

# High purity Neutral Pions selection using Genetic Programming Discriminate Function

James Cunha Werner  
University of Manchester  
jamwer2000@hotmail.com  
[www.geocities.com/jamwer2002/](http://www.geocities.com/jamwer2002/)

## Abstract

This paper introduces genetic programming discriminate function to discern between neutral pion and background. The target was to obtain a high purity sample that allows qualitative analysis in High Energy Physics. Genetic programming discriminate functions achieved 82% accuracy, 81% sensitivity, and 83% specificity. Background and contamination were studied in detail using Monte Carlo events generator and showed how the discriminate affects data. Several hadronic tau decays were analysed and the case worked out neutral pion and gamma energy distribution from  $\rho(770)$  decays showing good agreement with predicted values by Monte Carlo simulation.

## 1. Introduction

High Energy Physics (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 obtains a sample of events from specific decays with high purity, be able to perform a qualitative analysis (not only evaluate branch rates and widths), and understand the processes that explain data behaviour. 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.

There are models that foreseen the interaction between matter constituents and the detector. Experimental predictions can be made using these models with Monte Carlo Simulators. High Energy Physics goal is to fit data from the experiment with Monte Carlo Simulation.

Genetic programming (GP) [1]-[4] has been used to determinate cuts to maximize event selection [5]-[7]. Genetic algorithms can also be associated with neural networks to implement discriminate functions [8] for Higgs boson.

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 pion 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. (The neutral pion, with a mass of  $135 \text{ MeV}/c^2$ , decays to two photons.) Otherwise, the pair is deemed to come from another (background) source. Discriminate functions have been successfully applied to medical diagnostics [9]-[12].

This paper introduces HEP methodology (section 2) and describes genetic programming and the criteria adopted to analyse the results (section 3). Due its computational requirements, grid computing was used (section 4). Discriminate function was described in section 5, filter efficiency in section 6, and results filtering raw data in section 7. Our goals were obtain neutral pions energy distributions from rho decays and gammas energy distribution from neutral pions decays, showing agreement between Monte Carlo and experiment data using the discriminate function (section 8).

## 2. HEP experimental methodology

The BaBar experiment [13]-[15] 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 (tau leptons, pions, kaons, etc), 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 (Fig.1).

The decays we will use to analyse discriminate function are tau decaying in tau neutrino, charged pion, and N neutral pions (N=1, 2, 3, and 4). These decays were selected by its characteristic branch rates (25.86%, 9.36%, 1.21%, and 0.0016% respectively) [16]-[21] as benchmark for grid computing and discriminate approach. Electron and positron collide in BaBar detector and can produce events with a pair of tau particles. The first condition (cut) for event selection was only one muon in the event (tag decay). BaBar particle identification package provides several muon selectors. The second cut was for events with only one charged pion, and finally only gammas with energy bigger than 50 MeV should be considered to avoid noise from electronic modules.

*In this paper we will refer to "Tau decay in tau neutrino, rho(770) resonance which will decay in charged pion and 1 neutral pion" as "1 neutral pion" or "1 $\pi^0$ ". In case of two, three, and four neutral pion decays, the neutral pion of interest is always the one coming from rho(770) decay with one charged pion.*

Neutral pion invariant mass is reconstructed using combination of 2 gammas that have invariant mass between 130 and 150 MeV, given by:

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

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

The algorithm must take in account that every particle should be used only once. One gamma cannot come from 2 different neutral pions. The combination 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.

Combination of gammas (Fig. 2 a) produces real particles and *background with the correct invariant mass just by chance*. Usual kinematics equations (Fig 2b)  $\sqrt{2E_1E_2(1-\cos\mathbf{q}_{12})}-0.135$  are not able to discriminate between them. It is possible to fit functions for background and peak and obtain branch rates and widths without discriminate them. However, neutral pion energy distribution will require discriminating between real events and background from reconstructions with invariant mass inside the peak region. This is 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. Genetic Programming

GP [1]-[4] is an optimisation algorithm that mimics the evolution and improvement of life through reproduction. Each individual contributes with its own genetic information 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 was implemented [30] using trees in Reverse Polish Notation in a recursive algorithm.

**Chromosome representation.** The chromosome represents the model of the problem solution using trees. A tree is a model representation that contains nodes and leaves. Nodes are mathematical operators. We have used multiplication, addition, subtraction, and division. Leaves are terminals (the attributes of the dataset and random numbers). The discriminator function in a GP context is a tree using operators and leaves.

**Genetic operators.** Trees are manipulated through genetic operators. The crossover operator points a tree branch and exchanges it with another branch and obtains new trees. The mutation operator changes the branch for a random new branch. The length of the chromosome is variable.

The probability of crossover is 60% and the probability of mutation is 20%. We adopt a high value of the mutation probability to spread the population over all solution space.

**Fitness function.** The Fitness function defines the quality of chromosome as a solution to the problem. It is a numerical positive value. 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.

The fitness function evaluates how accurate the mathematical model coded in chromosome is, over all the training dataset counting the number of times the discriminator function is correct.

Receiver Operating Characteristics (ROC) evaluates the accuracy using the number of true negative ( $N_{TN}$ ), true positive ( $N_{TP}$ ), false negative ( $N_{FN}$ ), and false positive ( $N_{FP}$ ):

$$\mathbf{a} = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad \mathbf{b} = \frac{N_{TN}}{N_{TN} + N_{FP}} \quad \mathbf{g} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \quad (1)$$

where  $\mathbf{a}$  is the *Sensitivity*,  $\mathbf{b}$  is the *Specificity*, and  $\mathbf{g}$  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.

#### 4. LCG grid architecture and Genetic programming

The GridPP collaboration [22][23] and several other e-science projects have provided a distributed infrastructure and software middleware in UK. The LCG (Large Hadrons Collider Computing Grid) [24]-[26] software, developed by an international collaboration centred at CERN, provides a system for batch processing for High Energy Physics (HEP) through hundreds of computers connected by the Internet. Grid middleware, working over the Internet, provides the necessary hardware infrastructure grouped in functional modules. It can be seen as a homogeneous common ground in a heterogeneous platform.

The testbed platform available at University of Manchester [27] contains 6 worker nodes (WN) with 2 CPUs per node and 100Mbits/s network cards computers. The production farm contains 28 WN.

The gridification algorithm used to run GP in grid is a library with several functions to run conventional software on the grid doing functional parallelism, with minor changes in the source code. The algorithm implements a master / slave architecture. The master manages a task queue that contains elementary tasks each slave 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. The software was implemented using PVM commands [28][29].

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

## 5. Neutral Pion discriminator function

Experimental analysis uses Monte Carlo (MC) generators with particle decays + detector system transfer function. MC events contain all information from each track particle and gamma radiation, which allows event selection for training dataset without mistakes.

Three datasets were built. One for training with 57,992 records (dataset I) was used to obtain several discriminates and tests their accuracy with the second dataset of 302,374 records (dataset II). The next stage was training GP with dataset II, and applies the discriminate in a complete dataset of 4,890,000 events (dataset III) to evaluate the impact of contamination and filtering in the data (section 6). Events with one real neutral pion were selected and marked as 1. Events without real pions and invariant mass reconstruction in the same peak region of real neutral pions where also selected and marked 0.

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

Table I shows the results for training and test of 3 different runs. All results where in agreement and shows high specificity, fundamental to study observable variables from neutral pion particles. High specificity means there will be low background in the sample (less then 10%). Sensitivity of 74% means there will be a lost of 26% of real neutral pions from the sample, with decrease in total number and increase of statistical error.

If dataset II is used in training, the discriminator function obtained by genetic programming is:

$$D = 3*ener1+ener2+sinteta2+sinteta1-2.5428$$

and the analysis can be seen in table II. Accuracy was 82%, sensitivity 81%, and specificity 83% - equation (1), improving sensitivity and reducing lost to 17%. Fig 2c shows discriminate values for dataset I.

A better performance could perhaps be obtained by including knowledge of the kinematics of pion decay, but for this analysis we make no such prior assumptions and rely entirely on the training and the algorithm.

The next stage will apply the discriminate in Monte Carlo events (dataset III) to study how much contamination is in the sample.

## 6. Contamination and background

Neutral pions can decays in two gammas that are collected by the Electromagnetic Calorimeter (EMC). If one gamma is missing, it will not be possible reconstruct the neutral pion. The total neutral pions number will be wrong, and the event will be

classified in the wrong decay class. For example, if the real decay was 2 neutral pions, and one gamma was missed, the event will be classified as 1 neutral pion.

Fig 3 shows the contribution from other decays in the selected decay for all values of invariant mass (a) and the effect of discriminate function (b). Table III shows numerical values for each contamination and Table IV the filter impact increasing the accuracy of selection. Several events changed its classification going from 2 neutral pion to 1 pion, and other 3 pions to 2 pion, etc.

The more important effect is the elimination of background: 7% of 2 pion decay are background (Table III), which drops to 2% using the filter (Table IV). Three pion decays have 14% background without filter, which dropped to 3% after filtering. Finally, 4 pions decays reduce background from 21% to 3%. This is visible also in Fig 3.

There is an increase in the effect of missing gammas due the reduction in the total number of events selected for each decay (consequence of 83% sensitivity). For example, there was 17% of 3 pions decays classified as 2 pions without filter, and 19% with filter.

## 7. Raw Data filtering

The raw data available was 482,303,947 from BaBar's Runs 1, 2, 3, and 4. Fig. 4 shows filtering effect in each decay mode. There was a meaningful reduction in background (compare Figs. 3 and 4), consequence of how the filter was trained. The use of real neutral pions from MC tau decays provides a pure sample, and the filter will not accept any other source of neutral pions (or  $q\bar{q}$  background).

## 8. Neutral pions and gammas energy distribution

A cut in the  $\rho(770)$  invariant mass between 620 MeV and 930 MeV (width 150MeV) allows events selection in the resonance peak. Fig. 5 and 6 shows the final results for neutral pion and gammas energy distribution.

The agreement between MC and Raw data (Fig. 5b and 6b) shows the potential of discriminate functions obtained with genetic programming for qualitative analysis.

Oscillations in MC data are consequence of low statistics (only 4,890,000 MC events used in this analysis). These decays have 17% reduction due the tag decay and would require much more MC events to achieve a smooth fit.

Plots shifted left when there are more neutral pions from the decay because there are more particles to share total rho energy. There is not pion events bellow 150 MeV because the histogram's bin is 25 MeV and neutral pion mass is 134.9MeV, which means the energy must be at least 134.9MeV. Gamma energy is bigger than 100MeV due the 50MeV cut to avoid electronic noise.

Contamination effect can be obtained from MC energy distribution (Fig. 5a and 6a). Further analysis and comparison with theoretical values requires superpose all energy distributions values in a new plot for each decays (e.g.  $1\pi^0$  energy distribution is the sum of  $1\pi^0$  plot plus  $1\pi^0$  contamination in  $2\pi^0$  decay, plus  $1\pi^0$

contamination in  $3\pi^0$  decay, etc). Monte Carlo statistics is decisive factor for good quality results.

Four neutral pion decay was omitted due lack of statistics. There were real events in invariant mass plot, but its analysis was compromised due MC low statistics and the large number of bins in the energy distribution plot.

## 9. Conclusion

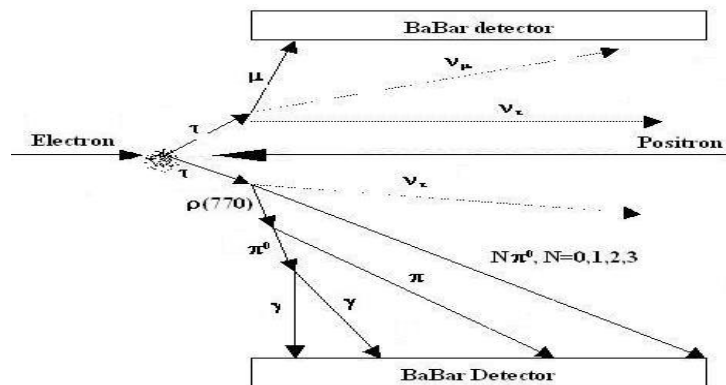
The paper described genetic programming approach to obtain neutral pion discriminate function to discern between background and real neutral pion particles. Background can produce a critical influence in systematic errors and constrain qualitative analysis.

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 analysed 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.

These results are important to check the validity of current theoretical models through the comparison between our energy distributions (and following the same procedure, of any other observable) and predictions from Standard Model of elementary particles.

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



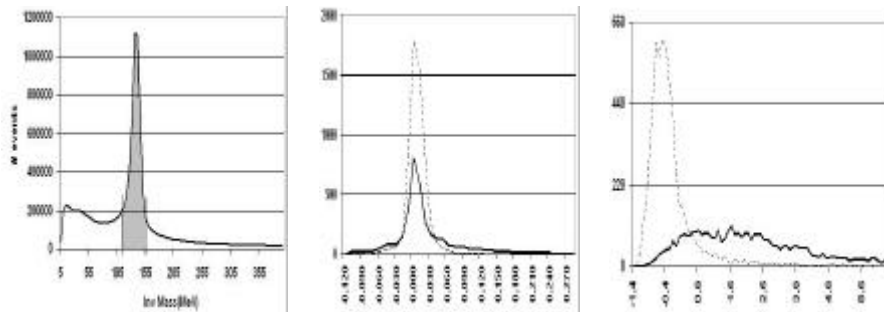
**Fig.1 Schematic diagram of Hadronic Tau decays from Electron – Positron collision.**

**Table I Training and tests results from discriminate function obtained using genetic programming with different datasets. a, b, and g defined in equation (1).**

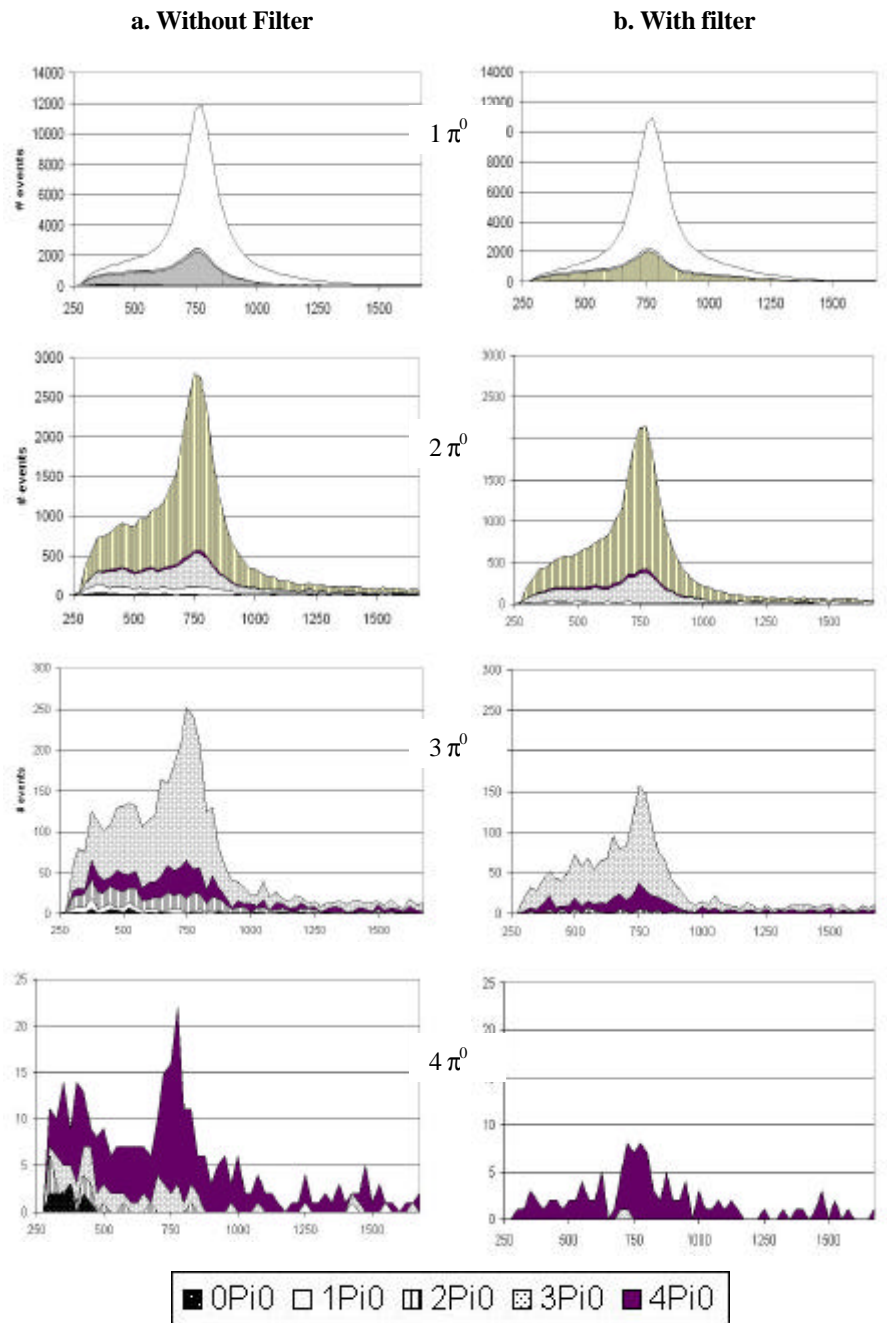
		Case 1: Forecast		Case 2: Forecast		Case 3: Forecast	
Training	Real	D>0	D<0	D>0	D<0	D>0	D<0
57992 records dataset I	1	23299	4819	23368	4750	22491	5627
	0	3093	26781	3040	26834	2731	27143
	$\gamma$	86		86		85	
	$\alpha$	82		83		80	
	$\beta$	89		89		90	
	Test	Real	D>0	D<0	D>0	D<0	D>0
302374 records Dataset II	1	117268	41037	117215	41090	111999	46306
	0	14153	129916	13870	130199	12543	131526
	$\gamma$	81		81		80	
	$\alpha$	74		74		70	
	$\beta$	90		90		91	

**Table II Training results from discriminate function obtained using genetic programming in dataset II.**

		Forecast	
		D>0	D<0
Real	1	129169	29136
	0	24110	119959



a) 2 gammas reconstruction      b) Kinematics discriminate      c) GP discriminate  
 Fig. 2 Neutral pion identification in dataset I using combination of 2 gammas, kinematics constrains, and discriminate function. (dashed lines background, bold real)



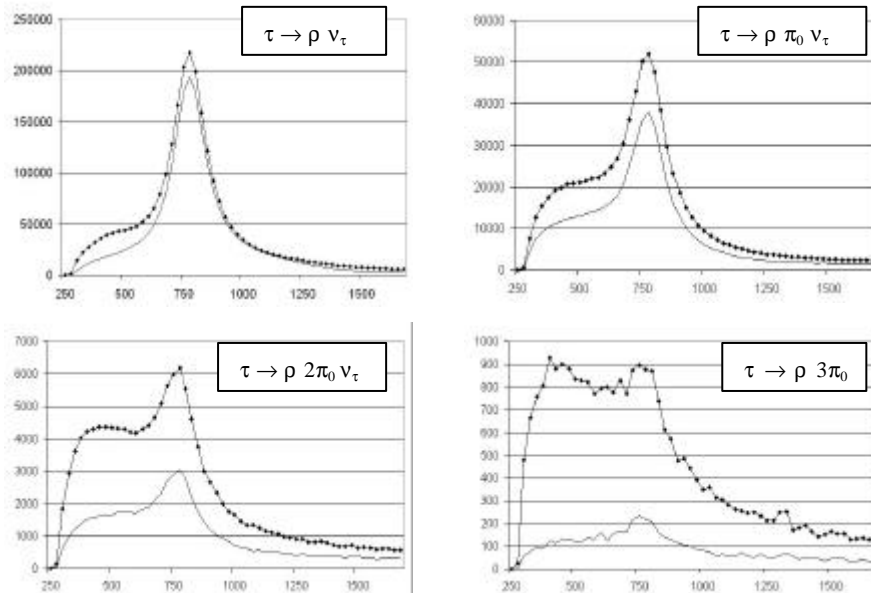
**Fig. 3 Rho invariant mass and filter effect in Monte Carlo data due background (cumulative plot).**

**Table III Contamination effect without discriminate function.**

Accuracy		Real decay.				
		0	1	2	3	4
Forecast	1	2	72	23	3	0
	2	1	6	74	17	2
	3	1	2	11	66	17
	4	2	1	3	15	66

**Table IV Contamination effect after discriminate function.**

Accuracy		Real decay				
		0	1	2	3	4
Forecast	1	1	70	25	4	0
	2	0	2	76	19	3
	3	0	0	3	73	20
	4	0	0	0	3	82



**Fig. 4 Raw data filtering using evolved discriminate function. Dashed line represents data without filtering, and continuous line data with filtering.**

a. Contamination effect (from MC)

b. MC x Raw data agreement

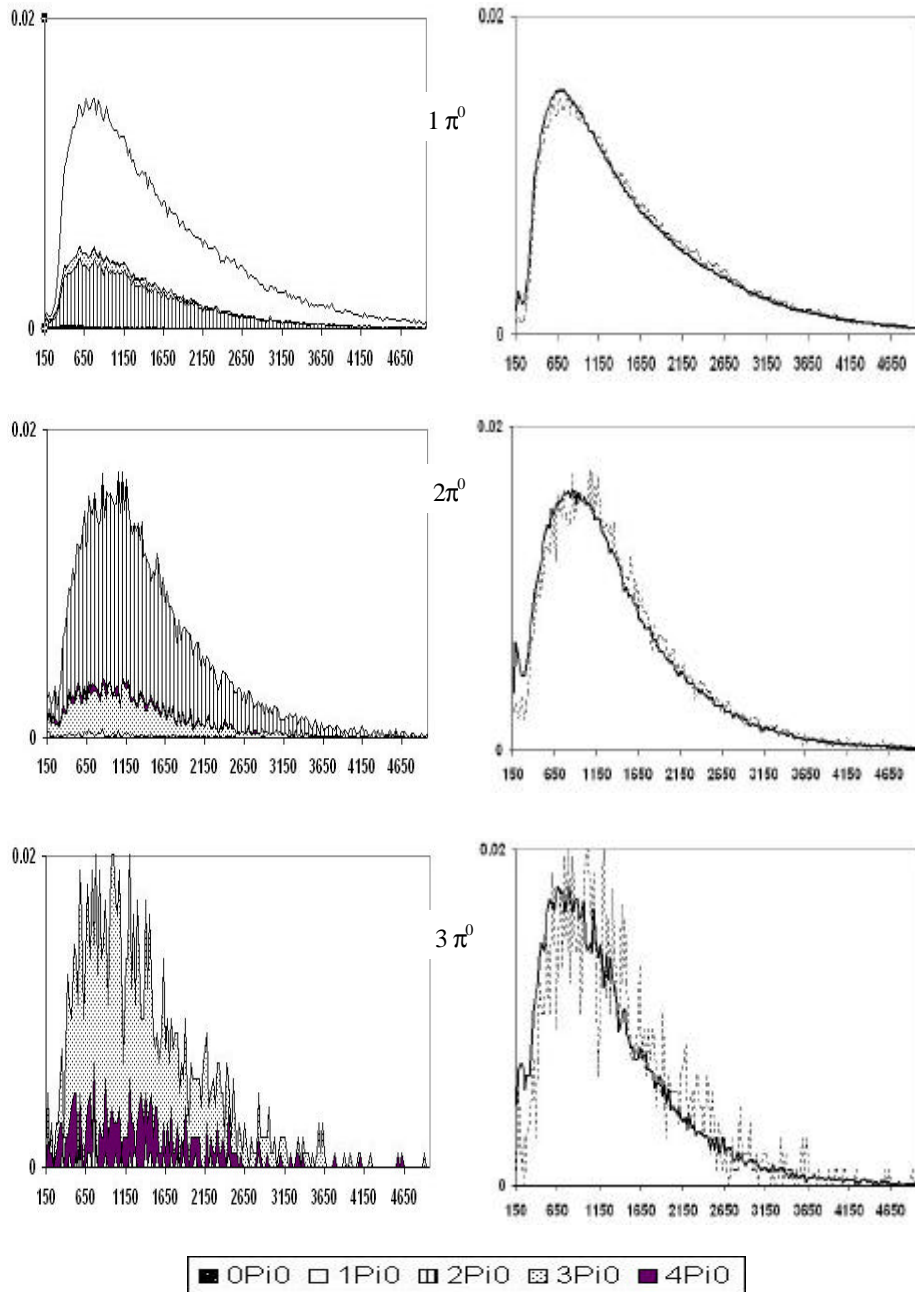
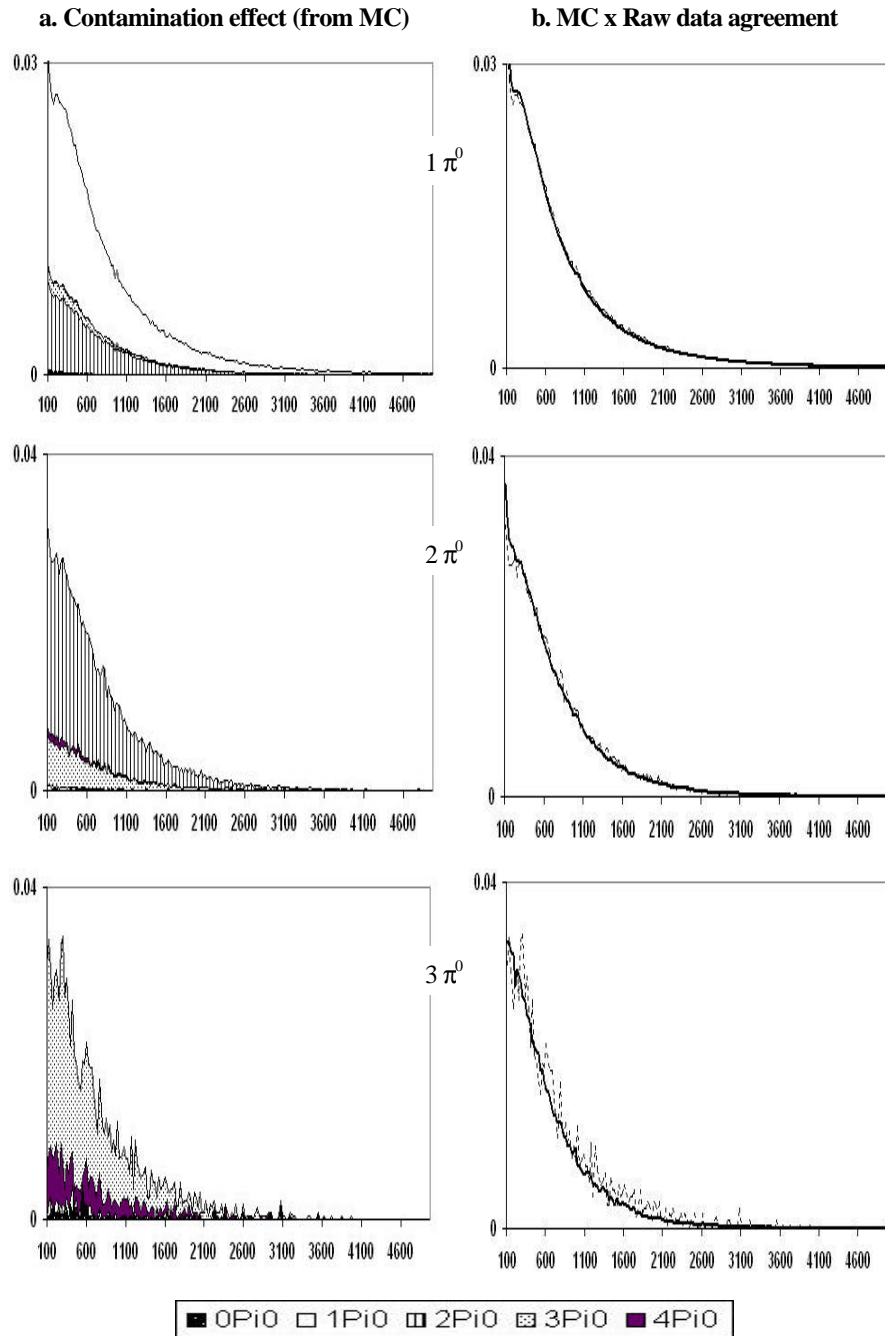


Fig. 5 Energy distribution (MeV) of neutral pion decayed from rho(770) (a) shows cumulative contributions. (b) Monte Carlo (dashed line) agrees with data (continuous line).



**Fig. 6** Gamma energy distribution (MeV) from neutral pion decayed from rho(770).  
 (a) shows cumulative contributions. (b) Monte Carlo (dashed line) agrees with data (continuous line).

## References

1. Holland, J.H. "Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence." Cambridge: Cambridge press 1992.
2. Goldberg, D.E. "Genetic Algorithms in Search, Optimisation, and Machine Learning." Reading, Mass.: Addison-Wesley, 1989.
3. Chambers, L.; "The practical handbook of Genetic Algorithms" Chapman & Hall/CRC, 2000.
4. Koza, J.R. "Genetic programming: On the programming of computers by means of natural selection." Cambridge, Mass.: MIT Press, 1992.
5. Cranmer, K.; Bowman, R.S.; "PhysicsGP: A genetic programming approach to event selection" Computer Physics Communications 167 (2005) 165-176.
6. Focus Collaboration, "Application of genetic programming to high energy physics event selection" Nuclear instruments and methods in physics research A 551 (2005) 504-527.
7. 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
8. Mjahed, M.; "Search for Higgs boson at LHC by using genetic algorithms" To be published in Nuclear Instruments and Methods in Physics Research.
9. Kalganova, T.; Karol, I.M.; Werner, J.C.; Silkou, N.I.; Lipnitskaya, N.G.; Probability prediction method of throat cancer with use of discriminate function (in Russian) 2nd International Belarusian-Polish Conference on Otorhinolaryngology: Actual Problems in Otorhinolaryngology, Grodno, 29-30 May 2003
10. Werner, J.C.; Kalganova, J.C.; Disease modeling using Evolved Discriminate Function. LNCS 2610, Proceedings 6th European Conference, EuroGP 2003, Essex, UK, April 14-16, 2003.
11. Werner, J.C.; Fogarty, T.C.; Severe diseases diagnostics using Genetic Programming. Intelligent Data Analysis in medicine and pharmacology IDAMAP2001; September 4th, 2001 London
12. Werner, J.C.; Fogarty, T.C.; Genetic programming applied to Collagen disease & thrombosis. PKDD 2001 Challenge on Thrombosis data Germany/ Freiburg September 3-7
13. BaBar Collaboration, "The BaBar experiment home page", <http://www.slac.stanford.edu/BFROOT/>
14. Harrison, P.F.; Quinn, H.R.; "The BaBar Physics Book" SLAC Report 504, October 1998, available at <http://www.slac.stanford.edu/pubs/slacreports/slacr-504.html>

15. 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>
16. Davier,M.; Hocker,A.; Zhang,Z.; "The Physics of Hadronic Tau Decays" arXiv:hep-ph/0507078
17. Naisbit,M.; Private communication, PhD Thesis.
18. Belle Collaboration; "A high statistics study of the decay  $\tau \rightarrow \pi \pi^0 \nu \tau$ " arXiv:hep-ex/0512071
19. Cleo collaboration; "Hadronic structure in the decay  $\tau \rightarrow \pi \pi^0 \nu \tau$ " Physical Review D, volume 61, 112002
20. Aleph collaboration; "Branching rates and spectral functions of Tau decays: final Aleph Measurements and Physics implications" arXiv:hep-ex/0506072
21. Cleo Collaboration; "Structure functions in the decay  $\tau \rightarrow \pi \pi^0 \nu \tau$ " Physical Review D, volume 61, 052004
22. GridPP site: <http://www.gridpp.ac.uk/>
23. 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
24. CERN site: <http://public.web.cern.ch/Public/Welcome.html>
25. LHC site: <http://lhc.web.cern.ch/lhc/>
26. LCG site: <http://lcg.web.cern.ch/LCG/>
27. Werner,J.C.; "HEP analysis, Grid and EasyGrid Job Submission Prototype: Babar/CM2 showcase" at <http://www.hep.man.ac.uk/u/jamwer/>
28. Parallel Virtual Machine site: [http://www.csm.ornl.gov/pvm/pvm\\_home.html](http://www.csm.ornl.gov/pvm/pvm_home.html)
29. Geist,A. et al; "PVM: Parallel Virtual Machine. A Users' Guide and Tutorial for Networked Parallel Computing" MIT Press, 1994 available from <http://www.netlib.org/pvm3/book/pvm-book.html>
30. Werner,J.C.; "Active noise control in ducts using genetic algorithm" PhD. Thesis - São Paulo University- São Paulo-Brazil-1999.