

Evaluating the Risk of Data Loss Due to Particle Radiation Damage in a DNA Data Storage System

Christopher N. Takahashi¹, David P. Ward¹, Carlo Cazzaniga², Christopher Frost², Paolo Rech³, Kumkum Ganguly⁴, Sean Blanchard⁴, Steve Wender⁴, Bichlien H. Nguyen^{1,5*} and Jake A. Smith^{1,5*}

¹Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA.

²Science and Technology Facilities Council, Swindon, United Kingdom.

³University of Trento, Trento, Italy.

⁴Los Alamos National Laboratory, Los Alamos, NM, USA.

^{5*}Microsoft Research, Redmond, WA, USA.

*Corresponding author(s). E-mail(s): bnguy@microsoft.com; jakesmith@microsoft.com;

Contents

1 Regression Reports

2 Supplementary Figures

3 Monte Carlo Simulation

001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046

2
4
5

047 **1 Regression Reports**

048

049

 σ_{nuc}^n Fitted With Sample Thickness Term

050

051

052

053

$$\log(\sigma^n) = \begin{cases} 0 & \log_{10}(E) < 6 \\ \beta_0 + \beta_1 \log_{10}(E) + \beta_2 \log_{10}(E)^2 + \beta_3 \log_{10}(E)^3 + \beta_4 \log_{10}(Lz) & 6 \leq \log_{10}(E) \leq 10 \\ 0 & \log_{10}(E) > 10 \end{cases}$$

054

055

056

057

058

059

060

061

Dep. Variable:	y	No. Observations:	160
Model:	Poisson	Df Residuals:	155
Method:	MLE	Df Model:	4
Date:	Fri, 15 Dec 2023	Pseudo R-squ.:	0.5226
Time:	12:40:57	Log-Likelihood:	-244.96
converged:	True	LL-Null:	-513.13
Covariance Type:	nonrobust	LLR p-value:	9.281e-115

062

063

064

065

066

067

068

	coef	std err	t	P > t	[0.025	0.975]
β_1	72.9389	10.428	6.995	0.000	52.340	93.538
β_2	-8.5132	1.332	-6.389	0.000	-11.145	-5.881
β_3	0.3223	0.056	5.715	0.000	0.211	0.434
β_4	-0.0192	0.035	-0.548	0.585	-0.089	0.050
β_0	-201.7467	27.038	-7.462	0.000	-255.158	-148.336

069

070

071

072

 σ_{H2O}^n Fitted With Sample Thickness Term

073

074

075

076

$$\log(\sigma^n) = \begin{cases} 0 & \log_{10}(E) < 6 \\ \beta_0 + \beta_1 \log_{10}(E) + \beta_2 \log_{10}(E)^2 + \beta_3 \log_{10}(E)^3 + \beta_4 \log_{10}(Lz) & 6 \leq \log_{10}(E) \leq 10 \\ 0 & \log_{10}(E) > 10 \end{cases}$$

077

078

079

080

081

082

083

084

Dep. Variable:	y	No. Observations:	160
Model:	Poisson	Df Residuals:	155
Method:	MLE	Df Model:	4
Date:	Fri, 15 Dec 2023	Pseudo R-squ.:	0.2322
Time:	13:44:04	Log-Likelihood:	-56.592
converged:	True	LL-Null:	-73.704
Covariance Type:	nonrobust	LLR p-value:	6.704e-07

085

086

087

088

089

090

091

092

	coef	std err	t	P > t	[0.025	0.975]
β_1	144.5223	59.905	2.413	0.017	26.188	262.857
β_2	-17.1974	7.464	-2.304	0.023	-31.942	-2.453
β_3	0.6710	0.308	2.178	0.031	0.062	1.280
β_4	0.0519	0.162	0.320	0.750	-0.269	0.373
β_0	-400.3496	159.323	-2.513	0.013	-715.074	-85.625

σ_{nuc}^n Fitted With Fraction GC Content Term	093
	094
$\log(\sigma^n) = \begin{cases} 0 & \log_{10}(E) < 6 \\ \beta_0 + \beta_1 \log_{10}(E) + \beta_2 \log_{10}(E)^2 + \beta_3 \log_{10}(E)^3 + \beta_4 F_{GC} & 6 \leq \log_{10}(E) \leq 10 \\ 0 & \log_{10}(E) > 10 \end{cases}$	095
	096
	097
<hr/>	098
Dep. Variable: y	No. Observations: 160
Model: Poisson	Df Residuals: 155
Method: MLE	Df Model: 4
Date: Fri, 15 Dec 2023	Pseudo R-squ.: 0.5234
Time: 13:48:33	Log-Likelihood: -244.55
converged: True	LL-Null: -513.13
Covariance Type: nonrobust	LLR p-value: 6.124e-115
	105
coef std err t P > t [0.025 0.975]	106
<hr/>	107
x1 31.8624 4.542 7.015 0.000 22.890 40.835	108
x2 -1.6153 0.252 -6.409 0.000 -2.113 -1.117	109
x3 0.0266 0.005 5.735 0.000 0.017 0.036	110
x4 1.9635 1.821 1.078 0.283 -1.633 5.560	111
const -203.6525 27.153 -7.500 0.000 -257.290 -150.015	112
	113
<hr/>	114
$\sigma_{H_2O}^n$ Fitted With Fraction GC Content Term	116
	117
$\log(\sigma^n) = \begin{cases} 0 & \log_{10}(E) < 6 \\ \beta_0 + \beta_1 \log_{10}(E) + \beta_2 \log_{10}(E)^2 + \beta_3 \log_{10}(E)^3 + \beta_4 F_{GC} & 6 \leq \log_{10}(E) \leq 10 \\ 0 & \log_{10}(E) > 10 \end{cases}$	118
	119
	120
<hr/>	121
Dep. Variable: y	No. Observations: 160
Model: Poisson	Df Residuals: 155
Method: MLE	Df Model: 4
Date: Fri, 15 Dec 2023	Pseudo R-squ.: 0.2341
Time: 13:51:05	Log-Likelihood: -56.453
converged: True	LL-Null: -73.704
Covariance Type: nonrobust	LLR p-value: 5.880e-07
	127
coef std err t P > t [0.025 0.975]	129
<hr/>	130
x1 61.3290 25.954 2.363 0.019 10.061 112.597	131
x2 -3.1682 1.405 -2.255 0.026 -5.943 -0.393	132
x3 0.0537 0.025 2.130 0.035 0.004 0.103	133
x4 -5.7263 9.559 -0.599 0.550 -24.609 13.157	134
const -389.2173 159.205 -2.445 0.016 -703.708 -74.726	135
	136
	137
	138

139 **2 Supplementary Figures**

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

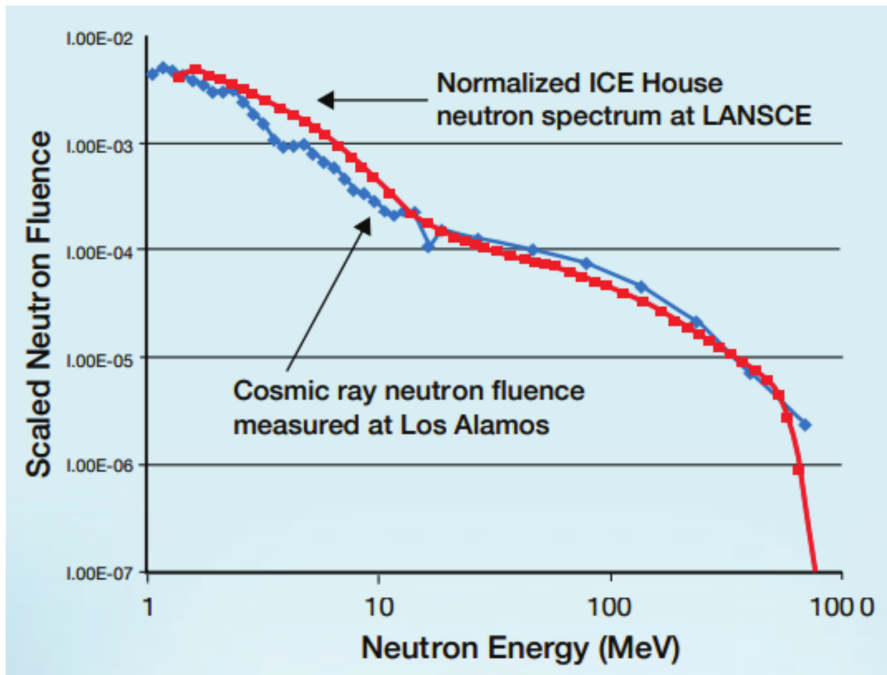
159

160

161

162

163



164 **Fig. 1** Plot of the neutron energy spectrum measured at LANSCE ICE as compared to the
 165 background cosmic ray neutron fluence measured at LANSCE. Retrieved from the LANSCE
 166 ICE House testing manual.

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

3 Monte Carlo Simulation

The supplementary software file `run_simulation.py` was used for approximation of σ_{nuc}^n and σ_{H2O}^n by Monte Carlo simulation. To replicate, please take the following steps.

- Prepare a Python environment that includes the packages `numpy`, `pandas`, and `scipy`.
- Divide the pre-formatted data files found in tabs of the provided Excel sheet into a series of `.csv` files.
- Set an environmental variable `CS_DATA_DIR` containing the path to the separated data files.
- For a given combination of neutron energy and target DNA depth, execute `run_simulation.py`.

```
1 python run_simulation.py <neutron_energy> <dna_depth> <output_path>
```

185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230