



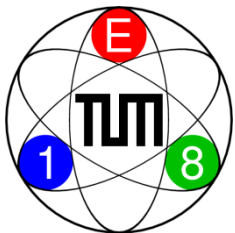
# An Evolutionary Algorithm for Model Selection in Partial-Wave Analyses

**Karl A. Bicker**, Sebastian Neubert,  
Suh-Urk Chung, Jan Friedrich, Boris Grube, Florian Haas,  
Bernhard Ketzer, Stephan Paul, Dimitry Ryabchikov

Technische Universität München

Physik-Department, E18

March 4, 2013



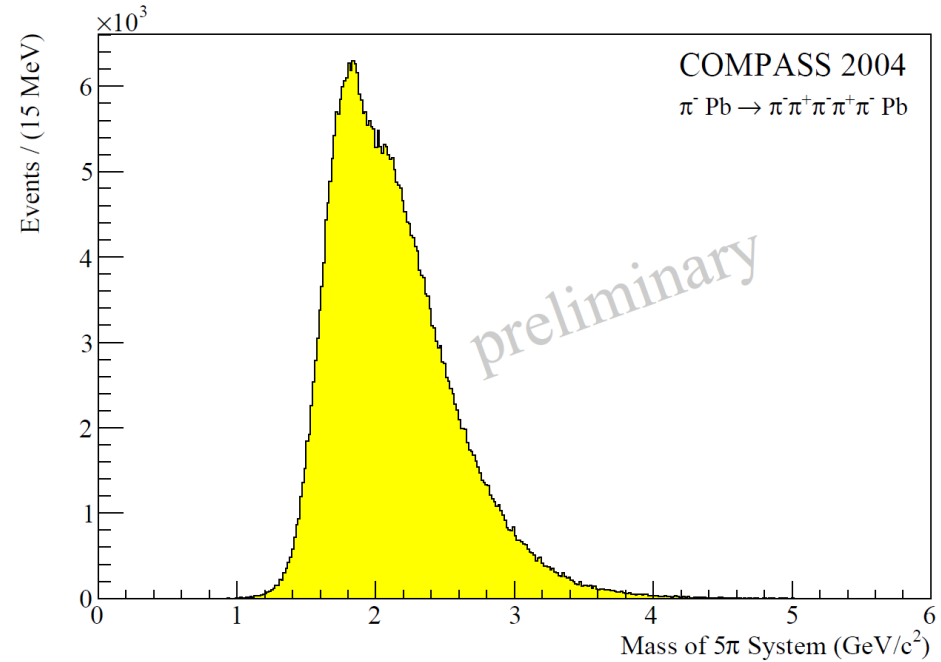
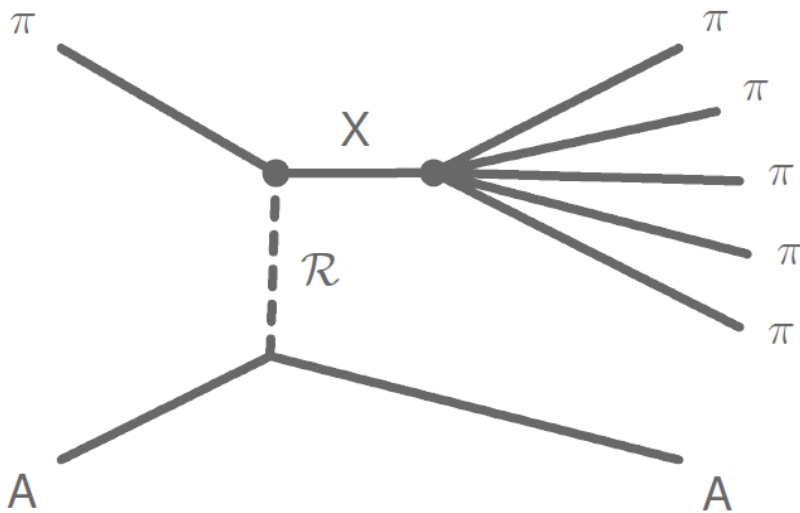
Federal Ministry  
of Education  
and Research



# Outline

- Introduction
- Genetic Algorithm
- Goodness-of-Fit Criterion
- Results
- Conclusion & Outlook

# Motivation



- Example case:  
Diffractive Dissociation  
into 5 Pions at COMPASS

- $J^P$ -Decomposition of mass  
spectrum using angular  
distribution:

Partial-Wave Analysis (PWA)

# Partial-Wave Analysis

- Parameterize cross section (simplified formula):

$$\sigma(\tau, m) = \sigma_0 \left| \sum_{i=0}^{\infty} T_i(m) \psi_i(\tau, m) \right|^2$$

- Multi-dimensional fit to experimental kinematic distributions
- Problem: infinite sum has to be truncated, i.e. limited wave set
- Truncation introduces systematic error



# Model selection

- Problem: find a suitable truncation of coherent sum over partial waves
- Traditional way:
  - Compare minimization criterion (i.e. log-likelihood or  $\chi^2$ ) for different choices of the truncation
  - Use physical arguments and/or preexisting knowledge
  - Trial and error
- May introduce bias, no methodical handle on systematic errors
- Proposal: Use algorithm to select truncation

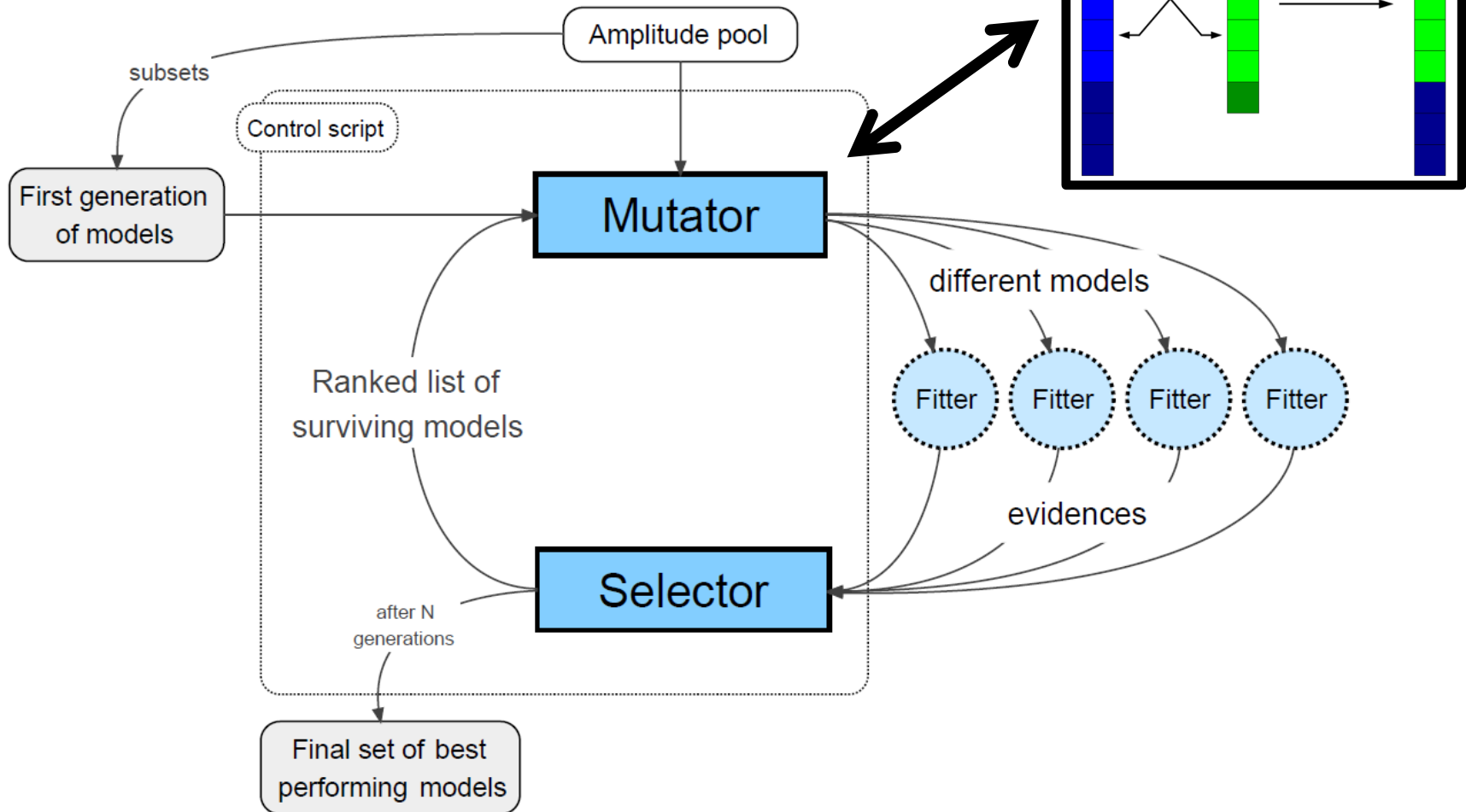
# Model Requirements

An optimal model should fulfill the following requirements:

1. The model should describe the data well.
2. The number of parameters should be as small as possible.
3. Correlations between parameters should be minimal.

Analogy between terms in the partial-wave expansion and genes in living organisms → **genetic algorithm**

# Genetic Algorithm



# Goodness-of-Fit Criterion

- Log-likelihood alone cannot be used to judge model quality, as more fit parameters tend to give a better log-likelihood
- Use Bayes' theorem to judge model quality:

$$\text{evidence} = P(\text{Model}_k | \text{Data}) = \frac{P(\text{Data} | \text{Model}_k)P(\text{Model}_k)}{\sum_j P(\text{Data} | \text{Model}_j)P(\text{Model}_j)}$$

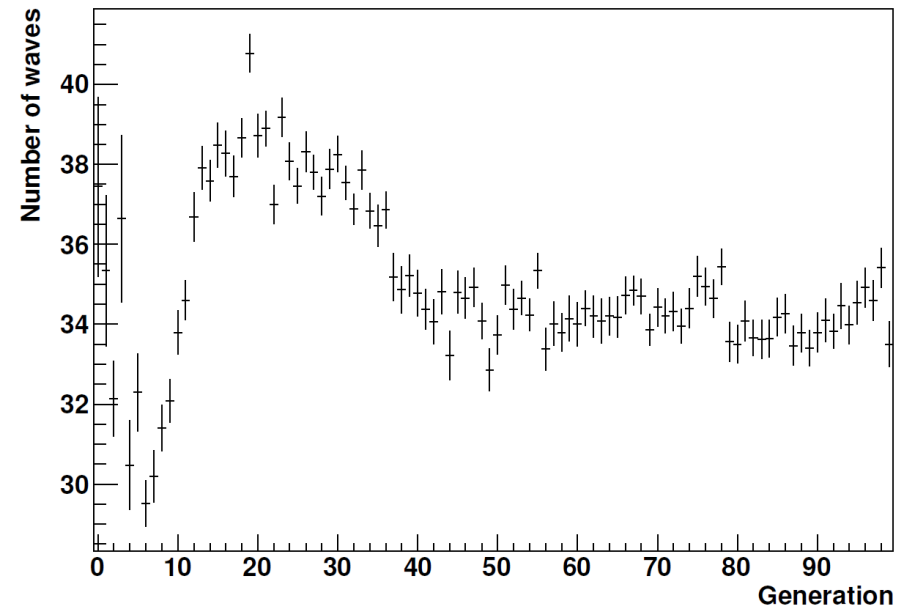
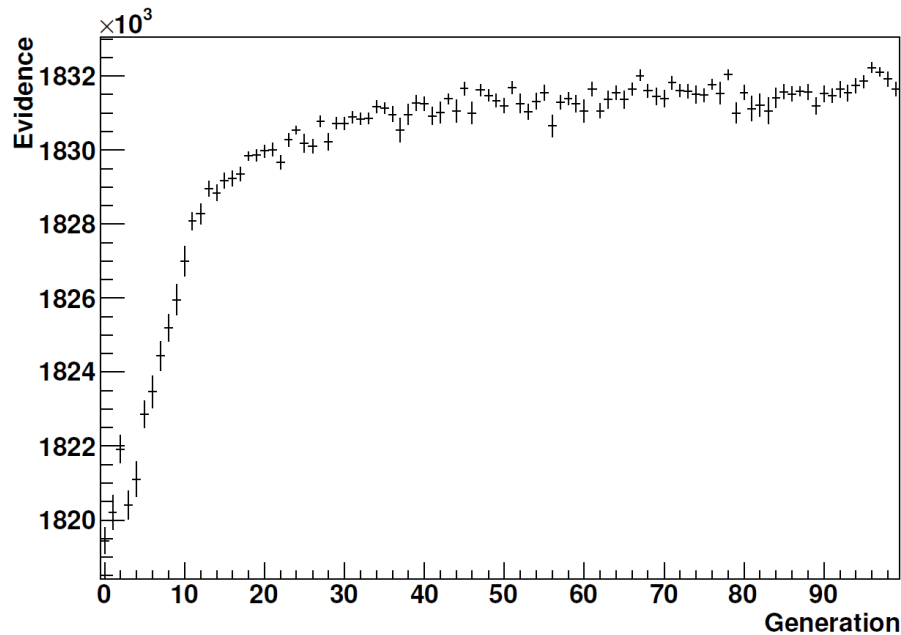
- A number of approximations are needed to calculate this
- Significant contribution of wave implemented as restriction on model parameters



# Results

- Data from COMPASS' 2004 hadron pilot run
- Use a pool of 284 waves
- Run 100 generations with 50 models each
- Parents stay alive if they meet fitness criterion
- Caveat: Breeding step still yielded a better model

# Results



- Final result: waveset of 31 Waves

# Conclusion and Outlook

- A genetic algorithm was implemented in the framework of the ROOTPWA toolkit (<http://sourceforge.net/projects/rootpwa/>)
- A first partial-wave analysis using the algorithm has been performed
- Further studies will have to be done:
  - Tune algorithm parameters
  - Tests with simulated dataset
  - Tests with other decay channels



# Backup

# Evidence

$$\text{evidence} = P(\text{Model}_k \mid \text{Data}) = \frac{P(\text{Data} \mid \text{Model}_k)P(\text{Model}_k)}{\sum_j P(\text{Data} \mid \text{Model}_j)P(\text{Model}_j)}$$

$$P(\text{Data} \mid M_k) = \int \underbrace{P(\text{Data} \mid A^k, M_k) P(A^k \mid M_k)}_{=\mathcal{L}} dA^k$$

# Approximating the Integral

$$P(\text{Data}|M_k) \approx P(\text{Data}|A_{\text{ML}}^k, M_k) \cdot \underbrace{P(A_{\text{ML}}^k|M_k) \cdot \sqrt{(2\pi)^d |\mathbf{C}_{A|D}|}}_{\text{Occam factor}}$$

$$P(A^k|M_k) \cdot \sqrt{(2\pi)^d |\mathbf{C}_{A|D}|} = \frac{\sqrt{(2\pi)^d |\mathbf{C}_{A^k|D}|}}{V_{A^k}} = \frac{V_{A^k|D}}{V_{A^k}}$$

$$\ln P(\text{Data}|M_k) \approx \ln P(\text{Data}|A_{\text{ML}}^k, M_k) + \ln P(A^k|M_k) + \ln \sqrt{(2\pi)^d |\mathbf{C}_{A|D}|}$$

# Prior Probabilities

$$P(A^k | M_k) = \frac{1}{V_A^k}$$

$$N_{\text{events}} = \sum_{\alpha, \beta} T_\alpha T_\beta^* I A_{\alpha\beta} \quad \sum_{\alpha} |T_\alpha|^2 \approx N_{\text{events}}$$

$$V_A^k = S_{m-1} = m \frac{\pi^{m/2}}{\Gamma\left(\frac{d}{2} + 1\right)} R^{d-1}$$

$$\ln V_A^k = \ln d + \frac{d}{2} \ln \pi + \frac{1}{2}(d-1) \ln N_{\text{events}} - \ln \Gamma\left(\frac{d}{2} + 1\right)$$

# Final Formula

$$S_\alpha = \int_{5\sigma_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{(x - |T_\alpha|^2)}{2\sigma_\alpha^2} \right] dx$$

$$\ln P(\text{Data}|M_k) \approx \ln P(\text{Data}|A_{\text{ML}}^k, M_k) + \ln \sqrt{(2\pi)^m |\mathbf{C}_{A|D}|} - \ln V_A^k + \sum_{\alpha} \ln S_\alpha$$

# Bayes Factors

$$B_{12} = \frac{P(\text{Data}|M_1)}{P(\text{Data}|M_2)}$$

$2 \ln B_{12}$	$B_{12}$	Evidence
0 to 2	1 to 3	Not worth mentioning
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong