294 fired and mesh does not work efficiently. When nodes are updated by the first (may be also by the  
2295 subsequent) layer, the new mesh features might be dominated by the original features in the second  
2296 column of tables 5.1, themselves a simple version of aggregation. Layer after layer, we might be  
2297 limited to that level of time information, lacking time redundancy.

2298 We have verified this by testing version of the GNN in which the link between fired and mesh was  
2299 cut, or in which no time info was included among the fired nodes features. It had only a small effect  
2300 which seems to confirm a problem in the way the full information, from all the individual PMTs, is  
2301 used by our GNN.

## 2302 Possible improvements

2303 It appears that the network is unable to aggregate the timing information correctly. While this could  
2304 be addressed by using a finer segmentation, with more mesh nodes, improvements might also arise  
2305 from refining the message-passing algorithm. The algorithm presented in this thesis is still quite  
2306 basic, relying on a simple linear combination of features. We have seen through examples in CNNs,  
2307 GNNs, and other architectures, both in research and industry, that specializing the network — for  
2308 instance, by incorporating convolutional filters — can lead to improvements that were previously

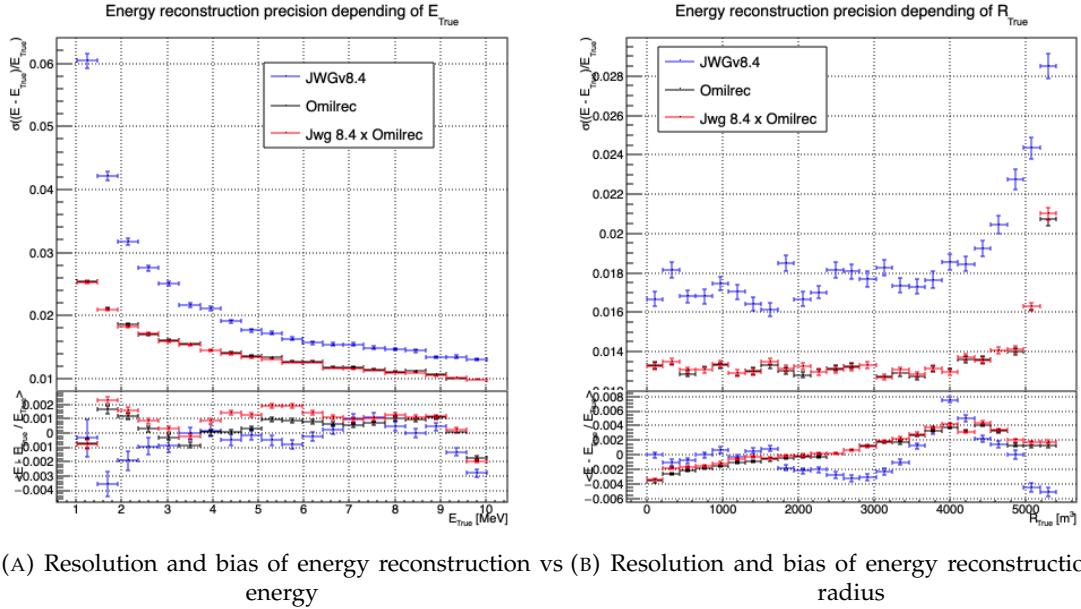


FIGURE 5.11 – Reconstruction performance of the Omilrec algorithm, JWGV8.4 and the combination between the two using the optimal variance estimator presented in annex A.2. The top part of each plot is the resolution and the bottom part is the bias.

unattainable with simpler FCDNNs. Applying this approach to the message-passing algorithm, by utilizing a GNN with a more advanced message-passing, could yield better results.

We could investigate alternative aggregation strategies, for example, by weighting the timing information more significantly during the message-passing phase. Additionally, testing a non-linear combination of features from fired to mesh nodes could help preserve more granular information. Another potential improvement would be to introduce attention mechanisms that dynamically assign more importance to relevant features in the fired nodes

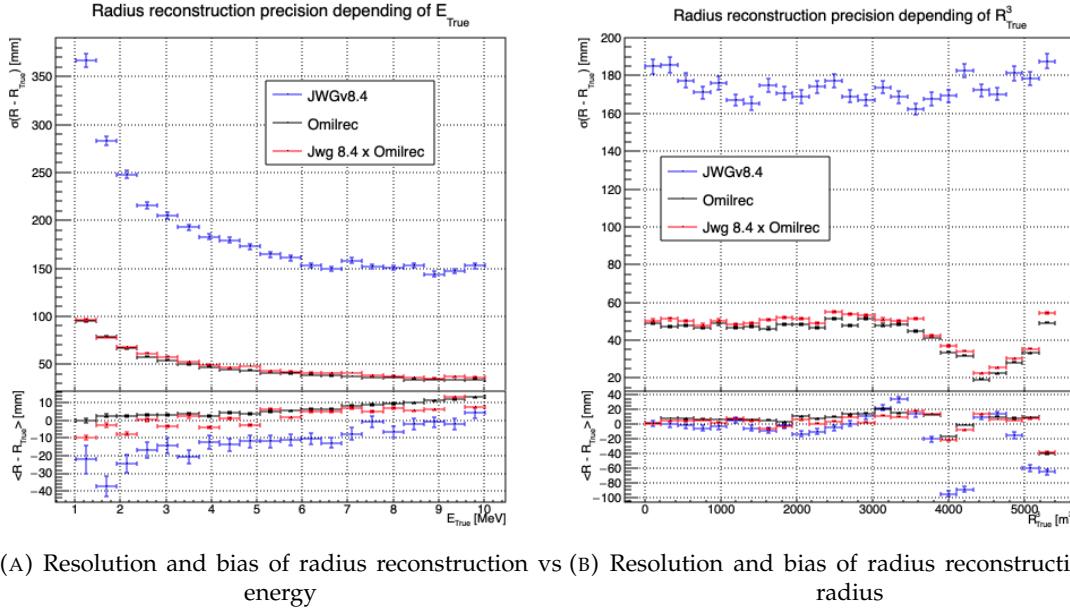
Regarding the timing information, we provided high-level features, assuming this would assist the neural network in converging to the solution. However, by offering such information upfront, the GNN might be taking the “easy” path, settling for a local and broader minimum, rather than extracting the features that could lead to better performance.

If there are difficulties in transferring information between the fired and mesh nodes, it may stem from the way we connected the fired nodes to the mesh nodes. By linking the fired nodes within the same mesh, or even connecting the fired nodes of neighboring mesh nodes, the GNN might be able to construct more meaningful information.

Finally, by providing directly the PMT waveform to the GNN, in the fired nodes, we could search for even finer precision and results. An idea would be to specialise the message function  $\phi_{m;F \rightarrow M}$  to be a 1D convolutional layer over the waveform. The resulting channels would be fed to the mesh nodes for their updates.

## 5.8 Conclusion

To achieve its scientific goals, JUNO requires a precise and well-understood reconstruction, as it needs an energy resolution of 3% at 1 MeV. Even small, unaccounted biases could make it impossible to determine the mass ordering, as explored in Chapter 7. A likelihood-based algorithm, designed to



(A) Resolution and bias of radius reconstruction vs energy (B) Resolution and bias of radius reconstruction vs radius

FIGURE 5.12 – Reconstruction performance of the Omilrec algorithm, JWGV8.4 and the combination between the two using the optimal variance estimator presented in annex A.2. The top part of each plot is the resolution and the bottom part is the bias.

meet JUNO's requirements and referred to as the classical algorithm, was developed and is detailed in Section 3.3.

Machine learning algorithms were developed to challenge this classical approach, and they are presented in Section 3.3.3. Although they achieve the precision of the classical algorithm, they do not offer significant improvements. The GNN previously developed is a convolutional GNN where nodes correspond to pixels, connected to their neighbors based on the Healpix [77] segmentation, with the  $(Q, t)$  information aggregated onto these pixels.

In this chapter, we introduce a novel and innovative architecture. In addition to the pixel segmentation represented by mesh nodes, we incorporate rawer information by directly representing the fired PMTs as nodes. We also fully connect the mesh nodes to each other, hoping to facilitate the transfer of information. Finally, we introduce a global node that holds global information about the detector.

These three types, or families, of nodes do not have the same number of features, resulting in a heterogeneous graph. Publicly available algorithms for graph processing are designed for homogeneous graphs, so we had to develop a custom algorithm adapted to heterogeneous graphs.

This GNN required significant technical development, but the results are not at the level of the classical algorithm. The tests we conducted suggest that the problem may lie in the aggregation of raw information from the fired nodes onto the mesh nodes, as removing the fired nodes does not degrade the results. Additionally, due to technical constraints, we had to reduce the number of pixels compared to the previous GNN. Other algorithms we developed, which use a higher pixel resolution, outperform this architecture, reinforcing our suspicion that the aggregation is the root of the issue.

The precision required for JUNO's scientific objectives, particularly in determining mass ordering, imposes stringent constraints on reconstruction algorithms. Small biases or errors in energy resolution could significantly affect the experiment's outcomes. Future improvements may involve refining the message-passing algorithm, incorporating additional detector-specific features, and experimenting with more advanced architectures such as attention-based GNNs to further reduce reconstruction

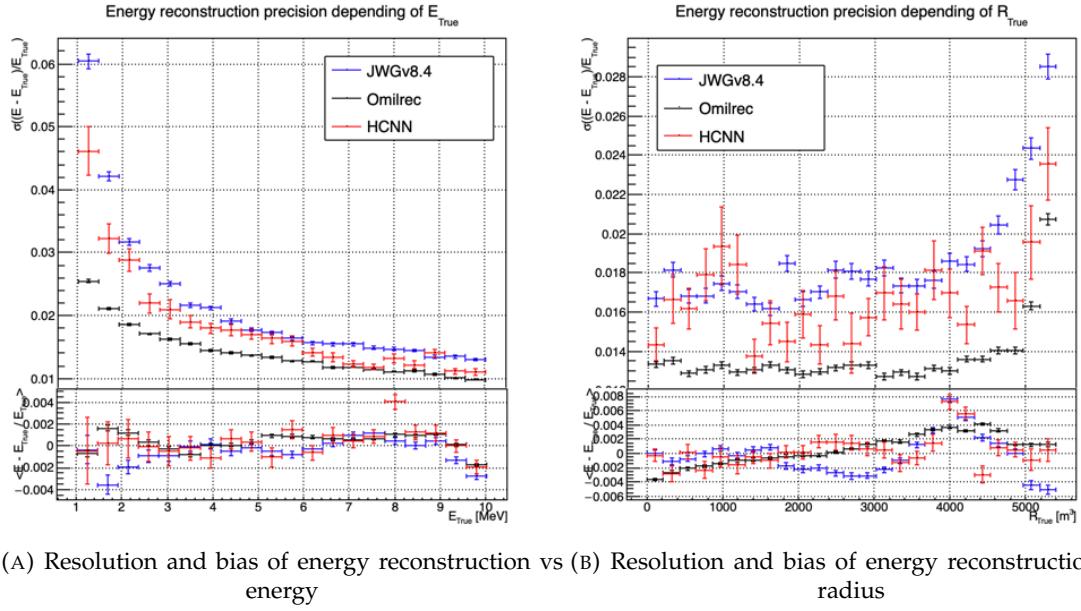


FIGURE 5.13 – Reconstruction performance of the Omilrec algorithm based on QTMLE presented in Section 3.3, JWGv8.4 presented in this chapter and the HCNN algorithm. The top part of each plot is the resolution and the bottom part is the bias.

2358 errors.

2359 Perhaps by incorporating rawer information, such as the waveform, refining the message-passing  
2360 algorithm, or adjusting the features on the different nodes, we could match the precision of the  
2361 classical algorithm. However, it is also possible that deeper, more radical changes are needed to  
2362 become competitive.

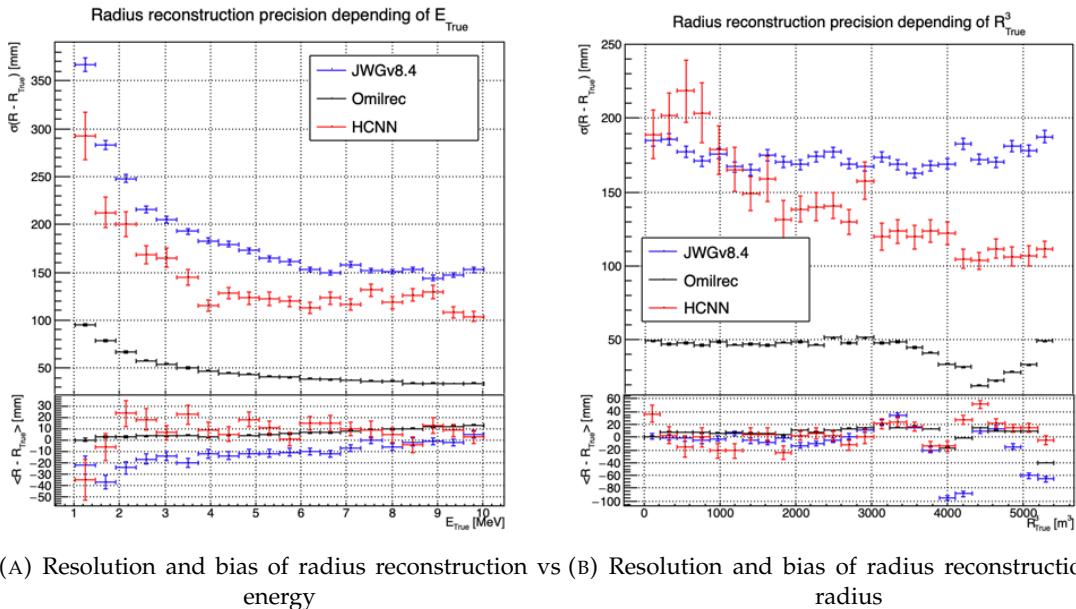


FIGURE 5.14 – Reconstruction performance of the Omilrec algorithm based on QTMLE presented in Section 3.3, JWGV8.4 presented in this chapter and the HCNN algorithm. The top part of each plot is the resolution and the bottom part is the bias.

<sup>2363</sup> **Chapter 6**

<sup>2364</sup> **Reliability of machine learning  
methods**

<sup>2365</sup>

<sup>2366</sup> *"Psychohistory was the quintessence of sociology; it was the science of  
human behavior reduced to mathematical equations. The individual  
human being is unpredictable, but the reactions of human mobs,  
Seldon found, could be treated statistically"*

*Isaac Asimov, Second Foundation*

<sup>2367</sup> **Contents**

---

<sup>2368</sup> <b>6.1</b>	<b>BDT for energy reconstruction (BDTE)</b>	<b>100</b>
<sup>2369</sup> <b>6.2</b>	<b>Adversarial method</b>	<b>102</b>
<sup>2370</sup> <b>6.3</b>	<b>ANN Architecture</b>	<b>104</b>
<sup>2371</sup> 6.3.1	Back-propagation problematic	105
<sup>2372</sup> 6.3.2	Reconstruction Network (FFNN)	106
<sup>2373</sup> 6.3.3	Adversarial Neural Network (ANN)	107
<sup>2374</sup> <b>6.4</b>	<b>Training of the ANN</b>	<b>110</b>
<sup>2375</sup> 6.4.1	First training phase: back to physics	110
<sup>2376</sup> 6.4.2	Second training phase: Breaking of the reconstruction	111
<sup>2377</sup> <b>6.5</b>	<b>Conclusion and prospect</b>	<b>114</b>

---

<sup>2378</sup> As explained in previous chapters, JUNO is a precision experiment where the complete understanding of the effects at hand is crucial. As it will be illustrated in Chapter 7, even small invisible biases or uncertainties could lead to the impossibility to run the measurements, or even worse, wrong NMO measurements. While the liquid scintillator technology is well known and straightforward, this is the first time it is deployed to such scale, and for such precision. This novelty bring its fair share of elements, effects or assumption, that, if they were to be overlooked, could cause issue.

<sup>2388</sup> We already shown a large variety of reconstruction algorithms, OMILREC for LPMT reconstruction in Section 3.3, numerous machine learning algorithms in Section 3.3.3 and our own work in Chapters 4 and 5. Those algorithms were compared to each others based on their performance as in [63] but we are the first that looked into the correlation between the reconstruction. The combinations of algorithms shown in Section 4.3.2 show that some information elude the algorithms. To efficiently compare algorithms between each other, they need to be publicly available to the collaboration to studies their differences event by event.

<sup>2395</sup> To achieve this goal, I implemented a BDT for energy reconstruction, named BDTE which was developed by another research team, in the JUNO official software. The details of this implementation and its combination with OMILREC are presented in Section 6.1.

Another way to ensure reliability is to challenge reconstruction algorithms with physically plausible perturbations in the PMT charge and time information. The search for such effects could be done by hand, but the process would be tedious. We propose in this thesis a machine learning method to probe for those effects. In Section 6.2, I describe the method behind the algorithm. In Section 6.3 I detail the architecture of our algorithm. The training and the results of our method are presented in Section 6.4. Finally, in section 6.5, I conclude and discuss the prospects and possible improvements to bring to this work.

## 6.1 BDT for energy reconstruction (BDTE)

To study the reliability of reconstruction algorithms it's necessary to be able to compare their reconstruction performance event by event. To ease the process, it is important that they are publicly available. JUNO's common software, discussed in Section 2.6, is based on the SNiPER framework [34] which allows the packaging of the different steps of JUNO's analysis, from Monte Carlo (MC) data generation to event reconstruction, including the simulation of the PMTs' waveform reconstruction, electronic effects and the trigger system.

This framework is modular, with each module being a C++ class bound in Python. The execution of successive algorithms is orchestrated via Python scripts.

We could have implemented the algorithms presented in Chapters 4 and 5, but since these are themselves not trivial, we chose to start with a simpler ML algorithm that presents similar energy reconstruction performances as OMILREC: a Boosted Decision Tree (BDT) for energy reconstruction developed by Gavrikov Arsenii et al. [64]. This BDT, named BDTE, is based on an aggregated feature approach where, from the set of  $(Q, t)$  in LPMTs, a set of higher-order variables is computed and then fed to the BDT. The list of the aggregated features used by the BDT is presented in Table 6.1. These higher-order variables are extracted from the charge  $Q$  and hit time  $t$  distribution. It also depends on two straightforward interaction vertex estimators.

The first one is the charge barycenter

$$\vec{r}_{cc} = \frac{\sum_i \vec{r}_{PMT,i} Q_i}{\sum_i Q_i} \quad (6.1)$$

where  $i$  index the fired PMT,  $\vec{r}_{PMT,i}$  is the position vector of the  $i$ th PMT and  $Q_i$  is the charge it collected.

The second estimator is the hit time barycenter

$$\vec{r}_{ht} = \frac{1}{\sum_i \frac{1}{t_i + c}} \sum_i \frac{\vec{r}_{PMT,i}}{t_i + c} \quad (6.2)$$

where  $t_i$  is the time of collection of the  $i$ th PMT and  $c = 50$  ns a constant to prevent divergence when  $t_i$  is 0.

The performance of this BDT, as published by Gavrikov Arsenii et. al, is reported in Figure 6.2a. This BDT was originally developed in Python and consisted of a collection of Python scripts for the training and the evaluation.

As stated before, JUNO software is composed of C++ modules orchestrated through Python scripts. The technical challenge was to extract the data from the internal representation of the event in JUNO software, the Event Data Model (EDM), into a comprehensible format for Python. This task, which was previously done via data pre-processing by Python scripts, had to be internalized within the software. The computation of the aggregated features was migrated from the Python scripts into

Feature	Description
AccumCharge	Sum of the charge collected by every LPMT
$R_{ht}$	Radius reconstructed by the hit time barycenter
$z_{cc}$	$z$ component of the vertex reconstructed by the charge barycenter
$\sigma_{PE}$	Standard deviation of the distribution of collected PE per PMTs
$N_{PMT}$	Number of fired PMTs
htKurtosis	Kurtosis of the hit time distribution
ht <sub>25%–20%</sub>	Difference between the 25% and 20% percentiles of the hit time distribution
$R_{cc}$	Radius reconstructed by the center of charge barycenter
ht <sub>5%–2%</sub>	Difference between the 5% and 2% percentiles of the hit time distribution
$\langle PE \rangle$	Mean number of PE collected per PMTs
$J_{ht}$	Jacobian of the hit time distribution
$\phi_{cc}$	$\phi$ component in spherical coordinate of the charge barycenter
ht <sub>35%–30%</sub>	Difference between the 25% and 20% percentiles of the hit time distribution
ht <sub>20%–15%</sub>	Difference between the 20% and 15% percentiles of the hit time distribution
PE <sub>35%</sub>	Value of the 35% percentile of the charge distribution
ht <sub>30%–25%</sub>	Difference between the 30% and 25% percentiles of the hit time distribution

TABLE 6.1 – Summary of the aggregated features used by the BDT to reconstruct the IBD energy. The charge barycenter and hit time barycenter vertex estimators are detailed in Eq. 6.1 and 6.2 respectively

2436 C++ modules. The final step was to fetch the reconstruction results of the algorithm into the C++  
 2437 framework to save the results in the EDM.

2438 We validated that the aggregated features were consistent between the original version and the  
 2439 implementation in JUNO software. With the help of Arsenii, we were able to compare over 1000  
 2440 events, and for the majority of the features, the relative difference between his and ours was either 0  
 2441 or on the order of  $10^{-15}$ , with the exception of three features:  $R_{cc}$ ,  $R_{ht}$ , and  $z_{cc}$ . For these three fea-  
 2442 tures, the relative difference is about  $10^{-6}$ , which, while small, is still surprisingly high for numerical  
 2443 computation. The distributions of the relative differences for these features are presented in Figure  
 2444 6.1.

2445 We investigated the source of those discrepancies. The difference in computation environments,  
 2446 Python using the Numpy [80] and C++ using the standard library in our cases, is most probably  
 2447 the source. As they are coming from the computation of the barycenter in Eq. 6.1 and 6.2, it could  
 2448 come from differences in the compiling optimization in the weighted sum. We consider that those  
 2449 difference are still small so that the performance of the BDT are unaffected.

2450 We investigated the source of these discrepancies. The difference in computation environments—Python  
 2451 using Numpy [80] and C++ using the standard library in our case—is most likely the cause. Since  
 2452 the discrepancies arise from the computation of the barycenter in Eq. 6.1 and 6.2, they may result  
 2453 from differences in compiler optimization during the weighted sum calculation. We consider that  
 2454 these differences are still small enough that the performance of the BDT is unaffected.

2455 The performance of our implementation of BDTE compared to the results presented in [64] are  
 2456 presented in figure 6.2b. The performance are compatibles.

2457 The reconstruction using BDTE was implemented in JUNO’s common software but Gavrikov et  
 2458 al. also detail the training and hyper-optimization. JUNO Monte Carlo is likely to evolve during  
 2459 the construction phase and will be further adjusted using calibration. The implementation of those  
 2460 procedures, the training and optimization, will be required as BDTE re-training and re-optimisation  
 2461 will be required with each JUNO software update.

2462 Figure 6.2b shows that the resolution of BDTE is very close to OMILREC. We measured the correla-

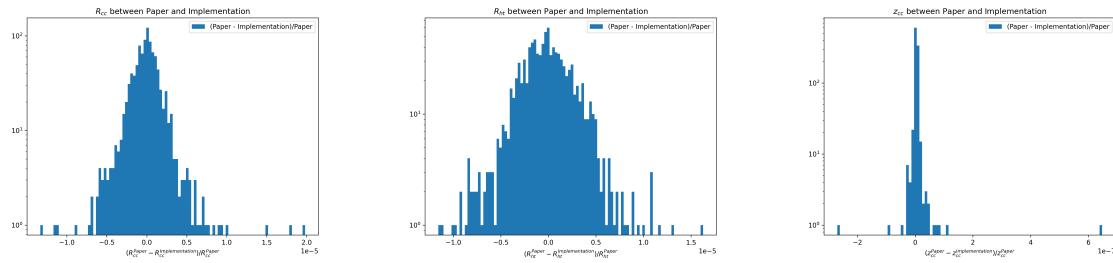


FIGURE 6.1 – Relative difference between the features computed by Gavrikov et. al (superscripted Paper) and our implementation (superscripted Implementation)

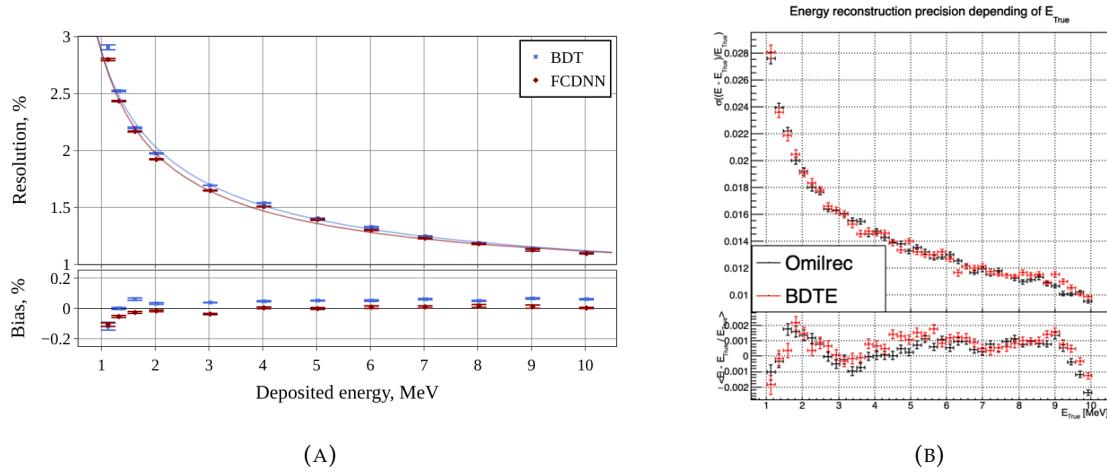


FIGURE 6.2 – Resolution of BDTE On the left: as reported by Gavrikov Arsenii et. al in [64], On the right: once implemented in JUNO common software. On the right plot is also reported the reconstruction performance of the OMILREC algorithm. The OMILREC algorithm  $E_{\text{vis}}$  has been corrected to  $E_{\text{dep}}$  following the procedure detailed in Annex C.

2463 tion between their errors:

$$\text{Corr}(E_{BDTE} - E_{\text{dep}}, E_{OMILREC} - E_{\text{dep}}) \quad (6.3)$$

2464 If the correlation is small enough, it hints at possible improvements in the IBD reconstruction. The  
2465 correlation between errors for different energy and event radius in the detector is presented in Figure  
2466 6.3. We see that for the vast majority of the  $(R^3, E)$  phase space, the correlation is  $> 0.995$ , down to  $\sim$   
2467 0.98 in the  $R \approx 9$  m and  $R > 17$  m regions. Such high correlations indicate that there is close to no  
2468 improvement that can be found in these algorithms.

## 2469 6.2 Adversarial method

2470 As introduced at the beginning of the chapter, JUNO needs a very good understanding of the biases  
2471 and effects affecting its reconstruction, as a small bias could distort the mass ordering measurement.  
2472 To calibrate those biases and effects, JUNO relies on multiple sources that can be located at various  
2473 points in the detector. The calibration strategy is already discussed in Section 2.4 and shows calibra-  
2474 tion sources of gammas, neutrons, and positrons (Table 2.3), with the catch that the positrons will  
2475 annihilate inside the encapsulation and only the two 511 keV gammas will deposit energy in the LS.

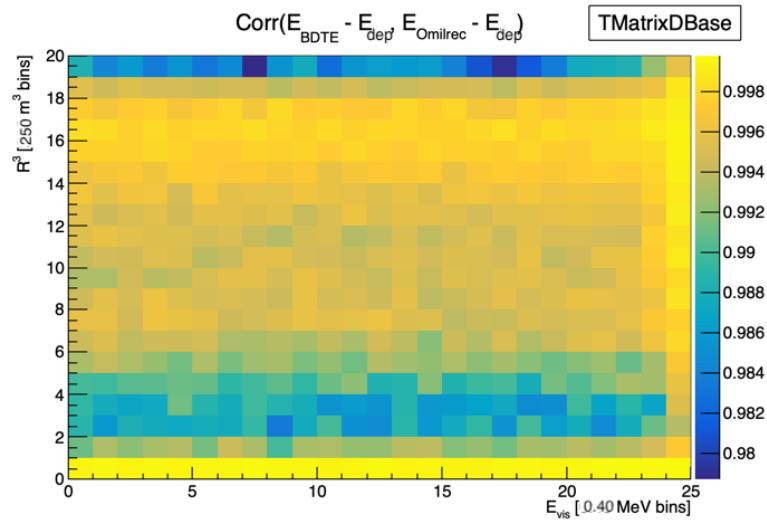


FIGURE 6.3 – Correlation between the errors in energy reconstruction between BDTE and OMILREC (Eq. 6.3). The correlation is computed in  $R^3$  bins of  $216 \text{ m}^3$  between 0 and  $5000 \text{ m}^3$ , 0 and 17 m in y axis, and in 0.40 MeV bins between 1.022 and 10.022 MeV of deposited energy.

None of the calibration sources considered are positron events. While electrons and positron events should be pretty similar in their interaction with the electronic cloud of the LS atoms, electron events are missing the two annihilation  $\gamma$  photons and the potential of forming a positronium [81]. The topology of the event is localized in a region of the order of magnitude of our reconstruction performance. A few nanoseconds between the energy deposition and the positronium annihilation against a time transit spread between 3 and 6 ns, depending on the PMT type [82–84]. The  $\gamma$  from the positron annihilation will travel distances of the order of magnitude of the typical LPMT resolution of 8 cm (see Section 3.3).

Another natural calibration source is the  $^{12}B$  spectrum. The  $^{12}B$  is a cosmogenically produced isotope through the passage of muons inside the LS. The  $^{12}B$  decays via  $\beta^-$  emissions with a Q value of 13.5 MeV, with more than 98% of the decay resulting in ground state  $^{12}C$ . The  $^{12}B$  events will be cleanly identified by looking for delayed high-energy  $\beta$  events after an energetic muon. Due to its natural causes, the  $^{12}B$  events will be uniformly distributed in the detector. The calibration strategy consists of fitting the energy spectrum of  $^{12}B$  with the results of the simulation to adjust the simulation parameters. Both sources will be used to *control* the response of the detector.

Unlike lasers and radioactive sources, from which the localization and energy will be well known, the individual events of  $^{12}B$  will be unknown, with only the localization loosely constrained by the muon track. Only higher-order observables such as the energy distribution will be accessible.

All of those considerations could hide potential unknown or undetected effects that could lead to issues in the mass ordering analysis. But, while we have an idea of where the issues could come from, the manual production of event perturbations that go unseen in the calibration would be tedious. That's why we propose to use an Adversarial Neural Network (ANN) to produce those perturbations if they exist. A schematic of the concept is presented in Figure 6.4.

This network should produce physically plausible perturbations that would not be seen by the calibration system but also by the visualization of the event. If the ANN manages to produce such perturbations, we can derive systematic uncertainties from it. If it fails to find any, it is a proof of robustness for the targeted reconstruction method.

For this study, we consider a “physics” dataset composed of 1M positron events from J23, uniformly

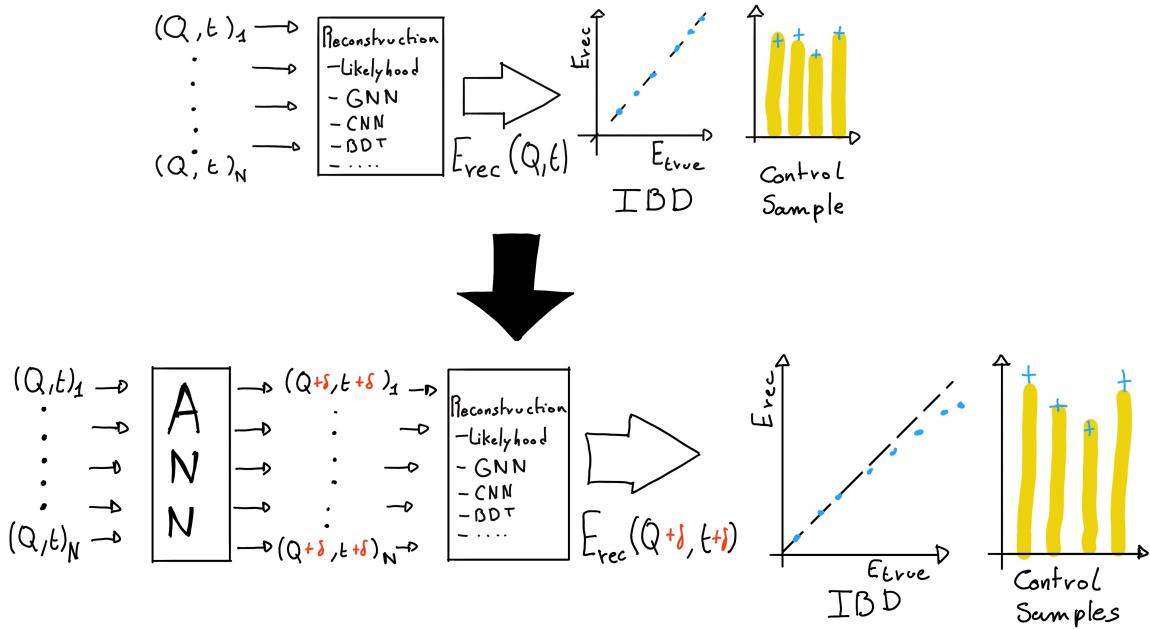


FIGURE 6.4 – Schema of the method to discover vulnerabilities in the reconstruction methods. **On the top** of the image, the standard data flow. The individual charge and times are fed to a reconstruction algorithm. From the reconstructed energies, we can produce an IBD spectrum and compute control observables from the control samples. **On the bottom**, the same data flow but we add an ANN between the input and the reconstruction. The ANN will slightly change the input charge and time so the reconstruction algorithm inaccurately reconstruct the IBD energy, but the perturbation is not visible in the control samples.

distributed in the Central Detector (CD) and in deposited energy between  $E_{dep} \in [1.022; 10.022]$ . This set represents the IBD events we want to the reconstruction to be fooled on.

We use a second “control” dataset of 1M electron events from J23, also uniformly distributed in the detector and over the same energy range. They mimic the energy deposition of  $^{12}B$  decay and are used as the sample to compute the control observables.

### 6.3 ANN Architecture

We can describe the goal of the ANN by using following loss function:

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{reg} \quad (6.4)$$

where  $\mathcal{L}_{adv}$  is the adversarial loss, which is minimal when the reconstruction is “broken”. We thus need to define what is a *wrong* reconstruction. We choose to define it through the correlation between the reconstructed and deposited energy

$$\mathcal{L}_{adv} = |\text{Corr}(E'_{rec}, E_{rec})| \quad (6.5)$$

where  $E'_{rec}$  and  $E_{rec}$  are the reconstructed energies after and before perturbation respectively. This loss is positive or null and is minimal when the reconstructed energy after perturbation is decorrelated with the original reconstruction.

The term  $\mathcal{L}_{reg}$  is the regularisation term, which is minimal when the control variables are correctly

2518 reconstructed

$$\mathcal{L}_{reg} = \sum_{\lambda} (O_{\lambda}^{rec} - O_{\lambda}^{th})^2 \quad (6.6)$$

2519 where  $\lambda$  index the different control observables that will be considered in this study. It's minimal  
 2520 when the control observables after perturbation  $O_{\lambda}^{rec}$  are coherent with their expected values  $O_{\lambda}^{th}$ . In  
 2521 this exploratory work, we choose as the control observable the difference between the reconstructed  
 2522 position and energy and the ground truth from the Monte Carlo simulation

$$\mathcal{L}_{reg} = \sum_{\lambda \in \{x,y,z,E\}} (\lambda_{rec} - \lambda_{true})^2 \quad (6.7)$$

2523 To these two loss, we adjoin a penalty term  $P$

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{reg} + P \quad (6.8)$$

2524 This penalty  $P$  is here to prevent the ANN from producing event too different from the initial event.  
 2525 It will be further detailed in Section 6.3.3.

2526 We see that the final loss is an equilibrium between the adversarial and regularisation loss.

### 2527 6.3.1 Back-propagation problematic

We would like this method to be applicable to any kind of reconstruction algorithm but this complicated considering standard training method through backward-propagation, discussed in details in Section 3.1.3. For explanation, let's define the application of the reconstruction algorithm as  $\mathcal{F}$  on an event  $X$ , resulting in the prediction  $Y$ , and the application of the ANN  $\mathcal{G}$  on  $X$  to give a perturbed event  $X'$ . We can parametrize the equation 6.4

$$Y = \mathcal{F}(X); Y' = \mathcal{F}(X') = \mathcal{F}(\mathcal{G}(X)) \quad (6.9)$$

$$\mathcal{L} \equiv \mathcal{L}(\mathcal{F}(\mathcal{G}(X)), Y_t) \quad (6.10)$$

2529 where  $Y_t$  is the reconstruction target of  $Y$ .

2530 Now if we consider the parameters  $\theta$  of the ANN on which we want to optimize  $\mathcal{L}$ , in the backward-  
 2531 propagation optimisation framework we need to compute

$$\frac{\partial \mathcal{L}(\mathcal{F}(\mathcal{G}(X)))}{\partial \theta} \quad (6.11)$$

2532 which, when using the chain rule, become

$$\frac{\partial \mathcal{L}(\mathcal{F}(\mathcal{G}(X)))}{\partial \theta} = \frac{\partial \mathcal{G}}{\partial \theta} \cdot \frac{\partial \mathcal{F}}{\partial \mathcal{G}} \cdot \frac{\partial \mathcal{L}}{\partial \mathcal{F}} \quad (6.12)$$

2533 The terms  $\frac{\partial \mathcal{G}}{\partial \theta}$  and  $\frac{\partial \mathcal{L}}{\partial \mathcal{F}}$  are easily computable but  $\frac{\partial \mathcal{F}}{\partial \mathcal{G}}$  depends on the nature of the reconstruction  
 2534 algorithm.

2535 While it comes naturally when using neural network algorithms, it's not so trivial for other types  
 2536 of algorithms like likelihood. Solutions exist to optimize networks that work in complex, non-  
 2537 differentiable environments, such as *Deep Reinforcement Learning* [85, 86], but as a first prototype,  
 2538 we will restrict ourselves to neural networks for the reconstruction algorithm.

2539 The choice to use gradient descent, and therefore neural networks, also allowed us to keep all  
 2540 technical software development wrapped in the same language and framework, PyTorch [49].

The backward-propagation introduce a second issue. At the beginning of the subsection we introduce  $X' = \mathcal{G}(\bar{X})$ , the event after perturbation. It's an input of the reconstruction  $\mathcal{F}$ , thus, let's say that the event, in its form  $X$ , is a list of tuples  $(id, Q, t)$  which are the hit on the PMT  $id$ . If  $\mathcal{F}$  require the information to be formatted in a specific way (graph, images, ...) via an algorithm  $\tau(X)$ , it means that

$$\frac{\partial \mathcal{L}(\mathcal{F}(\tau(\mathcal{G}(X))))}{\partial \theta} = \frac{\partial \mathcal{G}}{\partial \theta} \cdot \frac{\partial \tau}{\partial \mathcal{G}} \cdot \frac{\partial \mathcal{F}}{\partial \tau} \cdot \frac{\partial \mathcal{L}}{\partial \mathcal{F}} \quad (6.13)$$

which also requires that  $\frac{\partial \tau}{\partial \mathcal{G}}$  is differentiable.

On the other hand, if  $X$  is already formatted as the input of  $\mathcal{F}$ , it means that  $\mathcal{G}$  takes the same format as input, and we drop the requirement on  $\tau$  to be differentiable. Specifically, if  $\mathcal{F}$  takes an image as input, it means that  $\mathcal{G}$  will also take an image as input and output an image. Unfortunately, this also means that if some information is lost before  $\mathcal{G}$ , for example, during the charge and time aggregation in pixels, the ANN cannot retrieve and modify it.

A more elegant solution would that  $\mathcal{G}$  would also compute the transformation  $\tau$  in addition to finding relevant perturbation, but for the simplicity of this exploratory work, we use a  $\mathcal{G}$  that process transformed data.

### 6.3.2 Reconstruction Network (FFNN)

As introduced just before, we need a NN algorithm for IBD reconstruction. We could have used the GNN presented in Chapter 5 but we preferred a more simplistic approach to not be constrained by the memory consumption of the reconstruction network. This network is designated as FFNN.

This network takes as input a vector containing the results of the aggregation of charge and time on pixels, forming a vectorized image. We consider JUNO to be composed of 3072 pixels defined by the Healpix [77] pixelization. On each of these pixels, we sum the charges and keep the first time of hit, resulting in 3072  $(Q, t)$  tuples. To these tuples, we adjoin the position of the center of these pixels, resulting in 3072  $(Q, t, x, y, z)$  tuples. The data is finally represented as a  $3072 \times 5 = 15360$  vector. In the case where the charge in a pixel is 0, the time is set to 2048 ns, which is way after the closure of the trigger window.

The charge is expressed in  $N_{pe}$  and the time of hit in nanoseconds. The time is negative, meaning that 0 ns is the first hit time and -2048 ns is the latest hit time.

FFNN is a Fully Connected Neural Network (FCDNN) composed of the following layers: the input layer, providing the 15360-item vector, followed by fully connected linear layers with the respective number of neurons being  $[8192, 4096, 2048, 1024, 512, 256, 128, 64, 32]$ . These layers possess a Leaky ReLU activation function defined as

$$\text{LeakyReLU} = \begin{cases} x, & \text{if } x > 0 \\ 10^{-2} \cdot x, & \text{otherwise} \end{cases} \quad (6.14)$$

The last layer is a linear layer with 4 neurons, representing  $(x, y, z, E)$  without an activation function.

The loss used is the Mean Square Error (MSE)

$$\text{MSE}(\boldsymbol{\eta}, \boldsymbol{\eta}^{true}) = \sum_i (\eta_i - \eta_i^{true})^2 \quad (6.15)$$

where  $\eta$  takes the values of  $(x, y, z, E)$ .

2575 The optimizer used for its training is the Stochastic Gradient Descent with momentum

$$\theta_{t+1} = \theta_t - \Lambda \left( \sum_{i=0} \frac{\partial \mathcal{L}}{\partial \theta_{t-i}} \cdot 0.9^i \right) \quad (6.16)$$

2576 where  $\theta_t$  is vector of learnable parameters at step  $t$ .  $\Lambda$  is the learning rate set at  $10^{-3}$ . The difference  
 2577 with the classical SGD is the gradient term with  $i > 1$ . We save the gradient computed in the previous  
 2578 step and use them as momentum with a decaying weight. The factor 0.9 is an hyperparameter that  
 2579 has been selected for the training.

2580 Additionally, to prevent over-fitting, we introduce a weight decay. Each step, we reduce the amplitude  
 2581 of the parameters  $\theta$  by  $10^{-3}$ :

$$\theta_{t+1} = \theta_t \cdot (1 - 10^{-3}) \quad (6.17)$$

## 2582 Performances

2583 The FFNN is trained independently from the ANN. The dataset is comprised of 1M positrons events  
 2584 uniformly distributed in the detector and in energy over  $E_{dep} \in [1, 10]$  MeV. The training dataset  
 2585 account for 990'000 events with 10'000 events reserved for validation. The data are normalized,  
 2586 mean shifted to 0 and standard deviation scaled to 1, before being processed by the network.

2587 Each epochs goes trough the entire training datasets, with a batch size of 64. The training last for 25  
 2588 epochs. The performance the FFNN are presented in Figures 6.5 and 6.6. We remind that goal of this  
 2589 FFNN is not to have competitive performances against classical algorithms like OMILREC but more  
 2590 to have a simple, NN reconstruction algorithm to run the ANN against.

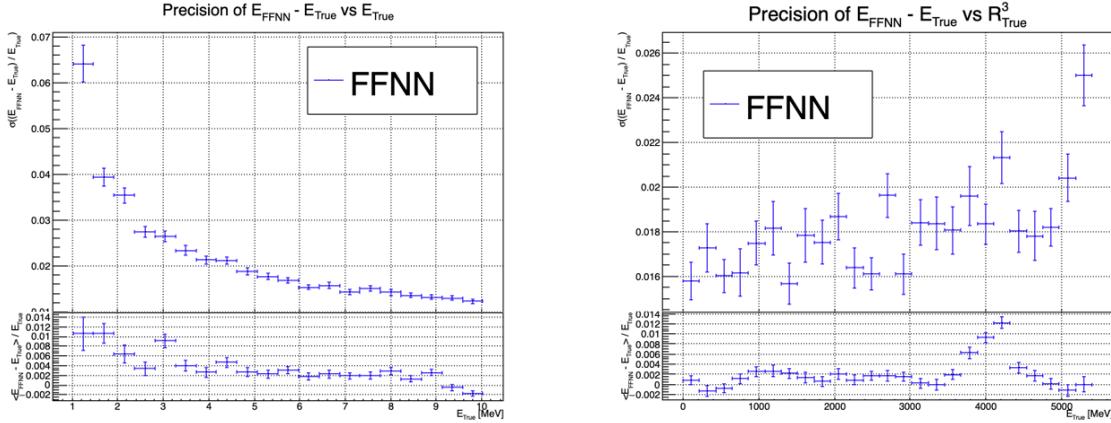


FIGURE 6.5 – Energy resolution of the FFNN with respect to the energy (On the right)  
 and the radius (On the left)

### 2591 6.3.3 Adversarial Neural Network (ANN)

2592 The ANN aims to introduce perturbations in the event data in such a way that these perturbations  
 2593 are not detectable in the control dataset while still degrading the energy reconstruction of the IBD  
 2594 dataset. For this purpose, and for the reasons detailed in Section 6.3.1, the ANN operates on the  
 2595 inputs of the reconstruction network presented above, namely the FFNN. During the training, the  
 2596 parameters of the FFNN are *frozen*, meaning they will not be updated during the ANN training. If  
 2597 they were free to be optimized, they would adapt to the perturbations of the ANN, which would go  
 2598 against the objective of this work.

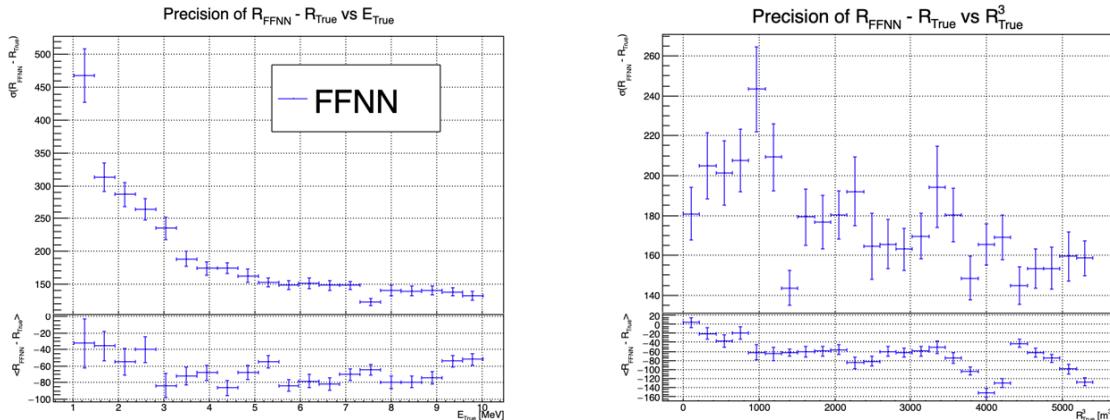


FIGURE 6.6 – Radial resolution of the FFNN with respect to the energy (On the right) and the radius (On the left)

2599 The FFNN takes as input a vector of  $5 \times 3072$  values, representing the  $(x, y, z, Q, t)$  of 3072 Healpix  
2600 pixels. Those values come from the aggregation of the PMTs belonging to those pixels.

2601 It seems unreasonable that the ANN would modify the pixel positions, as they are derived from a  
2602 mathematical construction. It could, however, perturb which PMTs are assigned to specific pixels,  
2603 introducing localization errors, but the position of the PMTs is carefully monitored during JUNO's  
2604 construction. Such aggregation errors would likely arise from PMTs located at the edges of the pixels,  
2605 yet this scenario seems unlikely. Moreover, due to the constraints mentioned in Section 6.3.1, the  
2606 ANN is required to work with the same format that the FFNN uses as input.

2607 At the start of the project, we attempted to have it operate on both time and charge information  
2608 simultaneously, but it struggled to converge. After discussions with colleagues in the collaboration,  
2609 we decided that the ANN would only introduce perturbations in the charge information, as most of  
2610 the energy information comes from the charge.

2611 Our ANN thus needs to output a 3200-dimensional vector, which represents the updated charges of  
2612 the detector.

2613 We decided on a Fully Connected Deep NN (DNN) “bottleneck” architecture for the ANN, illustrated  
2614 in Figure 6.7. This architecture places a 4096-neuron-wide layer after the input, followed by smaller  
2615 layers of sizes 2048, 1024, and 512 neurons, before finally reaching the 256-neuron layer. From this  
2616 layer, the size increases again to 512, 1024, and finally 2048 neurons before the output layer, which  
2617 consists of 3072 neurons.

2618 The idea behind this architecture is that, by reducing the number of neurons per layer, we force the  
2619 network to summarize the event in 256 parameters, that it will use to regenerate an event. This  
2620 architecture has also the advantage of keeping the number of learnable parameters relatively small,  
2621 as the connection between small layers do not require a lot of parameters.

## 2622 ANN loss

As it was mentioned in the introduction of Section 6.3, the loss of the ANN is composed of two losses, the adversarial loss  $\mathcal{L}_{adv}$  and the regularisation loss  $\mathcal{L}_{reg}$ . To those two losses, we adjoin a penalty term that prevent the ANN from producing non-physical events.

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{reg} + P$$

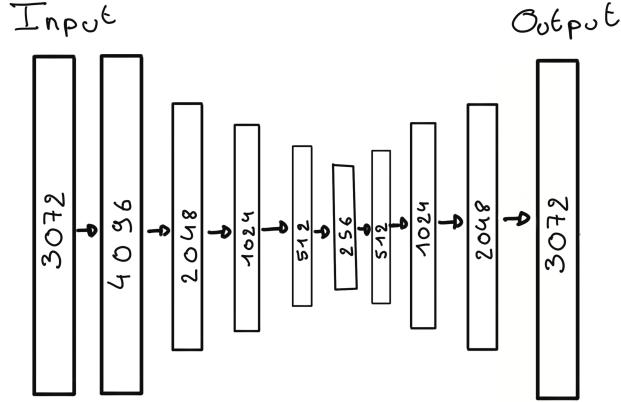


FIGURE 6.7 – Illustration of the “bottleneck” architecture of the ANN. Each block represent a fully connected layer with, on the left, the input layer and on the right the output layer. We see a first reduction of the number of neurons per layer, going from 4096 to 256, followed by an augmentation back to 4096 neurons, thus the “bottleneck”

2623 The adversarial loss  $\mathcal{L}_{adv}$  is defined as the absolute value correlation between the reconstructed  
 2624 energy and the energy deposit (Eq. 6.5). The regularisation loss  $\mathcal{L}_{reg}$  is the MSE of the true and  
 2625 reconstructed energy position vector  $(x, y, z, E)$  (Eq. 6.7).

2626 The penalty term is here to prevent the network from generating event that are too far from the initial  
 2627 event. The penalty  $P$  is a function that takes the pixelated event  $X$ , its transformation after the ANN  
 2628  $\mathcal{G}(X)$  and a constraint  $\epsilon$

$$P(X, \mathcal{G}(X), \epsilon) = \sum_{i=1}^{3072} (\text{ReLU}(-\mathcal{G}(X)_i) + D_i) \quad (6.18)$$

2629 with

$$D_i = \begin{cases} \frac{(X_i - \mathcal{G}(X)_i)^2}{X_i^2} & \text{if } \frac{|X_i - \mathcal{G}(X)_i|}{X_i} > \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (6.19)$$

2630 where  $i$  index the Healpix pixels. The term  $\text{ReLU}(-\mathcal{G}(X)_i)$  is minimal, equal 0, when the charge after  
 2631 perturbation is positive. This term prevent the ANN from producing negative charge, feat impossible  
 2632 for the PMTs.

2633 The second term  $D_i$  is equal to 0 when the relative charge between the original and perturbed pixel is  
 2634 less than  $\epsilon$ . Otherwise, it is the square of this relative charge difference. This term penalize the ANN  
 2635 from producing charges too different from the original event.

2636 When dealing with multiple losses like this, it is important keep them of the same order of magnitude,  
 2637 as we do not want one term to absorb the other.

2638 The loss  $\mathcal{L}_{adv}$  range from 0 to 1 while  $\mathcal{L}_{reg}$  is 0 when the vertex is perfectly reconstructed by it  
 2639 can theoretically go up to infinity. In practice we expect it to take value of the order of magnitude  
 2640 coherent with the reconstruction performances. In fact, if it would take higher value, it would mean  
 2641 that the reconstruction would reconstruct the event far away from the true vertex in comparison  
 2642 to the expected performance. This kind of issue would be immediately be detected, even with  
 2643 simplistic reconstructions such as the charge barycenter, which goes against the goal of producing  
 2644 subtle fluctuation.

2645 We evaluate  $\mathcal{L}_{reg}$  with  $(x, y, z)$  in meter and  $E$  in MeV. If the event is reconstructed with a precision  
 2646 of 15 cm and an energy resolution of 3% at 1 MeV, taking the reconstruction performance of the best

2647 reconstruction algorithm OMILREC (see Sections 3.3 and 5.7),  $\mathcal{L}_{reg} \approx 0.3^2 + 0.03^2 = 0.0909$ . We see  
2648 about an order of magnitude between  $\mathcal{L}_{adv}$  and  $\mathcal{L}_{reg}$ . To compensate for it we weight  $\mathcal{L}_{reg}$

$$\mathcal{L} = \mathcal{L}_{adv} + 60 \cdot \mathcal{L}_{reg} + P(\epsilon) \quad (6.20)$$

2649 The amplitude of  $P$  and the value of  $\epsilon$  will be further discussed in Section 6.4.

## 2650 Hyperparameter optimization

2651 All the ANN hyperparameters presented above have been optimized through the numerous iteration  
2652 the architecture went through. The training is computationally expensive as we need to host both  
2653 networks on the GPU card, reaching quickly the memory limit of the GPU. The training of the ANN  
2654 can takes up to 90h. The requirement of having a powerful GPU can be met locally, as Subatech  
2655 possess an available A100 [72] card with 40GB of memory. We could not port over computing center  
2656 as they only possess V100 [73] GPU with 20GB of memory.

2657 Those constraint made a random search optimization impossible. It is maybe possible, through  
2658 optimisation, to reduce the memory requirements to reach the threshold to run on V100 but the  
2659 challenge was deemed not worth it for an exploratory work.

## 2660 6.4 Training of the ANN

2661 The training of the ANN goes through two phases. The first one consist on reproducing physical  
2662 events, the second one into searching for physically sound perturbations. For both phases, we use  
2663 the both of the datasets presented in section 6.2. We use a batch size of 64 for both datasets meaning  
2664 that, for each steps, the network see 128 events.

2665 Each epochs goes through the entirety of the training dataset.

### 2666 6.4.1 First training phase: back to physics

2667 When the ANN is initialized, before any training has been done, its parameters are initialized with  
2668 random values. Multiple initialization methods exist. In this work, we use a common He initialization  
2669 [87], which is the default initialization in the PyTorch [49] library. If we were to ask for an event  
2670 from the ANN without training first, the results would be random noise. We thus first have the ANN  
2671 learn to reproduce physical events.

2672 For this, we conduct a training of 200 epochs where the loss consists only of the penalty term. For  
2673 scaling purposes, the penalty  $P$  is scaled by 0.25.

$$\mathcal{L}_1 = 0.25 \cdot P(\epsilon = 0.01) \quad (6.21)$$

2674 During this phase, the only objective of the network is to yield events that are the same as the original  
2675 events.

2676 The evolution of this loss  $\mathcal{L}_1$  during the training for the training dataset and the validation dataset is  
2677 presented in Figure 6.8. We see that the ANN converges to some stability in the loss.

2678 The time and charge channels of two events, after this training phase, are presented in Figures 6.9  
2679 and 6.10. We remind that the ANN only act on the charge channel of the event.

2680 We observe that for a localized event, Figure 6.9, the ANN correctly reproduces the event, while for  
2681 a more diffuse event, Figure 6.10, it produces a more uniform charge distribution. By looking at

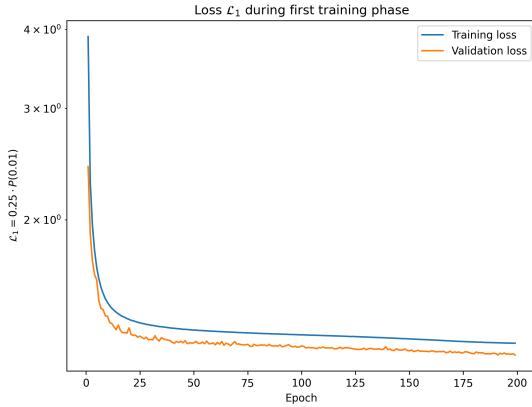


FIGURE 6.8 – Evolution of the loss  $\mathcal{L}_1 = 0.25 \cdot P(0.01)$  during the first phase of the training

the color scale in Figure 6.10, we observe that the ANN does not reproduce singular high numbers of  $N_{pe}$ . The highest pixel in the original was  $12 N_{pe}$ , whereas after the ANN, the highest pixel is  $5 N_{pe}$ . Furthermore, whereas in the original event the charge repartition, while diffuse, was still concentrated in specific pixels, the ANN spreads the charges in all the pixels.

In the next figures, we discuss the reconstruction of the FFNN ( $\mathcal{F}$ ) with and without the presence of the ANN ( $\mathcal{G}$ ) at the end of this Phase 1. The reconstruction by the FFNN of an event perturbed by the ANN is denoted  $(\mathcal{F} \circ \mathcal{G})$ . We differentiate the reconstruction between the two datasets, presented in Section ??: the physics dataset, designated as IBD, and the control dataset, designated  $^{12}B$ .

In Figure 6.11, we show the ratio between the reconstructed energy distribution before and after the application of the ANN. For the  $^{12}B$  dataset, the ratio is close to one except in the bin  $E_{rec} > 9.5$  MeV, where we see an excess of events after the ANN. For the IBD dataset, the ratio is close to 1 over the energy range.

In Figure ??, we present the distribution of the relative reconstruction errors  $(E_{rec}, E_{dep}) / E_{dep}$  with (light histogram) and without (dark histogram) the ANN. We see that without the ANN, the distribution was centered on 0, whereas with it, we observe a small positive bias. In the second row of the histogram, the ratio between the light and dark histograms, we see confirmation of the previous observation, with a deficit of events for  $-0.05 < (E_{rec}, E_{dep}) / E_{dep} < 0.02$  and an excess of events for  $(E_{rec}, E_{dep}) / E_{dep} > 0.02$ . This shift to higher energy explains the excess of events seen in the highest energy bins in Fig. 6.11.

The behavior between the  $^{12}B$  dataset (green histogram) and the IBD dataset (blue histogram) is similar.

#### 6.4.2 Second training phase: Breaking of the reconstruction

Once the ANN is able to reproduce physical events, we change the loss so that it starts to search for potential perturbations. For this we introduce the term  $\mathcal{L}_{adv}$  and  $\mathcal{L}_{red}$  producing a second loss  $\mathcal{L}_2$ . Adding those terms will significantly change the loss. The previous minima in the parameter phase space the ANN found minimizing  $\mathcal{L}_1$  will not be the minima  $\mathcal{L}_2$ . To prevent a gradient explosion, we introduce a growing factor  $\lambda$  in front of the term  $\mathcal{L}_{adv}$  and  $\mathcal{L}_{red}$ . This factor starts at  $\lambda = 0.01$  at epoch 201 and grows  $\lambda_{i+1} = \lambda_i + 0.01$  where  $i$  indexes the epoch. It caps at  $\lambda_{max} = 1$  at epoch 300 after which it stops growing.

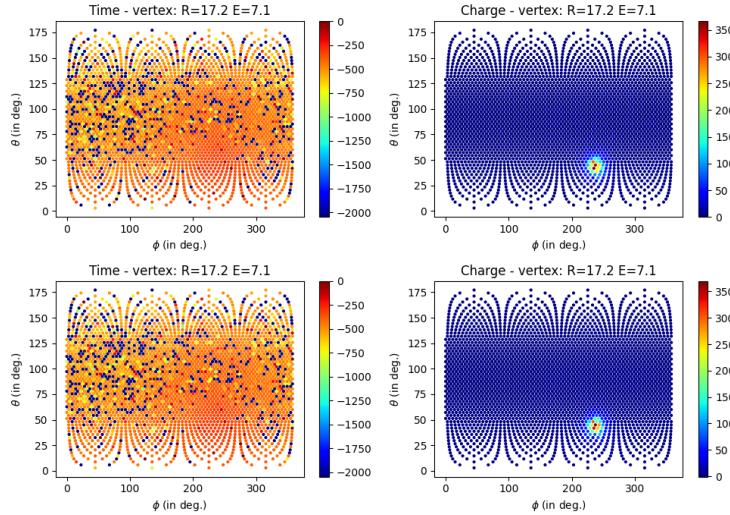


FIGURE 6.9 – Time channel (on the left) and charge channel (on the right) of a **radial, high energy event** ( $R = 17.2$  m,  $E_{dep} = 7.1$  MeV), **Top:** before the ANN perturbation, **Bottom:** after the ANN perturbation. The ANN have been trained for 200 epochs, just after Phase 1. Time channel in ns and charge channel in  $N_{pe}$ .

2711 Also to ease the task of the ANN, we relax the constraint in the penalty term  $P$  from  $P(0.01)$  to  
 2712  $P(0.15)$ .

2713 The expression of the phase 2 loss  $\mathcal{L}_2$  becomes:

$$\mathcal{L}_2 = \lambda (\mathcal{L}_{adv} + 60 \cdot \mathcal{L}_{reg}) + 0.25 \cdot P(0.15) \quad (6.22)$$

2714 This second phase of the training last for 200 more epochs, up to epoch 400.

2715 The profiles of  $\mathcal{L}_2$ ,  $\mathcal{L}_{adv}$ ,  $60 \cdot \mathcal{L}_{reg}$  and  $0.25 \cdot P(0.15)$  during this second phase of the training are  
 2716 presented in Figures 6.13 and 6.14. The profile of the loss  $\mathcal{L}$  over entirety of the training is presented  
 2717 in figure 6.15.

2718 We see in Figure 6.15 that the loss immediately grows at the start of Phase 2. Obviously, part of this  
 2719 effect is due to the term  $\mathcal{L}_{adv}$  as it tries to perturb the reconstruction but, interestingly, the term  $\mathcal{L}_{reg}$   
 2720 that ensures the reconstruction of the control sample is still correct continues to decrease over Phase  
 2721 2. It indicates that, while the penalty term  $P(0.01)$  in Phase 1 did some work in reproducing the input  
 2722 event, it was still not enough for the reconstruction to be at the same level as without the ANN.

2723 At the beginning of Phase 2,  $\mathcal{L}_{adv}$  is not equal to one due to the problem mentioned above, but as  
 2724  $\mathcal{L}_{reg}$  decrease, “correcting” the reconstruction,  $\mathcal{L}_{adv}$  grows close to 1. We see that after about 100  
 2725 epochs of stability in  $\mathcal{L}_{adv}$ , the correlation drops a bit from  $\sim 0.9985$  to  $\sim 0.9975$  while  $\mathcal{L}_{reg}$  continues  
 2726 to decrease, hinting at possible room for perturbation.

2727 After 200 epochs of Phase 2, the correlation in  $\mathcal{L}_{adv}$  is still at 0.998, the penalty term  $P(0.01)$  is stable  
 2728 and the regularisation loss  $reg$  is close to stability.

2729 For illustration, events produced by the ANN after 400 epochs are displayed in Figures 6.16 and 6.17.  
 2730 These are the same event as displayed in Figures 6.9 and 6.10.

2731 The same observations that were made after phase 1 still apply after phase 2. The ANN still spreads  
 2732 the charge over multiple pixels for central events, Figure 6.17, while for radial events it is able to  
 2733 reproduce the small localization of the event.

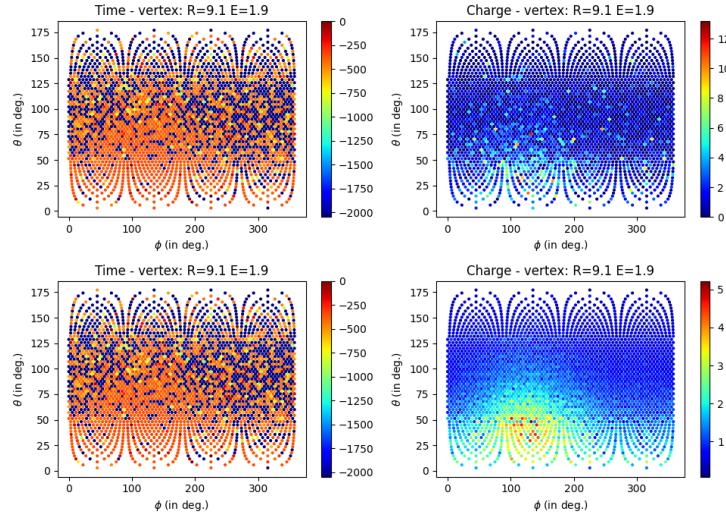


FIGURE 6.10 – Time channel (on the left) and charge channel (on the right) of a **central, low energy event** ( $R = 9.1$  m,  $E_{dep} = 1.9$  MeV), **Top:** before the ANN perturbation, **Bottom:** after the ANN perturbation. The ANN have been trained for 200 epochs, just after Phase 1. Time channel in ns and charge channel in  $N_{pe}$ .

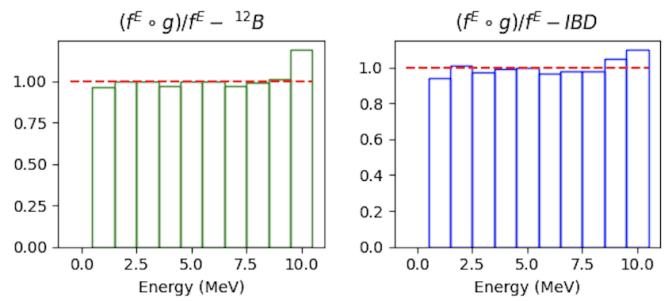


FIGURE 6.11 – Ratio of the reconstructed energy spectra between  $(\mathcal{F} \circ \mathcal{G})$  and  $\mathcal{F}$  at the end of Phase 1 of the training. **On the left:** For the  $^{12}\text{B}$  dataset. **On the right:** For the IBD dataset

When looking at the distribution of ratio between the reconstructed energy distribution before and after the application of the ANN, Figure 6.18, we observe this time a deficit of events in the high energy bin. This deficit is explained by the comparison between the distribution of relative reconstruction errors, Figure 6.19, in which we see a small negative bias. This same figure shows a wider loss in resolution when the ANN is present. This is the ANN working to degrade the resolution of the FFNN.

Figure 6.20 shows the ratio between the relative error on the reconstructed energy between the IBD and the  $^{12}\text{B}$  dataset with and without the ANN. We don't see any indicative difference, the ANN even seems to have harmonized the reconstruction error between the two datasets. The ANN is not capable of introducing perturbation that would target only the IBD dataset, it is not capable of distinguishing the physics dataset from the control dataset.

This leads us to one of two conclusions. Either this reconstruction algorithm is robust to ANN attacks, or the ANN is not powerful enough to find meaningful attack. We lean to the second proposition. The simple architecture of the ANN, the fact that it does not precisely reproduce events after Phase 1 of the training and the trouble we experienced to balance the loss between  $\mathcal{L}_{adv}$  and  $\mathcal{L}_{rec}$  are clues

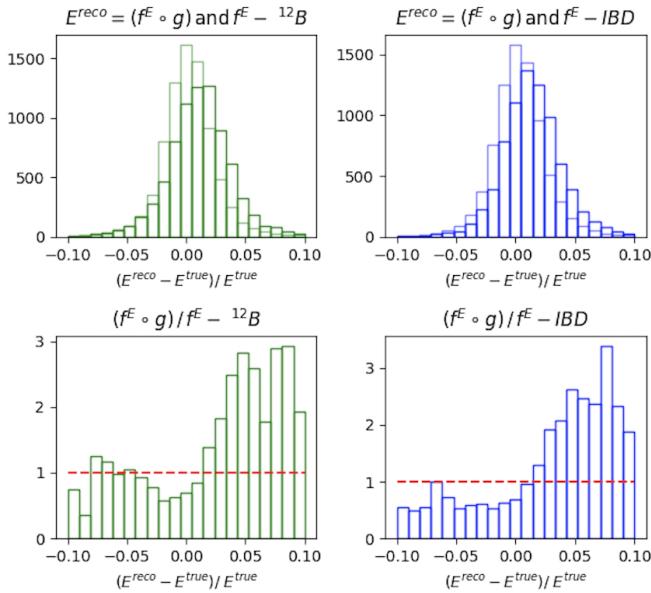


FIGURE 6.12 – **On the top :** Distribution of the relative energy reconstruction error between  $\mathcal{F}$  (light histogram) and  $(\mathcal{F} \circ \mathcal{G})$  (dark histogram) at then end of Phase 1 of the training. **On the bottom :** Ratio between the light and dark histogram of the top figure.

2749 indicating we can probably produce a more powerful ANN.

## 2750 6.5 Conclusion and prospect

2751 Reliability and knowledge of our reconstruction algorithms are crucial for the successful conduct  
 2752 of the experiment. The first step to testing and comparing the reconstruction algorithms is to have  
 2753 them publicly available. To this end, I have implemented a BDT for energy reconstruction in JUNO’s  
 2754 common software and compared its performance and behavior in detail to the classic likelihood algo-  
 2755 rithm OMILREC. The strong correlation between their errors indicates that close to no improvement  
 2756 can be made by combining the two algorithms, as they use the same information.

2757 In this chapter, we explore the relevance of using an ANN to produce physically sound perturbations  
 2758 that would distort the reconstructed energy spectrum while being invisible to the control sample. We  
 2759 present a simple architecture. I show the complexity of developing such an architecture; the gradient  
 2760 back-propagation technique poses a number of problems, namely the impossibility of backpropagat-  
 2761 ing through complex and external algorithms.

2762 We have developed a simple reconstruction method to run the ANN against. The current ANN  
 2763 architecture has trouble reproducing precisely the event, even before being tasked to introduce per-  
 2764 turbations, and once it tries, it is not able to find such perturbations.

2765 Multiple things can be implemented to explore further with the ANN. As discussed, the ANN has  
 2766 trouble reproducing events; it is possible that the loss  $\mathcal{L}_1 = P(0.01)$  is not sufficient for this task,  
 2767 but also that the architecture is not adapted to our problem. ResNet architectures [45] have already  
 2768 proven that the introduction of residual operations helps the network reach better performance. We  
 2769 can imagine a network where instead of  $X' = \mathcal{G}(X)$  we have  $X' = \mathcal{G}(X) + X$ , where the ANN  $\mathcal{G}$   
 2770 computes only the perturbation instead of a whole new event.

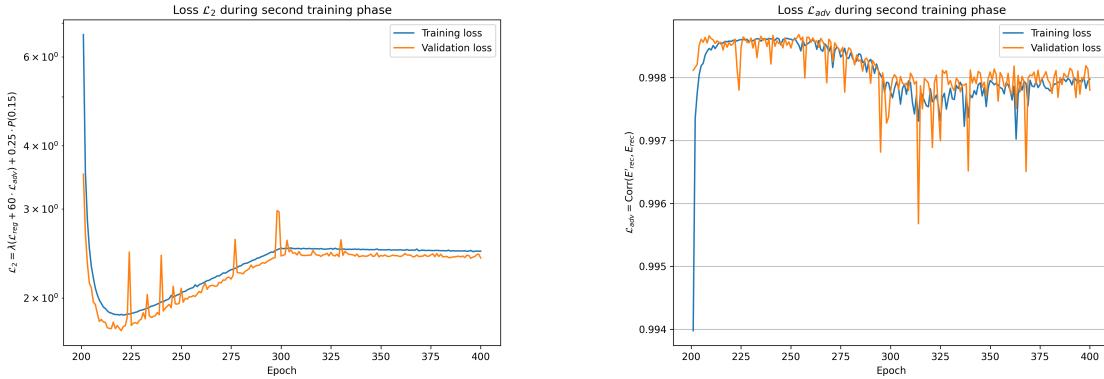


FIGURE 6.13 – Profile of the loss  $\mathcal{L}_2$  and  $\mathcal{L}_{adv}$  during the second phase of training. The linear increase of  $\mathcal{L}_2$  is due to the growing factor  $\lambda$  in Eq. 6.22.

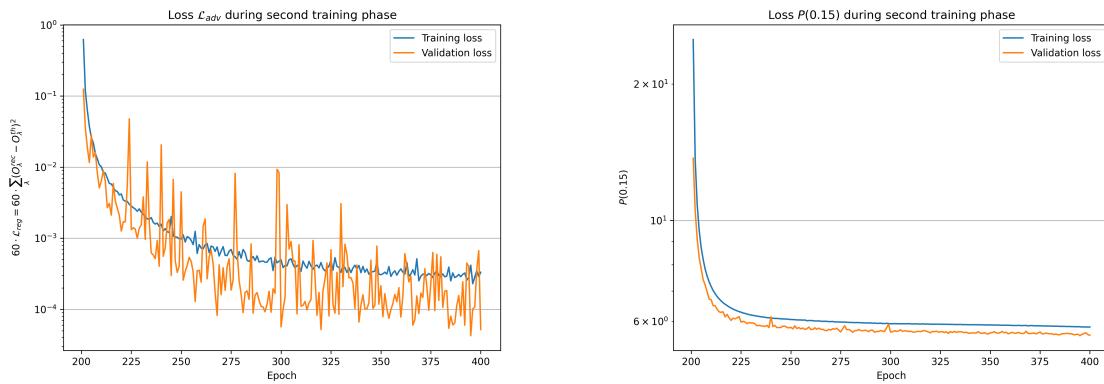


FIGURE 6.14 – Profile of the loss  $60 \cdot \mathcal{L}_{reg}$  and  $0.25 \cdot P(0.15)$  during the second phase of training

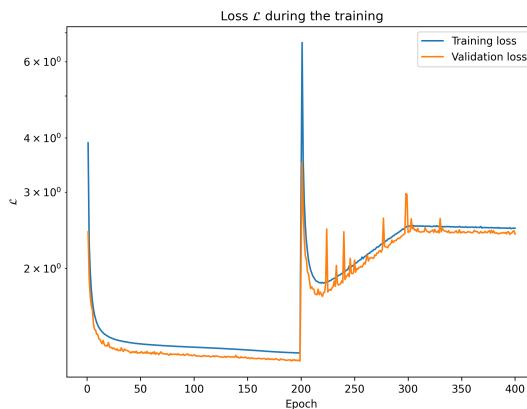


FIGURE 6.15 – Profile of the loss over the entirety of the training (Phase 1 and 2)

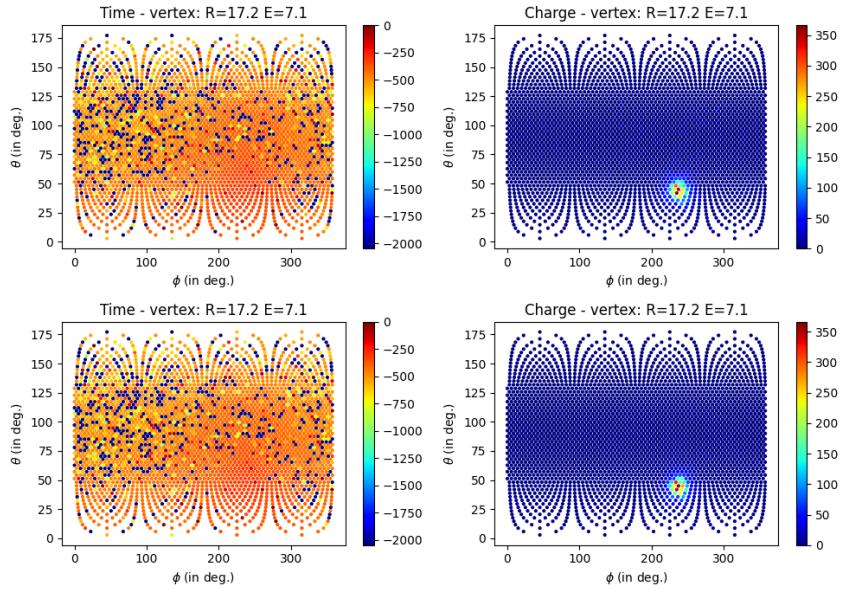


FIGURE 6.16 – Time channel (on the left) and charge channel (on the right) of a **radial, high energy event** ( $R = 17.2$  m,  $E_{dep} = 7.1$  MeV), **Top:** before the ANN perturbation, **Bottom:** after the ANN perturbation. The ANN have been trained for 400 epochs, just after Phase 2. Time channel in ns and charge channel in  $N_{pe}$ .

Further work on the loss is necessary. The equilibrium between  $\mathcal{L}_{adv}$  and  $\mathcal{L}_{reg}$  is crucial and should be further studied. A good way to study its effect would be to compare the performance of the ANN for a different set of weights. If one can determine an equilibrium rule between the two, it can be adjusted dynamically during the training, resulting in finer optimisation.

The architecture of the ANN is, for now, very simple; it's a Fully Connected Deep NN with a bottleneck architecture. Previous work in developing ML for reconstruction [63] and the algorithms presented in Chapters 4 and 5 show the relevance of convolutions in the reconstruction, and the work of Gavrikov et al. [64] presented at the beginning of this chapter hints at the importance of the time and charge distribution. A more complex and refined architecture can probably be more effective.

Another way to improve the ANN would be to find potential discrepancies between the IBD and the  $^{12}\text{B}$  datasets and “guide” it to produce those perturbations.

Finally, to use this method on every reconstruction algorithm, we must move away from the back-propagation method, for reasons detailed in Section 6.3.1, and use different methods such as Reinforcement Learning.

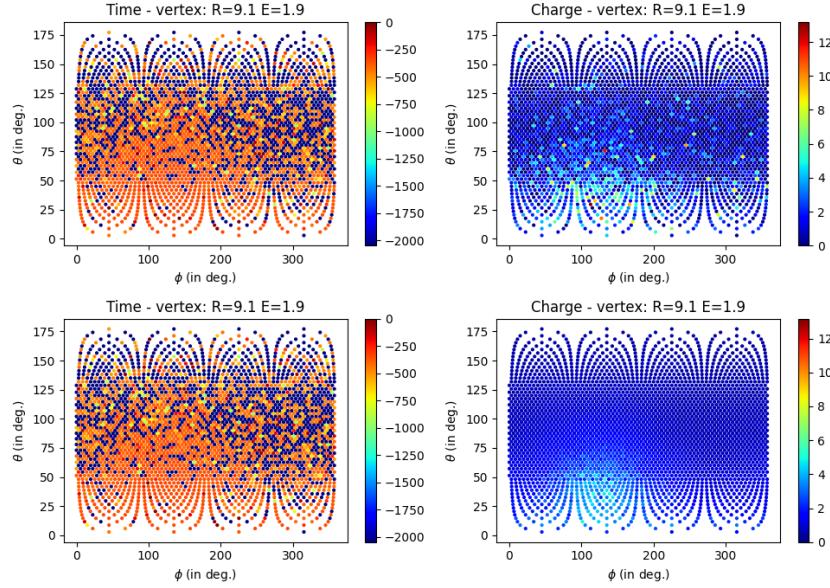


FIGURE 6.17 – Time channel (on the left) and charge channel (on the right) of a **central, low energy event** ( $R = 9.1$  m,  $E_{dep} = 1.9$  MeV), **Top:** before the ANN perturbation, **Bottom:** after the ANN perturbation. The ANN have been trained for 400 epochs, just after Phase 2. Time channel in ns and charge channel in  $N_{pe}$ .

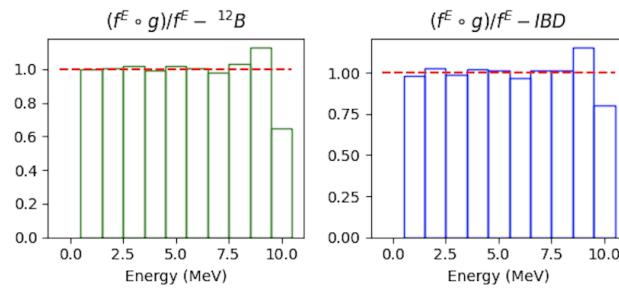


FIGURE 6.18 – Ratio of the reconstructed energy spectra between  $(\mathcal{F} \circ \mathcal{G})$  and  $\mathcal{F}$  at the end of Phase 2 of the training. **On the left:** For the  ${}^{12}B$  dataset. **On the right:** For the IBD dataset

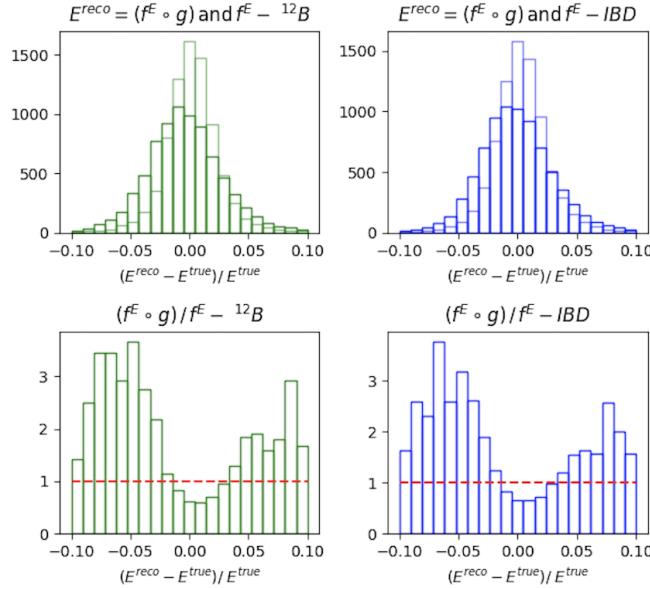


FIGURE 6.19 – **On the top :** Distribution of the relative energy reconstruction error between  $\mathcal{F}$  (light histogram) and  $(\mathcal{F} \circ \mathcal{G})$  (dark histogram) at then end of Phase 2 of the training. **On the bottom :** Ratio between the light and dark histogram of the top figure.

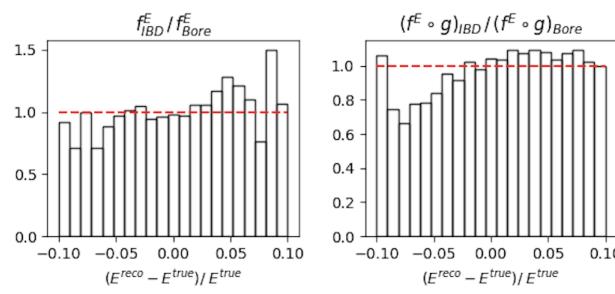


FIGURE 6.20 – Ratio between the relative error on the reconstructed energy between the IBD and the  ${}^{12}B$  dataset. **On the right :** without the ANN. **On the left :** with the ANN.

<sup>2785</sup> **Chapter 7**

<sup>2786</sup> **Dualcalorimetric analysis with  
neutrino oscillation for Precision  
Measurement**

<sup>2789</sup> “We demand rigidly defined areas of doubt and uncertainty!”  
*Douglas Adams, The Hitchhiker’s Guide to the Galaxy*

<sup>2790</sup> **Contents**

<sup>2791</sup> <b>7.1 Motivations</b>	<sup>2792</sup> . . . . .	<sup>2793</sup> <b>122</b>
7.1.1 Discrepancies between the SPMT and LPMT results	. . . . .	122
7.1.2 Charge Non-Linearity (QNL)	. . . . .	122
<sup>2794</sup> <b>7.2 Our approach to Dual Calorimetry with neutrino oscillation</b>	<sup>2795</sup> . . . . .	<sup>2796</sup> <b>124</b>
7.2.1 Toy experiments	. . . . .	126
7.2.2 Comparing the solar parameters from individual analyses : LPMT vs SPMT	. . . . .	127
7.2.3 Direct comparison between the SPMT and LPMT spectra	. . . . .	129
7.2.4 Joint fit of the SPMT and LPMT spectra : $\chi^2_{H_0} - \chi^2_{H_1}$	. . . . .	131
7.2.5 Joint fit of the SPMT and LPMT spectra : distribution of $\delta \sin^2(2\theta_{12})$ and $\delta \Delta m^2_{21}$	. . . . .	132
7.2.6 Limitations	. . . . .	132
<sup>2800</sup> <b>7.3 Fit software</b>	<sup>2801</sup> . . . . .	<sup>2802</sup> <b>133</b>
7.3.1 AveNue <sub>e</sub> Standalone Generators	. . . . .	134
7.3.2 AveNue <sub>e</sub> Fitting Package	. . . . .	134
7.3.3 Details of the IBD generator	. . . . .	135
<sup>2803</sup> <b>7.4 Technical challenges and development</b>	<sup>2804</sup> . . . . .	<sup>2805</sup> <b>136</b>
<sup>2806</sup> <b>7.5 Covariance matrix</b>	<sup>2807</sup> . . . . .	<sup>2808</sup> <b>137</b>
7.5.1 Analytical method	. . . . .	137
7.5.2 Empirical method	. . . . .	139
<sup>2809</sup> <b>7.6 Technical Validation</b>	<sup>2810</sup> . . . . .	<sup>2811</sup> <b>140</b>
<sup>2812</sup> <b>7.7 Results</b>	<sup>2813</sup> . . . . .	<sup>2814</sup> <b>143</b>
7.7.1 Effect of supplementary QNL on the LPMT spectrum	. . . . .	143
7.7.2 Comparison and statistical tests results	. . . . .	145
<sup>2815</sup> <b>7.8 Conclusion and perspectives</b>	<sup>2816</sup> . . . . .	<sup>2817</sup> <b>148</b>
7.8.1 Empirical correlation matrix from fully simulated event	. . . . .	149

<sup>2818</sup> JUNO is a high-precision neutrino oscillation experiment. To resolve the Neutrino Mass Ordering (NMO) with the required statistical significance, JUNO must be sensitive to the subtle spectral phase shift, on the order of a few percents, as illustrated in Figure 7.1. This phase shift manifests as a small

difference between the Normal Ordering (NO) and Inverted Ordering (IO) spectra, which becomes even smaller after accounting for detection effects such as energy resolution smearing, non-linear detector responses, and background contamination, as shown in Figure 7.2.

This chapter is based on simulated data due to the unavailability of real JUNO data, which will only be available in 2025. The purpose of this analysis is to validate the methods and tools developed for dual calorimetry and neutrino oscillation measurements, ensuring that they are robust and ready for future real data.

Among other condition, a precise and complete understanding of the reconstruction and detector effects is crucial. The challenge reside in the technology used in the detector, which, while based on well known technology: scintillator observed by PMT, is being deployed on a scale never seen before, in term of scintillator volume and PMT size. Understanding every effects that goes in the detector can become extremely complicated. Any method to help detecting problems is therefore welcome. Comparing the data and results obtained by two systems measuring the same events, but subject to different sources of error, is therefore precious. This is the purpose of the dual calorimetry techniques used in JUNO thanks to the existence of 2 PMT systems: the LPMT and SPMT systems.

The reconstruction of the IBD positron energy must be very performant: an unprecedented resolution of 3% at 1 MeV [41] is necessary to determine the NMO with the aimed significance.

Furthermore, an energy scale uncertainty below 1% is essential to accurately assess the likelihood of the NO and IO hypotheses. If this uncertainty exceeds 1%, systematic biases could distort the reconstructed spectra, potentially leading to the erroneous exclusion of the correct mass ordering hypothesis (NO or IO). For instance, a shift in the energy scale could mimic a phase shift between the spectra, making it possible to wrongly favor NO when IO is true, or vice versa. This effect has been studied in the introduction of Chapter 4 of [30].

Understanding all the effects influencing the detector response can be quite complex. Consequently, any methodologies that facilitate problem detection and validation of the reconstruction processes are essential for ensuring accurate results.

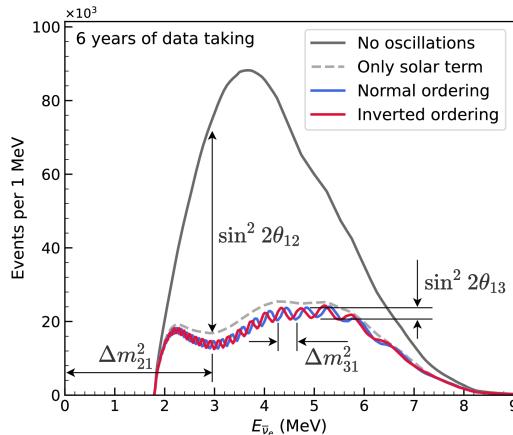


FIGURE 7.1 – Expected number of neutrinos event per MeV in JUNO after 6 years of data taking. The black curve shows the flux if there was no oscillation. The light gray curve shows the oscillation if only the solar terms are taken in account ( $\theta_{12}$ ,  $\Delta m_{21}^2$ ). The blue and red curve shows the spectrum in the case of, respectively, NO and IO. The dependency of the oscillation to the different parameters are schematized by the double sided arrows. We can see the NMO sensitivity by looking at the fine phase shift between the red and the blue curve.

One detector effect to take into account is the detector non linearity. Detector non-linearity can introduce significant biases in the energy reconstruction of events, compromising the precision of

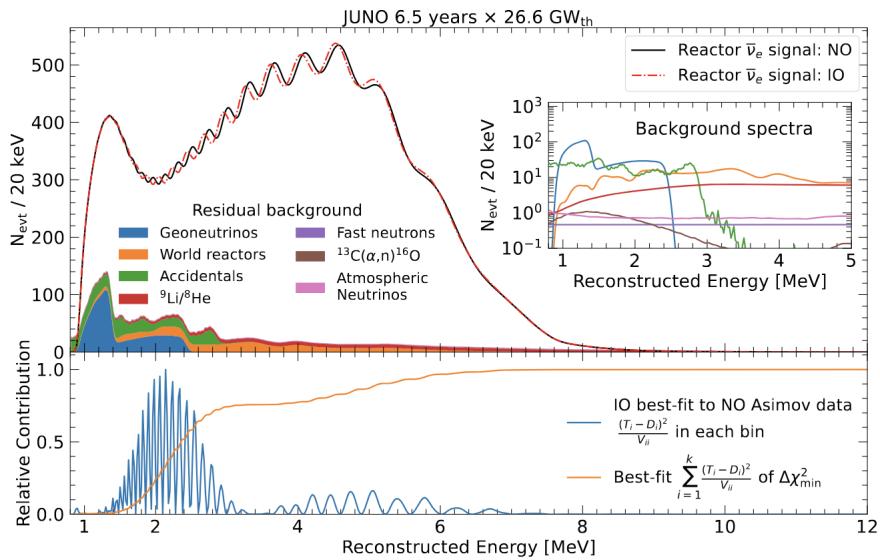


FIGURE 7.2 – Oscillated reactor  $\bar{\nu}_e$  spectra for the Normal Ordering (Black) and Inverted Ordering (Red) for 6.5 years data taking and a resolution of 3% without any statistical or systematic fluctuation. Figure from [38].

neutrino oscillation measurements and increasing systematic uncertainties, which could potentially distort the determination of the neutrino mass.

One of the possible source of non-linearity, which will be used as a reference in this chapter, is the charge non-linearity (QNL) that will be discussed in next section. Several dual calorimetry techniques can address this issue. Some are calibration techniques, that are also described in section 4.3 of [30]. More generally, comparing the results of the two systems will allow for the detection of potential issues on the calibration or reconstruction. This is done in this thesis by comparing directly the spectra and oscillation parameters measurements of the two PMT systems. We call this kind of dual calorimetry "Dual calorimetry with neutrino oscillation", since it is based on the visible energy spectra used by the oscillation analysis of reactor antineutrinos.

In this chapter, we explore several ways to perform this comparison. One of them relies on the difference between the values of  $\Delta m_{21}^2$ ,  $\sin^2(2\theta_{12})$  measured with the LPMT and the SPMT systems. Both systems measure them with similar uncertainties. For reasonable values of the QNL, we expect these differences to be smaller than the individual uncertainties. However, the significance of these differences might still be high. Indeed, both systems reconstruct the same events, therefore the same distribution of the true positron energy, as well as the same scintillation photon emission. Therefore, the energy spectra reconstructed by the two systems share a part of their fluctuations. This translates into correlated reconstructed spectra and consequently lead to correlations between the measurements of  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$ . The uncertainty on the SPMT-LPMT difference is largely decreased by this correlation. Other ways to perform the comparison (see next sections) all rely the reconstructed spectra, therefore on the evaluation of the correlation between the LPMT and SPMT spectra.

In the next section we will discuss the motivations behind this study. In Section 7.2, I present the methods we propose to implement Dual calorimetry with neutrino oscillation, and of the way we estimate their sensitivity. In section 7.3, I present the fit framework used, and then, in section 7.4 the technical improvement brought and the difficulties faced during the development. To end this chapter I present the results in 7.7 and discuss the conclusions and perspectives in 7.8.

## 2877 7.1 Motivations

### 2878 7.1.1 Discrepancies between the SPMT and LPMT results

2879 As mentioned earlier, the SPMT and LPMT systems are expected to detect the same events. Therefore,  
 2880 after proper calibration, any significant discrepancies between the two systems' results could  
 2881 indicate a calibration error, a systematic effect, or an unaccounted detector issue. Detecting such  
 2882 differences is critical, as even small deviations from the expected response could compromise the  
 2883 determination of the Neutrino Mass Ordering (MO) or introduce systematic biases in the oscillation  
 2884 parameter measurements, leading to incorrect conclusions about the true mass ordering.

2885 Both systems are anticipated to show similar sensitivity to the oscillation parameters  $\theta_{12}$  and  $\Delta m_{21}^2$   
 2886 [9]. Therefore, any detected discrepancies will be based on these parameter measurements. A sim-  
 2887 ple comparison of the values and independent uncertainties from the two systems could highlight  
 2888 discrepancies. However, we believe—and will demonstrate in this chapter—that an independent  
 2889 analysis of each system lacks critical information. By considering both statistical and systematic  
 2890 correlations between the two systems, we can design more robust and powerful statistical tests.

2891 Our work in this chapter is to develop such tools, which in practice implies to define test statistics. A  
 2892 first step will be to determine the distribution of these test statistics in the case when no unexpected  
 2893 problem affects the LPMT nor the SPMT problem. This will give us the distribution of those statistical  
 2894 test in absence of discrepancies. Later, the value of the test statistics that we will measure in real data  
 2895 can be compared to these distributions to produce p-values, to judge of the potential present of an  
 2896 unexpected effect.

2897 To evaluate the power of our methods, we need to simulate a concrete difference between the two  
 2898 spectra. We have chosen to study a specific potential effect, Charge Non-Linearity (QNL), which will  
 2899 be detailed in the following section. QNL affects the reconstructed energy spectrum by introducing  
 2900 a non-linear relationship between the true and measured charge in the PMTs. Our statistical tests  
 2901 are designed to detect such distortions, and they should be sensitive to unexpected effects—such as  
 2902 calibration errors or insufficient simulation precision—as long as the induced distortion exceeds a  
 2903 threshold of approximately 1-2% in the reconstructed energy spectrum.

### 2904 7.1.2 Charge Non-Linearity (QNL)

2905 The energy response of the Central Detector (CD) is influenced by two types of non-linearity. The  
 2906 first arises from the intrinsic properties of the Liquid Scintillator (LS), where the photon production  
 2907 is not linearly proportional to the deposited energy, as shown in Figure 2.12a. This non-linearity  
 2908 results from a combination of scintillation and Cherenkov light production. The scintillation yield is  
 2909 governed by Birk's law, which introduces a "quenching" effect that depends on the particle type and  
 2910 energy. Additionally, Cherenkov radiation, which constitutes less than 10% of the collected light, in-  
 2911 troduces a velocity-dependent non-linearity. These physical non-linearities in the LS contribute to the  
 2912 overall non-linearity of the energy response before any further distortions from the photomultiplier  
 2913 tubes (PMTs)

2914 The second type of non-linearity comes from the LPMT charge measurements. When photons hit a  
 2915 PMT and give rise to PEs, a current pulse is formed. In the photon counting regime, simply exceeding  
 2916 a certain threshold allows to conclude that a single photon hit the PMT. When several photons hit the  
 2917 PMT simultaneously, one enters the photon integration regime : the pulse is sampled and integrated  
 2918 over a certain time window to produce a reconstructed charge Q. Calibration methods are applied  
 2919 to determine the relationship between the charge Q and the number of PEs (which is the quantity  
 2920 proportional to the energy deposit one wants to measure). Several effects impact this procedure:  
 2921 the signal pulse can fluctuate and be distorted between two events where the same number PEs  
 2922 occurred; the PMT gain might not be linear as a function of the number of photons that hit the PMT;

the charge reconstruction algorithm is not supposed to be perfect, and its results are further affected by electronic noise and inter-channel cross-talk. The impact of these effects grows with the number of PEs.

Precedent studies, Section 4.2.3 of [30], suggest a model for the channel-wise QNL:

$$\frac{Q_{rec}}{Q_{true}} = \frac{-\gamma_{qnl}}{9} Q_{true} + \frac{\gamma_{qnl} + 9}{9} \quad (7.1)$$

where  $Q_{rec}$  is the reconstructed number of PE by the PMT,  $Q_{true}$  is true number of PE that hit the PMT, and  $\gamma_{qnl}$  is a factor representing the amplitude of the non-linearity.

Studies at previous experiments, like Daya Bay, concluded that the best reachable control of QNL in the 1-10 PEs range was  $\gamma_{qnl} = 0.01$  [88]. As already mentionned in Section 2.3.2, JUNO LPMTs operate in a larger range : 1-100 PEs (See also table 7.1). In such a case, a realistic value of  $\gamma_{qnl}$  is not known.

	1PE	2~5PE	5~10PE	10~20PE	20~50PE	50~100PE	>100PE
LPMT	42.56%	40.54%	8.74%	5.12%	2.80%	0.24%	0.003%
SPMT	95.19%	4.80%	0.01%	0%	0%	0%	0%

TABLE 7.1 – The charge fraction in terms of the number of PE collected at the single PMT for the reactor  $\bar{\nu}_e$  IBD events. Table taken from [30]

The event-wise impact resulting from the channel-wise QNL can be parameterised this way :

$$\frac{E_{vis}^{rec}}{E_{vis}^{true}} = \frac{-\alpha_{qnl}}{9} E_{vis}^{true} + \frac{\alpha_{qnl} + 9}{9} \quad (7.2)$$

In JUNO, the visible energy is proportional to the number of emitted photons per unit energy deposit. It includes the physical non linearities. In the equation above  $E_{vis}^{true}$  is this visible energy, while  $E_{vis}^{rec}$  is what it becomes when the reconstructed charges found in an event are modified according to Eq. 7.1.

An example is shown on Fig. 2.14, where we show the  $E_{vis}^{rec}/E_{vis}^{true}$  ratio for several samples of uniformly distributed electron events, generated with various values of  $E_{vis}^{true}$ . Here, an extreme value  $\gamma_{qnl} = 0.05$  was assumed. On can see on Fig. 2.14 that it corresponds to a 2% effect at 8 MeV, equivalent to  $\alpha_{qnl} = 0.025$ . The effect of Eq 7.2 is illustrated in Figure 7.3.

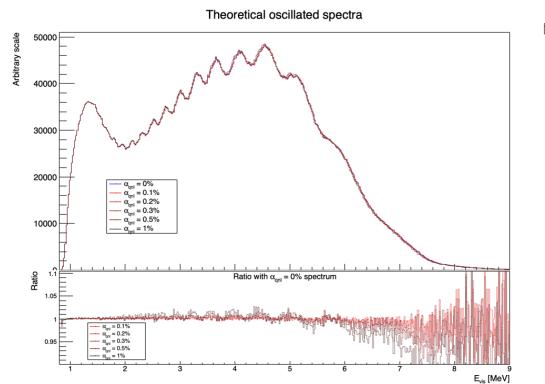


FIGURE 7.3 – On top: Oscillated spectra for different value of  $\alpha_{qnl}$ . On bottom: Ratio of the number of event with  $\alpha_{qnl} = 0\%$ .

This example is from references [30], which aimed at demonstrating the potential of the dual

2943 calorimetry calibration method mentioned in section 2.4.3. If it works as hoped, the residual event-  
 2944 wise QNL effect will be below 0.3%. In this chapter, we propose methods to detect residuals higher  
 2945 than this.

2946 Fig. 7.5 show several other examples with varying  $\gamma_{qnl}$  values, and the corresponding values of  $\alpha_{qnl}$ .  
 2947 Using 1M events from the JUNO official simulation J23.0.1-rc8.dc1 (released on 7th January 2024), we  
 2948 simulated events up to the photon collection in LPMTs and introduced an additional channel-wise  
 2949 QNL by using the equation 7.1 to modify the number of collected photons.

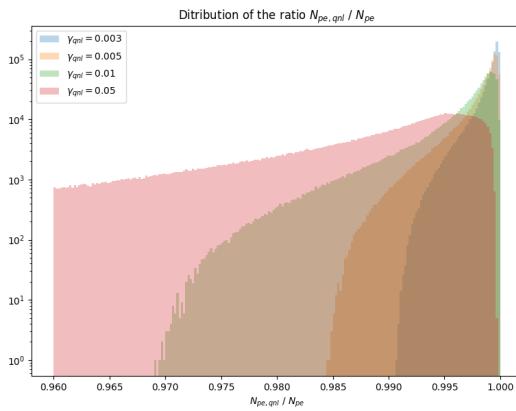


FIGURE 7.4 – Distribution the ratio reconstructed charge (in nPE equivalent) over the number of collected nPE for different value of  $\gamma_{qnl}$ . We use a sample of 1 million positron event uniformly distributed in the detector and in energy in the range  $E_{dep} \in [1, 10] MeV$

2950 In Figure 7.4 we show the distribution of the ratio of the reconstructed charge (in nPE equivalent)  
 2951 over the number of collected nPE for different values of  $\gamma_{qnl}$ . The right parts of those distribution,  
 2952 where the ratio is close to 1, are mostly central events. The charge is homogeneously distributed, the  
 2953 effect of the channel-wise QNL is reduced because the PMTs each collect a relatively small number  
 2954 of nPE. The left tail, with ratio < 1, are radial events, the photons are concentrated in a small number  
 2955 of PMTs, the effect of the channel wise QNL is greater.

2956 In Figure 7.5, we show the mean of the distributions of Figure 7.4 as a function of the energy. From  
 2957 the 8.5 MeV data point, we compute an effective  $\alpha_{qnl}$ . The effect of this effective  $\alpha_{qnl}$  is represented  
 2958 as the dashed line. On the bottom of Fig 7.5 is presented the charge ratio difference between the  
 2959 effective  $\alpha_{qnl}$  and the mean effect of a  $\gamma_{qnl}$ . We see that the event-wise QNL, described by Eq. 7.2,  
 2960 do not represent correctly the channel-wise QNL described by Eq. 7.1 at low energy. Indeed, Eq. 7.2  
 2961 assume no QNL effect at 1 MeV, where in reality some of the PMTs will still suffer from QNL.

2962 Despite this difference, the necessity to use the effective event-wise model expressed by Eq. 7.2,  
 2963 and consequently to find the correspondence between values of  $\gamma_{qnl}$  and  $\alpha_{qnl}$ , instead of directly the  
 2964 channel wise model of Eq. 7.1 will be explained in Section 7.2.1.

## 2965 7.2 Our approach to Dual Calorimetry with neutrino oscillation

2966 In this section, we describe 4 statistical tests that we propose to use to detect unexpected effects in  
 2967 one of the PMT systems. Each test is based on a particular test statistics. In practice, the main result  
 2968 we want to produce in this chapter is the distributions followed by these test statistics.

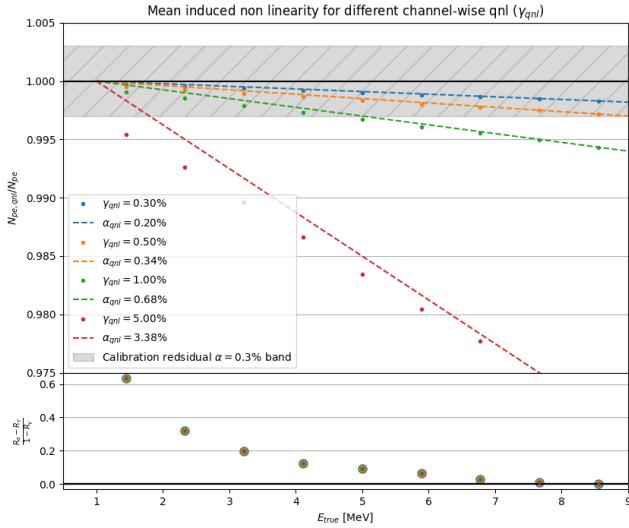


FIGURE 7.5 – **On top:** Ratio of the reconstructed charge (in nPE equivalent) over the number of collected nPE. The dots represent the mean of the distributions in Figure 7.4 and the dashed line are the equivalent event-wise non-linearity from eq 7.2. The hatched zone is the residual non-linearity expected after calibration [32]. **On bottom:** Difference between QNL induced by an event wise QNL and the mean QNL induced by a channel wise QNL. The value for  $\alpha_{qnl}$  and  $\gamma_{qnl}$  follow the color code of the top figure. For a given energy, all the data point have the same value.

In this section, we propose four distinct statistical tests designed to detect unexpected discrepancies between the LPMT and SPMT systems. Each test aims to evaluate different aspects of the reconstructed energy spectra:

1. Test 1 compares the measurements of solar oscillation parameters  $\sin^2 2\theta_{12}$  and  $\Delta m_{21}^2$  derived independently from each system.
2. Test 2 directly compares the LPMT and SPMT spectra bin by bin.
3. Test 3 involves a joint fit of the two spectra, with and without a hypothesis of discrepancy.
4. Test 4 examines the residuals in the fit of oscillation parameters  $\sin^2 2\theta_{12}$  and  $\Delta m_{21}^2$  after the joint fit. The primary objective of this analysis is to establish the distributions of these test statistics under both the null hypothesis (no unexpected effect) and the alternative hypothesis (presence of a discrepancy).

The distributions of these test statistics cannot be analytically determined and are instead generated empirically through toy experiments. In each toy experiment, we generate two spectra of the IBD visible energy: one from the LPMT system and the other from the SPMT system. Since both systems observe the same events, their statistical fluctuations are correlated. To account for this, we compute a  $(820 \times 820)$  covariance matrix that captures both the bin-to-bin correlations within each spectrum and the cross-correlations between the LPMT and SPMT spectra. Details of the sample generation process are provided in Section 7.3.3. Note that we use toy samples rather than samples produced by the full simulation of JUNO since the latter option would not be affordable in terms of computing time.

In the next subsection, we present the informations the reader must know about these spectra to understand the test statistics presented in the rest of the current section.

### 2991 7.2.1 Toy experiments

2992 The sensitivity of our tests depends on the sample size, which scales with the duration of exposure  
 2993 to the antineutrino flux: 100 days, 1 year, 2 years, and 6 years. For each exposure time, we generate  
 2994 1000 toy experiments, where the number of events in the LPMT and SPMT spectra is drawn from a  
 2995 Poisson distribution with the expected mean value for that exposure. Since the same physical events  
 2996 are reconstructed by both systems, their fluctuations are not independent, and we account for the  
 2997 statistical correlations between the LPMT and SPMT spectra in our toy generation process. It was  
 2998 recently evaluated in the recent reference paper on JUNO's sensitivity [38] that about 95000  
 2999 IBDs would be selected in 6 years.

3000 An example of pair of spectra is shown on Figure 2.16, in the form of two joint histogram of 410, 20  
 3001 keV wide bins each. This is the format used in the fit performed by the present version of the reactor  
 3002 oscillation analysis developed at Subatech. It is important to notice that the IBD events present in  
 3003 the LPMT spectrum of a toy experiment are the same as those in the SPMT spectrum: the same  
 3004 events are just reconstructed twice, by either system. The LPMT and SPMT spectra are therefore not  
 3005 independent : Their respective fluctuations in the number of entries per bin are correlated. These  
 3006 correlations stem from what is common between the LPMT et SPMT spectra, namely :

- 3007 — The statistical fluctuations of the true  $E_{vis}$  distribution (before any reconstruction).
- 3008 — The fluctuation of the number of photons produced by scintillation or Cherenkov effect.

3009  
 3010 When generating toy experiment, the fluctuations drawn in each bin around the average expected  
 3011 number of events must account for these correlations. We therefore evaluated the  $(820 \times 820)$  co-  
 3012 variance matrix describing the uncertainty on the number of entries in each of the 410 bins of the  
 3013 2 spectra, as well as the bin-to-bin correlations, especially those between the bins of the LPMT  
 3014 spectrum and those of the SPMT spectrum. This is described in Section 7.5. Here, we just want  
 3015 to emphasize the importance of this point, one of the original tasks to be carried out for the work  
 3016 presented in this chapter.

3017 As already stated earlier, toy experiments will be used to evaluate the distributions of the four test  
 3018 statistics. We will first produce reference distributions: the ones that rule the possible values of  
 3019 the test statistics if none of the PMT systems is affected by any unexpected effect. These references  
 3020 are sufficient to run a test once JUNO will take data: the values of the test statistics obtained in  
 3021 a real data sample can be compared with the reference distributions, to evaluate to which extent  
 3022 the null hypothesis (no unexpected effect) is credible (p-values, or any pertinent quantities, can be  
 3023 computed). This is true whatever the nature of the unexpected effect.

3024 To give an idea of the power of the method, an explicit scenario must be simulated for the unex-  
 3025 pected effect. For that purpose, we also generate sets of toy experiments where the  $E_{vis}$  spectrum  
 3026 reconstructed by the LPMT is distorted using Eq. 7.2. We will test the following levels of QNL:  $\alpha_{qnl} \in$   
 3027  $\{0.003, 0.002, 0.001\}$ . As a reminder, the calibration guarantees a residual event-wise non-linearity of  
 3028  $\alpha_{qnl} \leq 0.003$  [32].

3029 The most probable values in the distributions of the test statistics obtained in such cases will be  
 3030 compared with the reference distributions to derive a "median" predicted p-value. One can also  
 3031 compute the probability to observe in real data a p-value lower than a certain value, if the assumed  
 3032 QNL effect actually exists in these data.

3033 When we initiated this work, the best test statistics to use was not obvious to us. This is why we  
 3034 decided to test 4 test statistics, of growing complexity. We present them in the 4 next subsections.

### 3035 7.2.2 Comparing the solar parameters from individual analyses : LPMT vs SPMT

3036 The first test statistics is probably the most natural one: it's essentially a direct comparison of the  
 3037 values of  $\sin^2(2\theta_{12})$  and  $\Delta m_{21}^2$  measured by separate analyses of the LPMT and the SPMT spectra.  
 3038 These analyses are performed using the oscillation fit tool developed at Subatech, described in Sec-  
 3039 tions 2.7 and 7.3. A fit to the LPMT spectrum provides  $\sin^2(2\theta_{12})_L$  and  $\Delta m_{21,L}^2$ , while a separate fit to  
 3040 the SPMT spectrum provides  $\sin^2(2\theta_{12})_S$  and  $\Delta m_{21,S}^2$ .

The direct comparison proceeds in practice via the differences between the fit results :

$$\Delta\theta = \sin^2(2\theta_{12})_L - \sin^2(2\theta_{12})_S \quad (7.3)$$

$$\Delta D = \Delta m_{21,L}^2 - \Delta m_{21,S}^2 \quad (7.4)$$

3041

3042 A very simple test statistics would be for instance

$$S = \frac{|\Delta\theta|}{\sigma_{\Delta\theta}} \quad (7.5)$$

3043 directly related to the significance of the difference between the SPMT and LPMT results. This  
 3044 requires to determine the uncertainty  $\sigma_{\Delta\theta}$ . This cannot be considered as the mere quadratic sum  
 3045 of the uncertainties on  $\sin^2(2\theta_{12})_L$  and  $\sin^2(2\theta_{12})_S$  returned by the fitter. Indeed, because of the  
 3046 correlations, described in the previous subsection, between the LPMT and SPMT spectra, the fitted  
 3047 parameters are also correlated.

3048 The calculation of  $\sigma_{\Delta\theta}$  must account for it. Simple error propagation dictates :

$$\sigma_{\Delta\theta}^2 = \sigma_{\sin^2(2\theta_{12})_L}^2 + \sigma_{\sin^2(2\theta_{12})_S}^2 - 2\sigma_{\sin^2(2\theta_{12})_L}\sigma_{\sin^2(2\theta_{12})_S}C_{L,S} \quad (7.6)$$

3049 where  $C_{L,S}$  is the correlation between the SPMT and LPMT measurements. We expect it to be high  
 3050 (well above 0.9, see Figures 7.6, 7.7, 7.8 and 7.9). Consequently, we expect it to considerably lower  
 3051 the value of  $\sigma_{\Delta\theta}^2$ , and increase the significance  $S$ .

3052 This simple example can be seen as an illustration of the fact that the correlations between the LPMT  
 3053 and SPMT spectra boosts the sensitivity of our test statistics to unexpected effects. Indeed, with 6  
 3054 years of data, and counting only the statistical uncertainties, we expect the statistical uncertainties  
 3055  $\sigma_{\sin^2(2\theta_{12})_L}^2$  and  $\sigma_{\sin^2(2\theta_{12})_S}^2$  to both be around 0.15% [9]. A preliminary evaluation [66] of the impact  
 3056 of an uncorrected QNL effect with  $\alpha_{qnl} = 1\%$  on the value of  $\sin^2(\theta_{12})$  predicted a bias of 0.1%,  
 3057 therefore of 0.05% on  $\sin^2(2\theta_{12})$ . With no correlation, this would lead to a significance  $S$  far below 1.  
 3058 Accounting for the correlation allows far better.

3059 The test statistics we actually use for this direct comparison is a generalisation of the simple one  
 3060 above : it includes both the results on  $\sin^2(2\theta_{12})$  and  $\Delta m_{21}^2$  :

$$\chi_{ind}^2 = \Delta_{ind}^T U^{-1} \Delta_{ind} \quad (7.7)$$

3061 where  $\Delta_{ind}$  is a vector defined as

$$\Delta_{ind} = [\Delta\theta, \Delta D] \quad (7.8)$$

3062 using equations 7.3 and 7.4.

3063 The covariance matrix  $U$  is a  $(2 \times 2)$  matrix containing the uncertainties on the components of  $\Delta_{ind}$   
 3064 and the correlation between them. We derive this matrix from the  $(4 \times 4)$  covariance matrix  $V$ ,  
 3065 which contains the uncertainties on the fitted values of  $\sin^2(2\theta_{12})_L$ ,  $\sin^2(2\theta_{12})_S$ ,  $\Delta m_{21,L}^2$  and  $\Delta m_{21,S}^2$ ,  
 3066 as well as the correlations between these quantities. For that purpose, we simply use the linear error

propagation formalism, that can be found in section 40.2.6 of the statistical review of the PDG 2020 [40] :

$$U = A V A^T \quad (7.9)$$

where the transfer matrix  $A$  is obtained this way

$$A_{ij} = \frac{\partial \Delta_i^{ind}}{\partial \lambda_j} \quad (7.10)$$

where  $\lambda_j$  one of the parameters ( $\Delta m_{21,L}^2, \sin^2(2\theta_{12})_L, \Delta m_{21,S}^2, \sin^2(2\theta_{12})_S$ ). Assuming this indexing order for  $j$  and  $i$  ordering following Eq 7.8,  $A$  is expressed

$$A = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \quad (7.11)$$

We acknowledge that linear error propagation is valid when all fluctuations or uncertainties are gaussian. However, since our results will be based on distributions of  $\chi^2_{ind}$  produced with toy samples, this choice remains valid.

An important ingredient here is to determine the correlation coefficients in  $V$ . On a dedicated set of 1000 toy experiments, we perform fits to the LPMT and SPMT spectra, and compute the correlations empirically from the 1000 sets of best fit values of the solar parameters :  $\sin^2(2\theta_{12})_L$  vs.  $\sin^2(2\theta_{12})_S$ ,  $\Delta m_{21,L}^2$  vs  $\Delta m_{21,S}^2$ ,  $\sin^2(2\theta_{12})_L$  vs.  $\Delta m_{21,S}^2$ , etc. We need the correlations corresponding to the null hypothesis and therefore use toy experiments produced with no QNL effect.

The correlations between these parameters for 100 days, 1 year, 2 years and 6 years can be found in Figures 7.6, 7.7, 7.8 and 7.9 respectively.

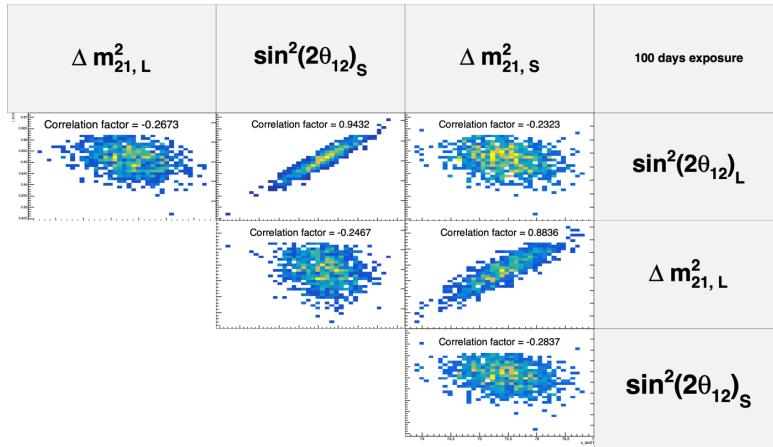


FIGURE 7.6 – Distribution and correlation between the best fit point of 1000 individual toys fit for 100 days exposure without supplementary QNL.

We observe strong correlation between the reconstructed  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$  of both systems as presented in Table 7.2, row one and two. As the relative statistical uncertainty decrease with exposure, the correlations grow ranging from 0.88 to 0.95 for  $\Delta m_{21}^2$  and from 0.94 to 0.98 for  $\sin^2(2\theta_{12})$ . We observe between parameters of the same fit, a small anti-correlation of about -0.25, line 4 and 5 of Table 7.2.

Because the parameters are heavily correlated between the LPMT and SPMT fit, and that  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$  are slightly anti-correlated in the same fit, the couples of different parameters from different fit,  $\text{Corr}(\sin^2(2\theta_{12})_L, \Delta m_{21,S}^2)$  and  $\text{Corr}(\sin^2(2\theta_{12})_S, \Delta m_{21,L}^2)$ , are also anti-correlated.

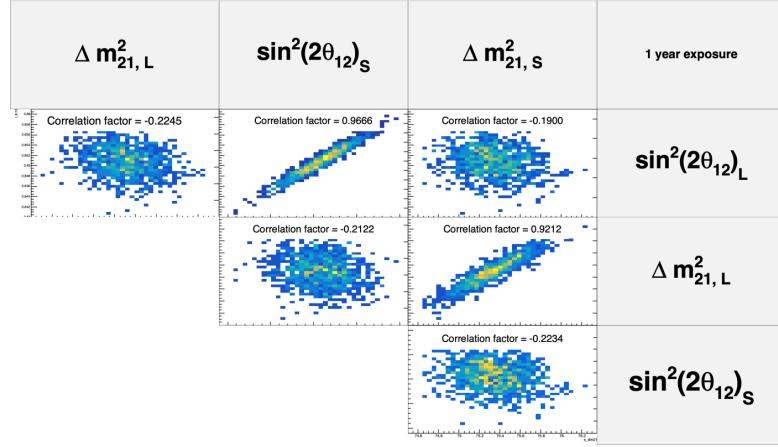


FIGURE 7.7 – Distribution and correlation between the best fit point of 1000 individual toys fit for 1 year exposure without supplementary QNL.

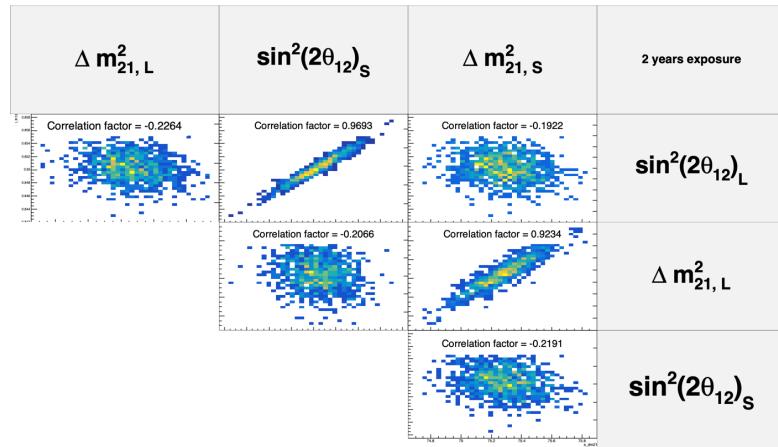


FIGURE 7.8 – Distribution and correlation between the best fit point of 1000 individual toys fit for 2 years exposure without supplementary QNL.

<sup>3090</sup> The distributions  $\chi^2_{ind}$  will be Shown in Section 7.7.

### <sup>3091</sup> 7.2.3 Direct comparison between the SPMT and LPMT spectra

<sup>3092</sup> In the second test, we perform a bin-by-bin comparison of the LPMT and SPMT spectra without fitting any oscillation parameters. Again, we use here a  $\chi^2$ -like statistics. We do not expect the reference <sup>3093</sup> distribution (for  $\alpha_{qnl} = 0$ ) to be centered around the number of degree of freedom (i.e. the number <sup>3094</sup> of bins of each spectrum in our case) as should be distributed (if the spectra contain enough events <sup>3095</sup> in each bin to assume a gaussian behavior of the number of entries) the  $\chi^2$  comparing 2 histograms <sup>3096</sup> when they are consistent with each other. Indeed, even in the absence of unexpected events, the <sup>3097</sup> LPMT and SPMT are quite different because of the very different reconstruction resolutions. We <sup>3098</sup> therefore need here again to establish this reference distributions with toys. And compare them later <sup>3099</sup> with the distributions obtained for the various tested values of  $\alpha_{qnl}$ . <sup>3100</sup>

<sup>3101</sup> Our test statistics is :

$$\chi^2_{spe} = \Delta_{spe}^T U^{-1} \Delta_{spe} \quad (7.12)$$

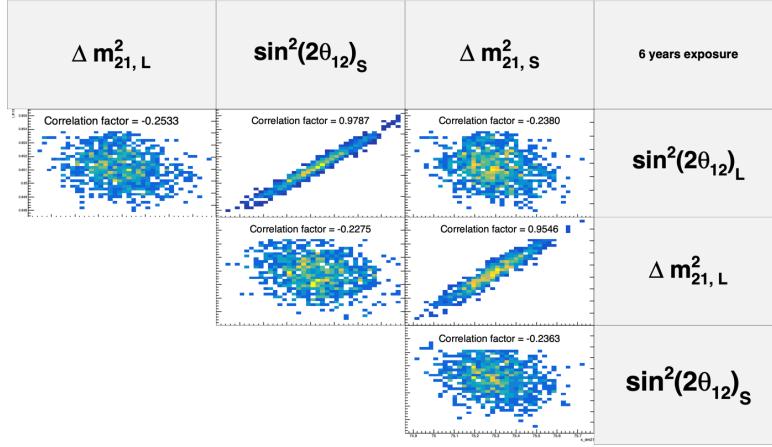


FIGURE 7.9 – Distribution and correlation between the best fit point of 1000 individual toys fit for 6 years exposure without supplementary QNL.

	100 days	1 year	2 years	6 years
Corr( $\Delta m_{21,L}^2, \Delta m_{21,S}^2$ )	0.8836	0.9212	0.9234	0.9546
Corr( $\sin^2(2\theta_{12})_L, \sin^2(2\theta_{12})_S$ )	0.9432	0.9666	0.9693	0.9787
Corr( $\sin^2(2\theta_{12})_L, \Delta m_{21,L}^2$ )	-0.2673	-0.2245	-0.2264	-0.2533
Corr( $\sin^2(2\theta_{12})_S, \Delta m_{21,S}^2$ )	-0.2837	-0.2234	-0.2191	-0.2363
Corr( $\sin^2(2\theta_{12})_L, \Delta m_{21,S}^2$ )	-0.2323	-0.19	-0.1922	-0.2380
Corr( $\sin^2(2\theta_{12})_S, \Delta m_{21,L}^2$ )	-0.2467	-0.2122	-0.2066	-0.2275

TABLE 7.2 – Correlations between the parameters BFP of the individual LPMT and SPMT fits for multiple exposures using 1000 toys.

3102 where

$$\Delta_i^{spe} = h_{L,i} - h_{S,i} \quad (7.13)$$

3103 and

$$U = AVA^T \quad (7.14)$$

3104 Here,  $i$  runs over the 410 bins of the individual spectra. Also,  $h_{L,i}$  and  $h_{S,i}$  are the contents of the  $i$ th  
 3105 bin of the LPMT and SPMT spectra respectively. We need to know the uncertainty on  $\Delta_i^{spe}$  and the  
 3106 correlations with  $\Delta_j^{spe}$ 's in other bins. We derive them from  $V$ , the  $(820 \times 820)$  covariance matrix  
 3107 introduced at the beginning of this section, which can be seen as the covariance matrix of a 820-bin  
 3108 double spectrum juxtaposing the LPMT and SPMT spectra. We remind its determination will be  
 3109 presented in Section 7.5. To obtain  $U$  from  $V$ , we again apply the linear error propagation, with the  
 3110 transfer matrix :

$$A_{ij} = \frac{\partial \Delta_i^{spe}}{\partial h_j} = \frac{\partial (h_{L,i} - h_{S,i})}{\partial h_j} \quad (7.15)$$

3111 Thus,  $A_{ij} = 1$  if  $i = j$ , and  $A_{ij} = -1$  if  $j$  is the SPMT bin corresponding to the  $i$  LPMT bin.

3112 We expect this statistics to have a certain power since  $\chi^2_{spe}$  can be increased for 2 reasons in case of  
 3113 unexpected problem: first, the LPMT spectrum (if the LPMT is affected) will be distorted and become  
 3114 less consistent with the SPMT spectrum; second, the correlations between the LPMT and SPMT might  
 3115 also modified. Since  $V$  present a peculiar correlation pattern (see Section 7.5), a departure from this  
 3116 pattern also has some valuable impact on  $\chi^2_{spe}$ .

### 3117 7.2.4 Joint fit of the SPMT and LPMT spectra : $\chi^2_{H_0} - \chi^2_{H_1}$

3118 This kind of fit has already been introduced in Section 2.7. As a reminder, it involves the minimisation  
 3119 of

$$\chi^2_{\text{joint}} = (\mathbf{T}(\boldsymbol{\theta}, \mathbf{h}) - \mathbf{D})^T V^{-1} (\mathbf{T}(\boldsymbol{\theta}, \mathbf{h}) - \mathbf{D}) + \ln(|V|) \quad (7.16)$$

3120 where  $\mathbf{T}(\boldsymbol{\theta}, \mathbf{h})$  is the predicted joint LPMT+SPMT spectrum and  $\mathbf{D}$  the corresponding data vector. The  
 3121 matrix  $V$  is the full  $(820 \times 820)$  covariance matrix which incorporate both the statistical uncertainties  
 3122 and the bin-to-bin correlations between the LPMT and SPMT spectra.

3123 In this fit, we include the usual oscillation parameters,  $\sin^2(2\theta_{12})$ ,  $\Delta m_{21}^2$ ,  $\sin^2(2\theta_{13})$  and  $\Delta m_{31}^2$  along  
 3124 with two additional parameters,  $\delta \sin^2(2\theta_{12})$  and  $\delta \Delta m_{21}^2$  which allow for a potential discrepancy in  
 3125 the LPMT reconstruction or calibration.

3126 Several remarks must be made here to better understand what we do precisely.

- 3127 — Given JUNO's lack of sensitivity to  $\sin^2(2\theta_{13})$ , this parameter is fixed in the fit to the PDG value  
 3128 (see table 7.3). In most of JUNO's fit procedures (see Section 2.7), it's allowed to float during  
 3129 the minimisation, but is treated like a nuisance parameter, by adding a penalty term based on  
 3130 the PDG central value and uncertainty.
- 3131 — The oscillation fit that we perform here does not really aim at the oscillation parameters in  
 3132 themselves, but is performed to detect a difference between the LPMT and SPMT spectra.  
 3133 JUNO is supposed to be very sensitive to  $\Delta m_{31}^2$  via the LPMT spectrum. However, it has been  
 3134 shown by studies carried out at Subatech (and confirmed since then by other groups in the  
 3135 Collaboration), that up to 2 years of data taking, the presence of multiple minima in  $\Delta m_{31}^2 \chi^2$   
 3136 profile can make its determination delicate. Since  $\Delta m_{31}^2$  is not the aim of our present study, we  
 3137 stabilize the fit by treating this parameter as a nuisance parameter, adding to  $\chi^2_{\text{joint}}$  the following  
 3138 penalty term :

$$\chi^2_{\Delta m_{31}^2} = \frac{(\Delta m_{31}^2 - \overline{\Delta m_{31}^2})^2}{\sigma_{\Delta m_{31}^2}^2} \quad (7.17)$$

3139  
 3140 We define two hypothesis. The hypothesis  $H_0$  assumes that no unexpected effect is present, meaning  
 3141 that  $\delta \sin^2(2\theta_{12}) = 0$  and  $\delta \Delta m_{21}^2 = 0$ , and the hypothesis  $H_1$  where  $\delta \sin^2(2\theta_{12}) \neq 0$  and  $\delta \Delta m_{21}^2 \neq 0$   
 3142 are needed to account for any potential calibration or reconstruction bias. The test statistic is then  
 3143 defined as the difference between the minimized  $\chi^2$  values under  $H_0$  and  $H_1$ :

$$\Delta \chi^2 = \chi^2_{\text{joint}, H_0} - \chi^2_{\text{joint}, H_1} \quad (7.18)$$

3144 where  $\chi^2_{\text{joint}, H_0}$  is the result of the minimisation when the fit assumed no unexpected effect (fixing  
 3145  $\delta \sin^2(2\theta_{12})$  and  $\delta \Delta m_{21}^2$  to 0), while  $\chi^2_{\text{joint}, H_1}$  assumes a possible effect, letting this parameters free  
 3146 to float. A large value of  $\Delta \chi^2$  would indicate a significant deviation from the null hypothesis (no  
 3147 discrepancy), suggesting the presence of an unexpected effect in the LPMT system.

3148 Distributions of  $\chi^2_{H_0} - \chi^2_{H_1}$  in the reference case and for various values of  $\alpha_{qnl}$  will be produced and  
 3149 studied in Section 7.7.

3150 The idea behind this joint fit is that by letting the oscillation parameters and  $\delta \sin^2(2\theta_{12})$  and  $\delta \Delta m_{21}^2$   
 3151 free to float, converging potentially to arbitrary, wrong values in the case of oscillation parameters,  
 3152 we add some flexibility to fully exploit the difference introduced by unexpected effects between the  
 3153 reference spectra and correlations.

3154 There were other reasons to develop this joint fit. The main one was that it required an update of  
 3155 our software framework so it's able to perform joint fit. It was not fully ready for that. This feature

$\sin^2(2\theta_{12})$	$\Delta m_{21}^2$	$\Delta m_{31}^2$	$\sin^2(2\theta_{13})$
$0.851^{+0.020}_{-0.018}$	$7.53 \pm 0.18 \times 10^{-5} \text{ eV}^2$	$2.5283 \pm 0.034 \times 10^{-3} \text{ eV}^2$	$0.08523 \pm 0.00268$

TABLE 7.3 – Nominal PDG2020 value [40]. All value are reported assuming Normal Ordering.

<sup>3156</sup> will be very useful when the Subatech team will include the TAO spectrum (via a joint fit) in the  
<sup>3157</sup> oscillation studies it will perform.

### <sup>3158</sup> 7.2.5 Joint fit of the SPMT and LPMT spectra : distribution of $\delta \sin^2(2\theta_{12})$ and <sup>3159</sup> $\delta \Delta m_{21}^2$

<sup>3160</sup> The last test statistics we will study might be complementary to  $\Delta\chi^2 = \chi^2_{joint,H0} - \chi^2_{joint,H1}$ .

<sup>3161</sup> These test statistics are simply the fitted values of  $\delta \sin^2(2\theta_{12})$  and  $\delta \Delta m_{21}^2$ . In the reference case, when  
<sup>3162</sup> no unexpected reconstruction problem is present, we expect them to be distributed in an approximate  
<sup>3163</sup> gaussian way, centered on 0. When QNL effect will be included, they will tend to converge to higher  
<sup>3164</sup> values, to compensate the bias introduced on the fitted  $\sin^2(2\theta_{12})$  and  $\Delta m_{21}^2$  due to the distortion of  
<sup>3165</sup> the LPMT spectra and of the correlations between the LPMT and SPMT spectra.

<sup>3166</sup> Again, these distributions will be studied in Section 7.7.

### <sup>3167</sup> 7.2.6 Limitations

#### <sup>3168</sup> QNL in backgrounds

<sup>3169</sup> The JUNO commons inputs provides background spectra that already have been smeared by the  
<sup>3170</sup> LPMT resolution. Because the resolution depends on  $E_{vis}$  (Eq. 7.19), to apply supplementary QNL we  
<sup>3171</sup> would need to de-convolute the LPMT resolution, apply the supplementary QNL then re-smear the  
<sup>3172</sup> spectra. This deconvolution is no trivial. Thus we ignore the background when produced distorted  
<sup>3173</sup> spectra.

<sup>3174</sup> This should not affect too much the power of our statistical tools, as the backgrounds are common to  
<sup>3175</sup> both spectra and should not have any effect on the statistical covariance matrix.

#### <sup>3176</sup> Systematics

<sup>3177</sup> It would be more rigorous to also include systematic uncertainties. However, in the present state of  
<sup>3178</sup> our fit framework, it would require the computation (often empirical, via the generation of thousands  
<sup>3179</sup> of toy samples) of  $(820 \times 820)$  covariance matrices, which was judge too time consuming with respect  
<sup>3180</sup> to the time we could devote to this chapter.

<sup>3181</sup> Moreover, it seems reasonable to think that the sensibilities evaluated with only statistical uncer-  
<sup>3182</sup> tainties would not be changed much by a full treatment. Indeed, all the systematic uncertainties  
<sup>3183</sup> affect the true visible energy spectrum, before reconstruction. This spectrum is a common input to  
<sup>3184</sup> both the LPMT and SPMT reconstructions. Therefore, observed differences between the oscillation  
<sup>3185</sup> parameters measured by one or the other system should not be due to these systematics effects, and  
<sup>3186</sup> remain of the same order as if these effects were absent.

### 3187 Correlation between LPMT and SPMT reconstruction

3188 Most of our results assume uncorrelated reconstruction uncertainties between the SPMT and LPMT  
 3189 systems. In practice, once the  $E^{vis}$  of a toy event is generated (see Section 7.3.3), we simulate the  
 3190 SPMT and LPMT reconstruction by adding a  $\delta E_{SPMT}^{rec}$  and a  $\delta E_{LPMT}^{rec}$ .

3191 The two latter increments are chosen randomly on Gaussian distribution. These two drawings are  
 3192 carried out independently. In reality, the reconstruction of  $E^{vis}$  is about proportional to the number of  
 3193 PE, therefore to the number of scintillation photons produced in the scintillator. Both the LPMT and  
 3194 SPMT reconstruction depend on the stochastic variation of this number event to event. Their results  
 3195 therefore vary in a correlated way. The correlation is kept low since it is shuffled by another source  
 3196 of variability, namely the sampling of photons : the SPMT indeed reconstruct only a few dozen PEs  
 3197 when more than 10000 photons are emitted.

3198 This correlation is higher when the interaction takes place close to the sphere's surface (ie close to  
 3199 some of the PMTS), the non-uniformity effect is correlated between the two systems. To account  
 3200 for it, when should ideal produce the simulated samples necessary to our studies by using the full  
 3201 simulation. However, it would be far too CPU intensive. The impact of neglecting this correlation  
 3202 will be discussed in Section 7.5.

### 3203 Realistic QNL

3204 The way we implement the QNL effect in toy samples is also simplified. The size of the QNL effect  
 3205 in a PMT depends on the number of photons hitting it, therefore on the position of the interaction.  
 3206 When generating toy events, we apply QNL event-wise, only as a function of the value of  $E^{vis}$  (Eq.  
 3207 7.2). As explained in Section 7.1.2, the full simulation has been used to find the average  $\alpha_{qnl}$  for a  
 3208 given  $\gamma_{qnl}$  which is considered sufficient for this exploration.

3209 Again, replacing toy samples with samples generated with the full simulation would yield more  
 3210 accurate results, but is prohibitive in terms of calculation time. For future studies, sophisticated  
 3211 solutions to this problem will have to be found, but are out of the scope of this thesis.

## 3212 7.3 Fit software

3213 In this section, I describe the fit framework that was used in this study. The AveNu<sub>e</sub> framework is  
 3214 the adaptation to JUNO of one of the frameworks, partly developed at Subatech, used by the Double  
 3215 Chooz [89] experiment. It is composed of two parts: the AveNu<sub>e</sub> Generators and the AveNu<sub>e</sub> Fitting  
 3216 Package. The Generators are a set of standalone macros, the Fitting Package is an C++ package, using  
 3217 the RooFit library.

3218 Both parts of the package are interfaced with what we call the JUNO inputs. These inputs comprise  
 3219 all the ingredients to build a  $T(\theta, \eta)$  prediction, among which :

- 3220 — Reactor antineutrino spectra for each isotope as predicted by Mueller [90].
- 3221 — The isotopes mean releases energy.
- 3222 — Reactors's thermal powers and fission fractions.
- 3223 — Various corrections to account for the contributions from the Non Equilibrium Regime and the  
 3224 Spent nuclear fuel.
- 3225 — A correction obtained by comparing these spectrum prediction in the case of the Daya Bay  
 3226 experiment with actual Daya Baya data [17].
- 3227 — The IBD differential cross section as function of the antineutrino energy.

- The assumed values of the oscillation and nuisance parameters at the start of the fit or for sensitivity studies.
- Parameters describing the non linearity of the photon emission as a function of the deposited energy.
- Energy reconstruction parameters (see equation 7.19 and Figure 7.10).
- The selected IBD and background expected yields per day, and the background spectra, all obtained from JUNO’s full simulation and studies to design the selection.
- Uncertainties on all these quantities for the computation of covariance matrices.

We describe in the next section the role of each part of the framework.

### 7.3.1 AveNu<sub>e</sub> Standalone Generators

The main macro here is the “IBD generator” macro. It is used to :

- Compute  $T_{no\ osc}(\eta)$  (unoscillated theoretical spectra) predictions. It is done by toy generating a spectrum. In order to not be affected by statistical fluctuations, it generates 100 times more statistics than JUNO’s expected yield after 6 years. It is provided in the form of a TTree. These predictions concern an non oscillated spectrum.
- Toy samples simulated data sets. It is essentially used to simulate data spectra altered by QNL effects (see below).
- The above productions are input to the Fitting Package, or to other macros from Standalone Generators, which compute the covariance matrices necessary to the Fitting Package. Some of the covariance matrices are computed from the  $T$ ’s, using linear error propagation, some other are empirical calculations based on sets of toy samples generated with varying parameters. This is also the case for one of the versions of the computation of the  $V_{stat}$  covariance matrix of the LPMT+SPMT double spectrum (see Section 7.5).

### 7.3.2 AveNu<sub>e</sub> Fitting Package

Its role is to perform fits to a single data samples, of to a set of toy samples. In practice :

- It loads TTrees containing the data to fit as well as the  $T_{no\ osc}(\eta)$  predictions, and create local objects representing the data spectrum and the pdf. For that purpose,  $T_{no\ osc}(\eta)$  are changed into predictions  $T(\theta, \eta)$  for the oscillated spectrum by weighting events in the TTree according to the oscillation probability.
- It loads the necessary covariance matrices.
- It creates from this a  $\chi^2$  object. The Pearson, Neyman, CNP and Pearson V versions are available
- It is interfaced with Minuit via RooFit classes to perform the minimisation. At each step,  $T(\theta, \eta)$  are re-weighted by the oscillation probability corresponding to the current value of the floating oscillation parameters.

Three kinds of data can be fitted with this Package : real data, Asimov simulated data and toy data.

When real data will be available at JUNO, we expect that the result of the IBD selection will be made available by the collaborations via TTrees.

3268 The principles of Asimov fits were described in Section 2.7.3. In practice, our Fit Package fill the  
 3269 local object representing the data spectrum with  $T(\theta, \eta)$ , assuming some values for the oscillation  
 3270 parameters.

3271 The toy data samples can have two origins. Some are produced by the IBD generator macro of the  
 3272 AveNue Generators. This is the case of the toy samples that we produce with QNL effects. It is  
 3273 also possible to generate toys directly with the Fitting Package. In that case, toy data spectra are  
 3274 produced by generating random fluctuations around each the values of  $T(\theta, \eta)$ . These fluctuations  
 3275 must be the reflect of both statistical and systematic uncertainties. Fluctuations between bins  $i$  and  $j$   
 3276 can be correlated. Such correlations are common in the case of systematic uncertainties. In general,  
 3277 they are 0 for the statistical uncertainties. In our case, as already explained earlier (see for instance  
 3278 the Sections where the test statistics are described), bins from the SPMT part of the LPMT+SPMT  
 3279 spectrum are correlated to bins of the LPMT part even for the statistical part.

3280 To generate correlated fluctuation we use, through Choleski decomposition, the covariance matrices.  
 3281 This way to generate toy is faster. We use it in this work in the reference case (no QNL). In the case  
 3282 where QNL effects are simulated, the corresponding statistical covariance matrix is not known, we  
 3283 therefore resort to the IBD generator.

### 3284 7.3.3 Details of the IBD generator

3285 The IBD generator is a standalone generator used to produce oscillated and non oscillated spectra as  
 3286 the one seen by the JUNO experiment. It is at the core of the fitting framework as it's used to generate  
 3287  $T(\theta, \eta)$ , the toy data and spectra to compute the covariances matrix.

3288 With thus have a flexible macro with options allow to enable or disable effects such as non-uniformity  
 3289 and non-linearity. It take as an argument the number of events to generate  $N_{evt}$ . Optionally, we  
 3290 generate an effective number of events  $N$  by drawing in a Poisson distribution of mean  $N_{evt}$ .

3291 Then for each event we:

- 3292 1. Choose randomly, following the reactor power fraction, the source reactor of the neutrino.
- 3293 2. Generate a random interaction position in the detector following a uniform distribution over  
 3294 the detector volume.
- 3295 3. Draw a random neutrino energy  $E_\nu$  from the expected neutrino emission spectrum of every  
 3296 reactor. This spectrum is computed by:
  - 3297 (a) Computing the power spectrum of each isotopes  $^{235}\text{U}$ ,  $^{238}\text{U}$ ,  $^{239}\text{Pu}$ ,  $^{241}\text{Pu}$  using the Huber-  
 3298 Mueller model [11, 14].
  - 3299 (b) Summing the contribution of each isotopes following the respective fission fraction [0.58,  
 3300 0.07, 0.30, 0.05] as reported in [91].
  - 3301 (c) The power of each reactor is then adjusted by their distances from the detector, the detector  
 3302 efficiency and their mean duty cycle (11 of 12 month).
  - 3303 (d) The total spectrum is then finally adjusted by taking into account the correction of the Day  
 3304 Bay bump [17], adjustment due to spent nuclear fuel and due to the non-equilibrium.
- 3305 4. (Optional) Compute the survival probability due to oscillation at nominal oscillation parameters  
 3306 value. If the neutrino does not survive, the event is rejected and the algorithm restart from step  
 3307 (1).
- 3308 5. Compute the emitted positron energy  $E_{pos}$  from the mass difference. If the neutrino does not  
 3309 have enough energy reject the event and start from step (1).
- 3310 6. Compute the deposited energy  $E_{dep}$  by incrementing  $E_{pos}$  by 511 keV to account for the positron  
 3311 annihilation. We do not consider cases where some of the energy leak outside of the detector  
 3312 (positron or annihilation gammas escaping the CD).

- 3313 7. Correct the deposited energy with the expected event-wise non-linearity from [32] to obtain the  
 3314 visible energy  $E_{vis}$ .
- 3315 8. (Optional) Add a custom non-linearity as described in Section 7.1.2. This non linearity is char-  
 3316 acterized by  $\alpha_{qnl}$  to obtain  $E_\alpha$ .
- 3317 9. Finally, using the expected resolution of the LPMT and SPMT systems, provided in the JUNO  
 3318 common inputs, we draw from a gaussian characterized by those resolution the reconstructed  
 3319 energy  $E_{rec}$  or  $E_{lpmt}$  and  $E_{spmt}$  for each systems. The resolutions are provided as ABC parame-  
 3320 ters using

$$\frac{\sigma E_{vis}}{E_{vis}} = \sqrt{\left(\frac{A}{\sqrt{E_{vis}}}\right)^2 + B^2 + \left(\frac{C}{E_{vis}}\right)^2} \quad (7.19)$$

3321 where A is the term driven by the Poisson statistics of the total number of detected photoelec-  
 3322 trons, C is dominated by the PMT dark noise, and B is dominated by the detector's spatial  
 3323 non-uniformity. The relative and absolute resolutions of the LPMT and SPMT systems are  
 3324 illustrated in Figure 7.10.

3325 The events are stored as n-tuples and are not yet binned at the end of the generator.

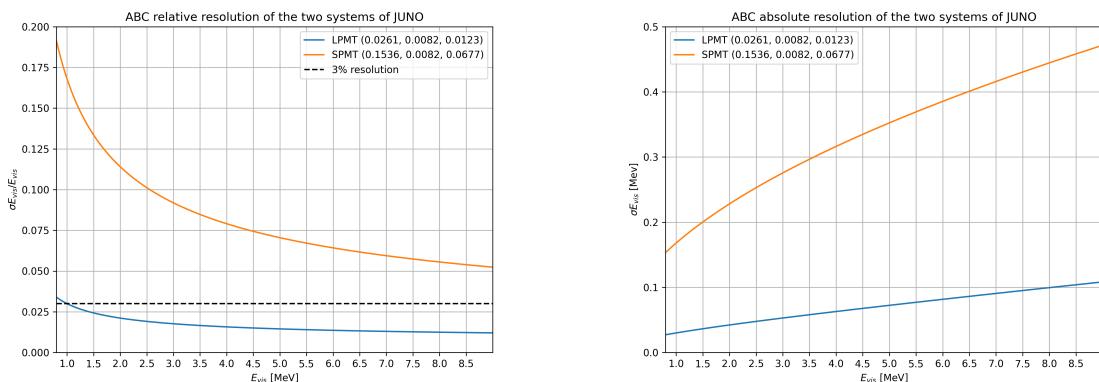


FIGURE 7.10 – Relative (On the left) and absolute (On the right) resolutions of the LPMT and SPMT systems used in this study. The number in parenthesis are the parameter A, B and C respectively for each systems.

## 3326 7.4 Technical challenges and development

3327 The fit framework Avenue was already partially developed with multispectra fitting in mind but  
 3328 a lot technical development was necessary to allow for a joint fit. This required a lot of work and  
 3329 constitute a good part of my total effort on this study. I remind that these development will be useful  
 3330 beyond this thesis and this subject. As already mentioned earlier, at some point, we should perform  
 3331 simultaneous fits of the JUNO and TAO spectra. It's also a potential starting point for combined  
 3332 analyses with other experiments, like long baseline experiments.

3333 The first step was to migrate the framework from ROOT5 (last release in March 2018) to ROOT6  
 3334 (v6.26.06 released in July 2022) to ensure compatibility with the data coming from the JUNO collab-  
 3335 oration, and benefiting of the improvement and corrections that came with ROOT6. This allow us to  
 3336 upgrade the C++ standard from C++11 to C++17. A substantial effort has been done to modernize  
 3337 the code, generalizing the functions and methods via templating to help readability and using smart  
 3338 pointer to prevent possible memory leaks.

3339 The Avenue framework had to be adapted, notably on the chi-square calculation and spectrum gen-  
 3340 eration to correctly take into account the correlation between the SPMT and LPMT spectra. The delta  
 3341 joint fit requiring two more parameters over a spectrum twice as large as before with LPMT takes  
 3342 much more time, around 15h for 6 years exposure, than the single LPMT fit. Thus the framework  
 3343 and the fit macro had to be updated for distributed computing. Notably the aggregation of fit results  
 3344 can now be done in a single file instead of managing a file per fit. In case of numerous toy, the hard  
 3345 drive access time could lead to long analysis time.

3346 While the IBD generator was already able to generate LPMT and SPMT spectrum, it was not designed  
 3347 for generating correlated spectrum. As detailed in Section 7.3.3, up to the reconstruction effect, the  
 3348 two spectrum need to share the same generation else the two spectrum would be decorrelated and it  
 3349 would be like we would run two different experiment.

## 3350 7.5 Covariance matrix

3351 The covariance matrix between the LPMT and SPMT spectra is at the heart of this study as it was  
 3352 already mentioned in Section 7.2. In this section we discuss the different approaches taken to estimate  
 3353 it. We remind that in this work, we consider only statistical effects and let to future works the  
 3354 task to include systematic uncertainties. We thus evaluate in this section the  $(820 \times 820)$  statistical  
 3355 covariance matrix  $V$  of the LMPT+SPMT spectrum.

3356 As already explained in previous Sections 7.2.6 and 7.3.3, we assume, in most of what follows, that  
 3357 the effect of the energy reconstruction resolution is independent between the LPMT system and the  
 3358 SPMT system, although this is an approximation. We therefore also briefly study the correlations  
 3359 between the two reconstructions.

### 3360 7.5.1 Analytical method

3361 The first method discussed is the analytical method where we propagate the resolution of the LPMT  
 3362 and SPMT spectra over a non-smeared spectrum. Following the approach used in the IBD generation  
 3363 in Section 7.3.3, we consider the system resolution  $\sigma(E)$  to be only dependent in energy. We do not  
 3364 consider the position of the event.

3365 Using the formalism of section 39.2.5 *Propagation of errors* of PDG2020 [40] and considering an ex-  
 3366 tended spectrum of 820 bins following the binning scheme introduced in 2.7.2, the first 410 for the  
 3367 LPMT and the last 410 for the SPMT, we consider

- 3368 —  $\mathbf{h} = (h_0, \dots, h_n)$  Is the n-dimensional vector ( $n=820$ ) containing the number of entries in each  
 3369 bin of the LPMT+SPMT true  $E^{vis}$  spectrum.
- 3370 —  $\zeta(\mathbf{h}) = (\zeta_0(\mathbf{h}), \dots, \zeta_n(\mathbf{h}))$  is the n dimensional vector containing the reconstructed  $E^{vis}$  LPMT+SPMT  
 3371 spectrum.

3372 Since, like in most sensitivities studies, resolution is simulated via a gaussian smearing,  $\zeta$  can be  
 3373 expressed this way :

$$\zeta_i = \sum_{j=0}^n G(j, \sigma(E_j))(i) \cdot h_j \quad (7.20)$$

3374 where  $G(j, \sigma(E_j))(i)$  is the smearing function defined as

$$G(j, \sigma(E_j))(i) = \int_{\lfloor E_i \rfloor}^{\lceil E_i \rceil} \frac{1}{\sigma(E_j)\sqrt{2\pi}} e^{-\frac{(E_j-E)^2}{2\sigma(E_j)^2}} dE \quad (7.21)$$

3375 where  $E_j$  is the mean energy in the bin  $j$  and  $\lfloor E_i \rfloor$  and  $\lceil E_i \rceil$  are the lower and higher energy bound of  
 3376 the  $j$ th bin respectively.

3377 According to 7.21, to evaluate  $V$ , the matrix describing the uncertainties on  $\zeta_i$ 's and the correlations  
 3378 between them, one has to consider uncertainties both on  $h_j$ 's and on  $G(j, \sigma(E_j))(i)$ 's. We use linear  
 3379 error propagation and split this problem in two steps :  $V = V_{inputs} + V_{rec}$ . The first matrix accounts  
 3380 for the uncertainties on the inputs from the true  $E^{vis}$  spectrum ( $h_i$ 's), while the second concerns the  
 3381 uncertainties due to  $G(j, \sigma(E_j))(i)$ 's.

3382 To evaluate  $V_{inputs}$ , we use  $V_{inputs} = AUA^T$  where  $U$  is the covariance matrix of the LPMT+SPMT  
 3383 true  $E^{vis}$  spectrum. Since before reconstruction the LPMT and SPMT spectra are the same, this  
 3384 LPMT+SPMT is the juxtaposition of two 410-bin identical spectra. Moreover we are interested only  
 3385 in statistical uncertainties. Therefore,  $U$  has the form :

$$U = \begin{cases} \sqrt{h_i h_j} & \text{if } i = j \text{ or } |i - j| = 410 \\ 0 & \text{otherwise} \end{cases} \quad (7.22)$$

3386 The condition  $|i - j| = 410$  express the fact that one  $h_i$  of the LPMT part of the spectrum is naturally  
 3387 100% correlated with the corresponding bin in the SPMT spectrum.

3388 We can then construct the transfer matrix  $A$  as

$$A_{ij} = \frac{\partial \zeta_i}{\partial h_j} = G(j, \sigma(E_j))(i) \quad (7.23)$$

3389 and then compute the first part of our covariance matrix

$$V_{inputs} = AUA^T \quad (7.24)$$

3390 Now we need to consider the uncertainty on the steaming from the resolution, ie to evaluate  $V_{rec}$ . It  
 3391 can be done considering no uncertainty on the true  $E^{vis}$  spectrum. The quantity  $G(j, u) \equiv G(j, \sigma(E_j))(i)$   
 3392 is the predicted probability for an event initially in bin  $j$  of the true  $E^{vis}$  spectrum to be reconstructed  
 3393 in bin  $i$ . In practice, the migration between these bins is a random process. Reconstructed many  
 3394 times the same event would not lead each time the same migrations. We need here to determine this  
 3395 variability. We consider that with 410 bins, migrations vary independently whatever  $i$  and  $j$ .

3396 This allows to consider  $V_{rec}$  as diagonal, thus we only need  $\sigma G(j, i)$ . We can derive this term from  
 3397 two equation:

- 3398 — The term  $G(j, i) \cdot h_j$  represent the number of event smeared from the bin  $j$  that end up in the  
 3399 bin  $i$ . This is a number, we thus assume poissonian statistic so that  $\sigma[G(j, i) \cdot h_j] = \sqrt{G(j, i) \cdot h_j}$ .
- 3400 — Using basic error propagation we can say that  $\sigma^2[G(j, i) \cdot h_j] = h_j^2 \sigma^2 G(j, i) + G(j, i)^2 \sigma^2 h_j$ .

Equating the above equations, and remembering that  $\sigma h_j = \sqrt{h_j}$  since  $h_j$  is also a number of events :

$$G(j, i)h_j = \sigma^2[G(j, i)h_j] = h_j^2 \sigma^2 G(j, i) + G(j, i)^2 h_j \quad (7.25)$$

$$\Rightarrow \sigma^2 G(j, i) = \frac{G(j, i)h_j - G(j, i)^2 h_j}{h_j^2} \quad (7.26)$$

$$= \frac{(1 - G(j, i))G(j, i)}{h_j} \quad (7.27)$$

3401 By summing the two covariance matrix  $V_{inputs}$  and  $V_{rec}$ , we can extract a correlation matrix presented  
 3402 in Figure 7.11. Typically, a bin in the SPMT part of the reconstructed spectrum is correlated up to a  
 3403 few percents to the corresponding bin in the LPMT spectrum and its neighbour. This might seem  
 3404 a small correlation. However, its concerns all bins. The global impact is therefore high. As an  
 3405 illustration, as seen in Section 7.2.2, the correlation between the value of  $\sin^2(2\theta_{12})$  measured with  
 3406 the LPMT spectrum and that measured with the SPMT spectrum are correlated at more than 95%.

3407 The correlation between the SPMT and LPMT spectra is greater at the start of the spectrum. This  
 3408 is expected since the absolute resolution is smaller in this region. For instance, at 1.5 MeV, the  
 3409 reconstruction by the SPMT re-distribute events with a sigma of more than 0.20 MeV. At 6 MeV, this  
 3410 is about twice more. Since the resolution reduces the initial correlations (true Evis spectra are share  
 3411 by both LPMT and SPMT, correlations are 100%), we therefore expect higher remaining correlations  
 3412 where the absolute resolution is smaller.

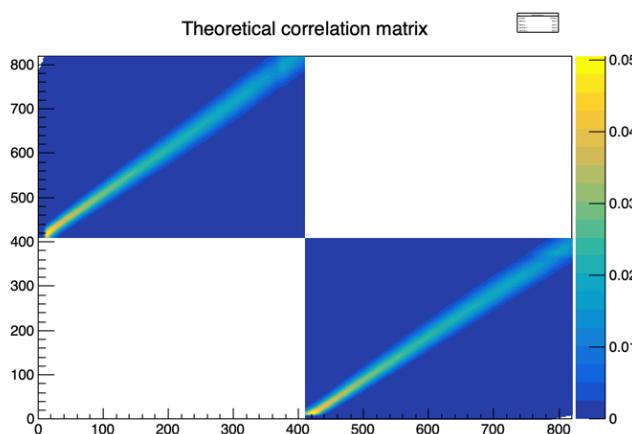


FIGURE 7.11 – Theoretical correlation matrix between the LPMT spectrum (bins 0-409) and the SPMT spectrum (410-819). The diagonal has been set to 0 (it was 1) for readability purpose.

### 3413 7.5.2 Empirical method

3414 The second method is the empirical way where we generate toys and just compute the empirical  
 3415 correlation between the bin contents.

$$\text{Corr}(h_i, h_j) = \frac{\mathbb{E}[h_i h_j] - \mathbb{E}[h_i] \mathbb{E}[h_j]}{\sigma_{h_i} \sigma_{h_j}} \quad (7.28)$$

3416 We thus generate  $10^7$  event using the IBD generator presented in Section 7.3.3, then produce spectra  
 3417 from this finite set of events, meaning we must choose a number  $N$  of toy each composed of  $M$  event  
 3418 in order to have the best estimate.

3419 It can be shown that empirical correlations are more precise when one maximises the number of  
 3420 samples, even at the price to have few events per sample. This effect is illustrated in Figure 7.12.

3421 The relative difference between the element of the theoretical matrix of Figure 7.11 and the empiric  
 3422 correlation matrix in Figure 7.12c is presented in Figure 7.13. Typically, correlations coefficient differ  
 3423 by 20% of their value. We have verified that differences larger than this are confined in the very low  
 3424 or high end of the energy spectrum, which carry no sensitivity to the solar oscillation parameters we

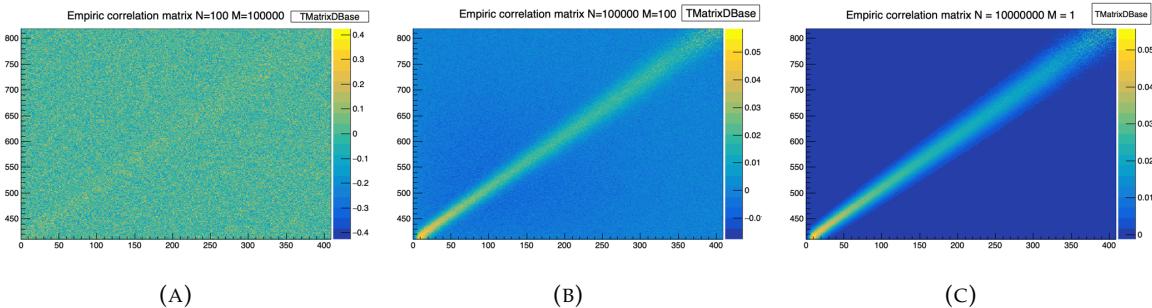


FIGURE 7.12 – Upper left corner of the estimated correlation matrix between the LPMT and SPMT spectrum for different configuration of  $N$  toy with different number of  $M$  events per toy. We observe that the statistical uncertainty, the noise effect, diminish with the number of toy considered.

3425 aim at. Therefore, for the statistical tests presented in this chapter we assume the correlations present  
3426 in the theoretical version of  $V$ . This should account for the effect of correlations well enough.

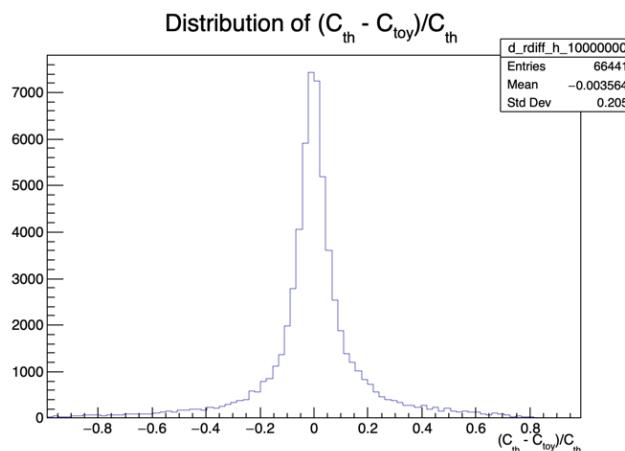


FIGURE 7.13 – Relative difference between the element of the theoretical and empiric correlation matrix

We chose to do so for practical reasons. Indeed, for the  $\chi^2$  computation and the Choleski decomposition (see Section 7.3.2) the matrix must be invertible and positive definite. The statistical uncertainty on the coefficient of the empiric matrix can prevent that, leading to complications.

## 3430 7.6 Technical Validation

3431 Standard Independent Joint Fit

We have already explained in Sections 7.2 and 7.5 that a correlation exist between the SPMT and LPMT spectra, and is accounted for in the LPMT+SPMT joint fit by the  $V$  covariance matrix, which determination is described in the previous section.

3435 We can, however, perform a test where we ignore these correlations, setting to 0 all off-diagonal  
 3436 elements of  $V$ . In this case, we implicitly assume that our data contains more information, and  
 3437 therefore expect the uncertainties on  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$  to be smaller than those obtained with

<sup>3438</sup> individual fits to the LPMT and SPMT spectra. Assuming a gaussian behavior of the number of  
<sup>3439</sup> entries per bin, these uncertainties should be close to the weighted average of the uncertainties with  
<sup>3440</sup> the individual fits :

$$\frac{1}{\sigma_{Weighted}^2} = \frac{1}{\sigma_{LPMT}^2} + \frac{1}{\sigma_{SPMT}^2} \quad (7.29)$$

<sup>3441</sup> These tests are performed using an Asimov sample. Indeed, if it was done via a toy study, then  
<sup>3442</sup> generating correlated toy spectra and fitting them assuming a diagonal  $V$  matrix would have led to  
<sup>3443</sup> biases, regardless of the quality of the technical implementation. Asimov spectra, on the other hand,  
<sup>3444</sup> are generated with no fluctuations. They are supposed to return fitted values of  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$   
<sup>3445</sup> exactly equal to the values assumed during the generation. This is, together with the comparison  
<sup>3446</sup> with  $\sigma_{weight}$ , a strong test of the technical implementation.

<sup>3447</sup> Note that we fix here the  $\delta m_{21}^2$  and  $\delta \sin^2(2\theta_{12})$  parameters to 0. Also, we assume 6 years of data  
<sup>3448</sup> taking, and the absence of unexpected instrumental effects (no supplementary QNL). A notable  
<sup>3449</sup> difference with the fit configuration used later in this chapter (and presented in Section 7.2) is that  
<sup>3450</sup> we do not treat  $\Delta m_{31}^2$  as a nuisance parameter. It is free to float.

	$\sigma(\Delta m_{21}^2)$ [eV <sup>2</sup> ]	$\sigma(\delta \Delta m_{21}^2)$ [eV <sup>2</sup> ]	$\sigma(\sin^2(2\theta_{12}))$	$\sigma(\delta \sin^2(2\theta_{12}))$	$\sigma(\Delta m_{31}^2)$ [eV <sup>2</sup> ]	$\chi^2$
LPMT	$1.29 \times 10^{-07}$		$1.33 \times 10^{-03}$		$4.39 \times 10^{-06}$	$3.23 \times 10^{-18}$
SPMT	$1.38 \times 10^{-07}$		$1.38 \times 10^{-03}$			$2.87 \times 10^{-18}$
Indep Standard joint	$9.48 \times 10^{-08}$		$9.86 \times 10^{-04}$		$4.39 \times 10^{-06}$	$6.10 \times 10^{-18}$
Standard joint	$1.29 \times 10^{-07}$		$1.18 \times 10^{-03}$		$4.39 \times 10^{-06}$	$3.38 \times 10^{-18}$
Weighted	$9.46 \times 10^{-08}$		$9.63 \times 10^{-04}$			
Delta joint	$1.35 \times 10^{-07}$	$3.43 \times 10^{-08}$	$1.38 \times 10^{-03}$	$1.46 \times 10^{-04}$	$4.39 \times 10^{-06}$	$3.38 \times 10^{-18}$
Indep Delta joint	$1.38 \times 10^{-07}$	$1.89 \times 10^{-07}$	$1.38 \times 10^{-03}$	$1.87 \times 10^{-03}$	$4.39 \times 10^{-06}$	$6.10 \times 10^{-18}$

TABLE 7.4 – Uncertainties on each parameters reported by Minuit on Asimov studies.  
LPMT and SPMT rows are the results on the individual fit on each spectra. The Weighted row correspond to the weighted average uncertainties between the LPMT and SPMT fits following Eq. 7.29. The Indep Standard joint row is the result of the joint LPMT+SPMT fit but the off-diagonal terms are set to 0. The Indep Standard joint and Standard joint fits both are LPMT+SPMT fit but the parameters  $\delta m_{21}^2$  and  $\delta \sin^2(2\theta_{12})$  are fixed to 0. The Delta joint and Indep Delta joint are LPMT+SPMT fit with  $\delta m_{21}^2$  and  $\delta \sin^2(2\theta_{12})$ , difference being that in the Indep version, the off-diagonal terms of the covariance matrix are set to 0.

<sup>3451</sup> The results are reported in Table 7.4. All those test are ran considering statistics error only, 6 years  
<sup>3452</sup> exposure with all backgrounds,  $\sin^2(2\theta_{13})$  fixed to its nominal value. For the SPMT individual  
<sup>3453</sup> fit  $\Delta m_{31}^2$  is fixed at its nominal value as the SPMT system is not sensitive to this parameter. We  
<sup>3454</sup> use here the simple Pearson  $\chi^2$ . Indeed, as explained above, an Asimov fit is supposed to find  
<sup>3455</sup> exactly the values of the parameters assumed for the generation of the spectrum, which implies a  
<sup>3456</sup> very low Pearson  $\chi^2$  (0 modulo numerical effects). This is also a strong indication that the technical  
<sup>3457</sup> implementation is correct. If we had used the usual Pearson V  $\chi^2$ , the  $\ln |V|$  term would have made  
<sup>3458</sup> the result more difficult to interpret.

<sup>3459</sup> When we performed the Standard Independent Joint Fit, as expected we observed that the fitted  
<sup>3460</sup> values of the parameters all matched the generation values. We can also see in table 7.4 that the  
<sup>3461</sup> uncertainty on  $\Delta m_{21}^2$  evaluated by the fit are equals the corresponding  $\sigma_{weight}$  up to 0.2%. In the  
<sup>3462</sup> case of  $\sin^2(2\theta_{12})$ , the agreement is up to 2.5%.

<sup>3463</sup> A slight difference exists in statistic between the SPMT and LPMT spectra. Indeed, due to a larger  
<sup>3464</sup> smearing in energy resolution, events that would be inside the spectrum range [0.8, 7.5] MeV are  
<sup>3465</sup> smeared outside it. The  $\sin^2(2\theta_{12})$  parameter being mainly driven by the amplitude of the spectrum  
<sup>3466</sup> (see illustration 7.1), it is more affected than  $\Delta m_{21}^2$ .

### 3467 Standard Joint Fit

3468 This case is similar to the previous one, with one difference : we now use the version of  $V$  that  
 3469 accounts for the correlations between the SPMT and LPMT spectra. The expected effect of this  
 3470 correlation is that the uncertainties on  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$  should see very little improvement with  
 3471 respect to individual fits.

3472 Moreover, the uncertainty on  $\Delta m_{31}^2$  should be very close to that obtained by the individual fit to  
 3473 the LPMT spectrum since only this one contains information on  $\Delta m_{31}^2$  (thanks to its high energy  
 3474 resolution). This is therefore a rather robust test.

3475 As can be seen in Table 7.4, these expectations are observed in practice.

### 3476 Delta Joint Fit

3477 It is the same fit as above, where we let the  $\delta \Delta m_{21}^2$  and  $\delta \sin^2(2\theta_{12})$  parameters free to float in the fit.  
 3478 A test assumes no correlations (diagonal  $V$ ), the other one assumes the usual  $V$ .

3479 A first test here is that the fitter should find these parameters at 0, since no QNL is introduced in these  
 3480 Asimov spectra. Also, in the correlated case, we expect the uncertainties on  $\delta \Delta m_{21}^2$  and  $\delta \sin^2(2\theta_{12})$   
 3481 to be far smaller than in the independent case. Indeed, when the  $\chi^2$  considers these two spectra are  
 3482 correlated, distorting only the LPMT part of the PDF without changing the SPMT part (remember:  
 3483  $\delta \Delta m_{21}^2$  and  $\delta \sin^2(2\theta_{12})$  appear only in  $T(\theta, \eta)$  for the 410 first bins, see Section 7.2) leads to a quick  
 3484 explosion of this  $\chi^2$  when profiling values of  $\delta \Delta m_{21}^2$  and  $\delta \sin^2(2\theta_{12})$  away from 0.

3485 Results in Table 7.4 are again consistent with these expectations.

### 3486 Toy studies

3487 The same tests as above have been repeated, using a set of 1000 toy samples instead of one Asimov  
 3488 sample. Only cases where we account for the correlations between the SPMT and LPMT spectra are  
 3489 carried out. The generation of the toy samples includes these correlations. We therefore also test that  
 3490 part.

3491 We can see on Figures 7.14 and 7.15 the distribution of the best fit values for all the parameters of  
 3492 interest. The mean values and standard deviations are in all cases consistent with the results of the  
 3493 Asimov tests (Table 7.4). Therefore, when realistic fluctuations are simulated, even with a peculiar  
 3494  $\chi^2$  computed with a complex covariance matrix and correlated data, the fit is stable and unbiased.

3495 These distributions also confirm that the uncertainties on  $\delta \Delta m_{21}^2$  and  $\delta \sin^2(2\theta_{12})$  are an order of  
 3496 magnitude smaller than the uncertainties on  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$ . This is an indication of the power  
 3497 of the test statistics used in this chapter.

### 3498 Conclusion of the technical validation

3499 All the tests carried out in this section are consistent with our expectation. We therefore conclude  
 3500 that the technical implementation of the tools used in this chapter is correct.

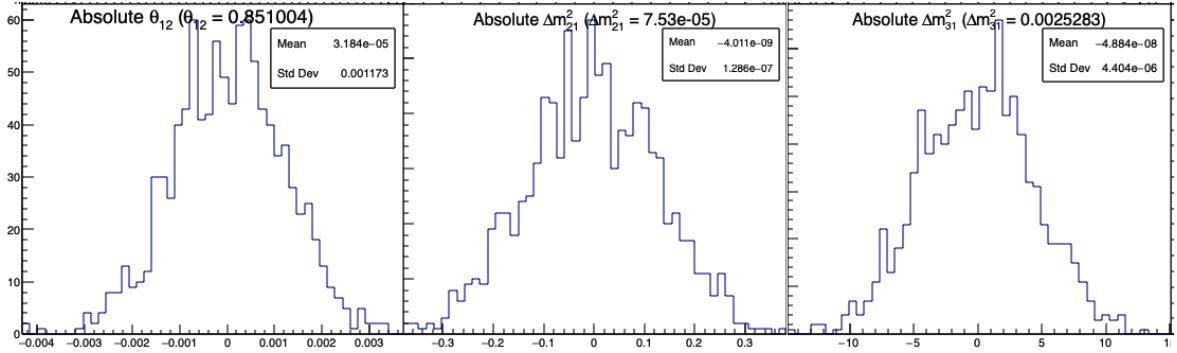


FIGURE 7.14 – Distribution of BFP - nominal value for 1000 toy Standard joint fit. 6 years exposure, all background, PearsonV  $\chi^2$ ,  $\theta_{13}$  fixed. In those plots,  $\theta_{12}$  stands for  $\sin^2(2\theta_{12})$

## 7.7 Results

### 7.7.1 Effect of supplementary QNL on the LPMT spectrum

In this first part of this Section 7.7, we will present the sensitivity of various test statistics to unexpected instrumental effects affecting the SPMT and LPMT differently. The latter effects will be illustrated by generated toy samples affected by the QNL effect.

Most of the tests involve either an individual fit to the LPMT spectrum or the SPMT spectrum, or a joint fit of these two spectra. To better understand why some test statistics turn out to be more powerful than others, we study briefly in the present subsection the results of these fits and interpret the differences.

We generate toy spectra, and fit them according to the default configuration described in Section 7.2. During the generation of the LPMT spectrum, we distort it to simulate a QNL effect, with an intensity of  $\alpha_{qnl} = 1\%$ . For reference, this is about three times the expected residual QNL after the application of dual calorimetric calibration methods ( $\alpha_{qnl} = 0.3\%$  [32]).

Backgrounds had to be ignored here: the JUNO inputs described in Section 7.3 provide a reconstructed spectrum, but not the event per event information about the true  $E_{vis}$ , which we need to apply the QNL effect (See Equation 7.19).

The effect of this QNL on the spectrum is illustrated in Figure 7.16. In Table 7.5 we report the results of the different kinds of fits.

We notice (1st line, first 3 columns) that the individual fit to the LPMT spectrum tends to find, as expected, biased values for  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$  and  $\Delta m_{31}^2$  (biased at about -1 sigma, -1.3 sigma and -2.2 sigmas respectively). When a joint fit is performed, with the  $\delta\Delta m_{21}^2$  and  $\delta\sin^2(2\theta_{12})$  fixed at 0, and ignoring in the computation of the  $\chi^2$  the correlations between the LPMT and SPMT spectra, the biases on  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$  (3rd line, first 3 columns) appear to be averages of the biases seen by the individual fits to these spectra, a logical result since the individual sensitivities to these parameters are similar. The bias on  $\Delta m_{31}^2$  (3rd column) remains the same as with the individual fit to the LPMT spectrum, however, which is expected since the SPMT spectrum carries no sensitivity to  $\Delta m_{31}^2$ .

When the joint fit is performed with the nominal covariance matrix (determined in Section 7.5 assuming no QNL), biases on  $\Delta m_{21}^2$  and  $\sin^2(2\theta_{12})$  explode: they are, respectively, about 6.5 and 2.5 times larger (4th line).

We explain it by the following mechanism : the fit tries to improve the agreement between the PDF and the data in the LPMT part of the spectrum by choosing biased values of the parameters. This

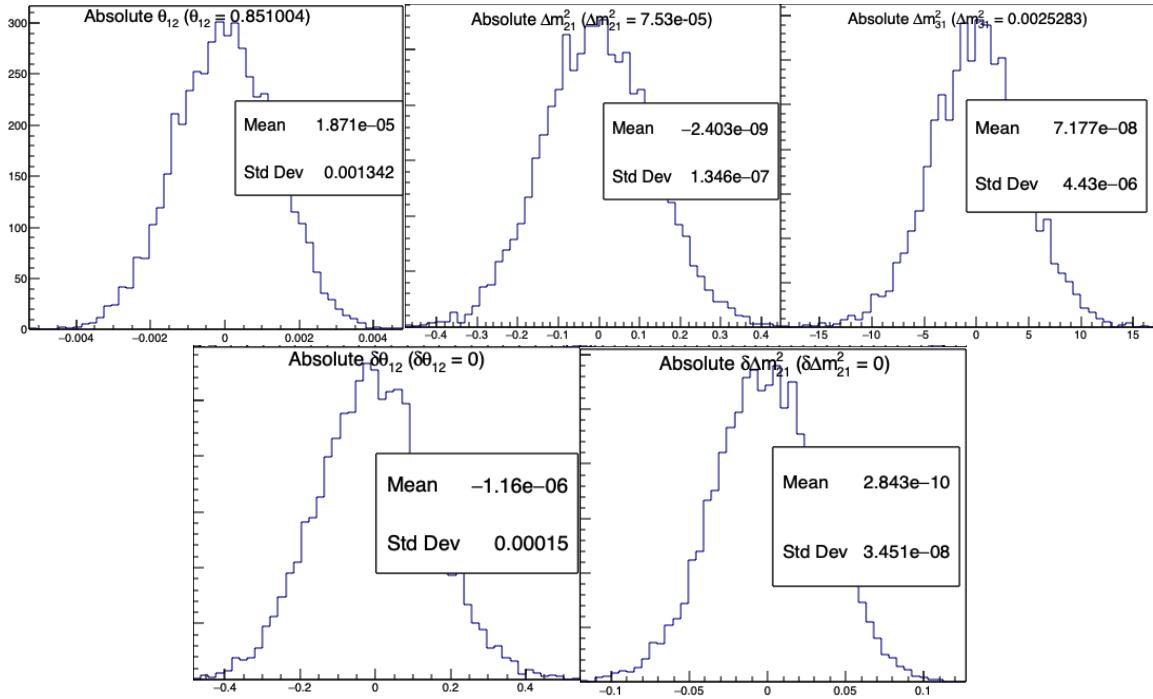


FIGURE 7.15 – Distribution of BFP - nominal value for 5000 toy Delta joint fit. 6 years exposure, all background, PearsonV  $\chi^2$ ,  $\theta_{13}$  fixed. In those plots,  $\theta_{12}$  stands for  $\sin^2(2\theta_{12})$  and  $\delta\theta_{12}$  for  $\delta \sin^2(\theta_{12})$

in turn tends to deteriorate the agreement between the PDF and the SPMT spectrum (not distorted by QNL). In the end, a discrepancy remains between data and PDF in at least one sector (LPMT or SPMT) if not both. When the  $\chi^2$  is built with a matrix which accounts for the correlations, this discrepancy can make the  $\chi^2$  explode.

For instance, in some bins of the LPMT spectrum, we can imagine the PDF overestimates the QNL-distorted data, while the contrary happens in the corresponding bins of the SPMT spectrum. If the expected correlation is positive between these two bins, the  $\chi^2$  will reach values accounting for a larger discrepancy than if no correlation existed and if only the raw agreement between the pdf and the spectra was important.

In reality, the consistency between the two can be judged only accounting for the correlations. This is the important role of the covariance matrix in this work. In other words, the spectra predictions are not only the  $T(\theta, \eta)$ 's, but also the correlations.

Another point must be noted : the correlation matrix  $V$  is evaluated assuming no QNL. With the QNL effect added, the actual correlations between the LPMT and SPMT generated toy spectra is a bit different, adding another source of discrepancy between the data and the predictions, and further increasing the  $\chi^2$ .

All in all, the minimisation of the  $\chi^2$  requires a larger scan of the oscillation parameters values than when correlations are ignored. Values can be chosen which are farther from the nominal ones, meaning larger biases.

This is actually an advantage. Indeed, we can see in table 7.5 that when  $\delta\Delta m^2_{21}$  and  $\delta \sin^2(2\theta_{12})$  are allowed to float in the fit, they "absorb" a large part of the bias. Notice in particular that adding the value of  $\delta\Delta m^2_{21}$  to the remaining bias on  $\Delta m^2_{21}$  (last line, columns 1 and 4) one retrieves the bias of the individual fit to the LPMT spectrum. The same applies to  $\sin^2(2\theta_{12})$ . Consequently, large values of  $\delta\Delta m^2_{21}$  and  $\delta \sin^2(2\theta_{12})$  are expected, hence high significances to help us to detect the distortion.

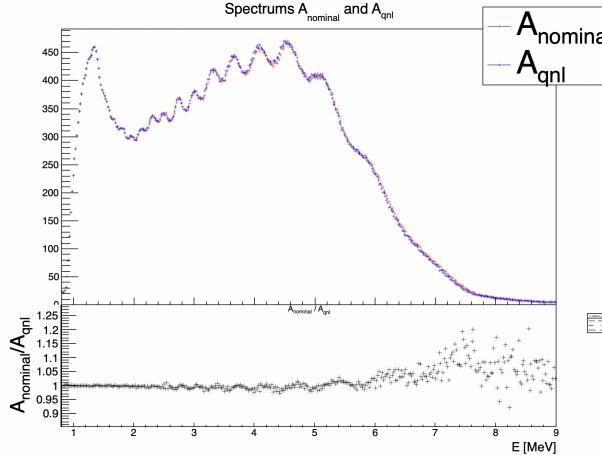


FIGURE 7.16 – **Top:** Theoretical spectrum without QNL (in red) and with  $\alpha_{\text{qnl}} = 1\%$  (in blue). **Bottom:** Ratio between the theoretical spectrum with and without QNL.

Mean (std dev)	$\theta_{12} [10^{-3}]$	$\Delta m_{21}^2 [10^{-7}\text{eV}^2]$	$\Delta m_{31}^2 [10^{-6}\text{eV}^2]$	$\delta\theta_{12} [10^{-3}]$	$\delta\Delta m_{21}^2 [10^{-7}\text{eV}^2]$
LPMT	-1.569 (1.171)	-0.957 (0.989)	-8.235 (3.898)	Irrelevant	Irrelevant
SPMT	-0.164 (1.191)	-0.603 (1.054)	Not sensitive	Irrelevant	Irrelevant
Indep Standard	-0.880 (1.174)	-0.786 (1.004)	-8.195 (3.900)	Irrelevant	Irrelevant
Standard	-8.106 (1.423)	-2.483 (1.018)	-6.649 (4.008)	Irrelevant	Irrelevant
Indep Delta	-0.169 (1.190)	-0.598 (1.054)	-8.234 (3.899)	-1.397 (0.259)	-0.361 (0.366)
Delta	-0.163 (1.183)	-1.532 (1.036)	-8.193 (3.934)	-1.441 (0.193)	0.654 (0.303)

TABLE 7.5 – In each column, the mean of the distribution of the 1000 best fit values found by fitting the 1000 toy samples with  $\alpha_{\text{qnl}} = 1\%$  is shown, from which we subtracted the value assumed when generating the toys. A value different from 0 indicates a bias. Between bracket, the average uncertainty of the fitted value is also shown. It allows to judge of the severity of the bias. For instance, the measurement of  $\sin^2(2\theta_{12})$  by fitting only the LPMT spectrum tends to be biased at the  $-1.569/1.171 = -1.34$  sigma.

In this case (last line, column 4 and 5), we see the most probable values of the fitted  $\delta\Delta m_{21}^2$  and  $\delta \sin^2(2\theta_{12})$  parameters differ from zero at about 7.46 sigma and 2.2 sigma.

Based on the above observations, we expect the " $\chi^2_{H_0} - \chi^2_{H_1}$ " and "Distributions of  $\delta\Delta m_{21}^2$  and  $\delta \sin^2(2\theta_{12})$ " test statistics described in sections 7.2.4 and 7.2.5 to have the highest power. The "Direct comparison between the SPMT and LPMT spectra" should perform in the same ballpark. Finally, the "Comparison of individual fits" is expected to be have less power.

### 7.7.2 Comparison and statistical tests results

I present in this following Subsection the results from the tests and comparison detailed in section 7.2. For each distribution we compute the median p-value with respect to the distribution  $\mathcal{D}(\alpha_{\text{qnl}} = 0\%)$ . For this, we compute the median value of the distribution of interest  $\mathcal{D}(\alpha_{\text{qnl}})$ , then compute the p value

$$p = \frac{N(\mathcal{D}(0) > \text{Median}[\mathcal{D}(\alpha_{\text{qnl}})])}{N_{\text{tot}}} \quad (7.30)$$

where  $N(\mathcal{D}(0) > \text{Median}[\mathcal{D}(\alpha_{\text{qnl}})])$  is the number of toy in the distribution  $\mathcal{D}(\alpha_{\text{qnl}} = 0\%)$  that have a greater value than the median of the  $\mathcal{D}(\alpha_{\text{qnl}})$ . The p-value represent the probability for a non perturbed event to do worse that the median perturbed event.

3570 The uncertainty on the p-value is computed using

$$\sigma p = \sqrt{\frac{p(1-p)}{N}} \quad (7.31)$$

3571 which do not account for all uncertainties but serves as indicator.

### 3572 Comparison of solar parameters from individual analysis: $\chi^2_{ind}$

3573 The results are presented in Figure 7.17. We see that the p-value are much less significant than the  
 3574 other tests, this is because this test possess much less information about the relation between the  
 3575 LPMT and SPMT systems.

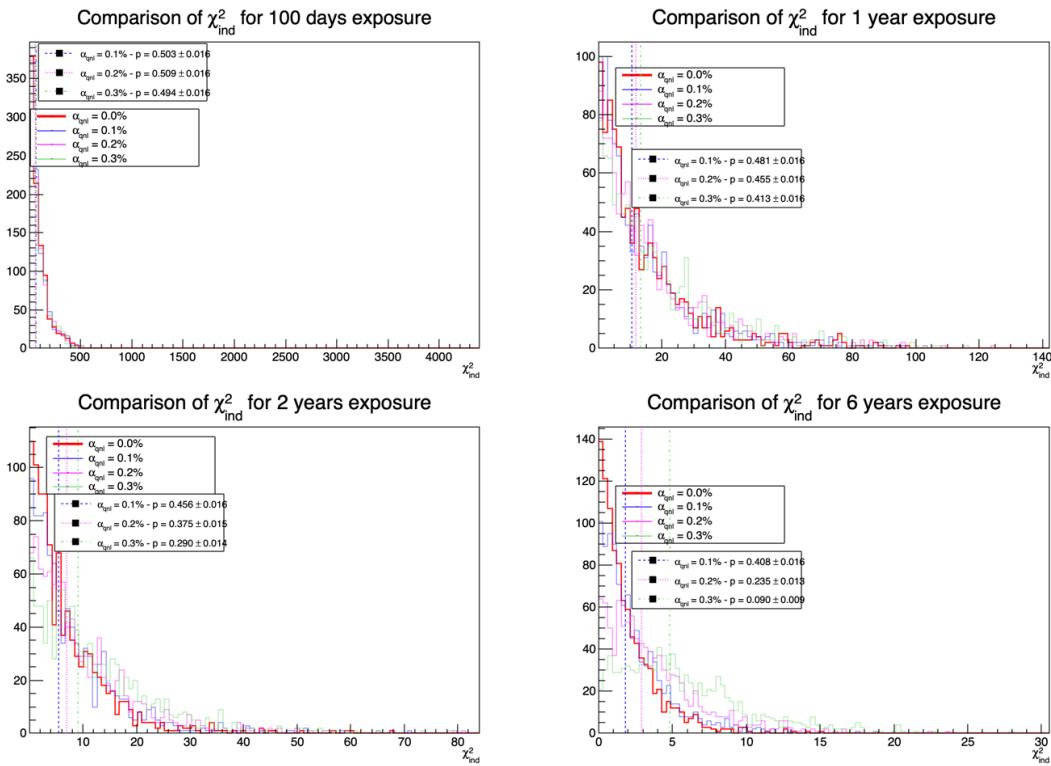


FIGURE 7.17 – Distribution of the  $\chi^2_{ind}$  for 1000 toys for different exposures. The dashed lines represent the median of the distributions and the p-value are the percentage of the  $\alpha_{qnl} = 0$  distribution that are greater than those medians.

3576 This test is the most straightforward as it require only the fit of the two spectra and the estimation of  
 3577 the parameters covariances, but is also the less powerful with a p value for  $\alpha_{qnl} = 0.3\%$  of  $0.09 \pm 0.009$   
 3578 at 6 years.

### 3579 Direct comparison between the LPMT and SPMT spectra: $\chi^2_{spe}$

3580 The results for different exposures can be found in Figure 7.18. To give an idea of the significance of  
 3581 this test, we provide the median p-value for each test  $\alpha_{qnl} \neq 0$ . As expected, the power of this test  
 3582 rises as the exposure does. We see significant discrimination at 6 years for  $\alpha_{qnl} \geq 0.3\%$  where the  
 3583 p-value for  $\alpha_{qnl} = 0.3\%$  is  $0.005 \pm 0.0022$ .

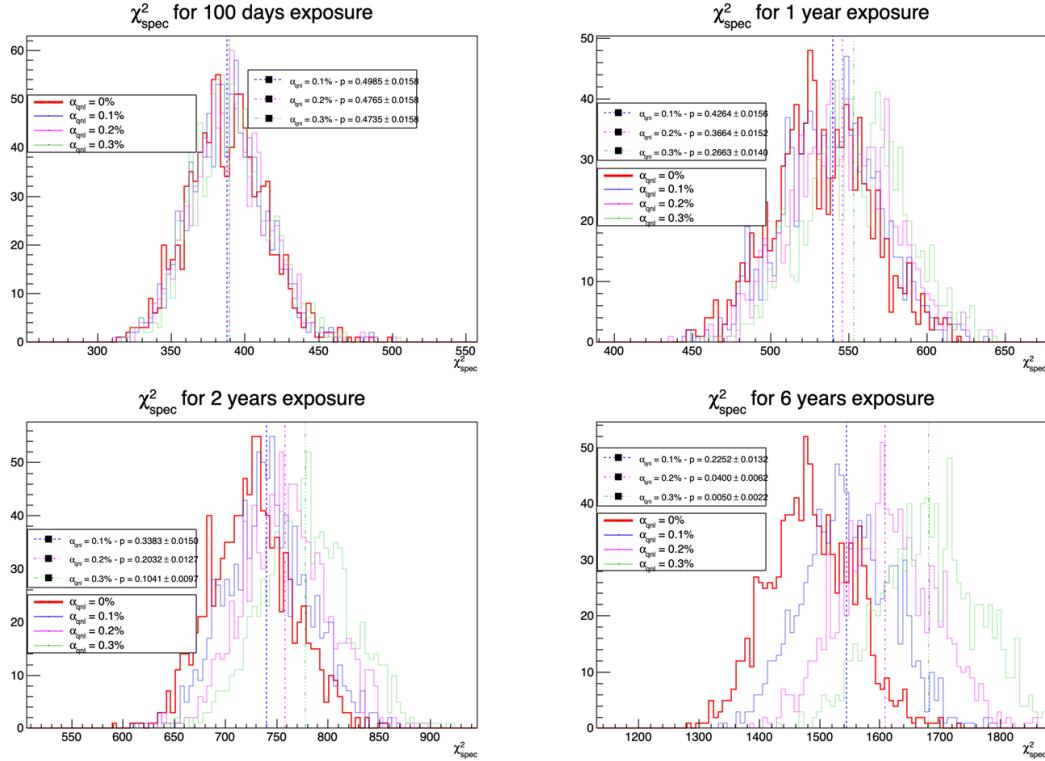


FIGURE 7.18 – Distribution of the  $\chi^2_{\text{spe}}$  for 1000 toys for different exposure. The dashed line represent the median of the distribution and the p-value are the percentage of the  $\alpha_{qnl} = 0$  distribution that are greater than those medians.

This test relies solely on the estimated covariance matrix between the two spectra, requiring no fitting. As a result, it is a very lightweight test that can still provide valuable indications of potential unknown distortions between the two spectra.

**Joint fit:**  $\chi^2_{H_0} - \chi^2_{H_1}$

This test is the most complex, requiring two fit and the covariance matrix between the LPMT and SPMT spectra. The results are presented in Figure 7.19.

The results are good, close to the  $\chi^2_{\text{spe}}$ , one with a p-value at 6 years for  $\alpha_{qnl} = 0.3\%$  of  $0.01 \pm 0.003$ . This sensitivity is consistent with that of  $\chi^2_{\text{spe}}$ .

### Comparison of the parameters $\delta \sin^2(2\theta_{12})$ and $\delta \Delta m_{21}^2$

We can see that the  $\delta \Delta m_{21}^2$  has a very small discriminative power (Figure 7.21) even at 6 years exposure with a p-value of  $0.34 \pm 0.01$  for  $\alpha_{qnl} = 0.3\%$ . On the other hand  $\delta \theta_{12}$  (Figure 7.20) has much more discriminative power with a p-value for  $\alpha_{qnl} = 0.3\%$  of  $0.025 \pm 0.005$ . This test with a single joint fit seems to be still less powerful than the  $\chi^2_{\text{spe}}$ . This can be explained as this method only get information through the oscillation parameters  $\theta_{12}$  and  $\Delta m_{21}^2$  missing potential informations contained in  $\Delta m_{31}^2$ .

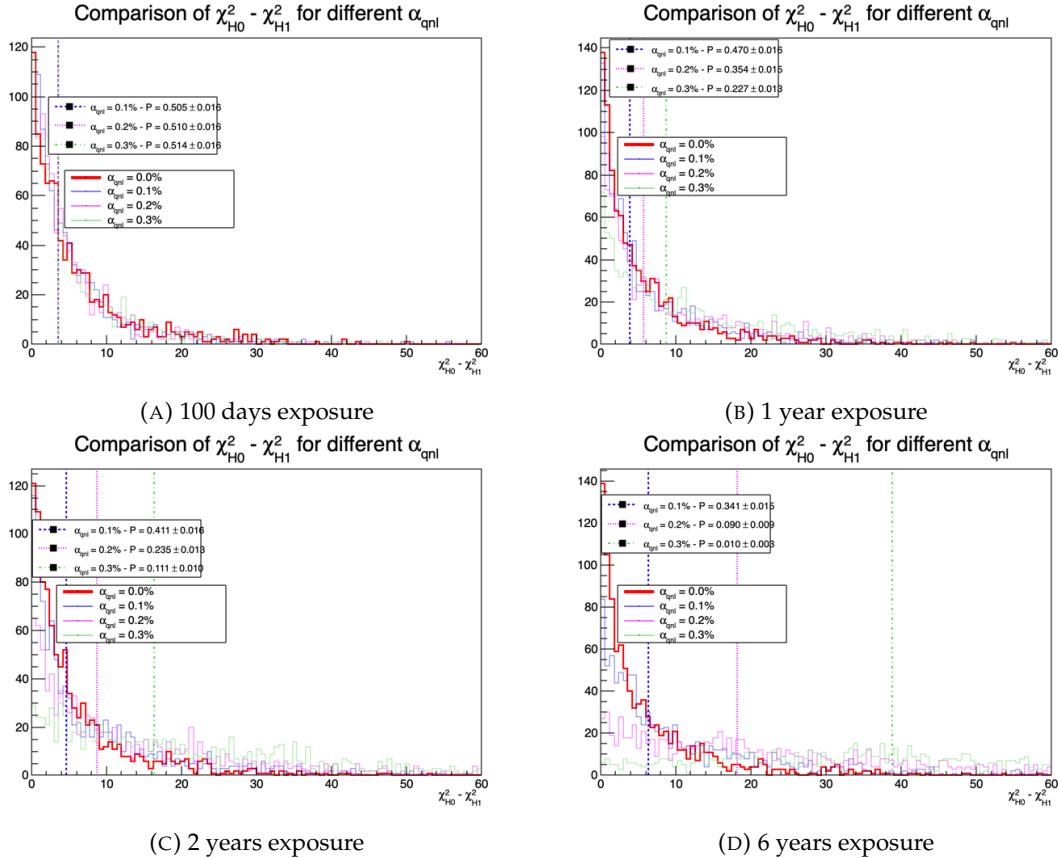


FIGURE 7.19 – Distribution of  $\chi^2_{H_0} - \chi^2_{H_1}$  for 1000 toys for different exposure. The dashed line represent the median of the distribution and the p-value are the percentage of the  $\alpha_{qnl} = 0$  distribution that are greater than those medians.

### 3599 Summary

The p-values from the different test and comparison for  $\alpha_{qnl} = 0.3\%$  are reported in Table 7.6.

	100 days	1 year	2 years	6 years
$\chi^2_{fid}$	0.49	0.41	0.29	0.090
$\chi^2_{spec}$	0.47	0.27	<b>0.10</b>	<b>0.005</b>
$\chi^2_{H_0} - \chi^2_{H_1}$	0.51	0.23	<b>0.11</b>	<b>0.010</b>
Comparison of $\delta \sin^2(2\theta_{12})$	0.39	0.2	0.14	0.025

TABLE 7.6 – Report of the p-value of the different tests and comparisons for  $\alpha_{qnl} = 0.3\%$  for the different exposures.

3600

## 3601 7.8 Conclusion and perspectives

3602 In this chapter, we present the development of a fit framework that allows us to fit multiple spectra  
 3603 simultaneously. We also introduce a set of tools that enable us to detect potential distortions in one of

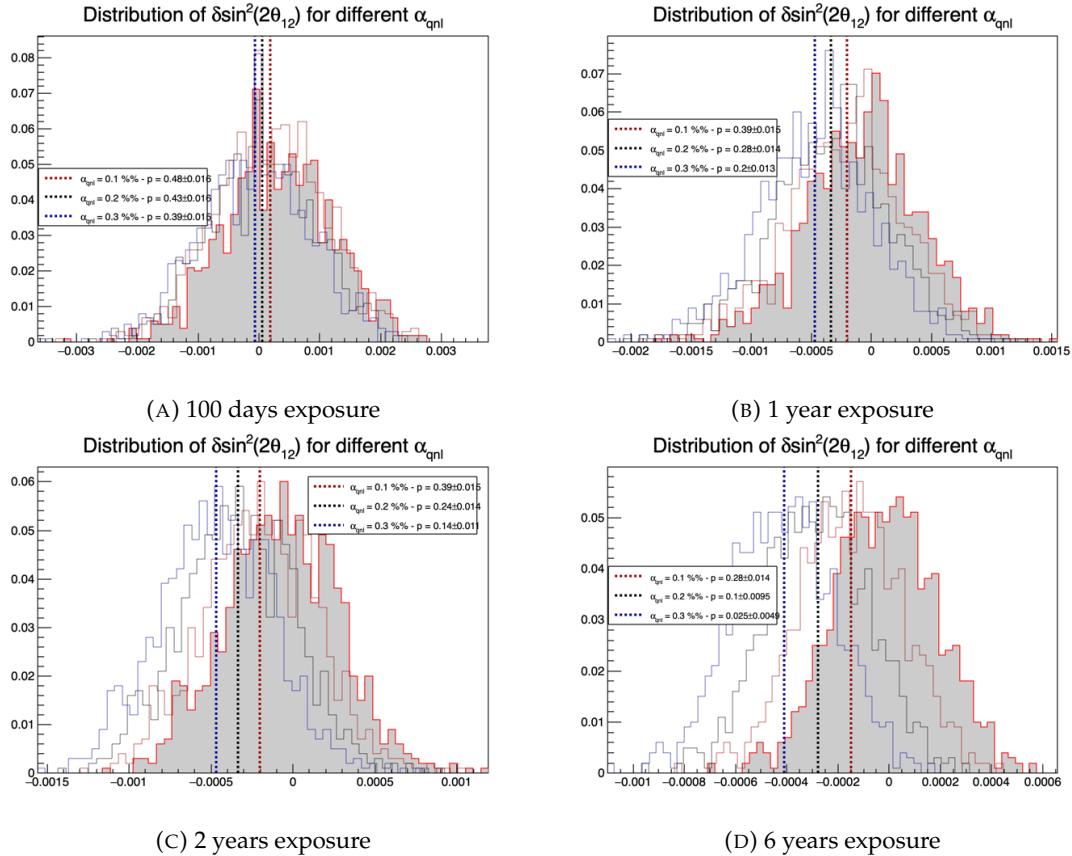


FIGURE 7.20 – Distribution of the  $\delta \sin^2(2\theta_{12})$  for 1000 toys for different exposure. The dashed line represent the median of the distribution and the p-value are the percentage of the  $\alpha_{qnl} = 0$  distribution that are greater than those medians.

the two spectra. As an illustration of the capability of these tools, we use supplementary event-wise non-linearity and compare it to the potential residual event-wise non-linearity after calibration.

Table 7.6 gives a synthetic view of the strength of our methods. As expected, two methods that exploit the knowledge of the correlations between the SPMT and LPMT spectra obtain the best results. At high exposures, if the QNL effects are not calibrated out as well as expected ( $> 0.3\%$ ), our best test statistics will be likely to detect them (median p-values below 10% after 2 years of data taking, about 1% after 6 years). In case of major effect (QNL or another unexpected instrumental effect) is worse, the detection will be even more likely. Below two years of data taking, only large unexpected instrumental effects can be detected.

One of JUNO most important goals is to determine the NMO independently of other experiments. This should not happen before 6 years of data taking. Our results demonstrate that dual calorimetry with neutrino oscillation can be a useful approach to help ensure the robustness of this result.

### 7.8.1 Empirical correlation matrix from fully simulated event

As already explained several times, one of the limitation of this work is that we assume the SPMT and LPMT energy reconstructions to be totally uncorrelated. In reality, this is not true. The  $V$  covariance matrix used in the test statistics should therefore be evaluated accounting for this. This involves complications that make the subject out of the scope of this thesis. We present here a brief study

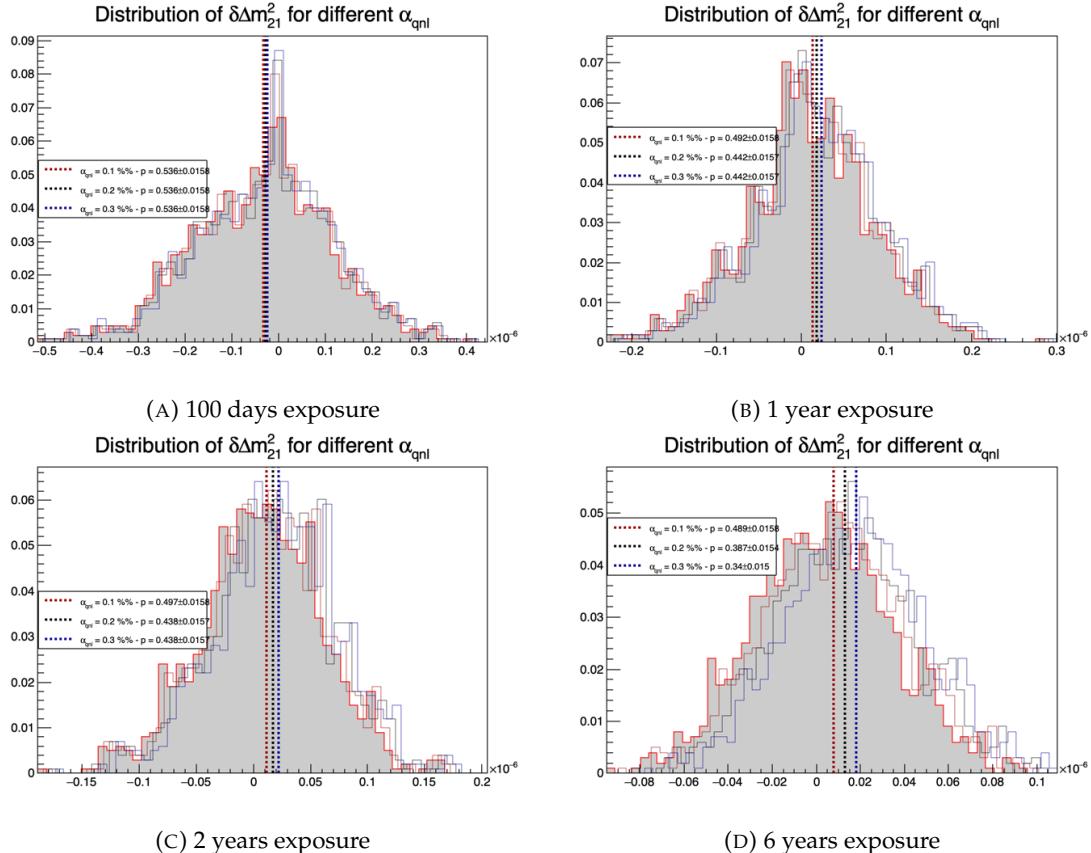


FIGURE 7.21 – Distribution of the  $\delta\Delta m_{21}^2$  for 1000 toys for different exposure. The dashed line represent the median of the distribution and the p-value are the percentage of the  $\alpha_{qnl} = 0$  distribution that are greater than those medians.

which goal is to get a rough idea of the impact of these reconstruction correlations.

The core of the idea is that the LPMT and SPMT reconstruction errors is bound to be correlated due to systematic effects. The first and most obvious one, for example, is energy escaping from the central detector. If the positron, or one of the two annihilation gamma, escape from the detector, less energy is deposited thus both of the systems will reconstruct a lower energy that was actually deposited. On a more subtle scale, the randomness in the production of scintillation photons is common for the two systems, if the liquid scintillator produces fewer scintillation photons for an event, both systems are likely to underestimate the energy.

We study those effects by computing from a dataset of IBD events, uniformly distributed in the CD, the correlation between the reconstruction errors on the energy

$$\text{Corr}(E_{rec}^{lpmt} - E_{vis}, E_{rec}^{spmt} - E_{vis}) \quad (7.32)$$

where  $E_{rec}^{lpmt}$  and  $E_{rec}^{spmt}$  are the reconstructed energies from both systems and  $E_{vis}$  the true visible energy. The OMILREC algorithm, presented in section 3.3, is used for the LPMT reconstruction  $E_{rec}^{lpmt}$ , and the CNN presented in Chapter 4 for the SPMT reconstruction  $E_{rec}^{spmt}$ .

The results of those correlations are presented in Figure 7.22 for the single energy and the interaction radius dependency, and Figure 7.23 for the dual energy and interaction radius dependencies.

The first observation here is that in most of the detector volume, the correlation between the SPMT

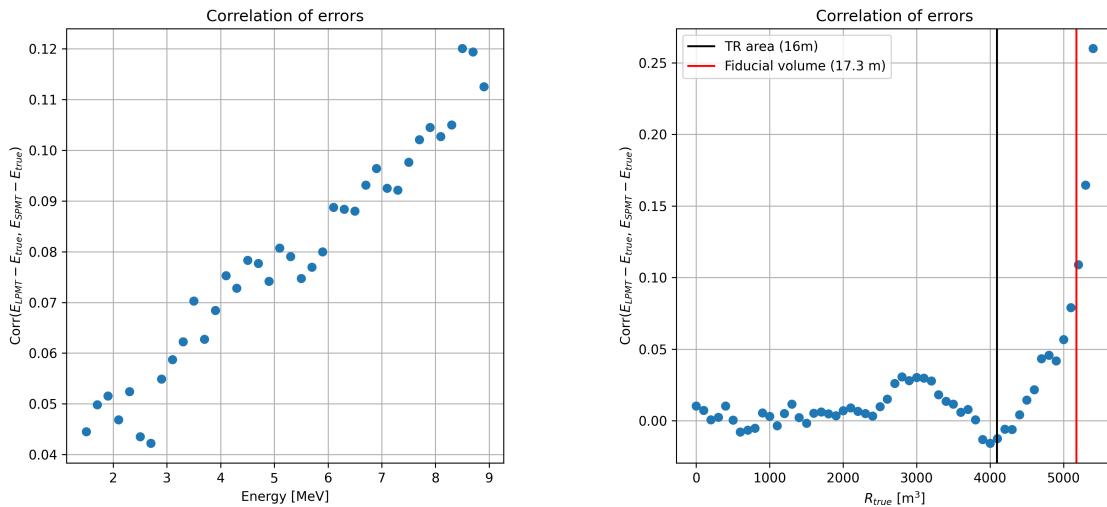


FIGURE 7.22 – Correlation on the reconstruction error between the LPMT and SPMT system as a function of (On the left) the energy, (On the right) the radius. The SPMT reconstruction comes from the NN presented in Chapter 4 and the LPMT reconstruction comes from OMILREC presented in Section 3.3. To prevent effect due to the CNN bad reconstruction, we select the event with  $1 < E_{dep} < 9$  MeV.

and LPMT energy reconstructions does not exceed a few percents, and is in general positive.

In principle, this correlation must be dominated by the fluctuations of the photon yield produced in the scintillator, which dominates the stochastic term of the resolution (see Equation 7.19). Indeed, in a given event, both the LPMT and SPMT reconstruct the energy from the same photon yield and both are affected in the same way by a fluctuation. The correlation is reduced by the fact that SPMT system, due to its low coverage, detect only a very small fraction of the photon. This sampling is also a random phenomenon : the corresponding fluctuations hide to some extent the fluctuations of the original photon yield, and are essentially independent of the random number of photons sampled by the LPMT.

When energy is deposited at high R, close to or in the total reflection area, the proximity of the PMTs increases the number of photons detected by LPMT, and therefore reduce the sampling fluctuations. In this case, the fluctuation of the original photon yield is less shuffled by the sampling fluctuations and the resulting correlation between the LPMT and SPMT reconstruction reaches high values, up to 25% (Fig. 7.22, right).

The original photon yield grows with the visible energy. For the same reason as above, the correlation grows as well, albeit far more slowly than as a function of  $R^3$ . On Fig. 7.23, one can see that cumulating the effects of high energy and high R, correlations can reach 35%. However, in the fiducial volume and at energies below 7 MeV (ie in a part of the spectrum containing the sensitivity to  $\Delta m_{12}^2$  and  $\sin^2(2\theta_{12})$ ), it never exceeds 15%.

To re-evaluate V with these reconstruction correlations accounted for, we should perform an empiric evaluation (like in Section 7.5.2). It would be based on toys generated with the IBD generator (see point 9 of Section 7.3.3), replacing the two independent random gaussian drawings by a drawing according to a 2 dimensional gaussian describing the  $(E_{rec}^{lpmt} - E_{vis})$  vs.  $(E_{rec}^{spmt} - E_{vis})$  distribution, and involving the correlations studied above.

A way must be found to include the variation of the correlation as a function of R and the  $E_{vis}$ . We have tried to define 2-dimensional regions in these variables, and defined each time the correspond-

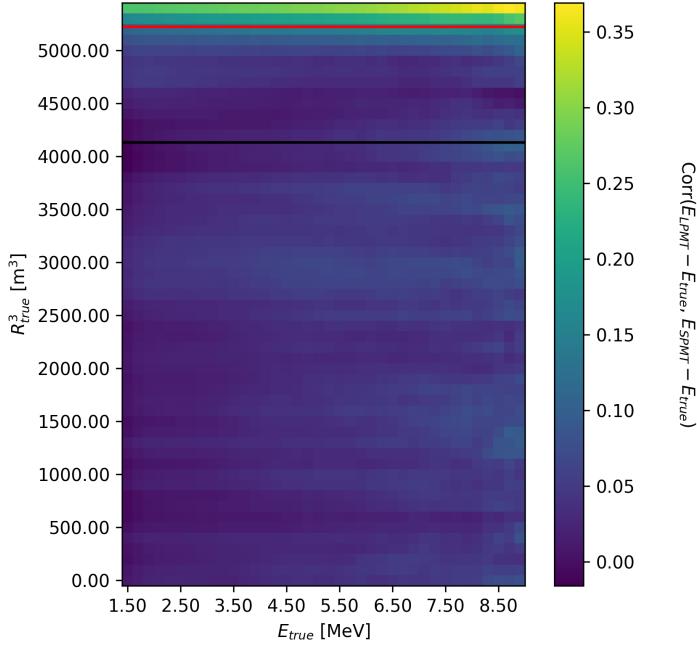


FIGURE 7.23 – Correlation on the reconstruction error between the LPMT and SPMT system as a function of the energy and the radius. The SPMT reconstruction comes from the NN presented in Chapter 4 and the LPMT reconstruction comes from OMILREC presented in Section 3.3. To prevent effect due to the CNN bad reconstruction, we select the event with  $1 < E_{dep} < 9$  MeV.

ing 2 dimensional gaussian. Then, we tuned the IBD generator to choose which of these gaussians to sample, based on the generated values of  $E_{vis}$  and  $R$ . Unfortunately, due to the limited statistics of the full simulation sample, the  $E_{vis} : R^3$  regions were too wide. It lead to sawtooth variations of the correlation and mean values of the gaussian between neighbouring regions. The reconstructed spectra finally showed irregularities instead of a normal, smooth aspect, making it improper for any oscillation analysis.

Before a solution can be found to this problem, we limit our conclusions to this :

- The correlation between the LPMT and SPMT energy reconstruction is positive. Therefore, the SPMT and LPMT spectra should be more correlated than assumed in the statistical tests presented in this chapter. With a proper treatment, we can therefore expect a higher sensitivity to unexpected instrumental effects like QNL.
- In 80% of the detector's volume, reconstruction correlations are low, and should not impact much the sensitivities of our test statistics. If the dependence of the correlation on  $E_{vis}$  and  $R$  proved too difficult to model, one could cut IBDs reconstructed in the Total reflection area. The loss in statistics would be limited, as well as the impact on the sensitivities of the statistical tests.

Additionally, this study is preliminary, as the background was neglected in the distortion test, and no systematic uncertainties were considered. Those points could be easily addressed by regenerating background spectra using the same reference as used by JUNO for the common inputs and by regenerating the systematic covariances matrix with both LPMT and SPMT spectra.

3683 The supplementary non-linearity was introduced event-wise but should be applied channel-wise to  
3684 account for the detector's non-uniformity. This can be addressed via generating oscillated spectra  
3685 through the JUNO official simulation. This process is very time consuming and require technical  
3686 development but could be achievable given enough time.

3687 The correlation matrix between the LPMT and SPMT spectra should also be further analyzed, as  
3688 indicated by the discrepancies between the theoretical and empirical correlation matrices.



# Conclusion

<sup>3689</sup> The field of neutrino physics still has a lot of unanswered questions, namely the mass of the mass states, the Neutrino Mass Ordering (NMO), the possible existence of CP violation in the lepton sector, the unitarity of the PMNS oscillation matrix, and even the nature of the neutrino—Dirac or Majorana—is still unknown. To answer all of these questions, neutrino physics must advance into an era of precision measurements, of which JUNO will be a part.

<sup>3695</sup> This thesis presents my contributions to the JUNO experiment. Its main goals are the measurement <sup>3696</sup> of the oscillation parameters  $\theta_{12}$ ,  $\Delta m_{21}^2$ , and  $\Delta m_{31}^2$  at the permille level, and to determine the Neutrino <sup>3697</sup> Mass Ordering with a significance that requires to reconstruct the energy of the reactor antineutrinos <sup>3698</sup> with a very high precision, and to understand this reconstruction very well. All my contributions are <sup>3699</sup> related to these goals.

<sup>3700</sup>

<sup>3701</sup> In the first two chapters, I gave a short introduction to Neutrino physics and presented the JUNO <sup>3702</sup> experiment. I presented both the detector and various fit approaches used at JUNO to perform the <sup>3703</sup> reactor antineutrino oscillation analysis. It's a base to understand the fit I developed in Chapter 7.

<sup>3704</sup> A large part of my thesis work was devoted to the development of Machine Learning algorithms <sup>3705</sup> for the reconstruction of reactor antineutrinos. In Chapter 3, I gave an introduction to a few types <sup>3706</sup> of algorithms (CNN, GNN) used at JUNO and in this thesis. I also present the existing antineutrino <sup>3707</sup> reconstruction methods, with and without machine learning, which are an important point of comparison <sup>3708</sup> with the methods I developed during this thesis. I showed that the performance of the ML <sup>3709</sup> algorithms developed before the beginning of this thesis did not exceed in a convincing way the <sup>3710</sup> performance of JUNO's canonical likelihood based reconstruction algorithms.

<sup>3711</sup>

<sup>3712</sup> In Chapter 4, I present the first algorithm I developed. It's a CNN reconstructing antineutrinos using <sup>3713</sup> only the SPMT system. Providing an alternative to classical methods in this context is interesting in <sup>3714</sup> its own right.

<sup>3715</sup> It was also for me a gallop of test to learn about JUNO's environment. Finally, classical algorithms <sup>3716</sup> not being available in JUNO's public software, I could use this CNN in Chapter 7, where the SPMT <sup>3717</sup> reconstruction was necessary. The performance reached by this tool is close to that of classical <sup>3718</sup> methods as far as the energy is concerned, but worse when it comes to the reconstruction of the <sup>3719</sup> interaction position.

<sup>3720</sup> One of the difficulties of my algorithm is that it has to train on a lot of pixels that have not been hit. <sup>3721</sup> This problem, partially due to the planar projection of a spherical experiment, is amplified by the <sup>3722</sup> specificities SPMTs (low coverage). The information these pixels carry is meaningless, which should <sup>3723</sup> cause problems in information aggregation. It could be solved by transforming the time information, <sup>3724</sup> a scalar, into a supplementary dimension in the image, resulting in the stacking of successive planar <sup>3725</sup> projections, each representing a time slice of the event. This would hopefully allow to match classical <sup>3726</sup> performances. I did not have enough time to implement such solutions, before I had to switch to <sup>3727</sup> my main thesis subjects. I also performed a combination of the CNN and the classical algorithm. Its <sup>3728</sup> performance exceeds that of the classical algorithm, demonstrating that there must exist an algorithm <sup>3729</sup> better using the input information.

3730  
 3731 In Chapter 5, we formulated the hypothesis according to which ML or DL methods might yield  
 3732 better performance than the classical one if they manage to use more of the information present in  
 3733 the detector, by starting from a rawer level of data (PMT waveforms). Dealing with such a quantity  
 3734 of data requires architectures that help the network to identify essential information and to converge  
 3735 toward the result. We studied the potential of a GNN with an innovative architecture (heterogeneous  
 3736 Graph). It required a lot of technical developments, and a lot of work on the optimisation of the  
 3737 architecture and hyperparameters. This is the ML related work on which I provided most my efforts.

3738 The best performance we obtained does not match that of the classical algorithm nor of other ML  
 3739 methods. We studied elements that suggest that when the GNN aggregates the signals from indi-  
 3740 vidual PMTs belonging to a certain region of the sphere, useful information, in particular temporal,  
 3741 is lost. This demonstrates the difficulty to find ML architectures that will actually improve recon-  
 3742 struction performance. Future versions of my GNN will have to work on this. We can look for new  
 3743 ways to link various regions of the detector, and spend further time refining and adapt the message  
 3744 passing algorithm.

3745  
 3746 In Chapter 6, we worked on ML reliability. We believe that the first step to ensure the reliability of the  
 3747 reconstruction is to benefit of a variety of algorithms. The combination method developed during  
 3748 this thesis allow to not only compare performance and behavior but also to probe in the difference  
 3749 in information used. This also underlines the interest of developing several algorithms for the same  
 3750 tasks, which are then useful even when they do not reach the best performance. However, this is  
 3751 possible only if all algorithms are available to any user. For that reason, my first work on reliability  
 3752 was to implement in JUNO's common software some tools necessary to include in the ML algorithms  
 3753 until then developed as standalone tools, available only to their authors. I also implemented one of  
 3754 these ML algorithms.

3755 We know it is crucial for JUNO not only to reconstruct very precisely the energy of antineutrinos,  
 3756 but also to understand the quality of this reconstruction, and the differences in this between real data  
 3757 and the models assumed by the fits employed to perform the oscillation analysis. We suspect that  
 3758 some subtle differences in the charge and time measured by individual PMTs could affect JUNO's  
 3759 results by distorting very slightly the energy spectrum, while being invisible to data/Monte Carlo  
 3760 comparisons carried out with calibration or signal free control samples. In this chapter 6, I also  
 3761 discuss the exploration of the usage of an Adversarial Neural Network which goal is to help identify  
 3762 the kind of discrepancies that could have this effect, by generating perturbations to the charge and  
 3763 time measured by individual PMTs.

3764 The conclusion of this part explains that this first ANN prototype does not manage to generate  
 3765 perturbations that affect IBD events more than control sample events. However, this exploration  
 3766 taught us several things, among which : it is very difficult to design an ANN able to introduce  
 3767 perturbations at the individual PMT level; some physics-informed guidance will be necessary to  
 3768 obtain an operational tool in the future.

3769  
 3770 The last chapter of this thesis is devoted to Dual calorimetry. There are several concrete applications  
 3771 of this technique. Generically, it is based on the comparison of quantities reconstructed individually  
 3772 by the LPMT and the SPMT systems. It will be used at calibration level. In this thesis, we explore  
 3773 another way, called Dual Calorimetry analysis with neutrino oscillation. It exploits the potential  
 3774 discrepancies between oscillation analyses carried out with either PMT systems.

3775 We designed four statistical tests to detect unexpected instrumental effects in one of the systems or  
 3776 both. We evaluated their sensitivity to a concrete problem: the Charge non linearity (QNL) that will  
 3777 plausibly affect LPMTs. These tests are : the direct comparison of the values of  $\sin^2(2\theta_{12})$  and  $\Delta m_{21}^2$   
 3778 obtained with the LPMT system or the SPMT system ; a direct comparison of the energy spectra

3779 reconstructed by either systems ; and two other tests based on a joint fit of these spectra. A crucial  
3780 ingredient there are the correlations between these spectra, which exist even at the level of statistical  
3781 uncertainties. We designed ways to evaluate them.

3782 We observe that the most powerful tests are those which indeed fully account for these correlations  
3783 : unexpected instrumental effects are not detected only because data spectra do not match the  
3784 predicted spectra but also because they are not consistent with the predicted correlations.

3785 JUNO's most important result will concern the determination of the NMO with JUNO's data only,  
3786 i.e. independent of other experiments. A 3 sigma result is possible with about 6 years of data taking.  
3787 With such statistics, our best statistical tests should detect with a p-value around 1% a QNL effect  
3788 if the calibration phase has not corrected it as well as expected. It proves the interest of the Dual  
3789 calorimetry analysis with neutrino oscillation.

3790 Several assumptions have been discussed concerning the impact of systematic uncertainties, of the  
3791 backgrounds or of the correlation between the SPMT and LPMT reconstructions. They will be the  
3792 subject of future works to make Dual Calorimetry with neutrino oscillation fully operational. We do  
3793 not expect the sensitivities observed here to change much after these refinements.

3794 This work was also the occasion of important technical developments which constitute a major  
3795 improvement of the analysis framework the Subatech group will use to contribute to JUNO's results.



<sup>3796</sup> **Appendix A**

<sup>3797</sup> **Calculation of optimal  $\alpha$  for estimator  
combination**

<sup>3799</sup> This annex the details of the determination of the optimal  $\alpha$  for estimator combination presented in  
<sup>3800</sup> section 4.3.2.

<sup>3801</sup> As a reminder, the combined estimator  $\hat{\theta}$  of  $X$  is defined as

$$\hat{\theta}(X) = \alpha\theta_N + (1 - \alpha)\theta_C; \alpha \in [0; 1] \quad (\text{A.1})$$

<sup>3802</sup> where  $\theta_N$  and  $\theta_C$  are both estimator of  $X$ .

<sup>3803</sup> **A.1 Unbiased estimator**

For the unbiased estimator, it is straight-forward. We search  $\alpha$  such as  $E[\hat{\theta}] = X$

$$E[\hat{\theta}] = E[\alpha\theta_N + (1 - \alpha)\theta_C] \quad (\text{A.2})$$

$$= E[\alpha\theta_N] + E[(1 - \alpha)\theta_C] \quad (\text{A.3})$$

$$= \alpha E[\theta_N] + (1 - \alpha)E[\theta_C] \quad (\text{A.4})$$

$$= \alpha(\mu_N + X) + (1 - \alpha)(\mu_C + X) \quad (\text{A.5})$$

$$X = \alpha\mu_N + \mu_C - \alpha\mu_C + X \quad (\text{A.6})$$

$$0 = \alpha(\mu_N - \mu_C) + \mu_C \quad (\text{A.7})$$

$$(A.8)$$

$$\Rightarrow \alpha = \frac{\mu_C}{\mu_C - \mu_N} \quad (\text{A.9})$$

<sup>3804</sup> **A.2 Optimal variance estimator**

The  $\alpha$  for this estimator is a bit more tricky. By expanding the variance we get

$$\text{Var}[\hat{\theta}] = \text{Var}[\alpha\theta_N + (1 - \alpha)\theta_C] \quad (\text{A.10})$$

$$= \text{Var}[\alpha\theta_N] + \text{Var}[(1 - \alpha)\theta_C] + \text{Cov}[\alpha(1 - \alpha)\theta_N\theta_C] \quad (\text{A.11})$$

$$= \alpha^2\sigma_N^2 + (1 - \alpha)^2\sigma_C^2 + 2\alpha(1 - \alpha)\sigma_N\sigma_C\rho_{NC} \quad (\text{A.12})$$

<sup>3805</sup> where, as a reminder,  $\rho_{NC}$  is the correlation factor between  $\theta_C$  and  $\theta_N$ .

Now we try to find the minima of  $\text{Var}[\hat{\theta}]$  with respect to  $\alpha$ . For this we evaluate the derivative

$$\frac{d}{d\alpha} \text{Var}[\hat{\theta}] = 2\alpha\sigma_N^2 - 2(1-\alpha)\sigma_C^2 + 2\sigma_N\sigma_C\rho_{NC}(1-2\alpha) \quad (\text{A.13})$$

$$= 2\alpha(\sigma_N^2 + \sigma_C^2 - 2\sigma_N\sigma_C\rho_{NC}) - 2\sigma_C^2 + 2\sigma_N\sigma_C\rho_{NC} \quad (\text{A.14})$$

then find the minima and maxima of this derivative by evaluating

$$\frac{d}{d\alpha} \text{Var}[\hat{\theta}] = 0 \quad (\text{A.15})$$

$$2\alpha(\sigma_N^2 + \sigma_C^2 - 2\sigma_N\sigma_C\rho_{NC}) - 2\sigma_C^2 + 2\sigma_N\sigma_C\rho_{NC} = 0 \quad (\text{A.16})$$

$$2\alpha(\sigma_N^2 + \sigma_C^2 - 2\sigma_N\sigma_C\rho_{NC}) = 2\sigma_C^2 - 2\sigma_N\sigma_C\rho_{NC} \quad (\text{A.17})$$

$$\alpha = \frac{\sigma_C^2 - \sigma_N\sigma_C\rho_{NC}}{\sigma_N^2 + \sigma_C^2 - 2\sigma_N\sigma_C\rho_{NC}} \quad (\text{A.18})$$

- 3806 This equation shows only one solution which is a minima. From Eq. A.18 arise two singularities:  
 3807 —  $\sigma_N = \sigma_C = 0$ . This is not a problem because as physicists we never measure with an absolute  
 3808 precision, neither us or our detectors are perfect.  
 3809 —  $\sigma_N = \sigma_C$  and  $\rho_{CN} = 1$ . In this case  $\theta_C$  and  $\theta_N$  are the same estimator in term of variance thus  
 3810 any value for  $\alpha$  yield the same result: an estimator with the same varianve as the original ones.

<sup>3811</sup> **Appendix B**

## <sup>3812</sup> Charge spherical harmonics analysis

<sup>3813</sup> When looking at JUNO events we can clearly see some pattern in the charge repartition based on  
<sup>3814</sup> the event radius as illustrated in figure B.4. When dealing with identifying features and pattern on a  
<sup>3815</sup> spherical plane, the astrophysics community have been using, with success, the spherical harmonic  
<sup>3816</sup> decomposition. The principle is similar to a frequency analysis via Fourier transform. It comes to  
<sup>3817</sup> saying that a function  $f(r, \theta, \phi)$ , here our charge repartition of the spherical plane constructed by our  
<sup>3818</sup> PMTs, can be expressed

$$f(r, \theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l a_l^m r^l Y_l^m(\theta, \phi) \quad (\text{B.1})$$

<sup>3819</sup> where  $a_l^m$  are constants complex factor,  $Y_l^m(\theta, \phi) = Ne^{im\phi} P_l^m(\cos \theta)$  are the spherical harmonics of  
<sup>3820</sup> degree  $l$  and order  $m$  and  $P_l^m$  their associated Legendre Polynomials. Those harmonics are illustrated  
<sup>3821</sup> in figure B.1. By reducing the problem to the unit sphere  $r = 1$ , we get rid of the term  $r^l$ . The Healpix  
<sup>3822</sup> library [77] offer function to efficiently find the  $a_l^m$  factor from a given Healpix map.

<sup>3823</sup> For the above decomposition, we will define the *Power* of an harmonic as

$$S_{ff}(l) = \frac{1}{2l+1} \sum_{m=-l}^l |a_l^m|^2 \quad (\text{B.2})$$

<sup>3824</sup> and the *Relative Power* as:

$$P_l^h = \frac{S_{ff}(l)}{\sum_l S_{ff}(l)} \quad (\text{B.3})$$

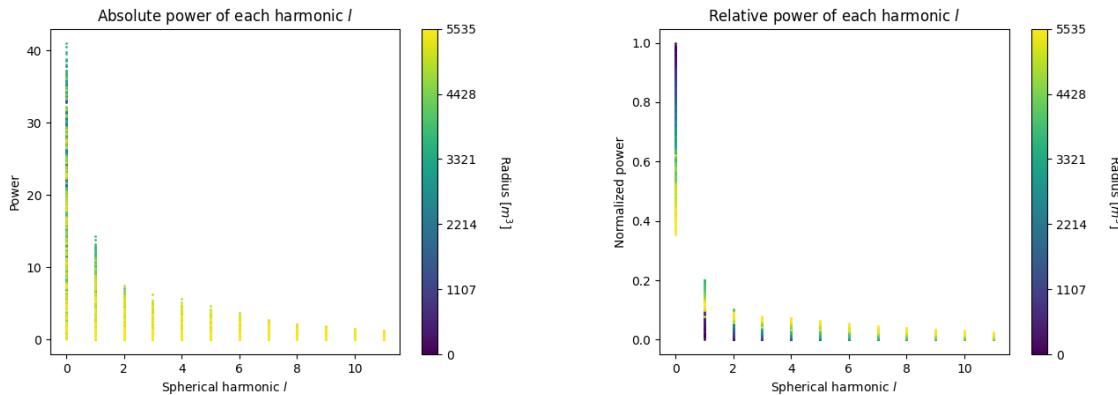
<sup>3825</sup> For this study we will use 10k positron events with  $E_{kin} \in [0; 9]$  MeV uniformly distributed in the  
<sup>3826</sup> CD from the JUNO official simulation version J23.0.1-rc8.dc1 (released the 7th January 2024). All the  
<sup>3827</sup> event are *calib* level, with simulation of the physics, electronics, digitizations and triggers. We first  
<sup>3828</sup> take a sub-set of 1k events and look at the power and relative power distribution depending on the  
<sup>3829</sup> radius and harmonic degree  $l$ . The results are shown in figure B.2. While don't see any pattern in  
<sup>3830</sup> absolute power, it is pretty clear that there is a correlation between the relative power of  $l = 0$  and  
<sup>3831</sup> the radius of the event.

<sup>3832</sup> When applying the same study but dependent on the energy, no clear correlation appear. The results  
<sup>3833</sup> for the  $l = 0$  harmonic are presented in the figure B.5. Thus, in this study we will focus on the radial  
<sup>3834</sup> dependency of the relative power of each harmonic.

<sup>3835</sup> In figures B.6 and B.7 are presented the distribution of the relative power of each harmonic for  $l \in$   
<sup>3836</sup>  $[0, 11]$ . The relation between the radius and the relative power become even more clear, especially  
<sup>3837</sup> for the first harmonics  $l \in [0, 4]$ . After that for  $l > 4$  their relative power is close to 0 for central event,  
<sup>3838</sup> thus loosing power. It also interesting to note the change of behavior in the TR area, clearly visible  
<sup>3839</sup> for  $l = 1$  and  $l = 2$ .

$l:$	$P_\ell^m(\cos \theta) \cos(m\varphi)$	$P_\ell^{ m }(\cos \theta) \sin( m \varphi)$
0 s		
1 p		
2 d		
3 f		
4 g		
5 h		
6 i		
$m:$	6 5 4 3 2 1 0	-1 -2 -3 -4 -5 -6

FIGURE B.1 – Illustration of the real part of the spherical harmonics

FIGURE B.2 – Scatter plot of the absolute and relative power, respectively on the left and right plot, of each harmonic degree  $l$ . The color indicate the radius of the event.

As an erzats of reconstruction algorithm, we fit each of those distribution with a 9th degree polynomial which give us the relation

$$F(R^3) \longmapsto P_l^h \quad (\text{B.4})$$

We do it this way because some of the distribution have multiple solution for a given relative power, for example  $l = 1$ , while each radius give only one power. We now just need to find

$$F^{-1}(P_l^h) \longmapsto R^3 \quad (\text{B.5})$$

Inverting a 9th degree polynomial is hard, if not impossible. The presence of multiple roots for the same power complexify the task even more. To circumvent this problem, we reconstruct the radius by locating the minima of  $(F(R^3) - \hat{P}_l^h)^2$  where  $\hat{P}_l^h$  is the measured power fraction.

To distinguish between multiple possible minima, we use as a starting point the radius given by the procedure on  $l = 0$  that, by looking at the fit in figure B.6, should only present one minima. For  $l > 0$  we also impose bound on the possible reconstructed  $R^3$  as  $R^3 \in [R_0^3 - 100, R_0^3 + 100]$  where  $R_0^3$  is the reconstructed  $R^3$  by the harmonic  $l = 0$ .

3851 The minimization algorithm used are the Bent algorithm for  $l = 0$  and the Bounded algorithm for  
 3852  $l > 0$  provided by the Scipy library [92]. We then do the mean of the reconstructed radius from  
 3853 the different harmonics. The reconstruction results are shown in figure B.3. The performance seems  
 3854 correct but we see heavy fluctuation in the bias. To really be used as a reconstruction algorithm, the  
 3855 method needs to be refined as discussed in the next section.

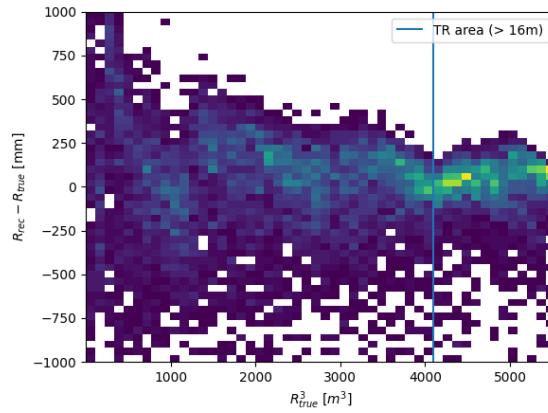


FIGURE B.3 – Error on the reconstructed radius vs the true radius by the harmonic method

## Conclusion

3856 We have clearly shown in this analysis the relevance the of relative harmonic power for radius  
 3857 reconstruction, and provided an erzats of a reconstruction algorithm. We will not delve further in  
 3858 this thesis but if we wanted to refine this algorithm multiple paths can be explored:  
 3859

- 3860 — No energy signature in the harmonics: This is surprising that there is no correlation between the  
 3861 energy and the amplitude of the harmonics. We know that the energy is heavily correlated with  
 3862 the total number of photoelectrons collected, it would be unintuitive that we see no relation.
- 3863 — Localization of the event: We shown here the relation between the relative power of the har-  
 3864 monic and the radius but don't get any information about the  $\theta$  and  $\phi$  spherical coordinates.  
 3865 This information is probably hidden in the individual power of each order  $m$  of the degree  $l$ .  
 3866 This intuition comes from the figure B.1 where in the higher degree  $l$  we see that the order  $m$   
 3867 are oriented. Intuitively, the order should be able to indicate a direction where the signal is  
 3868 more powerful.
- 3869 — Combination of the degree power: Here we combined the radius reconstructed by the different  
 3870 degree via a simple mean but we shown in section 4.3.2 and annex A that this is note the optimal  
 3871 way to combine estimator. A more refined algorithm probably exist to take into account the  
 3872 predicting power of each order.

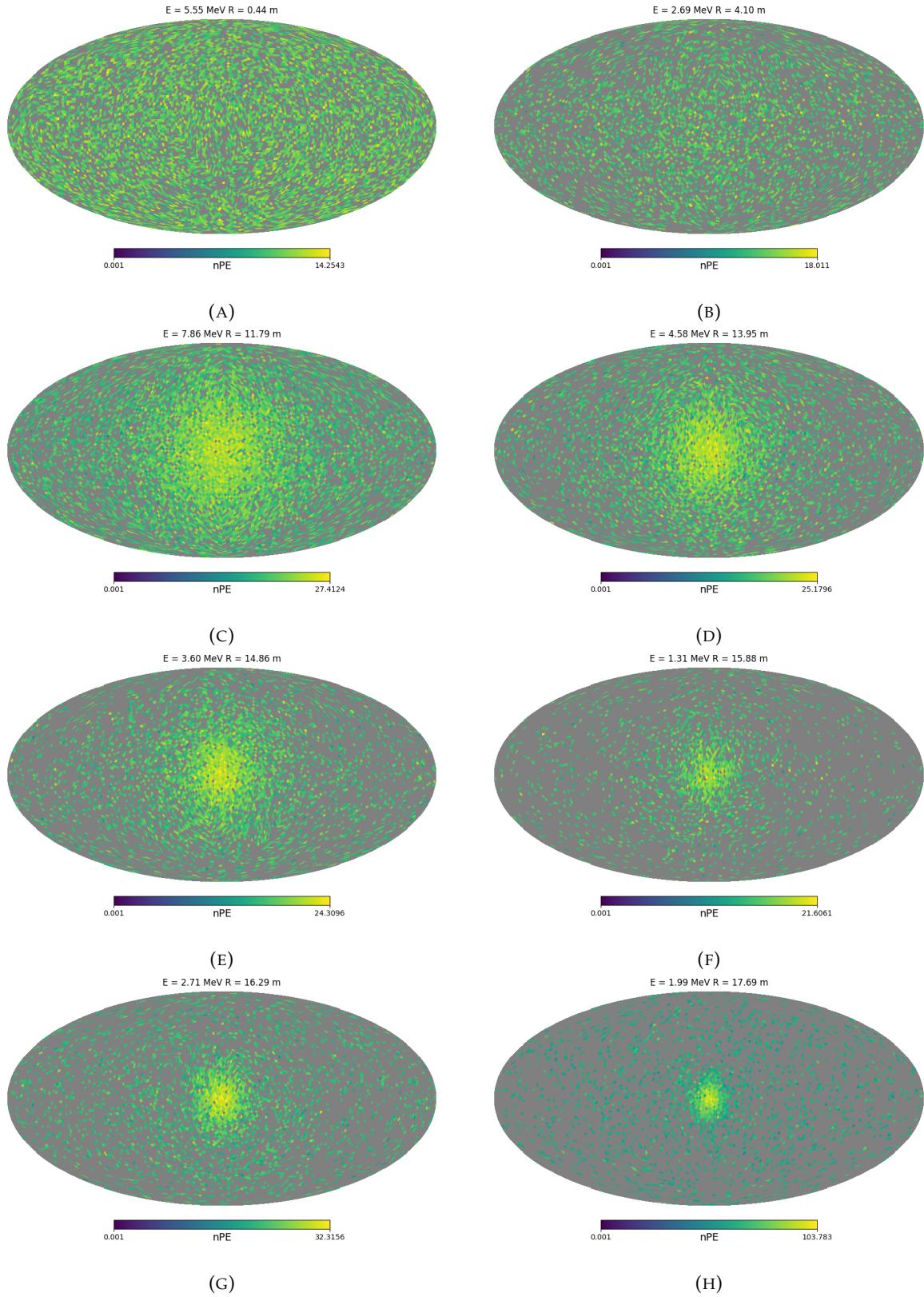


FIGURE B.4 – Charge repartition in JUNO as seen by the Healpix segmentation. Those are Healpix map of order 5 (i.e. 12288 pixels). The color represent the summed charge of the PMTs in each pixels. The color scale is logarithmic. The view have been centered to prevent event deformations.

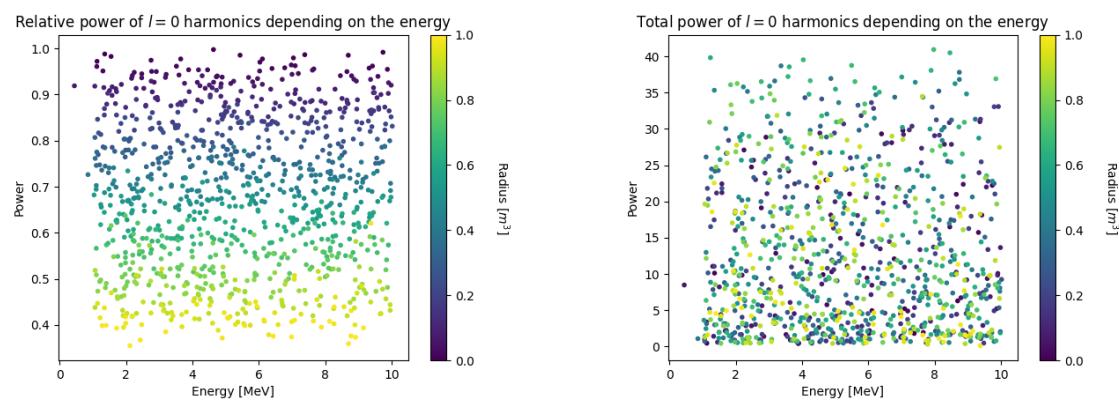


FIGURE B.5 – Scatter plot of the absolute and relative power, respectively on the left and right plot, of the  $l = 0$  harmonic. The color indicate the radius of the event.

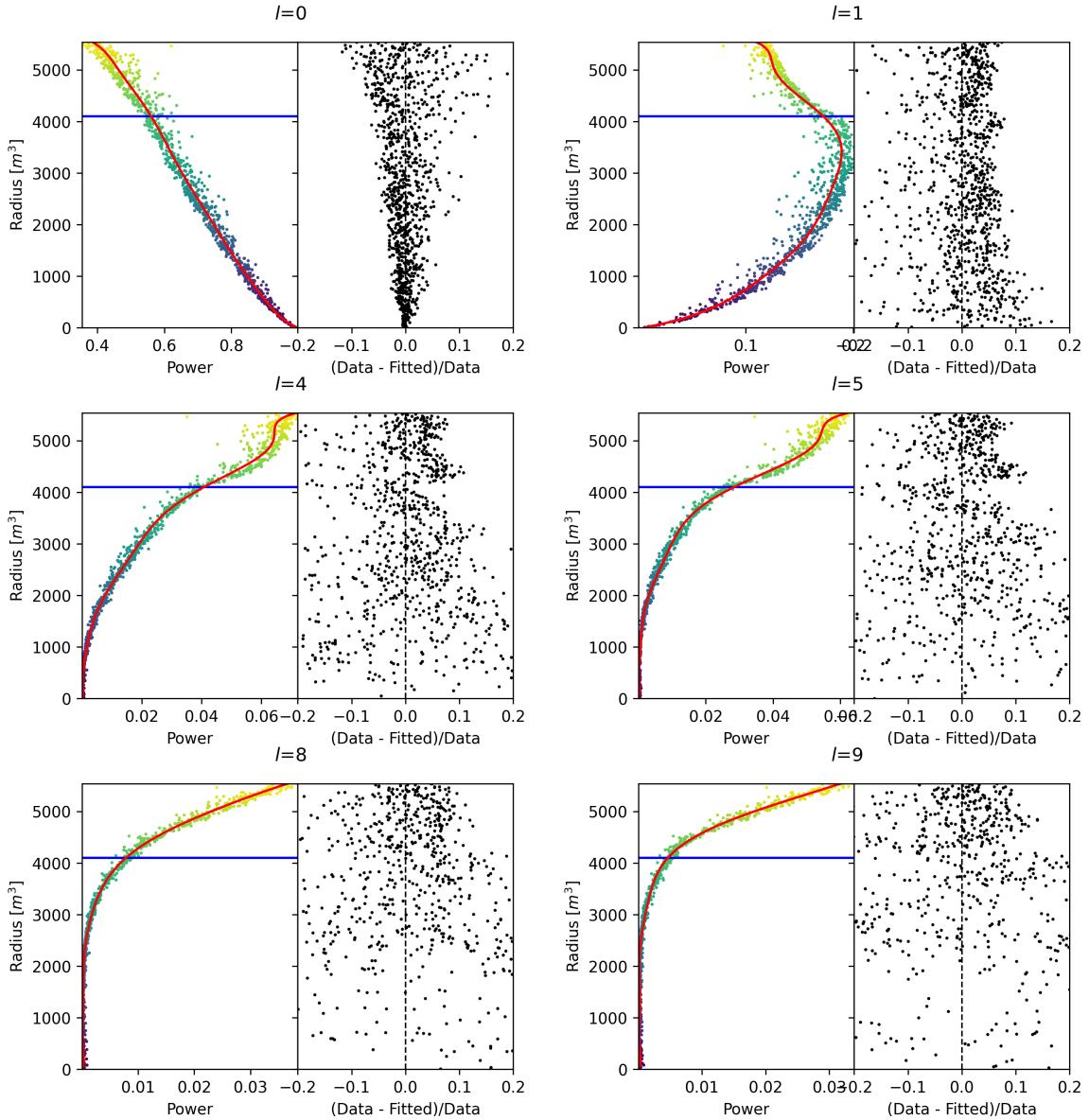


FIGURE B.6 – Plot of the distribution of the relative power of each harmonic dependent on  $R^3$  (on the left). The Total Reflection (TR) area is represented by the horizontal blue line. The distribution are fitted using a 9th degree polynomial (red curve). The relative power error between the distribution and the fit is represented on the left. **Part 1**

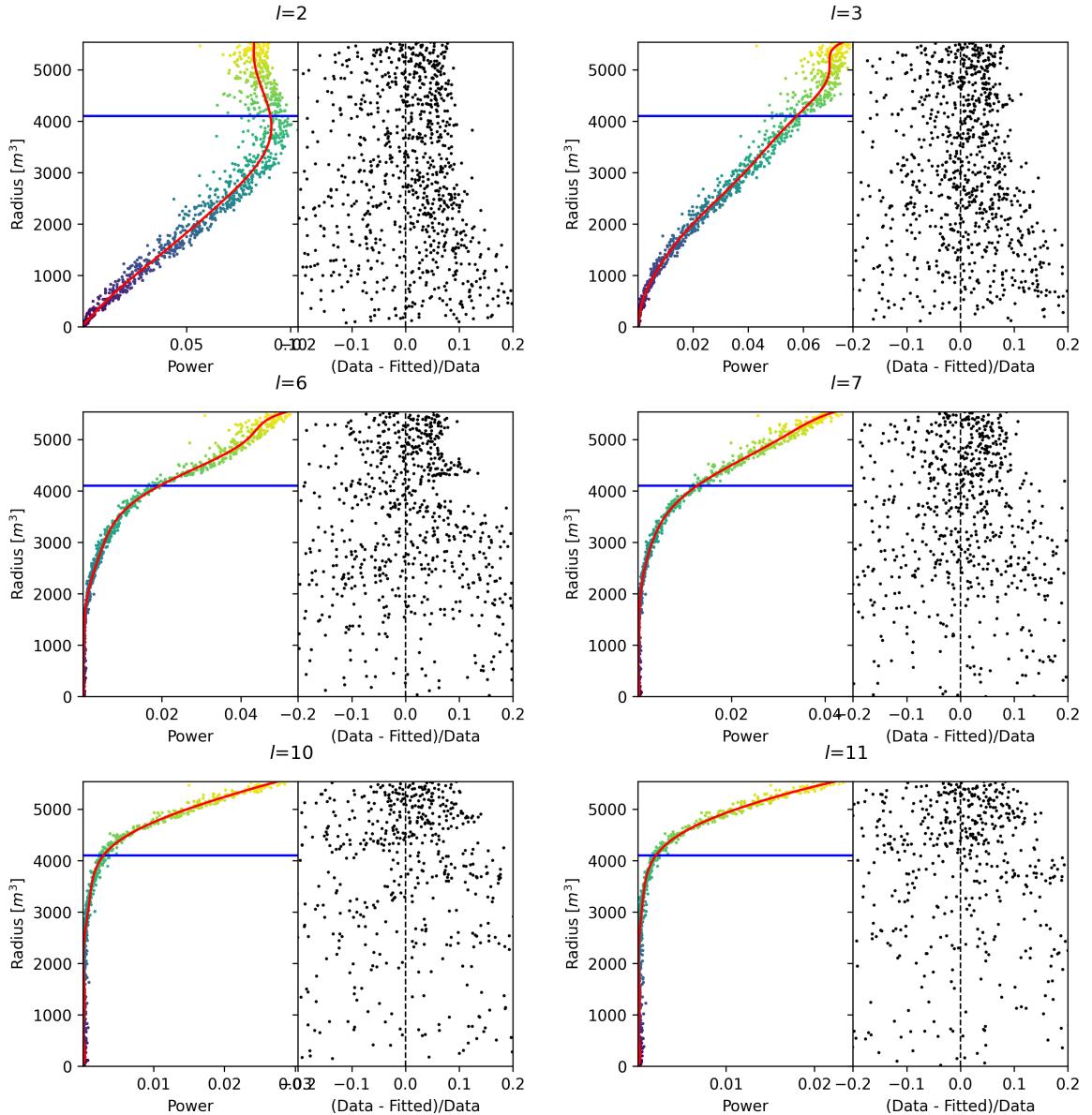


FIGURE B.7 – Plot of the distribution of the relative power of each harmonic dependent on  $R^3$  (on the left). The Total Reflection (TR) area is represented by the horizontal blue line. The distribution are fitted using a 9th degree polynomial (red curve). The relative power error between the distribution and the fit is represented on the left. **Part 2**



<sup>3873</sup> **Appendix C**

<sup>3874</sup> **Correction of  $E_{vis}$  bias**

<sup>3875</sup> The reconstruction algorithms that are presented in this thesis in Chapters 4 and 5 do not reconstruct  
<sup>3876</sup> the same energy as the classical algorithms presented in section 3.3. Our algorithms reconstruct the  
<sup>3877</sup> deposited energy  $E_{dep}$  while the classical algorithms reconstruct a visible energy  $E_{vis}$ .

To understand this phenomena, let's look at the equation 3.27:

$$\hat{\mu}(r, \theta, \theta_{pmt}, E_{vis}) = \frac{1}{E_{vis}} \frac{1}{M} \sum_i^M \frac{\frac{\bar{Q}_i}{\bar{Q}_i} - \mu_i^D}{DE_i}, \quad \mu_i^D = DNR_i \cdot L$$

<sup>3878</sup> which define the expected  $N_{pe}/E$ . This define a linear relation between the number of photoelectrons  
<sup>3879</sup> and the energy. However we discussed in sections 2.3.2 and 2.4 that the number of photoelectrons  
<sup>3880</sup> collected by the LPMT system do not follow a linear relationship. Thus this visible energy is not  
<sup>3881</sup> linear with the deposited energy. This effect is corrected in physics analysis and in Chapter 7 by  
<sup>3882</sup> applying the calibrated non-linearity profile the energy spectrum.

<sup>3883</sup> When we need to compare our algorithm that reconstruct the deposited energy to the classical  
<sup>3884</sup> algorithms we need to correct this non-linearity. For this we fit the systematic bias of the classical  
<sup>3885</sup> algorithm using a 5th degree polynomial

$$\frac{E_{dep}}{E_{vis}} = \sum_{i=0}^5 P_i E_{dep}^i \quad (C.1)$$

<sup>3886</sup> The fitted distribution and the corresponding fit is presented in figure C.1. The value fitted for this  
<sup>3887</sup> correction are presented in table C.1.

$P_0$	$1.24541 +/- 0.00585121$
$P_1$	$-0.168079 +/- 0.00716387$
$P_2$	$0.0489947 +/- 0.00312875$
$P_3$	$-0.00747111 +/- 0.000622003$
$P_4$	$0.000570998 +/- 5.7296e-05$
$P_5$	$-1.72588e-05 +/- 1.98355e-06$

TABLE C.1 – Parameters of the 5th degree polynomial used to correct Omilrec reconstructed energy.

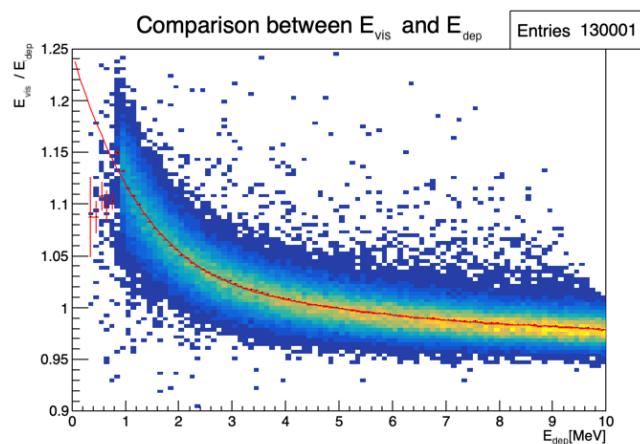


FIGURE C.1 – Comparison between Omilrec reconstructed  $E_{vis}$  and the deposited energy  $E_{dep}$ . The profile of the distribution  $E_{vis}/E_{dep}$  vs  $E_{dep}$  is fitted with a 5th degree polynomial.

# List of Tables

3888	2.1	Characteristics of the nuclear power plants observed by JUNO.	16
3890	2.2	Detectable neutrino signal in JUNO and the expected signal rates and major background sources	17
3891	2.3	List of sources and their process considered for the energy scale calibration	26
3893	2.4	Calibration program of the JUNO experiment	27
3894	2.5	Summary of cumulative reactor antineutrino selection efficiencies. The reported IBD rates (with baselines <300 km) refer to the expected events per day after the selection criteria are progressively applied. Table taken from [38]	31
3895	2.6	Expected background rates, background to signal ratio (B/S), and rate and shape uncertainties. The B/S ratio is calculated by using the IBD signal rate of 47.1/day. Table taken from [38]	31
3897	2.7	A summary of precision levels for the oscillation parameters. The reference value (PDG 2020 [40]) is compared with 100 days, 6 years and 20 years of JUNO data taking.	35
3901	3.1	Features used by the BDT for vertex reconstruction	59
3902	3.2	Features used by the BDTE algorithm. <i>pe</i> and <i>ht</i> reference the charge and hit-time distribution respectively and the percentages are the quantiles of those distributions. <i>cht</i> and <i>cc</i> reference the barycenters of hit time and charge respectively	60
3906	4.1	Sets of hyperparameters values considered in this study	66
3907	5.1	Features on the nodes of the graph. All charge are in [nPE], time in [ns] and position in [m]. $Q$ and $t$ are the reconstructed charge and time of the hit PMTs. $(x, y, z)$ is the position of the PMTs and the last parameter represent the type of the PMT. It's 1 for LPMT and -1 for SPMT $Q_m$ and $t_m$ is the set of charges and time of the PMT belonging the mesh $m$ . $(X_m, Y_m, Z_m)$ is the position of the center of the geometric region represented by the mesh $m$ ( $\langle X \rangle, \langle Y \rangle, \langle Z \rangle$ ) is the position of the charge barycenter, $\Sigma Q$ the sum of the collected charge in the detector and $P_l^h$ is the relative power of the $l$ th harmonic. See annex B for details.	84
3910	5.2	Features on the edges on the graph. It use the same notation as in table 5.1. $D_{m1 \rightarrow m2}^{-1}$ is the inverse of the distance between the mesh $m1$ and the mesh $m2$ . The features <b>A</b> and <b>B</b> are detailed in Section 5.1	85
3918	6.1	Summary of the aggregated features used by the BDT to reconstruct the IBD energy. The charge barycenter and hit time barycenter vertex estimators are detailed in Eq. 6.1 and 6.2 respectively	101
3921	7.1	The charge fraction in terms of the number of PE collected at the single PMT for the reactor $\bar{\nu}_e$ IBD events. Table taken from [30]	123
3922	7.2	Correlations between the parameters BFP of the individual LPMT and SPMT fits for multiple exposures using 1000 toys.	130
3924	7.3	Nominal PDG2020 value [40]. All value are reported assuming Normal Ordering.	132

3926	7.4	Uncertainties on each parameters reported by Minuit on Asimov studies. LPMT and SPMT rows are the results on the individual fit on each spectra. The Weighted row correspond to the weighted average uncertainties between the LPMT and SPMT fits following Eq. 7.29. The Indep Standard joint row is the result of the joint LPMT+SPMT fit but the off-diagonal terms are set to 0. The Indep Standard joint and Standard joint fits both are LPMT+SPMT fit but the parameters $\delta m_{21}^2$ and $\delta \sin^2(2\theta_{12})$ are fixed to 0. The Delta joint and Indep Delta joint are LPMT+SPMT fit with $\delta m_{21}^2$ and $\delta \sin^2(2\theta_{12})$ , difference being that in the Indep version, the off-diagonal terms of the covariance matrix are set to 0. . . . .	141
3927	7.5	In each column, the mean of the distribution of the 1000 best fit values found by fitting the 1000 toy samples with $\alpha_{qnl} = 1\%$ is shown, from which we subtracted the value assumed when generating the toys. A value different from 0 indicates a bias. Between bracket, the average uncertainty of the fitted value is also shown. It allows to judge of the severity of the bias. For instance, the measurement of $\sin^2(2\theta_{12})$ by fitting only the LPMT spectrum tends to be biased at the $-1.569/1.171 = -1.34$ sigma. . . . .	145
3928	7.6	Report of the p-value of the different tests and comparisons for $\alpha_{qnl} = 0.3\%$ for the different exposures. . . . .	148
3929	C.1	Parameters of the 5th degree polynomial used to correct Omilrec reconstructed energy. . . . .	169
3930			
3931			
3932			
3933			
3934			
3935			
3936			
3937			
3938			
3939			
3940			
3941			
3942			
3943			

# List of Figures

3945	1.1	List of the elementary particles in the Standard Model. The antiparticles are not displayed. . . . .	10
3946			
3947	2.1	<b>On the left:</b> Location of the JUNO experiment and its reactor sources in southern china. <b>On the right:</b> Aerial view of the experimental site . . . . .	14
3948			
3949	2.2	Expected number of neutrinos event per MeV in JUNO after 6 years of data taking. The black curve shows the flux if there was no oscillation. The light gray curve shows the oscillation if only the solar terms are taken in account ( $\theta_{12}$ , $\Delta m_{21}^2$ ). The blue and red curve shows the spectrum in the case of, respectively, NO and IO. The dependency of the oscillation to the different parameters are schematized by the double sided arrows. We can see the NMO sensitivity by looking at the fine phase shift between the red and the blue curve. . . . .	15
3950			
3951	2.3	Expected visible energy spectrum measured with the LPMT system with (grey) and without (black) backgrounds. The background amount for about 7% of the IBD candidate and are mostly localized below 3 MeV [9] . . . . .	16
3952			
3953	2.4	. . . . .	19
3954			
3955	a	Schematics view of the JUNO detector. . . . .	19
3956			
3957	b	Top down view of the JUNO detector under construction . . . . .	19
3958			
3959	2.5	Schematics of an IBD interaction in the central detector of JUNO . . . . .	20
3960			
3961	2.6	Schematics of the supporting node for the acrylic vessel . . . . .	21
3962			
3963	2.7	<b>On the left:</b> Quantum efficiency (QE) and emission spectrum of the LAB and the bis-MSB [23]. <b>On the right:</b> Sensitivity of the Hamamatsu LPMT depending on the wavelength of the incident photons [25]. . . . .	22
3964			
3965	2.8	Schematic of a PMT . . . . .	22
3966			
3967	2.9	The LPMT electronics scheme. It is composed of two part, the <i>wet</i> electronics on the left, located underwater and the <i>dry</i> electronics on the right. They are connected by Ethernet cable for data transmission and a dedicated low impedance cable for power distribution . . . . .	23
3968			
3969	2.10	Schematic of the JUNO SPMT electronic system ( <b>left</b> ), and exploded view of the main component of the UWB ( <b>right</b> ) . . . . .	24
3970			
3971	2.11	The JUNO top tracker . . . . .	25
3972			
3973	2.12	Fitted and simulated non linearity of gamma, electron sources and from the $^{12}\text{B}$ spectrum. Black points are simulated data. Red curves are the best fits. Figures taken from [32]. . . . .	26
3974			
3975	a	Gamma non-linearity . . . . .	26
3976			
3977	b	Boron spectrum . . . . .	26
3978			
3979	c	Electron non-linearity . . . . .	26
3980			
3981	2.13	Overview of the calibration system . . . . .	27

3982	2.14	Event-level instrumental non-linearity, defined as the ratio of the total measured LPMT charge to the true charge for events at the center of the detector. The solid red line represents event-level non-linearity without the channel-level correction in an extreme hypothetical scenario of 50% non-linearity over 100 PEs for the LPMTs. The dashed blue line represents that after the channel-level correction. The gray band shows the residual uncertainty of 0.3%, after the channel-level correction. Figure taken from [32].	28
3983	2.15	.....	29
3984	a	Schematic of the TAO satellite detector	29
3985	b	Schematic of the OSIRIS satellite detector	29
3986	2.16	Illustration of the spectrum considered when joint fitting	35
3987			
3988	3.1	Example of a BDT that determine if the given object is a duck	38
3989	3.2	Schema of a simple neural network	39
3990	3.3	Illustration of the training lifecycle	41
3991	3.4	.....	42
3992	a	Illustration of SGD falling into a local minima	42
3993	b	Illustration of the Adam momentum allowing it to overcome local minima	42
3994	3.5	Illustration of the SGD optimizer. In blue is the value of the loss function, orange, green and red are the path taken by the optimized parameter during the training for different LR.	43
3995	a	Illustration of the SGD optimizer on one parameter $\theta$ on the MAE Loss. We see here that it has trouble reaching the minima due to the gradient being constant.	43
3996	b	Illustration of the SGD optimizer on one parameter $\theta$ on the MSE Loss. We see two different behavior: A smooth one (orange and red) when the LR is small enough and a more chaotic one when the LR is too high.	43
3997	3.6	.....	44
3998	a	Illustration of overtraining. The task at hand is to determine depending on two input variable $x$ and $y$ if the data belong to the dataset $A$ or the dataset $B$ . The expected boundary between the two dataset is represented in grey. A possible boundary learnt by overtraining is represented in brown.	44
3999	b	Illustration of a very simple NN	44
4000	3.7	Illustration of the ResNet framework	45
4001	3.8	Illustration of the gradient explosion. Here it can be solved with a lower learning rate but its not always the case.	46
4002	3.9	.....	47
4003	a	Schema of a FCDNN	47
4004	b	Illustration of a composition of ReLU “approximating” a function. (1) No ReLU is taking effect (2) One ReLU is activating (3) Another ReLU is activating	47
4005	3.10	Illustration of the effect of a convolution filter. Here we apply a filter with the aim do detect left edges. We see in the resulting image that the left edges of the duck are bright yellow where the right edges are dark blue indicating the contour of the object. The convolution was calculated using [48].	48
4006	3.11	.....	49
4007	a	Example of images in the MNIST dataset	49
4008	b	Schema of the CNN used in Pytorch example to process the MNIST dataset	49
4009	3.12	Illustration of a graph and its tensor representation.	50
4010	3.13	Illustration of the message passing algorithm. The detailed explanation can be found in Section 3.2.3	50
4011	3.14	.....	52
4012	a	Illustration of the different optical photons reflection scenarios. 1 is the reflection of the photon at the interface LS-acrylic or acrylic-water. 2 is the transmission of the photons through the interfaces. 3 is the conduction of the photon in the acrylic.	52
4013			
4014			
4015			
4016			
4017			
4018			
4019			
4020			
4021			
4022			
4023			
4024			
4025			
4026			
4027			
4028			
4029			
4030			
4031			
4032			
4033			

4034	<b>b</b>	Heatmap of $R_{rec}$ and $R_{rec} - R_{true}$ as a function of $R_{true}$ for 4MeV prompt signals uniformly distributed in the detector calculated by the charge based algorithm	52
4035	<b>3.15</b>	.....	53
4036	<b>a</b>	$\Delta t$ distribution at different iterations step $j$	53
4037	<b>b</b>	Heatmap of $R_{rec}$ and $R_{rec} - R_{true}$ as a function of $R_{true}$ for 4MeV prompt signals uniformly distributed in the detector calculated by the time based algorithm	53
4038	<b>3.16</b>	Bias of the reconstructed radius R (left), $\theta$ (middle) and $\phi$ (right) for multiple energies by the time likelihood algorithm	53
4039	<b>3.17</b>	<b>On the left:</b> Resolution of the reconstructed R as a function of the energy in the TR area ( $R^3 > 4000\text{m}^3 \equiv R > 16\text{m}$ ) by the charge and time likelihood algorithms. <b>On the right:</b> Bias of the reconstructed R in the TR area for different energies by the charge likelihood algorithm	55
4040	<b>3.18</b>	Radial resolution of the different vertex reconstruction algorithms as a function of the energy	56
4041	<b>3.19</b>	.....	56
4042	<b>a</b>	Spherical coordinate system used in JUNO for reconstruction	57
4043	<b>b</b>	Definition of the variables used in the energy reconstruction	57
4044	<b>3.20</b>	.....	59
4045	<b>a</b>	Radial resolutions of the likelihood-based algorithm TMLE, QMLE and QTML	59
4046	<b>b</b>	Energy resolution of QMLE and QTML using different vertex resolutions	59
4047	<b>3.21</b>	Projection of the LPMTs in JUNO on a 2D plane. (a) Show the distribution of all PMTs and (b) and (c) are example of what the charge and time channel looks like respectively	60
4048	<b>3.22</b>	Radial (left) and energy (right) resolutions of different ML algorithms. The results presented here are from [63]. DNN is a deep neural network, BDT is a BDT, ResNet-J and VGG-J are CNN and GNN-J is a GNN.	61
4049	<b>4.1</b>	Graphic representation of the VGG-16 architecture, presenting the different kind of layer composing the architecture	64
4050	<b>4.2</b>	.....	69
4051	<b>a</b>	Spherical coordinate system used in JUNO for reconstruction	69
4052	<b>b</b>	Repartition of SPMTs in the image projection. The color scale is the number of SPMTs per pixel	69
4053	<b>4.3</b>	Example of a high energy, radial event. We see a concentration of the charge on the bottom right of the image, clear indication of a high radius event. <b>On the left:</b> the charge channel. The color is the charge in each pixel in NPE equivalent. <b>On the right:</b> The time channel in nanoseconds	69
4054	<b>4.4</b>	Example of a low energy, radial event. The signal here is way less explicit, we can kind of guess that the event is located in the top middle of the image. <b>On the left:</b> the charge channel. The color is the charge in each pixel in NPE equivalent. <b>On the right:</b> The time channel in nanoseconds	70
4055	<b>4.5</b>	Example of a high energy, central event. In this image we can see a lot of signal but uniformly spread, this is indicative of a central event. <b>On the left:</b> the charge channel. The color is the charge in each pixel in NPE equivalent. <b>On the right:</b> The time channel in nanoseconds	70
4056	<b>4.6</b>	Example of a low energy, central event. Here there is no clear signal, the uniformity of the distribution should make it central. <b>On the left:</b> the charge channel. The color is the charge in each pixel in NPE equivalent. <b>On the right:</b> The time channel in nanoseconds	71
4057	<b>4.7</b>	.....	72
4058	<b>a</b>	Distribution of PE/MeV in the J23 Dataset. This distribution is profiled and fitted using equation 4.6	72
4059	<b>b</b>	<b>On top:</b> Distribution of PE vs Energy. <b>On bottom:</b> Using the values extracted in 4.7a, we calculate the ration signal over background + signal	72

4086    4.8	Reconstruction performance of the Gen <sub>30</sub> model on J21 data and it's comparison to the performances of the classic algorithm "Classical algorithm" from [67]. The top part of each plot is the resolution and the bottom part is the bias. . . . .	73
4087    a	Resolution and bias of energy reconstruction vs energy . . . . .	73
4088    b	Resolution and bias of energy reconstruction vs radius . . . . .	73
4089    c	Resolution and bias of radius reconstruction vs energy . . . . .	73
4090    d	Resolution and bias of radius reconstruction vs radius . . . . .	73
4091    e	Resolution and bias of radius reconstruction vs $\theta$ . . . . .	73
4092    f	Resolution and bias of radius reconstruction vs $\phi$ . . . . .	73
4093    4.9	Residual distribution of the different component of the vertex by Gen <sub>30</sub> . The reconstructed component are $x$ , $y$ and $z$ but we see similar behavior in the error of $R$ , $\theta$ and $\phi$ . . . . .	74
4094    a	Distribution of the error on reconstructed $x$ by Gen <sub>30</sub> . . . . .	74
4095    b	Distribution of the error on reconstructed $y$ by Gen <sub>30</sub> . . . . .	74
4096    c	Distribution of the error on reconstructed $z$ by Gen <sub>30</sub> . . . . .	74
4097    d	Distribution of the error on reconstructed $R$ by Gen <sub>30</sub> . . . . .	74
4098    e	Distribution of the error on reconstructed $\theta$ by Gen <sub>30</sub> . . . . .	74
4099    f	Distribution of the error on reconstructed $\phi$ by Gen <sub>30</sub> . . . . .	74
4100    4.10	. . . . .	75
4101    a	Distribution of Gen <sub>30</sub> reconstructed energy and true energy of the analysis dataset (J21) . . . . .	75
4102    b	Distribution of Gen <sub>42</sub> reconstructed energy and true energy of the analysis dataset (J23) . . . . .	75
4103    4.11	Radius bias ( <b>on the left</b> ) and resolution( <b>on the right</b> ) of the classical algorithm in a $E$ , $R^3$ grid . . . . .	76
4104    4.12	Reconstruction performance of the Gen <sub>30</sub> model on J21, the classic algorithm "Classical algorithm" from [67] and the combination of both using weighted mean. The top part of each plot is the resolution and the bottom part is the bias. . . . .	77
4105    a	Resolution and bias of energy reconstruction vs energy . . . . .	77
4106    b	Resolution and bias of energy reconstruction vs radius . . . . .	77
4107    c	Resolution and bias of radius reconstruction vs energy . . . . .	77
4108    d	Resolution and bias of radius reconstruction vs radius . . . . .	77
4109    e	Resolution and bias of radius reconstruction vs $\theta$ . . . . .	77
4110    f	Resolution and bias of radius reconstruction vs $\phi$ . . . . .	77
4111    4.13	Correlation between CNN and classical method reconstruction ( <b>on the left</b> ) for energy and ( <b>on the right</b> ) for radius in a $E$ , $R^3$ grid . . . . .	78
4112    4.14	Reconstruction performance of the Gen <sub>42</sub> model on J23 data and it's comparison to the performances of the classic algorithm "Classical algorithm" from [67]. The top part of each plot is the resolution and the bottom part is the bias. . . . .	79
4113    a	Resolution and bias of energy reconstruction vs energy . . . . .	79
4114    b	Resolution and bias of energy reconstruction vs radius . . . . .	79
4115    c	Resolution and bias of radius reconstruction vs energy . . . . .	79
4116    d	Resolution and bias of radius reconstruction vs radius . . . . .	79
4117    e	Resolution and bias of radius reconstruction vs $\theta$ . . . . .	79
4118    f	Resolution and bias of radius reconstruction vs $\phi$ . . . . .	79
4119    5.1	. . . . .	83
4120    a	Illustration of the different nodes in our graphs and their relations. . . . .	83
4121    b	Illustration of what a dense adjacency matrix would looks like and the part we are really interested in. Because Fired $\rightarrow$ Mesh and Mesh $\rightarrow$ I/O relations are undirected, we only consider in practice the top right part of the matrix for those relations. . . . .	83

4137	5.2 Illustration of the Healpix segmentation. <b>On the left:</b> A segmentation of order 0. <b>On the right:</b> A segmentation of order 1 . . . . .	84
4138		
4139	5.3 Illustration of the different update function needed by our GNN . . . . .	86
4140		
4141	5.4 Distribution of the number of hits depending on the energy. <b>On the right:</b> for the LPMT system. <b>In the middle :</b> for the SPMT system. <b>On the left:</b> For both system . . . . .	87
4142	a . . . . .	87
4143	b . . . . .	87
4144	c . . . . .	87
4145		
4146	5.5 Distribution of the number of hits depending on the radius. <b>On the right:</b> for the LPMT system. <b>On the right :</b> for the SPMT system. To prevent the superposition of structure of different scales we limit ourselves to the energy range $E_{true} \in [0, 9]$ . . . . .	88
4147		
4148	a . . . . .	88
4149	b . . . . .	88
4150		
4151	5.6 Schema of the JWGv8.4.0 architecture, the colored triplet is the graph configuration after each JWG layers . . . . .	89
4152		
4153	5.7 Energy reconstruction depending on the true energy for samples of the different versions of the GNN . . . . .	90
4154		
4155	5.8 Reconstruction performance of the Omilrec algorithm based on QTML presented in Section 3.3, JWGv8.4 presented in this chapter. The top part of each plot is the resolution and the bottom part is the bias. . . . .	92
4156		
4157	a Resolution and bias of energy reconstruction vs energy . . . . .	92
4158	b Resolution and bias of energy reconstruction vs radius . . . . .	92
4159		
4160	5.9 Reconstruction performance of the Omilrec algorithm based on QTML presented in Section 3.3, JWGv8.4 presented in this chapter. The top part of each plot is the resolution and the bottom part is the bias. . . . .	93
4161		
4162	a Resolution and bias of radius reconstruction vs energy . . . . .	93
4163	b Resolution and bias of radius reconstruction vs radius . . . . .	93
4164		
4165	5.10 Reconstruction performance of the Omilrec algorithm based on QTML presented in Section 3.3, JWGv8.4 presented in this chapter. The top part of each plot is the resolution and the bottom part is the bias. . . . .	94
4166		
4167	a Resolution and bias of radius reconstruction vs $\theta$ . . . . .	94
4168	b Resolution and bias of radius reconstruction vs $\phi$ . . . . .	94
4169		
4170	5.11 Reconstruction performance of the Omilrec algorithm, JWGv8.4 and the combination between the two using the optimal variance estimator presented in annex A.2. The top part of each plot is the resolution and the bottom part is the bias. . . . .	95
4171		
4172	a Resolution and bias of energy reconstruction vs energy . . . . .	95
4173	b Resolution and bias of energy reconstruction vs radius . . . . .	95
4174		
4175	5.12 Reconstruction performance of the Omilrec algorithm, JWGv8.4 and the combination between the two using the optimal variance estimator presented in annex A.2. The top part of each plot is the resolution and the bottom part is the bias. . . . .	96
4176		
4177	a Resolution and bias of radius reconstruction vs energy . . . . .	96
4178	b Resolution and bias of radius reconstruction vs radius . . . . .	96
4179		
4180	5.13 Reconstruction performance of the Omilrec algorithm based on QTML presented in Section 3.3, JWGv8.4 presented in this chapter and the HCNN algorithm. The top part of each plot is the resolution and the bottom part is the bias. . . . .	97
4181		
4182	a Resolution and bias of energy reconstruction vs energy . . . . .	97
4183	b Resolution and bias of energy reconstruction vs radius . . . . .	97
4184		
4185	5.14 Reconstruction performance of the Omilrec algorithm based on QTML presented in Section 3.3, JWGv8.4 presented in this chapter and the HCNN algorithm. The top part of each plot is the resolution and the bottom part is the bias. . . . .	98
4186		
4187	a Resolution and bias of radius reconstruction vs energy . . . . .	98
4188	b Resolution and bias of radius reconstruction vs radius . . . . .	98

4189    6.1	Relative difference between the features computed by Gavrikov et. al (superscripted Paper) and our implementation (superscripted Implementation) . . . . .	102
4190		
4191    6.2	Resolution of BDTE <b>On the left:</b> as reported by Gavrikov Arsenii et. al in [64], <b>On the right:</b> once implemented in JUNO common software. On the right plot is also reported the reconstruction performance of the OMILREC algorithm. The OMILREC algorithm $E_{vis}$ has been corrected to $E_{dep}$ following the procedure detailed in Annex C.	102
4192    a	a . . . . .	102
4193    b	b . . . . .	102
4194		
4195    6.3	Correlation between the errors in energy reconstruction between BDTE and OMILREC (Eq. 6.3). The correlation is computed in $R^3$ bins of $216 \text{ m}^3$ between 0 and $5000 \text{ m}^3$ , 0 and 17 m in y axis, and in 0.40 MeV bins between 1.022 and 10.022 MeV of deposited energy. . . . .	103
4196		
4197    6.4	Schema of the method to discover vulnerabilities in the reconstruction methods. <b>On the top</b> of the image, the standard data flow. The individual charge and times are fed to a reconstruction algorithm. From the reconstructed energies, we can produce an IBD spectrum and compute control observables from the control samples. <b>On the bottom</b> , the same data flow but we add an ANN between the input and the reconstruc- tion. The ANN will slightly change the input charge and time so the reconstruction algorithm inaccurately reconstruct the IBD energy, but the perturbation is not visible in the control samples. . . . .	104
4198		
4199    6.5	Energy resolution of the FFNN with respect to the energy ( <b>On the right</b> ) and the radius ( <b>On the left</b> ) . . . . .	107
4200		
4201    6.6	Radial resolution of the FFNN with respect to the energy ( <b>On the right</b> ) and the radius ( <b>On the left</b> ) . . . . .	108
4202		
4203    6.7	Illustration of the “bottleneck” architecture of the ANN. Each block represent a fully connected layer with, on the left, the input layer and on the right the output layer. We see a first reduction of the number of neurons per layer, going from 4096 to 256, followed by an augmentation back to 4096 neurons, thus the “bottleneck” . . . . .	109
4204		
4205    6.8	Evolution of the loss $\mathcal{L}_1 = 0.25 \cdot P(0.01)$ during the first phase of the training . . . . .	111
4206		
4207    6.9	Time channel (on the left) and charge channel (on the right) of a <b>radial, high energy event</b> ( $R = 17.2 \text{ m}$ , $E_{dep} = 7.1 \text{ MeV}$ ), <b>Top:</b> before the ANN perturbation, <b>Bottom:</b> after the ANN perturbation. The ANN have been trained for 200 epochs, just after Phase 1. Time channel in ns and charge channel in $N_{pe}$ . . . . .	112
4208		
4209    6.10	Time channel (on the left) and charge channel (on the right) of a <b>central, low energy event</b> ( $R = 9.1 \text{ m}$ , $E_{dep} = 1.9 \text{ MeV}$ ), <b>Top:</b> before the ANN perturbation, <b>Bottom:</b> after the ANN perturbation. The ANN have been trained for 200 epochs, just after Phase 1. Time channel in ns and charge channel in $N_{pe}$ . . . . .	113
4210		
4211    6.11	Ratio of the reconstructed energy spectra between $(\mathcal{F} \circ \mathcal{G})$ and $\mathcal{F}$ at then end of Phase 1 of the training. <b>On the left:</b> For the $^{12}\text{B}$ dataset. <b>On the right:</b> For the IBD dataset .	113
4212		
4213    6.12	<b>On the top :</b> Distribution of the relative energy reconstruction error between $\mathcal{F}$ (light histogram) and $(\mathcal{F} \circ \mathcal{G})$ (dark histogram) at then end of Phase 1 of the training. <b>On the bottom :</b> Ratio between the light and dark histogram of the top figure. . . . .	114
4214		
4215    6.13	Profile of the loss $\mathcal{L}_2$ and $\mathcal{L}_{adv}$ during the second phase of training. The linear increase of $\mathcal{L}_2$ is due to the growing factor $\lambda$ in Eq. 6.22. . . . .	115
4216		
4217    6.14	Profile of the loss $60 \cdot \mathcal{L}_{reg}$ and $0.25 \cdot P(0.15)$ during the second phase of training . . .	115
4218    6.15	Profile of the loss over the entirety of the training (Phase 1 and 2) . . . . .	115
4219		
4220    6.16	Time channel (on the left) and charge channel (on the right) of a <b>radial, high energy event</b> ( $R = 17.2 \text{ m}$ , $E_{dep} = 7.1 \text{ MeV}$ ), <b>Top:</b> before the ANN perturbation, <b>Bottom:</b> after the ANN perturbation. The ANN have been trained for 400 epochs, just after Phase 2. Time channel in ns and charge channel in $N_{pe}$ . . . . .	116
4221		

4239	6.17 Time channel (on the left) and charge channel (on the right) of a <b>central, low energy event</b> ( $R = 9.1$ m, $E_{dep} = 1.9$ MeV), <b>Top:</b> before the ANN perturbation, <b>Bottom:</b> after the ANN perturbation. The ANN have been trained for 400 epochs, just after Phase 2. Time channel in ns and charge channel in $N_{pe}$ . . . . .	117
4240		
4241		
4242		
4243	6.18 Ratio of the reconstructed energy spectra between $(\mathcal{F} \circ \mathcal{G})$ and $\mathcal{F}$ at then end of Phase 2 of the training. <b>On the left :</b> For the $^{12}\text{B}$ dataset. <b>On the right :</b> For the IBD dataset . . . . .	117
4244		
4245	6.19 <b>On the top :</b> Distribution of the relative energy reconstruction error between $\mathcal{F}$ (light histogram) and $(\mathcal{F} \circ \mathcal{G})$ (dark histogram) at then end of Phase 2 of the training. <b>On the bottom :</b> Ratio between the light and dark histogram of the top figure. . . . .	118
4246		
4247		
4248	6.20 Ratio between the relative error on the reconstructed energy between the IBD and the $^{12}\text{B}$ dataset. <b>On the right :</b> without the ANN. <b>On the left :</b> with the ANN. . . . .	118
4249		
4250	7.1 Expected number of neutrinos event per MeV in JUNO after 6 years of data taking. The black curve shows the flux if there was no oscillation. The light gray curve shows the oscillation if only the solar terms are taken in account ( $\theta_{12}, \Delta m_{21}^2$ ). The blue and red curve shows the spectrum in the case of, respectively, NO and IO. The dependency of the oscillation to the different parameters are schematized by the double sided arrows. We can see the NMO sensitivity by looking at the fine phase shift between the red and the blue curve. . . . .	120
4251		
4252		
4253		
4254		
4255		
4256		
4257	7.2 Oscillated reactor $\bar{\nu}_e$ spectra for the Normal Ordering (Black) and Inverted Ordering (Red) for 6,5 years data taking and a resolution of 3% without any statistical or systematic fluctuation. Figure from [38]. . . . .	121
4258		
4259		
4260	7.3 <b>On top:</b> Oscillated spectra for different value of $\alpha_{qnl}$ . <b>On bottom:</b> Ratio of the number of event with $\alpha_{qnl} = 0\%$ . . . . .	123
4261		
4262	7.4 Distribution the ratio reconstructed charge (in nPE equivalent) over the number of collected nPE for different value of $\gamma_{qnl}$ . We use a sample of 1 million positron event uniformly distributed in the detector and in energy in the range $E_{dep} \in [1, 10] \text{ MeV}$ . . . . .	124
4263		
4264		
4265	7.5 <b>On top:</b> Ratio of the reconstructed charge (in nPE equivalent) over the number of collected nPE. The dots represent the mean of the distributions in Figure 7.4 and the dashed line are the equivalent event-wise non-linearity from eq 7.2. The hatched zone is the residual non-linearity expected after calibration [32]. <b>On bottom:</b> Difference between QNL induced by an event wise QNL and the mean QNL induced by a channel wise QNL. The value for $\alpha_{qnl}$ and $\gamma_{qnl}$ follow the color code of the top figure. For a given energy, all the data point have the same value. . . . .	125
4266		
4267		
4268		
4269		
4270		
4271		
4272	7.6 Distribution and correlation between the best fit point of 1000 individual toys fit for 100 days exposure without supplementary QNL. . . . .	128
4273		
4274	7.7 Distribution and correlation between the best fit point of 1000 individual toys fit for 1 year exposure without supplementary QNL. . . . .	129
4275		
4276	7.8 Distribution and correlation between the best fit point of 1000 individual toys fit for 2 years exposure without supplementary QNL. . . . .	129
4277		
4278	7.9 Distribution and correlation between the best fit point of 1000 individual toys fit for 6 years exposure without supplementary QNL. . . . .	130
4279		
4280	7.10 Relative ( <b>On the left</b> ) and absolute ( <b>On the right</b> ) resolutions of the LPMT and SPMT systems used in this study. The number in parenthesis are the parameter A, B and C respectively for each systems. . . . .	136
4281		
4282		
4283	7.11 Theoretical correlation matrix between the LPMT spectrum (bins 0-409) ans the SPMT spectrum (410-819). The diagonal has been set to 0 (it was 1) for readability purpose. . . . .	139
4284		
4285	7.12 Upper left corner of the estimated correlation matrix between the LPMT and SPMT spectrum for different configuration of $N$ toy with different number of $M$ events per toy. We observe that the statistical uncertainty, the noise effect, diminish with the number of toy considered. . . . .	140
4286		
4287		
4288	a . . . . .	140
4289	b . . . . .	140
4290		

4291	c	140
4292	7.13 Relative difference between the element of the theoretical and empiric correlation matrix	140
4293	7.14 Distribution of BFP - nominal value for 1000 toy Standard joint fit. 6 years exposure, all background, PearsonV $\chi^2$ , $\theta_{13}$ fixed. In those plots, $\theta_{12}$ stands for $\sin^2(2\theta_{12})$	143
4294	7.15 Distribution of BFP - nominal value for 5000 toy Delta joint fit. 6 years exposure, all background, PearsonV $\chi^2$ , $\theta_{13}$ fixed. In those plots, $\theta_{12}$ stands for $\sin^2(2\theta_{12})$ and $\delta\theta_{12}$ for $\delta \sin^2(\theta_{12})$	144
4295	7.16 Top: Theoretical spectrum without QNL (in red) and with $\alpha_{qnl} = 1\%$ (in blue). Bottom: Ratio between the theoretical spectrum with and without QNL.	145
4296	7.17 Distribution of the $\chi^2_{ind}$ for 1000 toys for different exposures. The dashed lines represent the median of the distributions and the p-value are the percentage of the $\alpha_{qnl} = 0$ distribution that are greater than those medians.	146
4297	7.18 Distribution of the $\chi^2_{spe}$ for 1000 toys for different exposure. The dashed line represent the median of the distribution and the p-value are the percentage of the $\alpha_{qnl} = 0$ distribution that are greater than those medians.	147
4298	7.19 Distribution of $\chi^2_{H_0} - \chi^2_{H_1}$ for 1000 toys for different exposure. The dashed line represent the median of the distribution and the p-value are the percentage of the $\alpha_{qnl} = 0$ distribution that are greater than those medians.	148
4299	a 100 days exposure	148
4300	b 1 year exposure	148
4301	c 2 years exposure	148
4302	d 6 years exposure	148
4303	7.20 Distribution of the $\delta \sin^2(2\theta_{12})$ for 1000 toys for different exposure. The dashed line represent the median of the distribution and the p-value are the percentage of the $\alpha_{qnl} = 0$ distribution that are greater than those medians.	149
4304	a 100 days exposure	149
4305	b 1 year exposure	149
4306	c 2 years exposure	149
4307	d 6 years exposure	149
4308	7.21 Distribution of the $\delta\Delta m_{21}^2$ for 1000 toys for different exposure. The dashed line represent the median of the distribution and the p-value are the percentage of the $\alpha_{qnl} = 0$ distribution that are greater than those medians.	150
4309	a 100 days exposure	150
4310	b 1 year exposure	150
4311	c 2 years exposure	150
4312	d 6 years exposure	150
4313	7.22 Correlation on the reconstruction error between the LPMT and SPMT system as a function of (On the left) the energy, (On the right) the radius. The SPMT reconstruction comes from the NN presented in Chapter 4 and the LPMT reconstruction comes from OMILREC presented in Section 3.3. To prevent effect due to the CNN bad reconstruction, we select the event with $1 < E_{dep} < 9$ MeV.	151
4314	7.23 Correlation on the reconstruction error between the LPMT and SPMT system as a function of the energy and the radius. The SPMT reconstruction comes from the NN presented in Chapter 4 and the LPMT reconstruction comes from OMILREC presented in Section 3.3. To prevent effect due to the CNN bad reconstruction, we select the event with $1 < E_{dep} < 9$ MeV.	152
4315	B.1 Illustration of the real part of the spherical harmonics	162
4316	B.2 Scatter plot of the absolute and relative power, respectively on the left and right plot, of each harmonic degree $l$ . The color indicate the radius of the event.	162
4317	B.3 Error on the reconstructed radius vs the true radius by the harmonic method	163

---

4341	B.4 Charge repartition in JUNO as seen by the Healpix segmentation. Those are Healpix 4342 map of order 5 (i.e. 12288 pixels). The color represent the summed charge of the PMTs 4343 in each pixels. The color scale is logarithmic. The view have been centered to prevent 4344 event deformations. . . . .	164
4345	a . . . . .	164
4346	b . . . . .	164
4347	c . . . . .	164
4348	d . . . . .	164
4349	e . . . . .	164
4350	f . . . . .	164
4351	g . . . . .	164
4352	h . . . . .	164
4353	B.5 Scatter plot of the absolute and relative power, respectively on the left and right plot, 4354 of the $l = 0$ harmonic. The color indicate the radius of the event. . . . .	165
4355	B.6 Plot of the distribution of the relative power of each harmonic dependent on $R^3$ (on 4356 the left). The Total Reflection (TR) area is represented by the horizontal blue line. The 4357 distribution are fitted using a 9th degree polynomial (red curve). The relative power 4358 error between the distribution and the fit is represented on the left. Part 1 . . . . .	166
4359	B.7 Plot of the distribution of the relative power of each harmonic dependent on $R^3$ (on 4360 the left). The Total Reflection (TR) area is represented by the horizontal blue line. The 4361 distribution are fitted using a 9th degree polynomial (red curve). The relative power 4362 error between the distribution and the fit is represented on the left. Part 2 . . . . .	167
4363	C.1 Comparison between Omilrec reconstructed $E_{vis}$ and the deposited energy $E_{dep}$ . The 4364 profile of the distribution $E_{vis}/E_{dep}$ vs $E_{dep}$ is fitted with a 5th degree polynomial. . . . .	170



# <sup>4365</sup> List of Abbreviations

<b>ACU</b>	Automatic Calibration Unit
<b>ANN</b>	Adversarial Neural Network
<b>BDT</b>	Boosted Decision Tree
<b>BFP</b>	Best Fit Point
<b>CD</b>	Central Detector
<b>CLS</b>	Cable Loop System
<b>CNN</b>	Convolutional NN
<b>DNN</b>	Deep NN
<b>DN</b>	Dark Noise
<b>EDM</b>	Event Data Model
<b>FCDNN</b>	Fully Connected Deep NN
<b>GNN</b>	Graph NN
<b>GT</b>	Guiding Tube
<b>IBD</b>	Inverse Beta Decay
<b>IO</b>	Inverse Ordering
<b>JUNO</b>	Jiangmen Underground Neutrino Observatory
<b>LPMT</b>	Large PMT
<b>LR</b>	Learning Rate
<b>LS</b>	Liquid Scintillator
<b>MC</b>	Monte Carlo simulation
<b>ML</b>	Machine Learning
<b>MSE</b>	Mean Squared Error
<b>NMO</b>	Neutrino Mass Ordering
<b>NN</b>	Neural Network
<b>NO</b>	Normal Ordering
<b>NPE</b>	Number of Photo Electron
<b>OSIRIS</b>	Online Scintillator Internal Radioactivity Investigation System
<b>PE</b>	Photo Electron
<b>PMT</b>	Photo-Multipliers Tubes
<b>PRelu</b>	Parametrized Rectified Linear Unit
<b>QNL</b>	Charge (Q) Non Linearity
<b>ROV</b>	Remotely Operated under-LS Vehicle
<b>ReLU</b>	Rectified Linear Unit
<b>ResNet</b>	Residual Network
<b>SGD</b>	Stochastic Gradient Descent
<b>SPMT</b>	Small PMT
<b>TAO</b>	Taishan Antineutrino Observatory
<b>TR Area</b>	Total Reflexion Area
<b>TTS</b>	Time Transit Spread
<b>TT</b>	Top Tracker
<b>UWB</b>	Under Water Boxes
<b>WCD</b>	Water Cherenkov Detector



# 4366 Bibliography

- 4367 [1] S. Navas et al. "Review of particle physics". *Phys. Rev. D* 110.3 (2024), 030001. DOI: [10.1103/PhysRevD.110.030001](https://doi.org/10.1103/PhysRevD.110.030001).
- 4368 [2] Daniel Carney, Valerie Domcke, and Nicholas L. Rodd. "Graviton detection and the quantization of gravity". *Physical Review D* 109.4 (Feb. 5, 2024). Publisher: American Physical Society. DOI: [10.1103/PhysRevD.109.044009](https://doi.org/10.1103/PhysRevD.109.044009). URL: <https://link.aps.org/doi/10.1103/PhysRevD.109.044009> (visited on 10/14/2024).
- 4369 [3] Emmy Noether. "Invariant variation problems". *Transport Theory and Statistical Physics* 1.3 (Jan. 1, 1971). Publisher: Taylor & Francis *eprint* : <https://doi.org/10.1080/00411457108231446>. ISSN: 0041-1450. DOI: [10.1080/00411457108231446](https://doi.org/10.1080/00411457108231446). URL: <https://doi.org/10.1080/00411457108231446> (visited on 10/14/2024).
- 4370 [4] T. D. Lee and C. N. Yang. "Question of Parity Conservation in Weak Interactions". *Physical Review* 104.1 (Oct. 1, 1956). Publisher: American Physical Society. DOI: [10.1103/PhysRev.104.254](https://doi.org/10.1103/PhysRev.104.254). URL: <https://link.aps.org/doi/10.1103/PhysRev.104.254> (visited on 10/14/2024).
- 4371 [5] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. "Experimental Test of Parity Conservation in Beta Decay". *Physical Review* 105.4 (Feb. 15, 1957). Publisher: American Physical Society. DOI: [10.1103/PhysRev.105.1413](https://doi.org/10.1103/PhysRev.105.1413). URL: <https://link.aps.org/doi/10.1103/PhysRev.105.1413> (visited on 10/14/2024).
- 4372 [6] J. H. Christenson, J. W. Cronin, V. L. Fitch, and R. Turlay. "Evidence for the  $2\pi$  Decay of the  $K_2^0$  Meson". *Physical Review Letters* 13.4 (July 27, 1964). Publisher: American Physical Society. DOI: [10.1103/PhysRevLett.13.138](https://doi.org/10.1103/PhysRevLett.13.138). URL: <https://link.aps.org/doi/10.1103/PhysRevLett.13.138> (visited on 10/14/2024).
- 4373 [7] Liang Zhan, Yifang Wang, Jun Cao, and Liangjian Wen. "Determination of the Neutrino Mass Hierarchy at an Intermediate Baseline". *Physical Review D* 78.11 (Dec. 10, 2008). ISSN: 1550-7998, 1550-2368. DOI: [10.1103/PhysRevD.78.111103](https://doi.org/10.1103/PhysRevD.78.111103). *eprint*: [0807.3203\[hep-ex, physics:hep-ph\]](https://arxiv.org/abs/0807.3203). URL: [http://arxiv.org/abs/0807.3203](https://arxiv.org/abs/0807.3203) (visited on 09/18/2023).
- 4374 [8] Fengpeng An et al. "Neutrino Physics with JUNO". *Journal of Physics G: Nuclear and Particle Physics* 43.3 (Mar. 1, 2016). ISSN: 0954-3899, 1361-6471. DOI: [10.1088/0954-3899/43/3/030401](https://doi.org/10.1088/0954-3899/43/3/030401). *eprint*: [1507.05613\[hep-ex, physics:physics\]](https://arxiv.org/abs/1507.05613). URL: [http://arxiv.org/abs/1507.05613](https://arxiv.org/abs/1507.05613) (visited on 07/28/2023).
- 4375 [9] JUNO Collaboration et al. "Sub-percent Precision Measurement of Neutrino Oscillation Parameters with JUNO". *Chinese Physics C* 46.12 (Dec. 1, 2022). ISSN: 1674-1137, 2058-6132. DOI: [10.1088/1674-1137/ac8bc9](https://doi.org/10.1088/1674-1137/ac8bc9). *eprint*: [2204.13249\[hep-ex\]](https://arxiv.org/abs/2204.13249). URL: [http://arxiv.org/abs/2204.13249](https://arxiv.org/abs/2204.13249) (visited on 08/11/2023).
- 4376 [10] A. A. Hahn, K. Schreckenbach, W. Gelletly, F. von Feilitzsch, G. Colvin, and B. Krusche. "Antineutrino spectra from  $^{241}\text{Pu}$  and  $^{239}\text{Pu}$  thermal neutron fission products". *Physics Letters B* 218.3 (Feb. 23, 1989). ISSN: 0370-2693. DOI: [10.1016/0370-2693\(89\)91598-0](https://doi.org/10.1016/0370-2693(89)91598-0). URL: <https://www.sciencedirect.com/science/article/pii/0370269389915980> (visited on 01/16/2024).
- 4377 [11] Th. A. Mueller et al. "Improved Predictions of Reactor Antineutrino Spectra". *Physical Review C* 83.5 (May 23, 2011). ISSN: 0556-2813, 1089-490X. DOI: [10.1103/PhysRevC.83.054615](https://doi.org/10.1103/PhysRevC.83.054615). *eprint*: [1101.2663\[hep-ex, physics:nucl-ex\]](https://arxiv.org/abs/1101.2663). URL: [http://arxiv.org/abs/1101.2663](https://arxiv.org/abs/1101.2663) (visited on 01/16/2024).
- 4378 [12] F. von Feilitzsch, A. A. Hahn, and K. Schreckenbach. "Experimental beta-spectra from  $^{239}\text{Pu}$  and  $^{235}\text{U}$  thermal neutron fission products and their correlated antineutrino spectra". *Physics*

- Letters B 118.1 (Dec. 2, 1982). ISSN: 0370-2693. DOI: [10.1016/0370-2693\(82\)90622-0](https://doi.org/10.1016/0370-2693(82)90622-0). URL: <https://www.sciencedirect.com/science/article/pii/0370269382906220> (visited on 01/16/2024).
- [13] K. Schreckenbach, G. Colvin, W. Gelletly, and F. Von Feilitzsch. "Determination of the antineutrino spectrum from  $^{235}\text{U}$  thermal neutron fission products up to 9.5 MeV". *Physics Letters B* 160.4 (Oct. 10, 1985). ISSN: 0370-2693. DOI: [10.1016/0370-2693\(85\)91337-1](https://doi.org/10.1016/0370-2693(85)91337-1). URL: <https://www.sciencedirect.com/science/article/pii/0370269385913371> (visited on 01/16/2024).
- [14] Patrick Huber. "On the determination of anti-neutrino spectra from nuclear reactors". *Physical Review C* 84.2 (Aug. 29, 2011). ISSN: 0556-2813, 1089-490X. DOI: [10.1103/PhysRevC.84.024617](https://doi.org/10.1103/PhysRevC.84.024617). eprint: [1106.0687 \[hep-ex, physics:hep-ph, physics:nucl-ex, physics:nucl-th\]](https://arxiv.org/abs/1106.0687). URL: [http://arxiv.org/abs/1106.0687](https://arxiv.org/abs/1106.0687) (visited on 01/16/2024).
- [15] P. Vogel, G. K. Schenter, F. M. Mann, and R. E. Schenter. "Reactor antineutrino spectra and their application to antineutrino-induced reactions. II". *Physical Review C* 24.4 (Oct. 1, 1981). Publisher: American Physical Society. DOI: [10.1103/PhysRevC.24.1543](https://doi.org/10.1103/PhysRevC.24.1543). URL: <https://link.aps.org/doi/10.1103/PhysRevC.24.1543> (visited on 01/16/2024).
- [16] D. A. Dwyer and T. J. Langford. "Spectral Structure of Electron Antineutrinos from Nuclear Reactors". *Physical Review Letters* 114.1 (Jan. 7, 2015). ISSN: 0031-9007, 1079-7114. DOI: [10.1103/PhysRevLett.114.012502](https://doi.org/10.1103/PhysRevLett.114.012502). eprint: [1407.1281 \[hep-ex, physics:nucl-ex\]](https://arxiv.org/abs/1407.1281). URL: [http://arxiv.org/abs/1407.1281](https://arxiv.org/abs/1407.1281) (visited on 01/16/2024).
- [17] Daya Bay Collaboration et al. "Measurement of the Reactor Antineutrino Flux and Spectrum at Daya Bay". *Physical Review Letters* 116.6 (Feb. 12, 2016). Publisher: American Physical Society. DOI: [10.1103/PhysRevLett.116.061801](https://doi.org/10.1103/PhysRevLett.116.061801). URL: <https://link.aps.org/doi/10.1103/PhysRevLett.116.061801> (visited on 09/06/2024).
- [18] G. Mention, M. Fechner, Th. Lasserre, Th. A. Mueller, D. Lhuillier, M. Cribier, and A. Letourneau. "Reactor antineutrino anomaly". *Physical Review D* 83.7 (Apr. 29, 2011). Publisher: American Physical Society. DOI: [10.1103/PhysRevD.83.073006](https://doi.org/10.1103/PhysRevD.83.073006). URL: <https://link.aps.org/doi/10.1103/PhysRevD.83.073006> (visited on 03/05/2024).
- [19] JUNO Collaboration et al. *TAO Conceptual Design Report: A Precision Measurement of the Reactor Antineutrino Spectrum with Sub-percent Energy Resolution*. May 18, 2020. DOI: [10.48550/arXiv.2005.08745](https://arxiv.org/abs/2005.08745). eprint: [2005.08745 \[hep-ex, physics:nucl-ex, physics:physics\]](https://arxiv.org/abs/2005.08745). URL: [http://arxiv.org/abs/2005.08745](https://arxiv.org/abs/2005.08745) (visited on 01/18/2024).
- [20] Super-Kamiokande Collaboration et al. "Diffuse Supernova Neutrino Background Search at Super-Kamiokande". *Physical Review D* 104.12 (Dec. 10, 2021). ISSN: 2470-0010, 2470-0029. DOI: [10.1103/PhysRevD.104.122002](https://doi.org/10.1103/PhysRevD.104.122002). eprint: [2109.11174 \[astro-ph, physics:hep-ex\]](https://arxiv.org/abs/2109.11174). URL: [http://arxiv.org/abs/2109.11174](https://arxiv.org/abs/2109.11174) (visited on 02/28/2024).
- [21] JUNO Collaboration et al. "JUNO Sensitivity on Proton Decay  $p \rightarrow \bar{\nu}K^+$  Searches". *Chinese Physics C* 47.11 (Nov. 1, 2023). ISSN: 1674-1137, 2058-6132. DOI: [10.1088/1674-1137/ace9c6](https://doi.org/10.1088/1674-1137/ace9c6). eprint: [2212.08502 \[hep-ex, physics:hep-ph\]](https://arxiv.org/abs/2212.08502). URL: [http://arxiv.org/abs/2212.08502](https://arxiv.org/abs/2212.08502) (visited on 08/09/2024).
- [22] Alessandro Strumia and Francesco Vissani. "Precise quasielastic neutrino/nucleon cross section". *Physics Letters B* 564.1 (July 2003). ISSN: 03702693. DOI: [10.1016/S0370-2693\(03\)00616-6](https://doi.org/10.1016/S0370-2693(03)00616-6). eprint: [astro-ph/0302055](https://arxiv.org/abs/astro-ph/0302055). URL: [http://arxiv.org/abs/astro-ph/0302055](https://arxiv.org/abs/astro-ph/0302055) (visited on 01/16/2024).
- [23] Daya Bay et al. *Optimization of the JUNO liquid scintillator composition using a Daya Bay antineutrino detector*. July 1, 2020. DOI: [10.48550/arXiv.2007.00314](https://arxiv.org/abs/2007.00314). eprint: [2007.00314 \[hep-ex, physics:physics\]](https://arxiv.org/abs/2007.00314). URL: [http://arxiv.org/abs/2007.00314](https://arxiv.org/abs/2007.00314) (visited on 07/26/2023).
- [24] J. B. Birks. "CHAPTER 3 - THE SCINTILLATION PROCESS IN ORGANIC MATERIALS—I". *The Theory and Practice of Scintillation Counting*. Ed. by J. B. Birks. International Series of Monographs in Electronics and Instrumentation. Jan. 1, 1964, 39–67. ISBN: 978-0-08-010472-0. DOI: [10.1016/B978-0-08-010472-0.50008-2](https://doi.org/10.1016/B978-0-08-010472-0.50008-2). URL: <https://www.sciencedirect.com/science/article/pii/B9780080104720500082> (visited on 02/07/2024).
- [25] Photomultiplier tube R12860 | Hamamatsu Photonics. URL: [https://www.hamamatsu.com/eu/en/product/optical-sensors/pmt/pmt\\_tube-alone/head-on-type/R12860.html](https://www.hamamatsu.com/eu/en/product/optical-sensors/pmt/pmt_tube-alone/head-on-type/R12860.html).

- [26] Yan Zhang, Ze-Yuan Yu, Xin-Ying Li, Zi-Yan Deng, and Liang-Jian Wen. "A complete optical model for liquid-scintillator detectors". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 967 (July 2020). ISSN: 01689002. DOI: [10.1016/j.nima.2020.163860](https://doi.org/10.1016/j.nima.2020.163860). eprint: [2003.12212\[physics\]](https://arxiv.org/abs/2003.12212). URL: <http://arxiv.org/abs/2003.12212> (visited on 02/07/2024).
- [27] Hai-Bo Yang et al. "Light Attenuation Length of High Quality Linear Alkyl Benzene as Liquid Scintillator Solvent for the JUNO Experiment". *Journal of Instrumentation* 12.11 (Nov. 27, 2017). ISSN: 1748-0221. DOI: [10.1088/1748-0221/12/11/T11004](https://doi.org/10.1088/1748-0221/12/11/T11004). eprint: [1703.01867\[hep-ex, physics:physics\]](https://arxiv.org/abs/1703.01867). URL: <http://arxiv.org/abs/1703.01867> (visited on 07/28/2023).
- [28] JUNO Collaboration et al. *The Design and Sensitivity of JUNO's scintillator radiopurity pre-detector OSIRIS*. Mar. 31, 2021. DOI: [10.48550/arXiv.2103.16900](https://doi.org/10.48550/arXiv.2103.16900). eprint: [2103.16900\[physics\]](https://arxiv.org/abs/2103.16900). URL: <http://arxiv.org/abs/2103.16900> (visited on 02/07/2024).
- [29] Angel Abusleme et al. "Mass Testing and Characterization of 20-inch PMTs for JUNO". *The European Physical Journal C* 82.12 (Dec. 24, 2022). ISSN: 1434-6052. DOI: [10.1140/epjc/s10052-022-11002-8](https://doi.org/10.1140/epjc/s10052-022-11002-8). eprint: [2205.08629\[hep-ex, physics:physics\]](https://arxiv.org/abs/2205.08629). URL: <http://arxiv.org/abs/2205.08629> (visited on 02/08/2024).
- [30] Yang Han. "Dual Calorimetry for High Precision Neutrino Oscillation Measurement at JUNO Experiment". AstroParticule et Cosmologie, France, Paris U. VII, APC, June 2021.
- [31] R. Acquaferro et al. "The OPERA experiment in the CERN to Gran Sasso neutrino beam". *Journal of Instrumentation* 4.4 (Apr. 2009). ISSN: 1748-0221. DOI: [10.1088/1748-0221/4/04/P04018](https://doi.org/10.1088/1748-0221/4/04/P04018). URL: <https://dx.doi.org/10.1088/1748-0221/4/04/P04018> (visited on 02/29/2024).
- [32] JUNO collaboration et al. "Calibration Strategy of the JUNO Experiment". *Journal of High Energy Physics* 2021.3 (Mar. 2021). ISSN: 1029-8479. DOI: [10.1007/JHEP03\(2021\)004](https://doi.org/10.1007/JHEP03(2021)004). eprint: [2011.06405\[hep-ex, physics:physics\]](https://arxiv.org/abs/2011.06405). URL: <http://arxiv.org/abs/2011.06405> (visited on 08/10/2023).
- [33] Hans Th J. Steiger. *TAO – The Taishan Antineutrino Observatory*. Sept. 21, 2022. DOI: [10.48550/arXiv.2209.10387](https://doi.org/10.48550/arXiv.2209.10387). eprint: [2209.10387\[physics\]](https://arxiv.org/abs/2209.10387). URL: <http://arxiv.org/abs/2209.10387> (visited on 01/16/2024).
- [34] Tao Lin et al. "The Application of SNiPER to the JUNO Simulation". *Journal of Physics: Conference Series* 898.4 (Oct. 2017). Publisher: IOP Publishing. ISSN: 1742-6596. DOI: [10.1088/1742-6596/898/4/042029](https://doi.org/10.1088/1742-6596/898/4/042029). URL: <https://dx.doi.org/10.1088/1742-6596/898/4/042029> (visited on 02/27/2024).
- [35] S. Agostinelli et al. "Geant4—a simulation toolkit". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (July 1, 2003). ISSN: 0168-9002. DOI: [10.1016/S0168-9002\(03\)01368-8](https://doi.org/10.1016/S0168-9002(03)01368-8). URL: <https://www.sciencedirect.com/science/article/pii/S0168900203013688> (visited on 02/27/2024).
- [36] J. Allison et al. "Geant4 developments and applications". *IEEE Transactions on Nuclear Science* 53.1 (Feb. 2006). Conference Name: IEEE Transactions on Nuclear Science. ISSN: 1558-1578. DOI: [10.1109/TNS.2006.869826](https://doi.org/10.1109/TNS.2006.869826). URL: <https://ieeexplore.ieee.org/document/1610988?isnumber=33833&arnumber=1610988&count=33&index=7> (visited on 02/27/2024).
- [37] J. Allison et al. "Recent developments in Geant4". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 835 (Nov. 1, 2016). ISSN: 0168-9002. DOI: [10.1016/j.nima.2016.06.125](https://doi.org/10.1016/j.nima.2016.06.125). URL: <https://www.sciencedirect.com/science/article/pii/S0168900216306957> (visited on 02/27/2024).
- [38] Angel Abusleme et al. "Potential to Identify the Neutrino Mass Ordering with Reactor Antineutrinos in JUNO" (May 2024). eprint: [2405.18008](https://arxiv.org/abs/2405.18008).
- [39] Xiangpan Ji, Wenqiang Gu, Xin Qian, Hanyu Wei, and Chao Zhang. *Combined Neyman-Pearson Chi-square: An Improved Approximation to the Poisson-likelihood Chi-square*. Mar. 17, 2019. URL: <https://arxiv.org/abs/1903.07185v3> (visited on 10/03/2024).
- [40] Particle Data Group et al. "Review of Particle Physics". *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 14, 2020). ISSN: 2050-3911. DOI: [10.1093/ptep/ptaa104](https://doi.org/10.1093/ptep/ptaa104). URL: <https://doi.org/10.1093/ptep/ptaa104> (visited on 12/04/2023).

- [41] JUNO Collaboration et al. "JUNO Physics and Detector". *Progress in Particle and Nuclear Physics* 123 (Mar. 2022). ISSN: 01466410. DOI: [10.1016/j.ppnp.2021.103927](https://doi.org/10.1016/j.ppnp.2021.103927). eprint: 2104.02565 [hep-ex]. URL: <http://arxiv.org/abs/2104.02565> (visited on 09/18/2023).
- [42] Leo Breiman, Jerome Friedman, R. A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. New York: Chapman and Hall/CRC, Oct. 25, 2017. ISBN: 978-1-315-13947-0. DOI: [10.1201/9781315139470](https://doi.org/10.1201/9781315139470).
- [43] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics* 29.5 (Oct. 2001). Publisher: Institute of Mathematical Statistics. ISSN: 0090-5364, 2168-8966. DOI: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451). URL: <https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full> (visited on 04/29/2024).
- [44] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. Jan. 29, 2017. DOI: [10.48550/arXiv.1412.6980](https://doi.org/10.48550/arXiv.1412.6980). eprint: 1412.6980[cs]. URL: <http://arxiv.org/abs/1412.6980> (visited on 05/13/2024).
- [45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. June 2016, 770–778. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90). URL: <https://ieeexplore.ieee.org/document/7780459> (visited on 07/17/2024).
- [46] Olga Russakovsky et al. *ImageNet Large Scale Visual Recognition Challenge*. Jan. 29, 2015. DOI: [10.48550/arXiv.1409.0575](https://doi.org/10.48550/arXiv.1409.0575). eprint: 1409.0575[cs]. URL: <http://arxiv.org/abs/1409.0575> (visited on 05/17/2024).
- [47] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Apr. 10, 2015. DOI: [10.48550/arXiv.1409.1556](https://doi.org/10.48550/arXiv.1409.1556). eprint: 1409.1556[cs]. URL: <http://arxiv.org/abs/1409.1556> (visited on 05/17/2024).
- [48] generic-github-user/Image-Convolution-Playground. original-date: 2018-09-28T22:42:55Z. July 15, 2024. URL: <https://github.com/generic-github-user/Image-Convolution-Playground> (visited on 07/16/2024).
- [49] Jason Ansel et al. *PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation*. Publication Title: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24) original-date: 2016-08-13T05:26:41Z. Apr. 2024. DOI: [10.1145/3620665.3640366](https://doi.org/10.1145/3620665.3640366). URL: <https://pytorch.org/assets/pytorch2-2.pdf> (visited on 07/16/2024).
- [50] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE* 86.11 (Nov. 1998). Conference Name: Proceedings of the IEEE. ISSN: 1558-2256. DOI: [10.1109/5.726791](https://doi.org/10.1109/5.726791). URL: <https://ieeexplore.ieee.org/document/726791> (visited on 07/16/2024).
- [51] NVIDIA T4 Tensor Core GPUs for Accelerating Inference. URL: <https://www.nvidia.com/en-gb/data-center/tesla-t4/> (visited on 07/16/2024).
- [52] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. *Neural Message Passing for Quantum Chemistry*. June 12, 2017. DOI: [10.48550/arXiv.1704.01212](https://doi.org/10.48550/arXiv.1704.01212). eprint: 1704.01212[cs]. URL: <http://arxiv.org/abs/1704.01212> (visited on 05/22/2024).
- [53] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. *Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering*. Feb. 5, 2017. DOI: [10.48550/arXiv.1606.09375](https://doi.org/10.48550/arXiv.1606.09375). eprint: 1606.09375[cs, stat]. URL: <http://arxiv.org/abs/1606.09375> (visited on 04/04/2024).
- [54] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. *Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting*. Feb. 22, 2018. DOI: [10.48550/arXiv.1707.01926](https://doi.org/10.48550/arXiv.1707.01926). eprint: 1707.01926[cs, stat]. URL: <http://arxiv.org/abs/1707.01926> (visited on 05/22/2024).
- [55] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. *Generative Adversarial Networks*. June 10, 2014. DOI: [10.48550/arXiv.1406.2661](https://doi.org/10.48550/arXiv.1406.2661). eprint: 1406.2661[cs, stat]. URL: <http://arxiv.org/abs/1406.2661> (visited on 05/29/2024).

- [56] Wenjie Wu, Miao He, Xiang Zhou, and Haoxue Qiao. "A new method of energy reconstruction for large spherical liquid scintillator detectors". *Journal of Instrumentation* 14.3 (Mar. 8, 2019). ISSN: 1748-0221. DOI: [10 . 1088 / 1748 - 0221 / 14 / 03 / P03009](https://doi.org/10.1088/1748-0221/14/03/P03009). eprint: [1812 . 01799 \[hep-ex, physics:physics\]](https://arxiv.org/abs/1812.01799). URL: <http://arxiv.org/abs/1812.01799> (visited on 07/28/2023).
- [57] Guihong Huang et al. "Improving the energy uniformity for large liquid scintillator detectors". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1001 (June 11, 2021). ISSN: 0168-9002. DOI: [10 . 1016 / j . nima . 2021 . 165287](https://doi.org/10.1016/j.nima.2021.165287). URL: <https://www.sciencedirect.com/science/article/pii/S0168900221002710> (visited on 03/01/2024).
- [58] Ziyuan Li et al. "Event vertex and time reconstruction in large volume liquid scintillator detector". *Nuclear Science and Techniques* 32.5 (May 2021). ISSN: 1001-8042, 2210-3147. DOI: [10 . 1007 / s41365 - 021 - 00885 - z](https://doi.org/10.1007/s41365-021-00885-z). eprint: [2101 . 08901 \[hep-ex, physics:physics\]](https://arxiv.org/abs/2101.08901). URL: <http://arxiv.org/abs/2101.08901> (visited on 07/28/2023).
- [59] Gioacchino Ranucci. "An analytical approach to the evaluation of the pulse shape discrimination properties of scintillators". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 354.2 (Jan. 30, 1995). ISSN: 0168-9002. DOI: [10 . 1016 / 0168 - 9002\(94\)00886 - 8](https://doi.org/10.1016/0168-9002(94)00886-8). URL: <https://www.sciencedirect.com/science/article/pii/0168900294008868> (visited on 03/07/2024).
- [60] C. Galbiati and K. McCarty. "Time and space reconstruction in optical, non-imaging, scintillator-based particle detectors". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 568.2 (Dec. 1, 2006). ISSN: 0168-9002. DOI: [10 . 1016 / j . nima . 2006 . 07 . 058](https://doi.org/10.1016/j.nima.2006.07.058). URL: <https://www.sciencedirect.com/science/article/pii/S0168900206013519> (visited on 03/07/2024).
- [61] M. Moszyński and B. Bengtson. "Status of timing with plastic scintillation detectors". *Nuclear Instruments and Methods* 158 (Jan. 1, 1979). ISSN: 0029-554X. DOI: [10 . 1016 / S0029 - 554X\(79\)90170 - 8](https://doi.org/10.1016/S0029-554X(79)90170-8). URL: <https://www.sciencedirect.com/science/article/pii/S0029554X79901708> (visited on 03/07/2024).
- [62] Gui-Hong Huang, Wei Jiang, Liang-Jian Wen, Yi-Fang Wang, and Wu-Ming Luo. "Data-driven simultaneous vertex and energy reconstruction for large liquid scintillator detectors". *Nuclear Science and Techniques* 34.6 (June 17, 2023). ISSN: 2210-3147. DOI: [10 . 1007 / s41365 - 023 - 01240 - 0](https://doi.org/10.1007/s41365-023-01240-0). URL: <https://doi.org/10.1007/s41365-023-01240-0> (visited on 08/17/2023).
- [63] Zhen Qian et al. "Vertex and Energy Reconstruction in JUNO with Machine Learning Methods". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1010 (Sept. 2021). ISSN: 01689002. DOI: [10 . 1016 / j . nima . 2021 . 165527](https://doi.org/10.1016/j.nima.2021.165527). eprint: [2101 . 04839 \[hep-ex, physics:physics\]](https://arxiv.org/abs/2101.04839). URL: <http://arxiv.org/abs/2101.04839> (visited on 07/24/2023).
- [64] Arsenii Gavrikov, Yury Malyshkin, and Fedor Ratnikov. "Energy reconstruction for large liquid scintillator detectors with machine learning techniques: aggregated features approach". *The European Physical Journal C* 82.11 (Nov. 14, 2022). ISSN: 1434-6052. DOI: [10 . 1140 / epjc / s10052 - 022 - 11004 - 6](https://doi.org/10.1140/epjc/s10052-022-11004-6). eprint: [2206 . 09040 \[physics\]](https://arxiv.org/abs/2206.09040). URL: <http://arxiv.org/abs/2206.09040> (visited on 07/24/2023).
- [65] K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. "HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere". *The Astrophysical Journal* 622 (Apr. 1, 2005). ADS Bibcode: 2005ApJ...622..759G. ISSN: 0004-637X. DOI: [10 . 1086 / 427976](https://doi.org/10.1086/427976). URL: <https://ui.adsabs.harvard.edu/abs/2005ApJ...622..759G> (visited on 04/04/2024).
- [66] Anatael Cabrera et al. *Multi-Calorimetry in Light-based Neutrino Detectors*. Dec. 20, 2023. DOI: [10 . 48550 / arXiv . 2312 . 12991](https://arxiv.org/abs/2312.12991). eprint: [2312 . 12991 \[hep-ex, physics:physics\]](https://arxiv.org/abs/2312.12991). URL: <http://arxiv.org/abs/2312.12991> (visited on 08/19/2024).
- [67] Victor Lebrin. "Towards the Detection of Core-Collapse Supernovae Burst Neutrinos with the 3-inch PMT System of the JUNO Detector". These de doctorat. Sept. 5, 2022. URL: <https://theses.fr/2022NANU4080> (visited on 05/22/2024).

- [68] Dan Cireşan, Ueli Meier, and Juergen Schmidhuber. *Multi-column Deep Neural Networks for Image Classification*. version: 1. Feb. 13, 2012. DOI: [10.48550/arXiv.1202.2745](https://doi.org/10.48550/arXiv.1202.2745). eprint: [1202.2745\[cs\]](https://arxiv.org/abs/1202.2745). URL: <http://arxiv.org/abs/1202.2745> (visited on 06/27/2024).
- [69] R. Abbasi et al. “A Convolutional Neural Network based Cascade Reconstruction for the Ice-Cube Neutrino Observatory”. *Journal of Instrumentation* 16.7 (July 1, 2021). ISSN: 1748-0221. DOI: [10.1088/1748-0221/16/07/P07041](https://doi.org/10.1088/1748-0221/16/07/P07041). eprint: [2101.11589\[hep-ex\]](https://arxiv.org/abs/2101.11589). URL: <http://arxiv.org/abs/2101.11589> (visited on 06/27/2024).
- [70] D. Maksimović, M. Nieslony, and M. Wurm. “CNNs for enhanced background discrimination in DSNB searches in large-scale water-Gd detectors”. *Journal of Cosmology and Astroparticle Physics* 2021.11 (Nov. 2021). Publisher: IOP Publishing. ISSN: 1475-7516. DOI: [10.1088/1475-7516/2021/11/051](https://doi.org/10.1088/1475-7516/2021/11/051). URL: <https://dx.doi.org/10.1088/1475-7516/2021/11/051> (visited on 06/27/2024).
- [71] Taco S. Cohen, Mario Geiger, Jonas Koehler, and Max Welling. *Spherical CNNs*. Feb. 25, 2018. DOI: [10.48550/arXiv.1801.10130](https://doi.org/10.48550/arXiv.1801.10130). eprint: [1801.10130\[cs, stat\]](https://arxiv.org/abs/1801.10130). URL: <http://arxiv.org/abs/1801.10130> (visited on 07/13/2024).
- [72] NVIDIA A100 GPUs Power the Modern Data Center. URL: <https://www.nvidia.com/en-gb/data-center/a100/> (visited on 08/06/2024).
- [73] NVIDIA V100. URL: <https://www.nvidia.com/en-gb/data-center/v100/> (visited on 08/06/2024).
- [74] *leonard-IMBERT/datamo*. original-date: 2023-10-17T12:37:38Z. Aug. 9, 2024. URL: <https://github.com/leonard-IMBERT/datamo> (visited on 08/09/2024).
- [75] “IEEE Standard for Floating-Point Arithmetic”. *IEEE Std 754-2019 (Revision of IEEE 754-2008)* (July 2019). Conference Name: IEEE Std 754-2019 (Revision of IEEE 754-2008). DOI: [10.1109/IEEEESTD.2019.8766229](https://doi.org/10.1109/IEEEESTD.2019.8766229). URL: <https://ieeexplore.ieee.org/document/8766229> (visited on 07/03/2024).
- [76] Chuanya Cao et al. “Mass production and characterization of 3-inch PMTs for the JUNO experiment”. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1005 (July 2021). ISSN: 01689002. DOI: [10.1016/j.nima.2021.165347](https://doi.org/10.1016/j.nima.2021.165347). eprint: [2102.11538\[hep-ex, physics:physics\]](https://arxiv.org/abs/2102.11538). URL: <http://arxiv.org/abs/2102.11538> (visited on 02/08/2024).
- [77] K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelman. “HEALPix – a Framework for High Resolution Discretization, and Fast Analysis of Data Distributed on the Sphere”. *The Astrophysical Journal* 622.2 (Apr. 2005). ISSN: 0004-637X, 1538-4357. DOI: [10.1086/427976](https://doi.org/10.1086/427976). eprint: [astro-ph/0409513](https://arxiv.org/abs/astro-ph/0409513). URL: <http://arxiv.org/abs/astro-ph/0409513> (visited on 08/10/2023).
- [78] Teng Li, Xin Xia, Xing-Tao Huang, Jia-Heng Zou, Wei-Dong Li, Tao Lin, Kun Zhang, and Zi-Yan Deng. “Design and development of JUNO event data model\*”. *Chinese Physics C* 41.6 (June 2017). Publisher: IOP Publishing. ISSN: 1674-1137. DOI: [10.1088/1674-1137/41/6/066201](https://doi.org/10.1088/1674-1137/41/6/066201). URL: <https://dx.doi.org/10.1088/1674-1137/41/6/066201> (visited on 08/16/2024).
- [79] *Ducc0*. original-date: 2021-04-12T15:35:50Z. Aug. 9, 2024. URL: <https://gitlab.mpcdf.mpg.de/mtr/ducc> (visited on 08/16/2024).
- [80] Charles R. Harris et al. “Array programming with NumPy”. 585.7825 (Sept. 2020). Publisher: Nature Publishing Group. ISSN: 1476-4687. DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2). URL: <https://www.nature.com/articles/s41586-020-2649-2> (visited on 10/12/2024).
- [81] Measurements of the Lifetime of Orthopositronium in the LAB-Based Liquid Scintillator of JUNO. Apr. 25, 2018. DOI: [10.1016/j.nima.2018.12.068](https://doi.org/10.1016/j.nima.2018.12.068). eprint: [1804.09456\[physics\]](https://arxiv.org/abs/1804.09456). URL: <http://arxiv.org/abs/1804.09456> (visited on 09/17/2024).
- [82] Narongkiat Rodphai, Zhimin Wang, Narumon Suwonjandee, and Burin Asavapibhop. “20-inch photomultiplier tube timing study for JUNO”. *Journal of Physics: Conference Series* 2145.1 (Dec. 2021). Publisher: IOP Publishing. ISSN: 1742-6596. DOI: [10.1088/1742-6596/2145/1/012017](https://doi.org/10.1088/1742-6596/2145/1/012017). URL: <https://dx.doi.org/10.1088/1742-6596/2145/1/012017> (visited on 09/17/2024).

- 4673 [83] Dong-Hao Liao et al. "Study of TTS for a 20-inch dynode PMT\*". *Chinese Physics C* 41.7 (July  
 4674 2017). Publisher: IOP Publishing. ISSN: 1674-1137. DOI: [10.1088/1674-1137/41/7/076001](https://doi.org/10.1088/1674-1137/41/7/076001). URL: <https://dx.doi.org/10.1088/1674-1137/41/7/076001> (visited on 09/17/2024).
- 4675 [84] Nan Li et al. "Characterization of 3-inch photomultiplier tubes for the JUNO central detector".  
 4676 *Radiation Detection Technology and Methods* 3.1 (Nov. 22, 2018). ISSN: 2509-9949. DOI: [10.1007/s41605-018-0085-8](https://doi.org/10.1007/s41605-018-0085-8). URL: <https://doi.org/10.1007/s41605-018-0085-8> (visited on  
 4677 09/17/2024).
- 4678 [85] B. Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil  
 4679 Yogamani, and Patrick Pérez. *Deep Reinforcement Learning for Autonomous Driving: A Survey*.  
 4680 Jan. 23, 2021. eprint: [2002.00444\[cs\]](https://arxiv.org/abs/2002.00444). URL: <http://arxiv.org/abs/2002.00444> (visited on  
 4681 10/02/2024).
- 4682 [86] Oriol Vinyals et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning".  
 4683 575.7782 (Nov. 2019). Publisher: Nature Publishing Group. ISSN: 1476-4687. DOI: [10.1038/s41586-019-1724-z](https://doi.org/10.1038/s41586-019-1724-z). URL: <https://www.nature.com/articles/s41586-019-1724-z>  
 4684 (visited on 10/02/2024).
- 4685 [87] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Delving Deep into Rectifiers: Surpassing  
 4686 Human-Level Performance on ImageNet Classification*. Feb. 6, 2015. URL: <https://arxiv.org/abs/1502.01852v1> (visited on 10/08/2024).
- 4687 [88] Daya Bay Collaboration et al. *A high precision calibration of the nonlinear energy response at Daya  
 4688 Bay*. Feb. 21, 2019. URL: <https://arxiv.org/abs/1902.08241v2> (visited on 10/01/2024).
- 4689 [89] Double Chooz Collaboration et al. "The Double Chooz antineutrino detectors". *The European  
 4690 Physical Journal C* 82.9 (Sept. 8, 2022). ISSN: 1434-6052. DOI: [10.1140/epjc/s10052-022-10726-x](https://doi.org/10.1140/epjc/s10052-022-10726-x). eprint: [2201.13285\[physics\]](https://arxiv.org/abs/2201.13285). URL: <http://arxiv.org/abs/2201.13285> (visited on  
 4691 10/07/2024).
- 4692 [90] Th. A. Mueller et al. "Improved predictions of reactor antineutrino spectra". *Physical Review C*  
 4693 83.5 (May 23, 2011). Publisher: American Physical Society. DOI: [10.1103/PhysRevC.83.054615](https://doi.org/10.1103/PhysRevC.83.054615). URL: <https://link.aps.org/doi/10.1103/PhysRevC.83.054615> (visited on 09/06/2024).
- 4694 [91] X. B. Ma, W. L. Zhong, L. Z. Wang, Y. X. Chen, and J. Cao. "Improved calculation of the energy  
 4695 release in neutron-induced fission". *Physical Review C* 88.1 (July 12, 2013). Publisher: American  
 4696 Physical Society. DOI: [10.1103/PhysRevC.88.014605](https://doi.org/10.1103/PhysRevC.88.014605). URL: <https://link.aps.org/doi/10.1103/PhysRevC.88.014605> (visited on 09/06/2024).
- 4697 [92] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python".  
 4698 *Nature Methods* 17.3 (Mar. 2020). Publisher: Nature Publishing Group. ISSN: 1548-7105. DOI:  
 4699 [10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2). URL: <https://www.nature.com/articles/s41592-019-0686-2> (visited on 08/14/2024).

4708

4709

4710 **Titre :** Méthode Deep Learning and analyse Double Calorimétrique pour la mesure de haute  
 4711 précision des paramètres d'oscillation des neutrinos dans JUNO  
 4712

4713 **Mot clés :** Neutrinos; expérience JUNO; Deep Learning; reconstruction d'IBD; oscillations des  
 4714 neutrinos; double calorimetrie

4715 **Résumé :** JUNO est un observatoire de  
 4716 neutrinos à scintillateur liquide, polyvalent et  
 4717 medium baseline (environ 52 km), situé en  
 4718 Chine. Ses principaux objectifs sont de  
 4719 mesurer les paramètres d'oscillation  $\theta_{12}$ ,  $\Delta m_{21}^2$   
 4720 et  $\Delta m_{31}^2$  avec une précision au pour-mille  
 4721 et de déterminer l'ordre des masses des  
 4722 neutrinos avec un niveau de confiance de  
 4723  $3\sigma$ . Atteindre ces objectifs nécessite une  
 4724 résolution énergétique sans précédent de  
 4725  $3\%/\sqrt{E(\text{MeV})}$  avec cette technologie. Cela  
 4726 demande une compréhension approfondie  
 4727 des divers effets au sein du détecteur. Le

système de double calorimetrie, composé de deux systèmes de mesure distincts observant le même événement, permet non seulement une calibration mais aussi une détection des effets du détecteur avec une grande précision, comme démontré dans cette thèse. Le Deep Learning, un outil de plus en plus utilisé en physique expérimentale, joue un rôle crucial dans cet effort. Dans cette thèse, je présente le développement, l'application et l'analyse des techniques de Deep Learning pour la reconstruction d'évènements dans l'expérience JUNO.

4741

4742 **Title:** Deep learning methods and Dual Calorimetric analysis for high precision neutrino oscil-  
 4743 lation measurements at JUNO  
 4744

4745 **Keywords:** Neutrinos; JUNO experiment; Deep learning; IBD reconstruction; neutrinos Oscil-  
 4746 lation; dual Calorimetry

4747 **Abstract:** JUNO is a multipurpose, medium  
 4748 baseline ( $\sim 52$  km) liquid scintillator neutrino  
 4749 observatory located in China. Its primary  
 4750 objectives are to measure the oscillation  
 4751 parameters  $\theta_{12}$ ,  $\Delta m_{21}^2$ , and  $\Delta m_{31}^2$  with per mil  
 4752 precision and to determine the neutrino mass  
 4753 ordering at a  $3\sigma$  confidence level. Achiev-  
 4754 ing these goals requires an unprecedented  
 4755 energy resolution of  $3\%/\sqrt{E(\text{MeV})}$  with this  
 4756 technology. This demands a comprehensive  
 4757 understanding of the various effects within the

detector. The Dual Calorimetry system—two distinct measurement systems observing the same event—enables not only high-precision calibration but also detection of detector effects, as demonstrated in this thesis. Deep learning, an increasingly powerful tool in physics, plays a critical role in this effort. In this thesis, I present the development, application, and analysis of Deep Learning techniques for reconstruction in the JUNO experiment.

