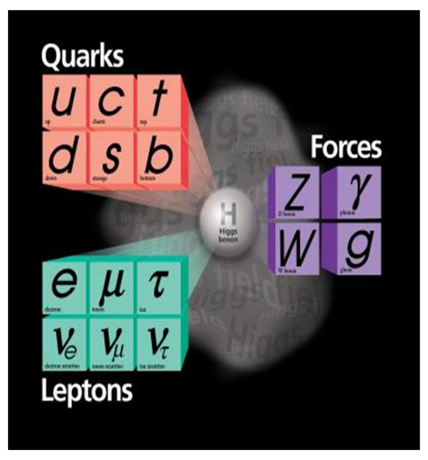
Seeing the Light From the Tevatron: Improving Photon Identification at Fermilab's CDF

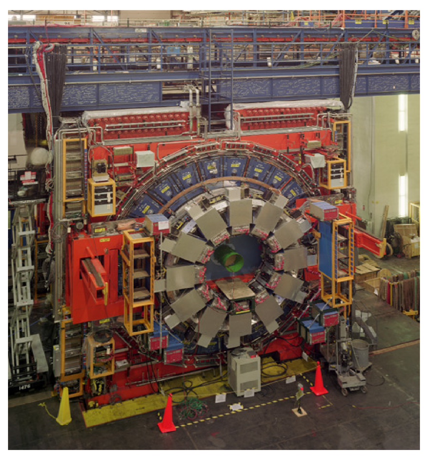
Jamie Ray (Stanford University), Craig Group (CDF/University of Virginia), Ray Culbertson (CDF), Karen Bland (Baylor)

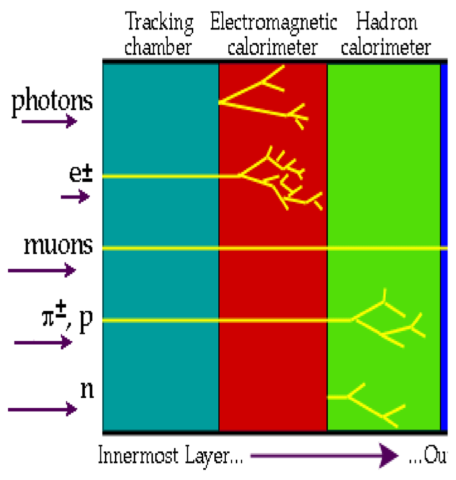
Particle Physics, Fermi National Accelerator Laboratory, Batavia, IL

Introduction

•The Standard Model (SM) of particle physics predicts 17 fundamental particles (See Fig. 1), with countless combinations. The Tevatron is a particle accelerator that collides protons and antiprotons at an energy of 1.96 TeV. To understand high-energy collisions, we need to know not only final-state kinematics, but also what particles were produced!
•The Collider Detector at Fermilab (CDF, seen in Fig. 2) measures variables like position and momenta with great precision. Unfortunately, particles can behave differently depending on mass, spin, flavor, etc. There is no single variable whose value definitively identifies a particle.
•Photons, or γ particles, are ubiquitous and important to identify. However, they can be difficult to distinguish from two 'backgrounds:' an electron, or a pair of photons resulting from quark decay. Better identification against these backgrounds yields improved sensitivity to photons – resulting in more powerful searches or more precise measurements.
•The problem: Can we improve our ID so that more photons pass the requirements? Can we improve our ID so that fewer fakes are included?

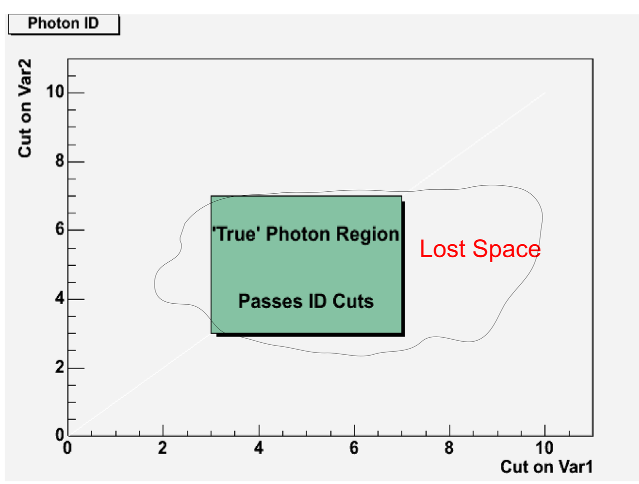


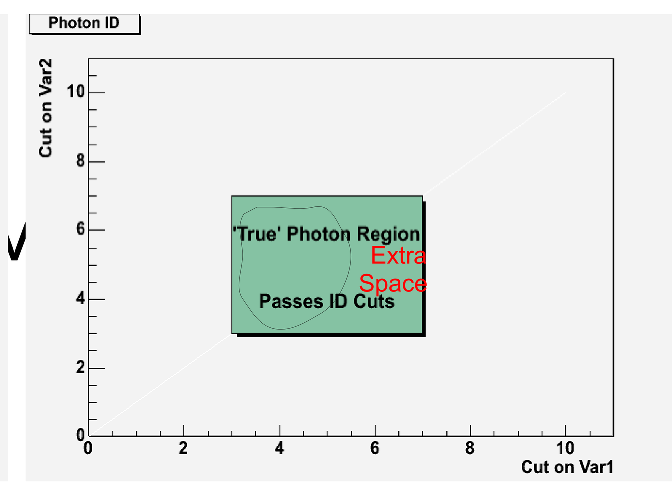




Background

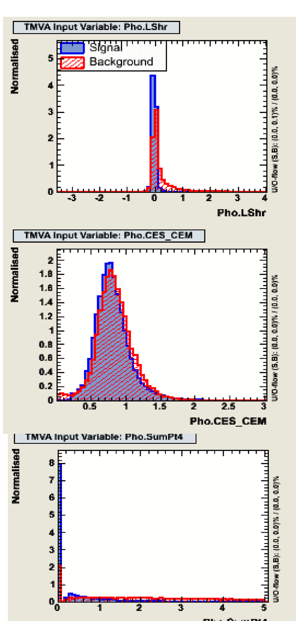
•CDF detector has inner layers to track charged particles, and calorimetry layers to determine energy of interacting particles. Fig. 3 shows detection of several common particles.
•Photons have predictable detector signature. No charge → no track, and strong electromagnetic (EM) interaction → most energy in EM calorimeter.
•CDF currently uses a series of 'ID cuts' restricting track and calorimeter variables to expected photon ranges. This defines a (hyper)-rectangular region of parameter space for positive photon ID. Figures 5 and 6 show a 2D rendering of the cut-space (shaded in green) compared with scenarios of actual photon variable-space (inside black line).
•Cuts ignore correlations by checking each variable independently, but complete independence is unlikely. They may also miss important regions of variable space (reducing ID efficiency, Fig. 5) or include unnecessary space (allowing more 'fakes', Fig. 6).

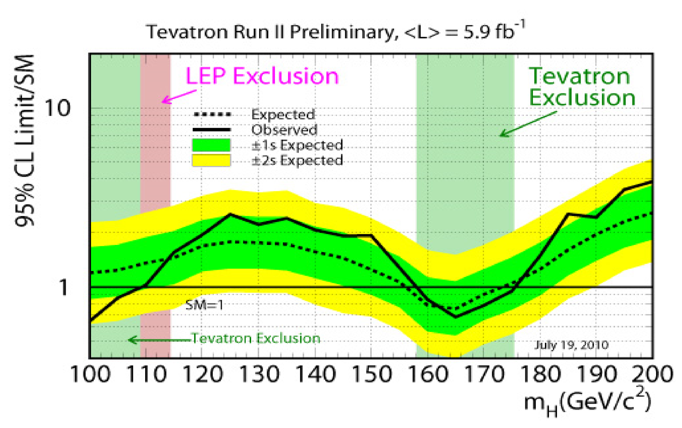


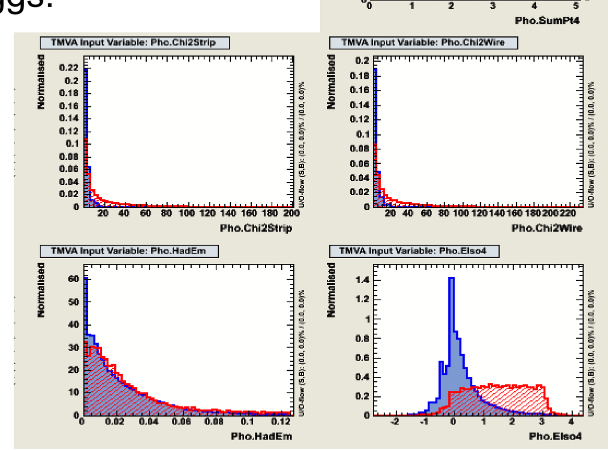


Motivation

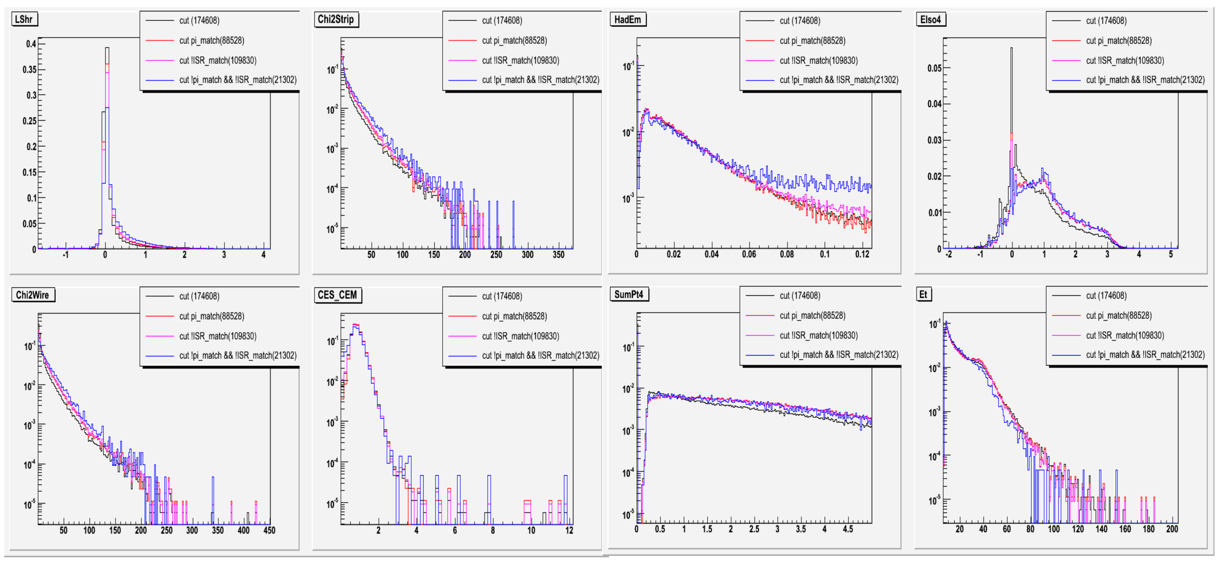
•Electron 'fakes' can be removed by cutting on track variables. Quark decays into jets can imitate isolated photons,so cut-based ID may not be as effective. Fig. Compares variable distributions for photons (blue) and jets (red). Overlap → 'fakes' that cuts can't remove.
•Improved ID increases signal efficiency (fraction of real photons passing ID) and/or increases background rejection (fraction of 'fake' candidates failing). Either is unquestionably valuable, but the optimal proportion depends on context.
•In particular, the Higgs boson can decay into a pair of photons. This decay is especially important in the low-mass range currently being searched (See Fig. , non-excluded mass range). Any improvement in sensitivity to photons will translate into improved sensitivity to the Higgs.







Methods and Training

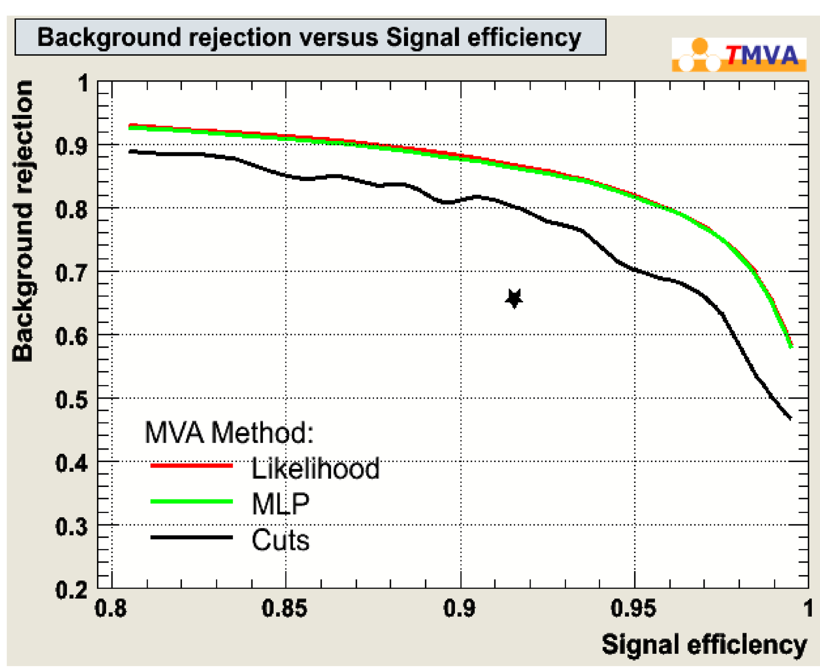


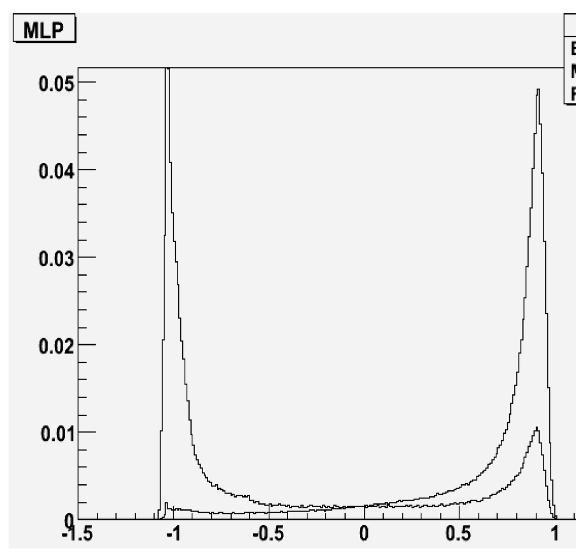
•Cuts pick a window that 'should' describe photon behavior. Instead, we train a computer to discriminate 'signal' (true photons) and 'background' (other particles that look like photons). Training is complex and requires careful consideration, including the following steps:
•Picking method – Likelihood and Neural Net (NN) methods popular, perform well.
•Training samples – Monte Carlo (MC) simulated photons for 'signal,' MC jets with 'loose' photon signature for 'background.' Electrons ignored as they are simpler to remove with cuts.
•Simulation matching – ID based on simulated detector response, but were detected particles actually photons? We carefully match between 'generator-level' and 'detector-level' simulation. Fig. Shows variable distributions for different matching options. We use all jet 'fakes' (pink line) to be more robust than commonly used pi/eta decay matching (red line).
•Variable choice – seven variables (most from standard cuts) describing shape of energy deposition (See Fig. _ and _ for shapes).
•Energy matching – 'fake' energy spectrum differs from true photons, so we weight the background sample to match signal energy spectrum (See Fig. _).
•Data processing – used the Tools for Multivariate Analysis (TMVA) package, which requires specialized formatting and extra code development.

Obstacles

•Long training times (NN requires many iterations to optimize).
•Multivariate techniques are complex, less transparent than standard cut ID. We must avoid 'overtraining' and check performance on other data samples.
•Relies on simulation's ability to model both signal and background signatures effectively. MC simulation might create opportunities for discrimination that don't exist in data. We assess this ability using a 'Data/MC Scale Factor' comparing efficiencies for data and simulation

Results



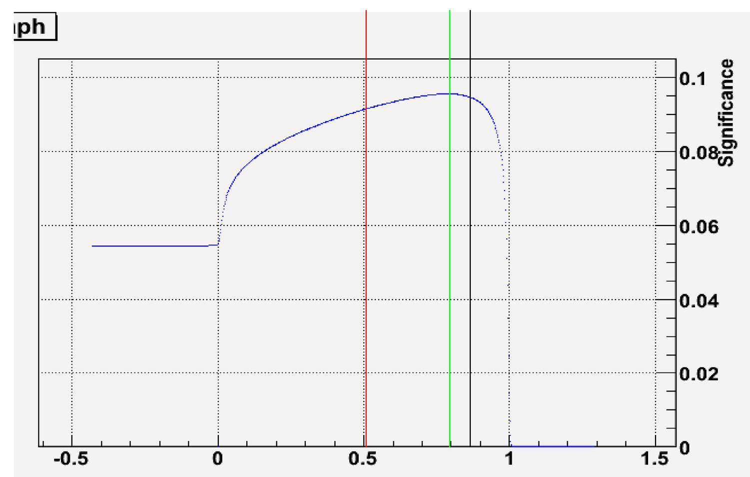


•Some variables not in current ID have significant discriminating power (Lshr, CES_CEM, Chi2).
•ID efficiency for signal and rejection of background compared for NN, Likelihood and standard 'photon ID cuts.' Training options tailored to increase performance.
•Both Likelihood and Neural Net methods (fully trained) suggest significant improvement over the standard ID (Fig. _, current cuts shown as black star). Table _ quantifies improvements.
• Actual gain depends on how much of background consists of jets faking photons – a substantial fraction in most cases.
•'Scale Factors' (SF) assessing simulation suggest that data efficiency is close to (>94%) simulated efficiency. Table _ indicates that SF's vary little (~1%) from current cut ID for the standard working points chosen.
•Developed trained methods into standalone package, then used to evaluate simulated signal and background samples (See output distributions in Fig. _) and determine optimal working point for Higgs search. Package will be included in CDF's standard set of photon variables for easy access in any analysis.

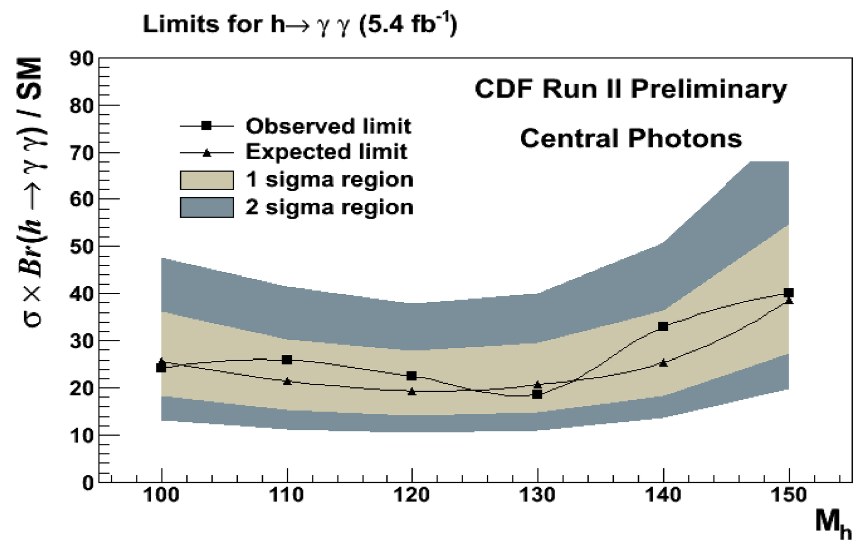
ID Method	Rejection (const. signal eff)	Efficiency (const. background rej.)
Neural Net (MLP, Green)	+30%	+9%
Likelihood (Red)	+30%	+9%
Optimized cuts (Black)	+18%	+6%

ID choice	Scale Factor = Data eff / MC eff
Standard cuts	95.7%
Const. signal	94.5%
Const. background	95.6%
Optimal Higgs point	94.6%

Higgs Potential



•In low-mass regime of Higgs search, we expect ~75% of photon background as jet fakes
•Calculated an 'optimal' working point (cut on classifier output) for NN based on this background composition. Optimal significance is between constant-signal and constant-background points (Fig. _, Table _).
•SF at this point = 94.6%, very close to SF of 95.7% for cut-based ID currently used. Significant sensitivity gain (20%?) is predicted.
•New photon ID will be implemented in next iteration of the Higgs search. With current limits at ~20x SM expectation (Fig. _), photon decay contributes to combined limits (Fig. _).



Conclusions

•'Fake' photons from quark jets are expected to comprise a large fraction of any background when trying to identify photons.
•Identification currently relies on a series of independent 'cuts' that are significantly outperformed by multivariate techniques.
•Use of a multivariate method also allows optimization of ID stringency to suit the context of a particular analysis – previously, the only option had been the standard photon cuts.
•Calculated Scale Factors are close to the SF for standard cuts, indicating that the predicted gain in efficiency can be realized.
•Will be applied in next version of the Higgs search and is predicted to provide a valuable increase in sensitivity – two photons to ID and large jet background.
•Can be valuable to any analysis requiring photon ID and will be made available through CDF's photon data block (integrated into standard data processing procedure).

Acknowledgments

•I thank Dr. Craig Group, Dr. Ray Culbertson, Karen Bland, and many other CDF physicists who have given suggestions and provided resources for my work on this project. I also thank the CDF Collaboration, Fermilab, and the Department of Energy for their support.

•For reference, more information will be forthcoming in a CDF Public Note. Please search the Notes database for author 'Jamie Ray'