

Elementary Particle Identification using Evolvable Discriminate Function and Grid

James Cunha Werner

University of Manchester, Schuster Laboratory, Brunswick Street, Manchester M13 9PL,
United Kingdom
jamwer2000@hotmail.com

Abstract. This paper addresses the use of genetic programming evolving discriminate functions to distinguish neutral pions from background in hadronic tau decays using grid farms and EasyGrid job submission system (an intermediate layer between grid and user's software). Accurate discrimination is important to obtain high purity samples with small systematic errors that allows ratify theoretical predictions with experimental results. Four different data selections criteria according to gammas energies and conventional cut were used to validate our method. A comparison between real data using the same data selection criteria and the discriminate shows our approach is better to discriminate real neutral pions from background than the conventional approach. Monte Carlo and experiment data adherence on neutral pion energy distribution proofs method's effectiveness.

1. INTRODUCTION

The GridPP collaboration [1],[2] and several other e-science projects have provided a distributed infrastructure and software middleware in UK. The LCG (Large Hadrons Collider Computing Grid) [3]-[5] software, developed by an international collaboration centered at CERN, provides a batch processing system for High Energy Physics (HEP) through thousands of computers connected by the Internet.

HEP experimental analysis is very complex due independent sources of background and contamination from different decays with same final products, which can be misclassified. It is a challenge select a sample of events from specific decays with high purity, be able to perform a qualitative analysis (not only evaluate branch rates and widths from histograms), and understand the processes that explain data behavior.

Events were selected imposing constraints in data variables (called "cuts"). If the criteria are too narrow the sample will have high purity, however, the number of selected events will be low and statistical error will increase, which depreciates possible findings.

Genetic programming (GP) [6]-[10] has been used to determinate cuts to maximize event selection [11]-[13]. Genetic algorithms can also be associated with neural networks to implement discriminate functions [14] for Higgs boson.

The problem we will be dealing with is discriminate neutral pion reconstruction with two gammas decays from background with same $135 \text{ MeV}/c^2$ invariant mass. Our approach is innovative because the mathematical model obtained with GP maps the variables hyperspace to a real value through the discriminator function, an algebraic function of kinematics variables. Applying the discriminator to a given pair of gammas, if the discriminate value is bigger than zero, the pair of gammas is deemed to come from pion decay. Otherwise, the pair is deemed to come from another (background) source.

This paper starts describing BaBar experiment (section 2), LCG architecture (section 3), and our grid-GP code (section 4). Four different Neutral Pion Discriminate Functions (NPDF) were obtained with different cuts and described in section 5. Their convergence to the same results is shown in section 7 to support our claim about method's stability and robustness.

Section 6 shows their accuracy compared with the same cuts performed in real data and Monte Carlo events. Finally, the NPDF model without cut in energy is applied to data and Monte Carlo to produce neutral pion energy distributions.

2. The BaBar High Energy Experiment and elementary particles.

The BaBar experiment [15]-[17] studies the differences between matter and antimatter, to throw light on the problem, posed by Sakharov, of how the matter-antimatter symmetric Big Bang can have given rise to today's matter-dominated universe. High energy collisions between electrons (matter) and positrons (antimatter) produce other elementary particles, giving tracks and clusters which are recorded by several high granularity detectors and from which the properties of the short-lived particles can be deduced. Tau leptons pairs can decay in rho resonance, neutral pions (NP), charged pions, muons, etc. The decays we will use to evolve discriminate function using GP are tau decaying in tau neutrino, rho(770) resonance, plus N neutral pions (N=0, 1, 2, and 3). The neutral pion we will study is the one from rho decay in charged pion + neutral pion, and not the others neutral pions deriving from tau decay. These decays were selected by its characteristic branch rates (25.86%, 9.36%, 1.21%, and 0.0016% respectively) to evaluate the effect of number of events in the method.

Neutrinos will not be mentioned in decay labels because they do not interact with the detector, but they carry a reasonable amount of energy and have to be considered in other analysis such as reconstruct tau leptons – not our case study.

Each tau from collision electron-positron could decay producing neutral pions, and would be difficult to classify the event properly, e.g. combining charged pion from one tau decay with neutral pion from the other tau decay to reconstruct rho(770).

To overcome this, the first condition (called "cut") for event selection was one muon must be present in the event (called "tag decay"), because tau decays in muon, anti-muon neutrino and tau neutrino without neutral pions and pions, avoiding any combination from different tau particle origins. Despite BaBar particle identification package provides several muon selectors that are 80%+ accurate, this is a source of error that can distort energy distribution.

The second cut was events with the other tau decaying only with one charged pion. BaBar charged pion selector accuracy is also 80%+.

Neutral pion invariant mass is reconstructed using combination of 2 gammas that have invariant mass between 90 and 165 MeV, given by:

$$M^2 = \left(\sum E_i\right)^2 - \left(\sum P_i\right)^2 \quad (1)$$

where M is the invariant mass, E is the energy and P is the total momentum of each gamma considered in the reconstruction.

Reconstruction algorithm took in account that every gamma should be used only once (one gamma cannot come from 2 different neutral pions) and should minimize error in the invariant mass. If a pair of gammas is an invariant mass candidate, there would not be a better pair of available gammas with lower invariant mass error.

Rho(770) is a broad resonance with large invariant mass width, and its reconstruction requires the charged pion combined with one neutral pion must have invariant mass between 500 MeV and 1.0GeV and minimize error with invariant mass 770 MeV in the same way described for gammas combination.

Energy distribution was adopted due its classical and easy to understand characteristics to evaluate the method instead some obscure quantum relativistic observable.

Combination of gammas produces real particles and *background with the correct invariant mass just by chance*, a major source of systematic error.

Another source of errors are *events with missing gammas (contamination)*. If the event has 3 neutral pions, but one gamma is missing, the event will be classified as 2 neutral pions. They are real neutron pions, with correct observable variables, but they are in the wrong group and will change distribution shape.

3. LCG e-Science grid architecture and gridification techniques.

Grid middleware, working over the Internet, provides the necessary hardware infrastructure grouped in functional modules published through catalogues and tags. LCG grid implementation has been used in HEP experiments due its characteristics of independent parallel processing and modularity, with more than 200 independent sites and thousands of CPUs around the world. It can be seen as a homogeneous common ground in a heterogeneous platform, available under user's request.

There are worker nodes (WN) with access to storage elements (SE), managed by Computer Elements (CE), running jobs distributed by resource brokers (RB). RB manages resources available against policies previously defined by the management board for each experiment, according to limits and priorities. CE runs batch system managers such as PBS or Condor using queues, and have the mission of allocate efficiently resources in the farm and deliver software's results back to the RB after processing.

To access these resources, users submit their jobs using EasyGrid[18],[19] software. It is an intermediate layer between Grid middleware and user's software. It

integrates data, parameters, software, and grid middleware doing all submission and management of several users' software copies to grid.

The gridification algorithm is a library with several functions to run GP software on the grid distributing fitness evaluation. The algorithm implements a master / slave architecture. The master manages a task queue that contains tasks each WN can perform independently. One task can store data for a set of individual cases (service string) to overcome problems with communication delays between master/slaves.

Table 1 shows the time expended running standalone and with several numbers of WN, with good performance: 10 WN should reduce the time in ideal conditions to 10%, and our implementation achieved 24% despite all necessary communication overheads, security validation, and access control.

Table 1. Execution time for the same software with different number of slaves and nodes.

	Standalone	1node / 2 slaves	5 nodes / 10 slaves
Time(1,000s)	80	47	19
Improvement		58%	24%

The testbed platform available at University of Manchester has 1 RB; 1 CE; 6 WN; 1 SE; 1 BDII, Proxy manager (PX), and Grid monitor database; 2 NFS file servers with 1.7 Terabytes each. The production farm contains 1 CE and 28 WNs, sharing other systems with the testbed.

4. Genetic Programming implementation for grid.

GP is an optimization algorithm that mimics the evolution and improvement of life through reproduction. Each individual contributes with its own genetic information (chromosome) to the building of new ones (offspring) adapted to the environment with higher chances of surviving. This is the basis of genetic algorithms and programming. Specialized Markov Chains underline the theoretical bases of this algorithm, changes of states and searching procedures.

GP implementation uses Reverse Polish Notation to store its trees. The population size is 500 individuals; crossover and mutation probabilities are 60% and 20% respectively. Every generation, 20 best individuals are copied as they are (without crossover and mutation) and half population is generated randomly and replace the worse individuals. Algebraic operators have been used with kinematics data.

The Fitness function evaluates the quality of chromosome as a solution to the problem, obtained with datasets with necessary information to rank the solution. The dataset is divided in two parts: one is for training and the second for validation. The training dataset is used to obtain the model and the validation dataset is used to measure the accuracy of the model with data that was not used in training.

Genetic programming expends most computational effort evaluating fitness functions when dealing with complex systems. Each generation hundreds of individuals have their chromosome decoded into the problem solution that is tested against data. The service we have distributed in grid was fitness evaluation, in parallel by many WN, using Monte Carlo events.

The number of true negative (NTN), true positive (NTP), false negative (NFN), and false positive (NFP) are used to rank the final solution:

$$\mathbf{a} = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad \mathbf{b} = \frac{N_{TN}}{N_{TN} + N_{FP}} \quad \mathbf{g} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \quad (2)$$

where α is the Sensitivity, β is the Specificity, and γ is the accuracy. Sensitivity is the probability that a test result will be positive when the condition is true (true positive rate, expressed as a percentage). Specificity is the probability that a test result will be negative when the condition is false (true negative rate, expressed as a percentage). Accuracy is the probability of correct forecasts.

5. Evolving Neutral Pion Discriminator Functions.

Neutral pion discriminator (NPDF) can be obtained using genetic programming with training datasets with real and fake neutral pions, aiming to maximize NPDF total number of correct predictions (accuracy).

These datasets were obtained from Monte Carlo (MC) generators that integrate particle decays models with detector's system transfer function. MC events contain all information from each track particle and gamma radiation, which allows select high purity training datasets (96%+). The information is structured in decays' data lists without detector interaction (MC truth), and data simulation as if the event has been read from the detector.

Events with real neutral pion were selected and marked as "1". Events without real pions into MC truth and invariant mass reconstruction in the same region of real neutral pions where also selected and marked as "0".

Kinematics data from each gamma used in the reconstruction were written in the datasets: angles of the gamma rays, 3-vector momentum, total momentum, and energy in the calorimeter. To avoid unit problems and the need for typed-GP, we use sine and cosine values for each angle measured in the genetic trees. All other attributes are measured in GeV (1,000 million electron-volts).

There are four different conditions we have obtained NPDF taking in account gammas total energy: a. all gammas without energy cut (60,000 real and background records for training, and 60,000 real and 44527 background for test), b. more energetic than 30 MeV electronics' noise threshold (32,000 real and background records for training and test), c. more energetic than 50 MeV (15,000 real and background records for training and test), and finally, d. more energetic than 30MeV, lateral moment between 0.0 and 0.8, and have hit more than one crystal in the electromagnetic calorimeter - the conventional cut for neutral pion(16,000 real and background records for training and test).

Table 2 shows the results for training and test. All results were in agreement and show high specificity, fundamental to study observable variables from neutral pion particles. The best result was without energy cuts, with specificity $\beta=89\%$ (only 11% of contamination) and sensitivity $\alpha=72\%$ (lost of 28% of good data, which increase of statistical error).

Table 2. Training and test results of discriminate functions from different gamma energy threshold. α , β and γ defined in equation (I).

Steps	Real	a. all gammas		b. E>30MeV		c. E>50 MeV		d. Usual Cuts	
		D<0	D>0	D<0	D>0	D<0	D>0	D<0	D>0
Train	0	53559	6441	28143	3857	12147	2853	13620	2380
	1	15834	44166	6316	25684	2180	12820	2400	13600
	a	74		80		85		85	
	b	89		88		81		85	
	g	81		84		83		85	
	Test	Real	D<0	D>0	D<0	D>0	D<0	D>0	D<0
0		39531	4996	27510	4490	11799	3201	13319	2681
1		16480	43520	6726	25274	2222	12778	2566	13434
a		72		79		85		84	
b		89		86		79		83	
g		79		82		82		84	

6. NPDF Proof-of-case with real data.

Lets consider the case where neutral pions were reconstructed with gammas more energetic than 50MeV, to avoid any noise from electronics and detector. Table 3 shows comparative results between real tau decays to neutral pions data (482,303,947 BaBar's detector events and 20,489,668 MC events), and discriminator's result obtained with all gammas data and NPDF obtained with training data with gammas more energetic than 50 MeV (Table 2, column c).

Each event classified as 1, 2, 3, or 4 neutral pions decays (following the cuts described in section 2) have their MC list checked and classified again in columns *a* and *b* (Table 3) according to its real number of neutral pions. For example, one neutral pion decays forecasts are 73% of times correct in real data with cuts and 78% correct using NPDF. They are actually 2 NP 22% and 17% of time, respectively.

Figs 1 and 2 show adherence between NPDF using MC and experiment data. The results in low energy are more realistic using NPDF than the energy cut, despite there still have some background at very low energy. The peaks in low energy also bias the distribution down.

Adherence between the two different approaches in high energy shows NPDF can be reproduced with same cuts in the dataset. The low energy peak in experiment data that is significantly reduced in the NPDF curve shows the benefit of the method over conventional cuts.

7. Solution robustness under different energy cuts.

The comparison between the different results obtained by each NPDF can check experimental results for unsuspected systematic effects. If the values agree despite having been found by differing techniques, that argues that the result is stable and free of unsuspected errors [20].

Table 3. Comparison between real data with energy cut and NPDF applied in neutral pion reconstruction for background discrimination. Values in percentage per forecasted value.

Real # decays		a. Real data gamma E>50MeV					b. NPDF gamma E>50MeV				
		0	1	2	3	4	0	1	2	3	4
Fore Cast	1	1	73	22	2	0	2	78	17	1	0
	2	0	5	74	16	2	1	6	73	16	1
	3	0	2	11	67	16	0	3	10	68	15
	4	0	0	2	14	66	0	0	2	6	76

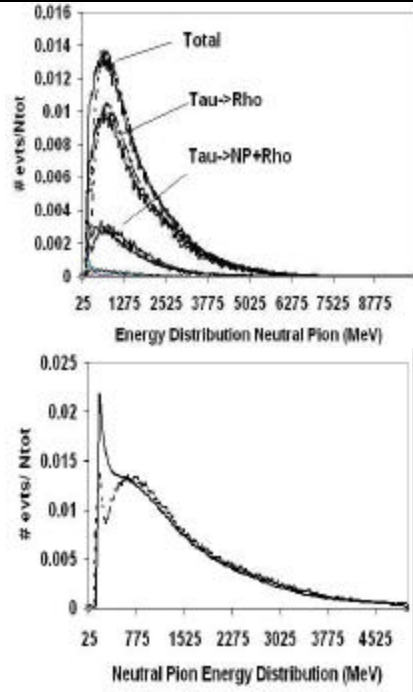
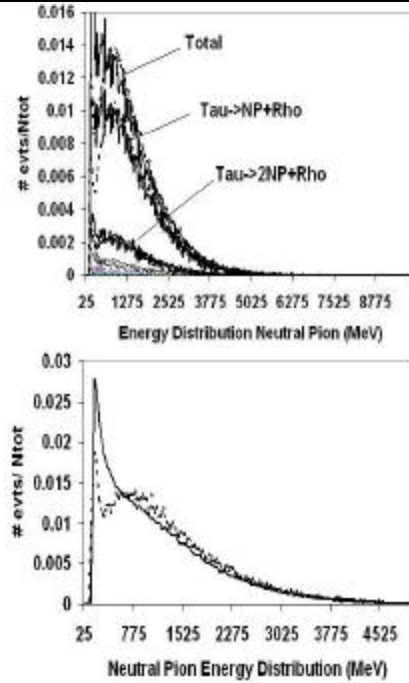

Fig. 1. Comparison of each decay contribution in Neutral pion energy distribution between MC (up) and experimental data (down) with energy cut (dark line) and using NPDF (dashed line) from tau -> rho(770)

Fig. 2. Idem as Fig 1 from tau -> rho(770) + neutral pion.

Fig. 3 and 4 show neutral pion energy distribution for tau decaying in 1 and 2 neutral pions from MC data for all NPDF. Contributions from fake neutral pions (background) were added in one plot, real decays in other, and decays with more neutral pions that have been misclassified due gamma missing are added together in another plot.

The plots show concordance between the four NPDF, even in the low energy peak.

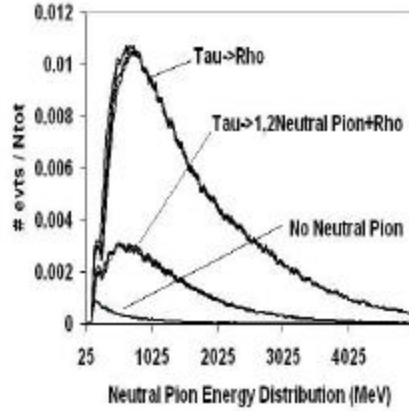


Fig. 3. Superposition of all four NPDF in Tau -> Rho(770).

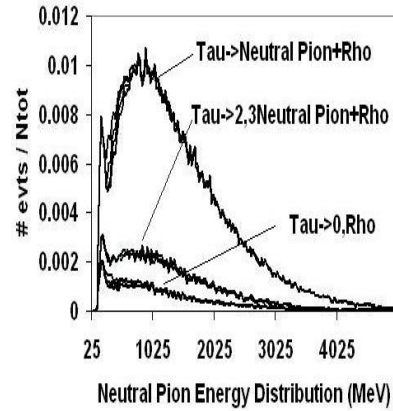


Fig. 4. Superposition of all four NPDF in Tau -> Neutral Pion+Rho(770).

8. Neutral Pion Energy distribution with NPDF.

Figs 5 shows cumulative plot of energy distribution for 1, 2, and 3 neutral pion decays using all gammas NPDF (Table 2,a). Contamination effect can be seen from MC energy distribution, and its effect in energy distribution. The abrupt cut in data's low energy can be result of electronics' threshold, and can bias the curve up in comparison with MC data.

Conclusion.

The paper described genetic programming approach to obtain neutral pion discriminate function to discern between background and real neutral pion particles. Background can produce a critical influence in systematic errors and constrain qualitative analysis.

Usually, a complex analysis and deep understanding are required to design selection criteria (cuts) with high purity. Genetic programming can be applied to refine preliminary trivial cuts, with expressive research time saving and good quality results.

Results from hadronic tau decays analyzed in this paper showed genetic programming discriminate function has an important role in background reduction, improving analysis quality. High purity samples were selected from massive datasets, and grid computing was used to host the development with high efficiency and availability.

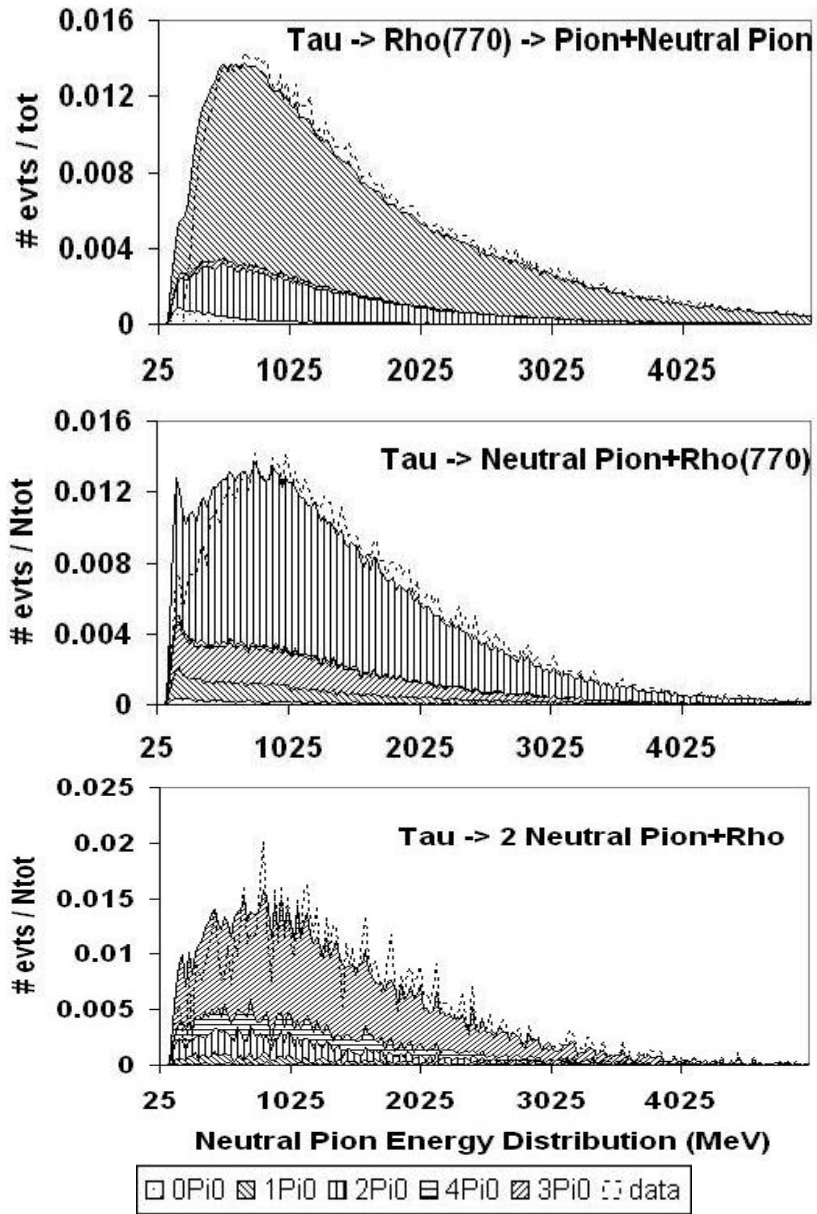


Fig. 5. Neutral pion energy distribution cumulative plots show adherence between MC and experiment data for hadronic tau decays.

The adherence between selection using NPDP shows this is a convenient and reliable technique to discriminate between real neutral pions and background with 80% accuracy, 87% sensitivity, and 84% specificity. The use of NPDP will allow the study of observable and ratify values obtained from theoretical Standard Model, with a high purity sample of events.

The author thanks GridPP and PPARC for funding this project, and the BaBar collaboration for granting access to their data.

REFERENCE

- [1] GridPP site: <http://www.gridpp.ac.uk/>
- [2] The GridPP Collaboration: P J W Faulkner et al "GridPP: development of the UK computing Grid for particle physics" 2006 J. Phys. G: Nucl. Part. Phys. 32 N1-N20 doi:10.1088/0954-3899/32/1/N01
- [3] CERN site: <http://public.web.cern.ch/Public/Welcome.html>
- [4] LHC site: <http://lhc.web.cern.ch/lhc/>
- [5] LCG site: <http://lcg.web.cern.ch/LCG/>
- [6] Holland, J.H. "Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence." Cambridge: Cambridge press 1992.
- [7] Goldberg, D.E. "Genetic Algorithms in Search, Optimisation, and Machine Learning." Reading, Mass.: Addison-Wesley, 1989.
- [8] Chambers, L.; "The practical handbook of Genetic Algorithms" Chapman & Hall/CRC, 2000.
- [9] Koza, J.R. "Genetic programming: On the programming of computers by means of natural selection." Cambridge, Mass.: MIT Press, 1992.
- [10] Werner, J.C.; "Active noise control in ducts using genetic algorithm" PhD. Thesis- São Paulo University- São Paulo -Brazil-1999.
- [11] Cranmer, K.; Bowman, R.S.; "PhysicsGP: A genetic programming approach to event selection" Computer Physics Communications 167 (2005) 165-176.
- [12] Focus Collaboration, "Application of genetic programming to high energy physics event selection" Nuclear instruments and methods in physics research A 551 (2005) 504-527.
- [13] Focus Collaboration; "Search for $L+c \rightarrow pK+p-$ and $D+s \rightarrow K+K+p-$ using genetic programming event selection" Physics letters B 624 (2005) 166-172
- [14] Mjahed, M.; "Search for Higgs boson at LHC by using genetic algorithms" To be published in Nuclear Instruments and Methods in Physics Research.
- [15] BaBar Collaboration, "The BaBar experiment home page", <http://www.slac.stanford.edu/BFROOT/>
- [16] Harrison, P.F.; Quinn, H.R.; "The BaBar Physics Book" SLAC Report 504, October 1998, available at <http://www.slac.stanford.edu/pubs/slacreports/slac-r-504.html>
- [17] BaBar Collaboration; "The BaBar detector", Nuclear Instruments and Methods in Physics Research A479(2002) 1-116 available at <http://www.hep.man.ac.uk/u/jamwer/babarnucl.pdf>
- [18] Werner, J.C.; "HEP analysis, Grid and EasyGrid Job Submission Prototype: Babar/CM2 showcase" at <http://www.hep.man.ac.uk/u/jamwer/>
- [19] Werner, J. C.; "Grid computing in High Energy Physics using LCG: the BaBar experience" e-Science All Hands Meeting AHM2006, 18-21 September 2006, Nottingham, UK
- [20] Barlow, R.; "Evaluating Systematic Errors" Manchester Particle Physics Internal Report MAN/HEP/93/9, November 29, 1993.