Building an infrastructure for accessing and analysing figures in scholarly publications

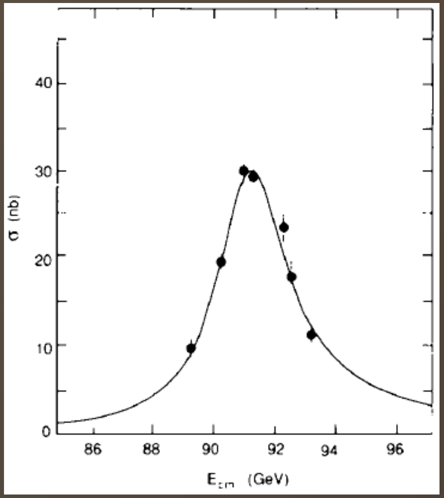
Piotr Praczyk - CERN, 10.03.2011

Usage of graphics in scholarly communication

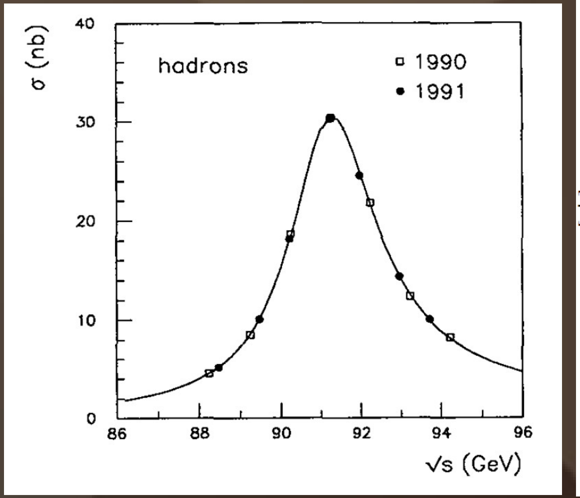
- Describe experiments
- Summarise large amounts of data
- Illustrate relations between results
- Present ideas in a schematic manner

Different measurements of the same quantity

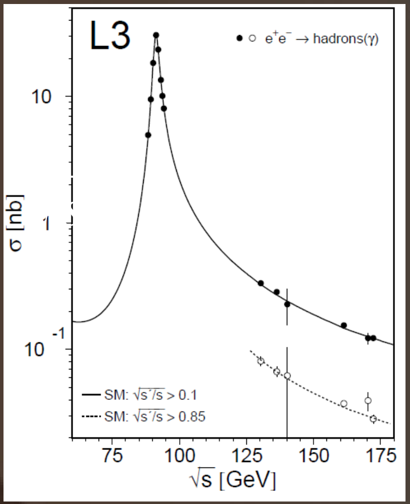
Measured cross section for $e^+e^- \rightarrow$ hadrons as a function of $\sqrt{s}$



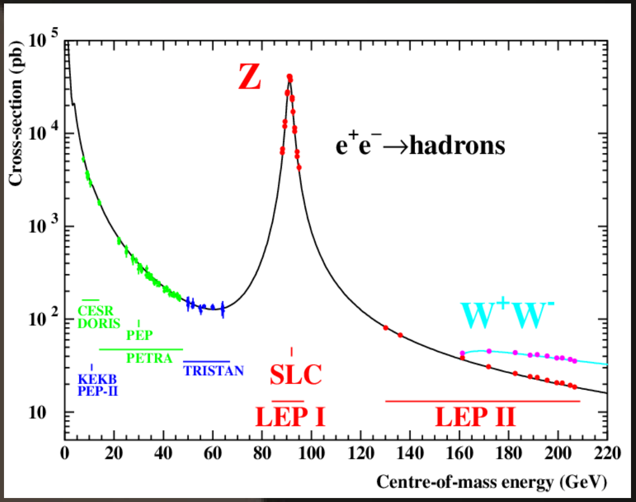
1989



1991



1997



2005

Current understanding of the work

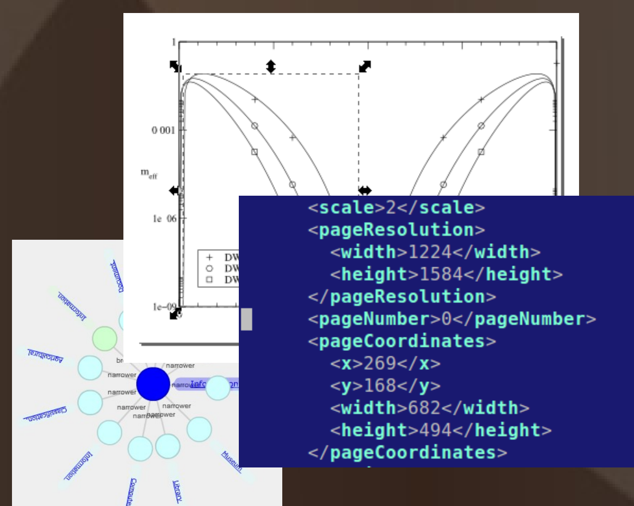
Extraction

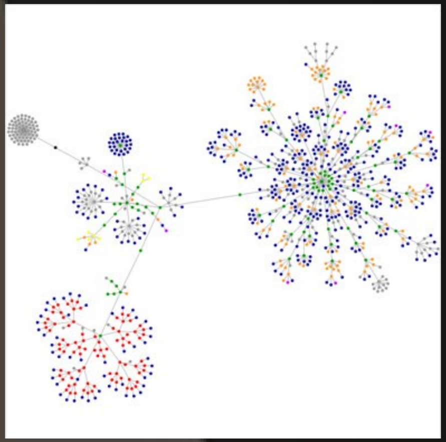
Indexing

Scholarly publications

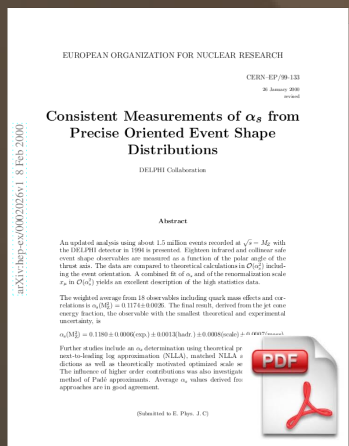
Description of figures as separate entities

Collective description Of figures





Automatic extraction of figures


EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN–EP/99-133

26 January 2000
revised

Consistent Measurements of α_s from Precise Oriented Event Shape Distributions

DELPHI Collaboration

Abstract

An updated analysis using about 1.5 million events recorded at $\sqrt{s} = M_Z$ with the DELPHI detector in 1994 is presented. Eighteen infrared and collinear safe event shape observables are measured as a function of the polar angle of the thrust axis. The data are compared to theoretical calculations in $\mathcal{O}(\alpha_s^2)$ including the event orientation. A combined fit of α_s and of the renormalization scale x_μ in $\mathcal{O}(\alpha_s^2)$ yields an excellent description of the high statistics data.

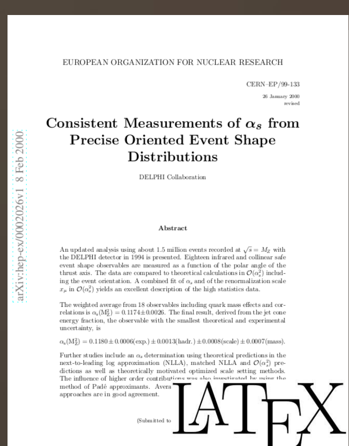
The weighted average from 18 observables including quark mass effects and correlations is $\alpha_s(M_Z^2) = 0.1174 \pm 0.0026$. The final result, derived from the jet cone energy fraction, the observable with the smallest theoretical and experimental uncertainty, is

$\alpha_s(M_Z^2) = 0.1180 \pm 0.0006(\text{exp.}) \pm 0.0013(\text{hadr.}) \pm 0.0008(\text{scale})$

Further studies include an α_s determination using theoretical pr
next-to-leading log approximation (NLLA), matched NLLA a
dictions as well as theoretically motivated optimized scale se
The influence of higher order contributions was also investigat
method of Padé approximants. Average α_s values derived fro
approaches are in good agreement.

(Submitted to E. Phys. J. C)

arXiv:hep-ex/0002026v1 8 Feb 2000



EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN–EP/99-133

26 January 2000
revised

Consistent Measurements of α_s from Precise Oriented Event Shape Distributions

DELPHI Collaboration

Abstract

An updated analysis using about 1.5 million events recorded at $\sqrt{s} = M_Z$ with the DELPHI detector in 1994 is presented. Eighteen infrared and collinear safe event shape observables are measured as a function of the polar angle of the thrust axis. The data are compared to theoretical calculations in $\mathcal{O}(\alpha_s^2)$ including the event orientation. A combined fit of α_s and of the renormalization scale x_μ in $\mathcal{O}(\alpha_s^2)$ yields an excellent description of the high statistics data.

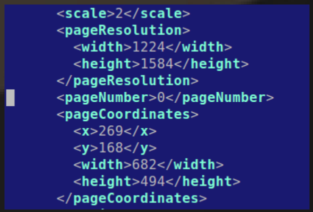
The weighted average from 18 observables including quark mass effects and correlations is $\alpha_s(M_Z^2) = 0.1174 \pm 0.0026$. The final result, derived from the jet cone energy fraction, the observable with the smallest theoretical and experimental uncertainty, is

$\alpha_s(M_Z^2) = 0.1180 \pm 0.0006(\text{exp.}) \pm 0.0013(\text{hadr.}) \pm 0.0008(\text{scale}) \pm 0.0007(\text{mass})$.

Further studies include an α_s determination using theoretical predictions in the next-to-leading log approximation (NLLA), matched NLLA and $\mathcal{O}(\alpha_s^2)$ predictions as well as theoretically motivated optimized scale setting methods. The influence of higher order contributions was also investigated by using the method of Padé approximants. Avera
approaches are in good agreement.

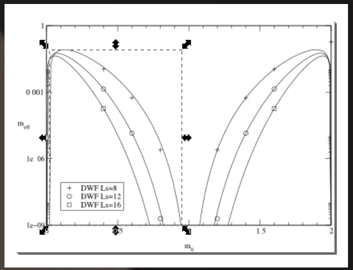
(Submitted to

arXiv:hep-ex/0002026v1 8 Feb 2000



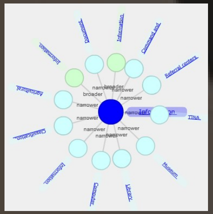
```
<scale>2</scale>
<pageResolution>
  <width>1224</width>
  <height>1584</height>
</pageResolution>
<pageNumber>0</pageNumber>
<pageCoordinates>
  <x>269</x>
  <y>168</y>
  <width>682</width>
  <height>494</height>
</pageCoordinates>
```


Meta-data



Vector + Raster images

- Evaluation of the extraction quality
- Merging of results
- Acquisition of additional data

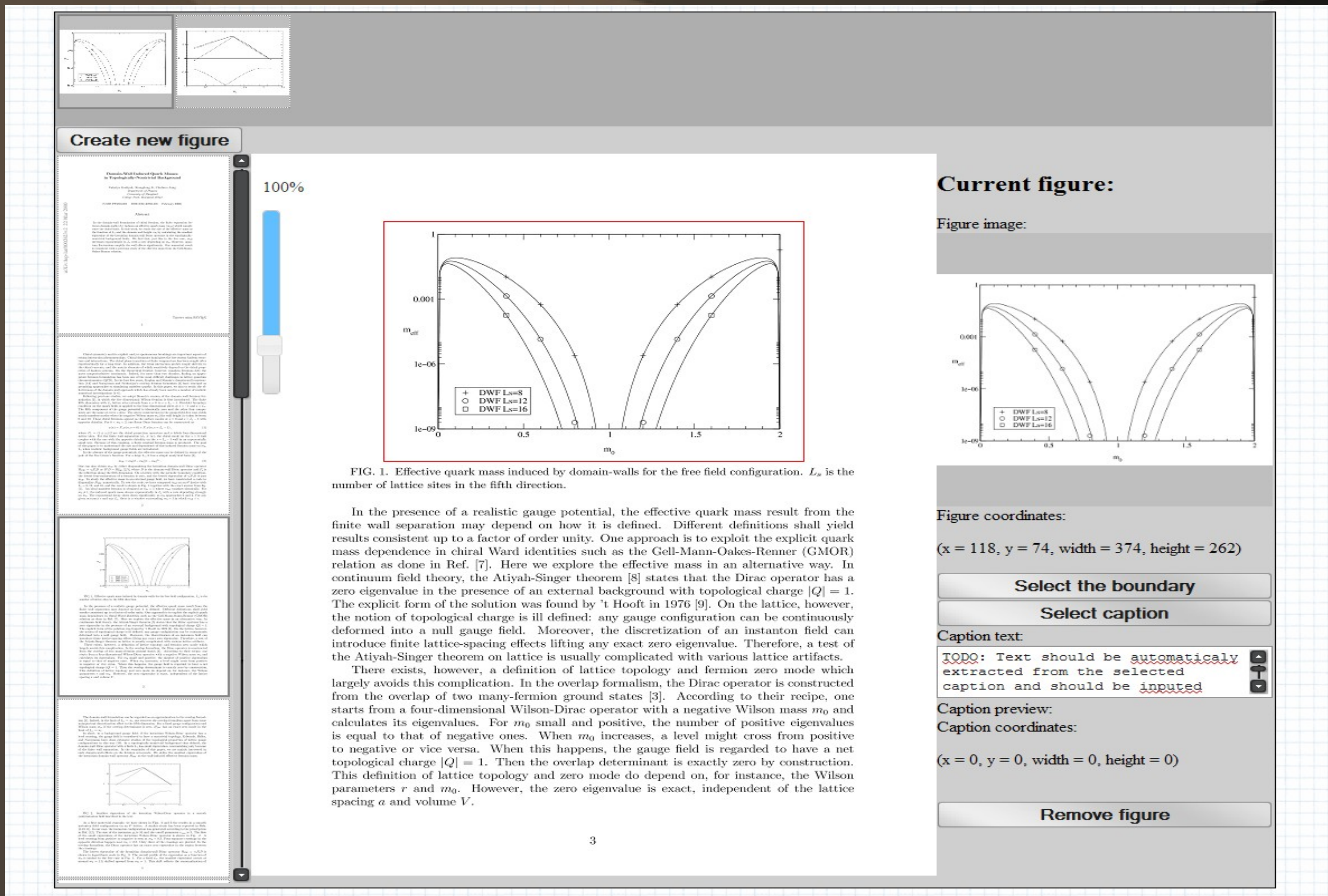


(in the future) Semantic description

Types of extracted meta-data

- Boundaries of figures
- Boundaries of captions
- Text of captions
- Graphics in PNG and SVG formats
- Places, where figure is referenced
- Name of the figure inside a document
- Text present inside the figure

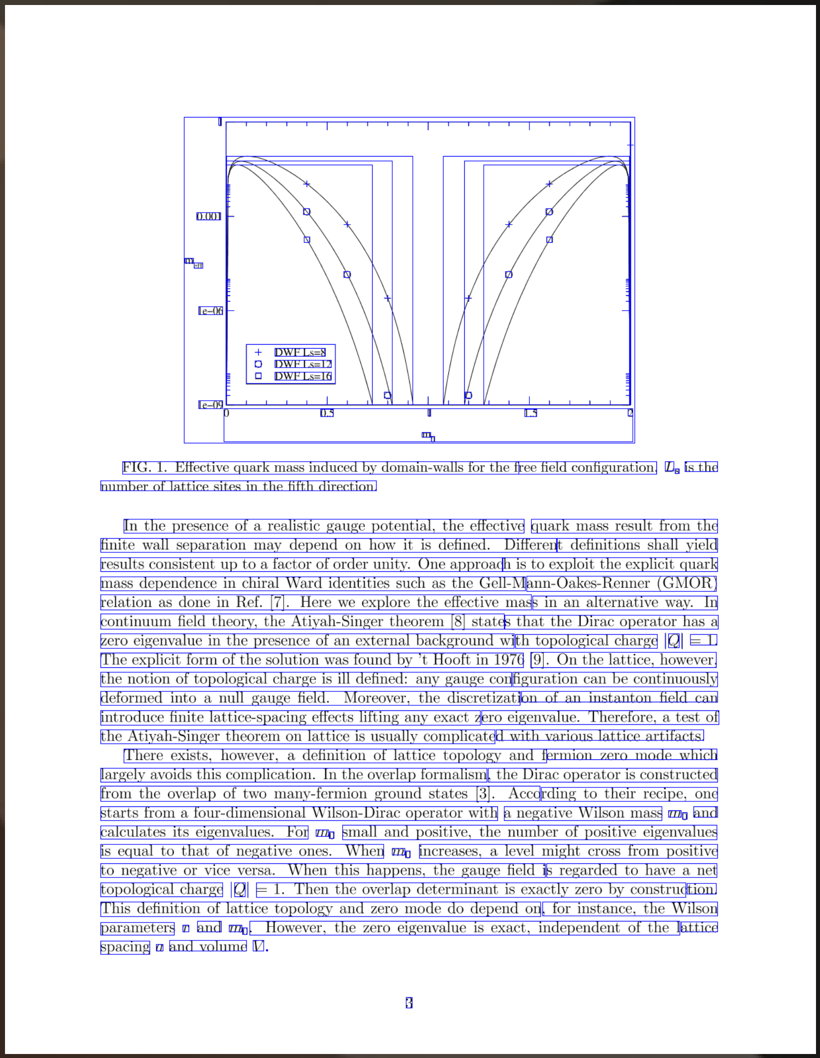
Select Your Figure



Extracting data from PDF

PDF:

- Stream of instructions
- Embeded objects
  - Fonts
  - External objects
- Meta-description



DWF Ls=8
DWF Ls=12
DWF Ls=16

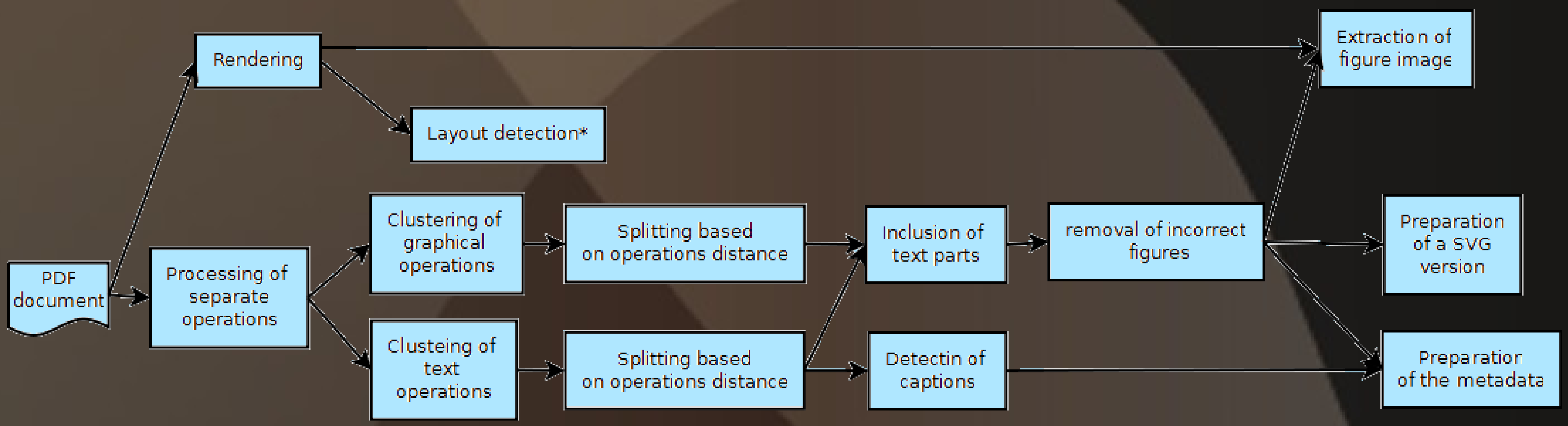
FIG. 1. Effective quark mass induced by domain-walls for the free field configuration. L_s is the number of lattice sites in the fifth direction.

In the presence of a realistic gauge potential, the effective quark mass result from the finite wall separation may depend on how it is defined. Different definitions shall yield results consistent up to a factor of order unity. One approach is to exploit the explicit quark mass dependence in chiral Ward identities such as the Gell-Mann-Oakes-Renner (GMOR) relation as done in Ref. [7]. Here we explore the effective mass in an alternative way. In continuum field theory, the Atiyah-Singer theorem [8] states that the Dirac operator has a zero eigenvalue in the presence of an external background with topological charge $|Q| = 1$. The explicit form of the solution was found by 't Hooft in 1976 [9]. On the lattice, however, the notion of topological charge is ill defined: any gauge configuration can be continuously deformed into a null gauge field. Moreover, the discretization of an instanton field can introduce finite lattice-spacing effects lifting any exact zero eigenvalue. Therefore, a test of the Atiyah-Singer theorem on lattice is usually complicated with various lattice artifacts.

There exists, however, a definition of lattice topology and fermion zero mode which largely avoids this complication. In the overlap formalism, the Dirac operator is constructed from the overlap of two many-fermion ground states [3]. According to their recipe, one starts from a four-dimensional Wilson-Dirac operator with a negative Wilson mass m_0 and calculates its eigenvalues. For m_0 small and positive, the number of positive eigenvalues is equal to that of negative ones. When m_0 increases, a level might cross from positive to negative or vice versa. When this happens, the gauge field is regarded to have a net topological charge $|Q| = 1$. Then the overlap determinant is exactly zero by construction. This definition of lattice topology and zero mode do depend on, for instance, the Wilson parameters r and m_0. However, the zero eigenvalue is exact, independent of the lattice spacing a and volume V.

3

Schema of the PDF extraction process



Intermediate steps of the algorithm

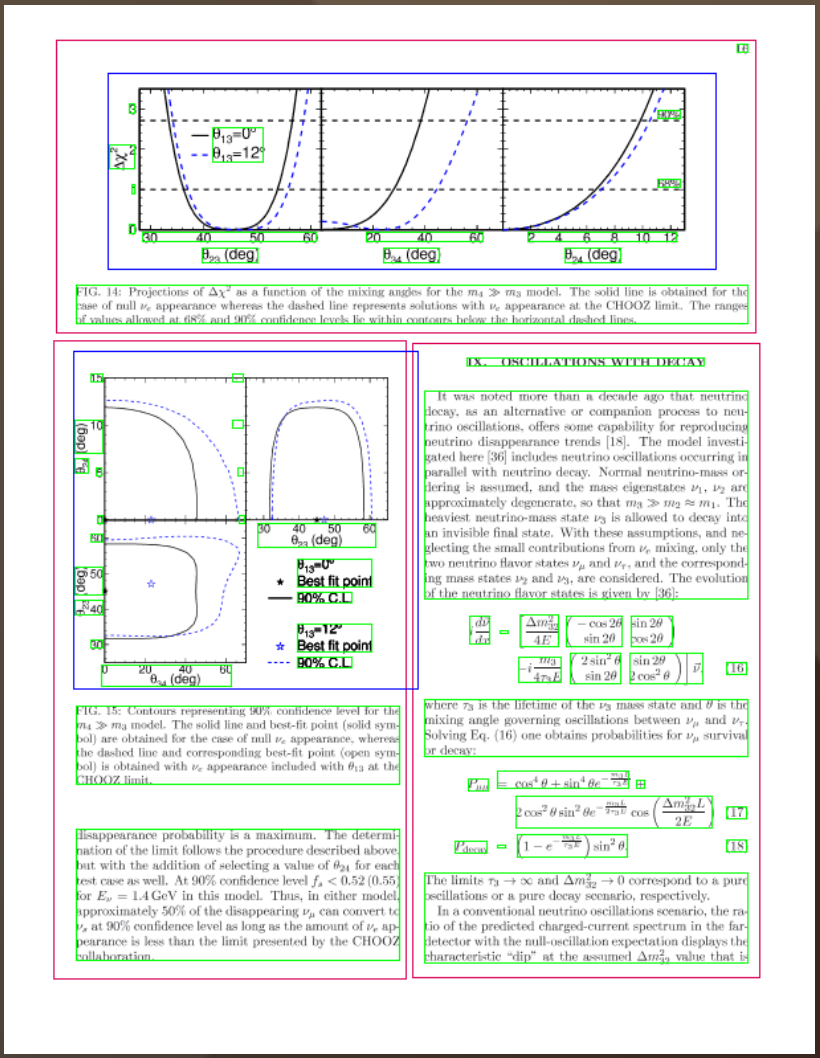


FIG. 14: Projections of $\Delta\chi^2$ as a function of the mixing angles for the $m_4 \gg m_3$ model. The solid line is obtained for the case of null ν_e appearance whereas the dashed line represents solutions with ν_e appearance at the CHOOZ limit. The ranges of values allowed at 68% and 90% confidence levels lie within contours below the horizontal dashed lines.

FIG. 15: Contours representing 90% confidence level for the $m_4 \gg m_3$ model. The solid line and best-fit point (solid symbol) are obtained for the case of null ν_e appearance, whereas the dashed line and corresponding best-fit point (open symbol) is obtained with ν_e appearance included with θ_{13} at the CHOOZ limit.

disappearance probability is a maximum. The determination of the limit follows the procedure described above, but with the addition of selecting a value of θ_{24} for each test case as well. At 90% confidence level $f_s < 0.52\,(0.55)$ for $E_\nu = 1.4\,\mathrm{GeV}$ in this model. Thus, in either model, approximately 50% of the disappearing ν_μ can convert to ν_s at 90% confidence level as long as the amount of ν_e appearance is less than the limit presented by the CHOOZ collaboration.

IX. OSCILLATIONS WITH DECAY

It was noted more than a decade ago that neutrino decay, as an alternative or companion process to neutrino oscillations, offers some capability for reproducing neutrino disappearance trends [18]. The model investigated here [36] includes neutrino oscillations occurring in parallel with neutrino decay. Normal neutrino-mass ordering is assumed, and the mass eigenstates ν_1, ν_2 are approximately degenerate, so that $m_3 \gg m_2 \approx m_1$. The heaviest neutrino-mass state ν_3 is allowed to decay into an invisible final state. With these assumptions, and neglecting the small contributions from ν_e mixing, only the two neutrino flavor states ν_μ and ν_τ, and the corresponding mass states ν_2 and ν_3, are considered. The evolution of the neutrino flavor states is given by [36]:

$$i\frac{d\vec{\nu}}{dx} = \left[\frac{\Delta m^2_{32}}{4E}\begin{pmatrix} -\cos 2\theta & \sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{pmatrix} - i\frac{m_3}{4\tau_3 E}\begin{pmatrix} 2\sin^2\theta & \sin 2\theta \\ \sin 2\theta & 2\cos^2\theta \end{pmatrix}\right]\vec{\nu}, \quad (16)$$

where τ_3 is the lifetime of the ν_3 mass state and θ is the mixing angle governing oscillations between ν_μ and ν_τ. Solving Eq. (16) one obtains probabilities for ν_μ survival or decay:

$$P_{\mu\mu} = \cos^4\theta + \sin^4\theta e^{-\frac{m_3 L}{\tau_3 E}} + 2\cos^2\theta\sin^2\theta e^{-\frac{m_3 L}{2\tau_3 E}}\cos\left(\frac{\Delta m^2_{32} L}{2E}\right) \quad (17)$$

$$P_{decay} = \left(1 - e^{-\frac{m_3 L}{\tau_3 E}}\right)\sin^2\theta. \quad (18)$$

The limits $\tau_3 \to \infty$ and $\Delta m^2_{32} \to 0$ correspond to a pure oscillations or a pure decay scenario, respectively.

In a conventional neutrino oscillations scenario, the ratio of the predicted charged-current spectrum in the far-detector with the null-oscillation expectation displays the characteristic "dip" at the assumed Δm^2_{32} value that is

- Regions of graphics (blue)
  - Clustered graphic operations
- Regions of text (green)
  - Clustered text operations
- Elements of page layout (red)

Future work

- Finishing work on the PDF extractor and selction interface

  Extracting + tagging semantics of images
- Similarity measure based on semantic and graphical properties of figures
- Extraction of data described by figures
- Improvements in extractor algorithms
  - usage of different algorithms
  - usage of different types of data that are produced)

Thank You

Questions ?