FERMILAB-PUB-13-579-E

Improved b quark jet identification at the D0 experiment

V.M. Abazov,[31] B. Abbott,[67] B.S. Acharya,[25] M. Adams,[46] T. Adams,[44] J.P. Agnew,[41] G.D. Alexeev,[31] G. Alkhazov,[35] A. Alton[a],[56] A. Askew,[44] S. Atkins,[54] K. Augsten,[7] C. Avila,[5] F. Badaud,[10] L. Bagby,[45] B. Baldin,[45] D.V. Bandurin,[73] S. Banerjee,[25] E. Barberis,[55] P. Baringer,[53] J.F. Bartlett,[45] U. Bassler,[15] V. Bazterra,[46] A. Bean,[53] M. Begalli,[2] L. Bellantoni,[45] S.B. Beri,[23] G. Bernardi,[14] R. Bernhard,[19] I. Bertram,[39] M. Besançon,[15] R. Beuselinck,[40] P.C. Bhat,[45] S. Bhatia,[58] V. Bhatnagar,[23] G. Blazey,[47] S. Blessing,[44] K. Bloom,[59] A. Boehnlein,[45] D. Boline,[64] E.E. Boos,[33] G. Borissov,[39] M. Borysova[l],[38] A. Brandt,[70] O. Brandt,[20] R. Brock,[57] A. Bross,[45] D. Brown,[14] X.B. Bu,[45] M. Buehler,[45] V. Buescher,[21] V. Bunichev,[33] S. Burdin[b],[39] C.P. Buszello,[37] E. Camacho-Pérez,[28] B.C.K. Casey,[45] H. Castilla-Valdez,[28] S. Caughron,[57] S. Chakrabarti,[64] K.M. Chan,[51] A. Chandra,[72] E. Chapon,[15] G. Chen,[53] S.W. Cho,[27] S. Choi,[27] B. Choudhary,[24] S. Cihangir,[45] D. Claes,[59] J. Clutter,[53] M. Cooke[k],[45] W.E. Cooper,[45] M. Corcoran,[72] F. Couderc,[15] M.-C. Cousinou,[12] D. Cutts,[69] A. Das,[42] G. Davies,[40] S.J. de Jong,[29,30] E. De La Cruz-Burelo,[28] R.T. de Lima,[1] F. Déliot,[15] R. Demina,[63] D. Denisov,[45] S.P. Denisov,[34] S. Desai,[45] C. Deterre[c],[20] K. DeVaughan,[59] H.T. Diehl,[45] M. Diesburg,[45] P.F. Ding,[41] A. Dominguez,[59] A. Dubey,[24] L.V. Dudko,[33] A. Duperrin,[12] S. Dutt,[23] M. Eads,[47] D. Edmunds,[57] J. Ellison,[43] V.D. Elvira,[45] Y. Enari,[14] H. Evans,[49] V.N. Evdokimov,[34] L. Feng,[47] T. Ferbel,[63] F. Fiedler,[21] F. Filthaut,[29,30] W. Fisher,[57] H.E. Fisk,[45] M. Fortner,[47] H. Fox,[39] S. Fuess,[45] P.H. Garbincius,[45] A. Garcia-Bellido,[63] J.A. García-González,[28] V. Gavrilov,[32] W. Geng,[12,57] C.E. Gerber,[46] Y. Gershtein,[60] G. Ginther,[45,63] G. Golovanov,[31] P.D. Grannis,[64] S. Greder,[16] H. Greenlee,[45] G. Grenier,[17] Ph. Gris,[10] J.-F. Grivaz,[13] A. Grohsjean[c],[15] S. Grünendahl,[45] M.W. Grünewald,[26] T. Guillemin,[13] G. Gutierrez,[45] P. Gutierrez,[67] J. Haley,[68] L. Han,[4] K. Harder,[41] A. Harel,[63] J.M. Hauptman,[52] J. Hays,[40] T. Head,[41] T. Hebbeker,[18] D. Hedin,[47] H. Hegab,[68] A.P. Heinson,[43] U. Heintz,[69] C. Hensel,[1] I. Heredia-De La Cruz[d],[28] K. Herner,[45] G. Hesketh[f],[41] M.D. Hildreth,[51] R. Hirosky,[73] T. Hoang,[44] J.D. Hobbs,[64] B. Hoeneisen,[9] J. Hogan,[72] M. Hohlfeld,[21] J.L. Holzbauer,[58] I. Howley,[70] Z. Hubacek,[7,15] V. Hynek,[7] I. Iashvili,[62] Y. Ilchenko,[71] R. Illingworth,[45] A.S. Ito,[45] S. Jabeen,[69] M. Jaffré,[13] A. Jayasinghe,[67] M.S. Jeong,[27] R. Jesik,[40] P. Jiang,[4] K. Johns,[42] E. Johnson,[57] M. Johnson,[45] A. Jonckheere,[45] P. Jonsson,[40] J. Joshi,[43] A.W. Jung,[45] A. Juste,[36] E. Kajfasz,[12] D. Karmanov,[33] I. Katsanos,[59] R. Kehoe,[71] S. Kermiche,[12] N. Khalatyan,[45] A. Khanov,[68] A. Kharchilava,[62] Y.N. Kharzheev,[31] I. Kiselevich,[32] J.M. Kohli,[23] A.V. Kozelov,[34] J. Kraus,[58] A. Kumar,[62] A. Kupco,[8] T. Kurča,[17] V.A. Kuzmin,[33] S. Lammers,[49] P. Lebrun,[17] H.S. Lee,[27] S.W. Lee,[52] W.M. Lee,[45] X. Lei,[42] J. Lellouch,[14] D. Li,[14] H. Li,[73] L. Li,[43] Q.Z. Li,[45] J.K. Lim,[27] D. Lincoln,[45] J. Linnemann,[57] V.V. Lipaev,[34] R. Lipton,[45] H. Liu,[71] Y. Liu,[4] A. Lobodenko,[35] M. Lokajicek,[8] R. Lopes de Sa,[64] R. Luna-Garcia[g],[28] A.L. Lyon,[45] A.K.A. Maciel,[1] R. Madar,[19] R. Magaña-Villalba,[28] S. Malik,[59] V.L. Malyshev,[31] J. Mansour,[20] J. Martínez-Ortega,[28] R. McCarthy,[64] C.L. McGivern,[41] M.M. Meijer,[29,30] A. Melnitchouk,[45] D. Menezes,[47] P.G. Mercadante,[3] M. Merkin,[33] A. Meyer,[18] J. Meyer[i],[20] F. Miconi,[16] N.K. Mondal,[25] M. Mulhearn,[73] E. Nagy,[12] M. Narain,[69] R. Nayyar,[42] H.A. Neal,[56] J.P. Negret,[5] P. Neustroev,[35] H.T. Nguyen,[73] T. Nunnemann,[22] J. Orduna,[72] N. Osman,[12] J. Osta,[51] A. Pal,[70] N. Parashar,[50] V. Parihar,[69] S.K. Park,[27] R. Partridge[e],[69] N. Parua,[49] A. Patwa[j],[65] B. Penning,[45] M. Perfilov,[33] Y. Peters,[41] K. Petridis,[41] G. Petrillo,[63] P. Pétroff,[13] M.-A. Pleier,[65] V.M. Podstavkov,[45] A.V. Popov,[34] M. Prewitt,[72] D. Price,[41] N. Prokopenko,[34] J. Qian,[56] A. Quadt,[20] B. Quinn,[58] P.N. Ratoff,[39] I. Razumov,[34] I. Ripp-Baudot,[16] F. Rizatdinova,[68] M. Rominsky,[45] A. Ross,[39] C. Royon,[15] P. Rubinov,[45] R. Ruchti,[51] G. Sajot,[11] A. Sánchez-Hernández,[28] M.P. Sanders,[22] A.S. Santos[h],[1] G. Savage,[45] L. Sawyer,[54] T. Scanlon,[40] R.D. Schamberger,[64] Y. Scheglov,[35] H. Schellman,[48] C. Schwanenberger,[41] R. Schwienhorst,[57] J. Sekaric,[53] H. Severini,[67] E. Shabalina,[20] V. Shary,[15] S. Shaw,[57] A.A. Shchukin,[34] V. Simak,[7] P. Skubic,[67] P. Slattery,[63] D. Smirnov,[51] G.R. Snow,[59] J. Snow,[66] S. Snyder,[65] S. Söldner-Rembold,[41] L. Sonnenschein,[18] K. Soustruznik,[6] J. Stark,[11] D.A. Stoyanova,[34] M. Strauss,[67] L. Suter,[41] P. Svoisky,[67] M. Titov,[15] V.V. Tokmenin,[31] Y.-T. Tsai,[63] D. Tsybychev,[64] B. Tuchming,[15] C. Tully,[61] L. Uvarov,[35] S. Uvarov,[35] S. Uzunyan,[47] R. Van Kooten,[49] W.M. van Leeuwen,[29] N. Varelas,[46] E.W. Varnes,[42] I.A. Vasilyev,[34] A.Y. Verkheev,[31] L.S. Vertogradov,[31] M. Verzocchi,[45] M. Vesterinen,[41] D. Vilanova,[15] P. Vokac,[7] H.D. Wahl,[44] M.H.L.S. Wang,[45] J. Warchol,[51] G. Watts,[74] M. Wayne,[51] J. Weichert,[21] L. Welty-Rieger,[48] M.R.J. Williams,[49] G.W. Wilson,[53] M. Wobisch,[54] D.R. Wood,[55] T.R. Wyatt,[41] Y. Xie,[45] R. Yamada,[45] S. Yang,[4] T. Yasuda,[45] Y.A. Yatsunenko,[31] W. Ye,[64] Z. Ye,[45] H. Yin,[45] K. Yip,[65] S.W. Youn,[45] J.M. Yu,[56]

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

J. Zennamo,[62] T.G. Zhao,[41] B. Zhou,[56] J. Zhu,[56] M. Zielinski,[63] D. Zieminska,[49] and L. Zivkovic[14]
(The D0 Collaboration*)

[1]*LAFEX, Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro, Brazil*
[2]*Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil*
[3]*Universidade Federal do ABC, Santo André, Brazil*
[4]*University of Science and Technology of China, Hefei, People's Republic of China*
[5]*Universidad de los Andes, Bogotá, Colombia*
[6]*Charles University, Faculty of Mathematics and Physics, Center for Particle Physics, Prague, Czech Republic*
[7]*Czech Technical University in Prague, Prague, Czech Republic*
[8]*Institute of Physics, Academy of Sciences of the Czech Republic, Prague, Czech Republic*
[9]*Universidad San Francisco de Quito, Quito, Ecuador*
[10]*LPC, Université Blaise Pascal, CNRS/IN2P3, Clermont, France*
[11]*LPSC, Université Joseph Fourier Grenoble 1, CNRS/IN2P3, Institut National Polytechnique de Grenoble, Grenoble, France*
[12]*CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France*
[13]*LAL, Université Paris-Sud, CNRS/IN2P3, Orsay, France*
[14]*LPNHE, Universités Paris VI and VII, CNRS/IN2P3, Paris, France*
[15]*CEA, Irfu, SPP, Saclay, France*
[16]*IPHC, Université de Strasbourg, CNRS/IN2P3, Strasbourg, France*
[17]*IPNL, Université Lyon 1, CNRS/IN2P3, Villeurbanne, France and Université de Lyon, Lyon, France*
[18]*III. Physikalisches Institut A, RWTH Aachen University, Aachen, Germany*
[19]*Physikalisches Institut, Universität Freiburg, Freiburg, Germany*
[20]*II. Physikalisches Institut, Georg-August-Universität Göttingen, Göttingen, Germany*
[21]*Institut für Physik, Universität Mainz, Mainz, Germany*
[22]*Ludwig-Maximilians-Universität München, München, Germany*
[23]*Panjab University, Chandigarh, India*
[24]*Delhi University, Delhi, India*
[25]*Tata Institute of Fundamental Research, Mumbai, India*
[26]*University College Dublin, Dublin, Ireland*
[27]*Korea Detector Laboratory, Korea University, Seoul, Korea*
[28]*CINVESTAV, Mexico City, Mexico*
[29]*Nikhef, Science Park, Amsterdam, the Netherlands*
[30]*Radboud University Nijmegen, Nijmegen, the Netherlands*
[31]*Joint Institute for Nuclear Research, Dubna, Russia*
[32]*Institute for Theoretical and Experimental Physics, Moscow, Russia*
[33]*Moscow State University, Moscow, Russia*
[34]*Institute for High Energy Physics, Protvino, Russia*
[35]*Petersburg Nuclear Physics Institute, St. Petersburg, Russia*
[36]*Institució Catalana de Recerca i Estudis Avançats (ICREA) and Institut de Física d'Altes Energies (IFAE), Barcelona, Spain*
[37]*Uppsala University, Uppsala, Sweden*
[38]*Taras Shevchenko National University of Kyiv, Kiev, Ukraine*
[39]*Lancaster University, Lancaster LA1 4YB, United Kingdom*
[40]*Imperial College London, London SW7 2AZ, United Kingdom*
[41]*The University of Manchester, Manchester M13 9PL, United Kingdom*
[42]*University of Arizona, Tucson, Arizona 85721, USA*
[43]*University of California Riverside, Riverside, California 92521, USA*
[44]*Florida State University, Tallahassee, Florida 32306, USA*
[45]*Fermi National Accelerator Laboratory, Batavia, Illinois 60510, USA*
[46]*University of Illinois at Chicago, Chicago, Illinois 60607, USA*
[47]*Northern Illinois University, DeKalb, Illinois 60115, USA*
[48]*Northwestern University, Evanston, Illinois 60208, USA*
[49]*Indiana University, Bloomington, Indiana 47405, USA*
[50]*Purdue University Calumet, Hammond, Indiana 46323, USA*
[51]*University of Notre Dame, Notre Dame, Indiana 46556, USA*
[52]*Iowa State University, Ames, Iowa 50011, USA*
[53]*University of Kansas, Lawrence, Kansas 66045, USA*
[54]*Louisiana Tech University, Ruston, Louisiana 71272, USA*
[55]*Northeastern University, Boston, Massachusetts 02115, USA*
[56]*University of Michigan, Ann Arbor, Michigan 48109, USA*
[57]*Michigan State University, East Lansing, Michigan 48824, USA*
[58]*University of Mississippi, University, Mississippi 38677, USA*
[59]*University of Nebraska, Lincoln, Nebraska 68588, USA*

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

[60]Rutgers University, Piscataway, New Jersey 08855, USA
[61]Princeton University, Princeton, New Jersey 08544, USA
[62]State University of New York, Buffalo, New York 14260, USA
[63]University of Rochester, Rochester, New York 14627, USA
[64]State University of New York, Stony Brook, New York 11794, USA
[65]Brookhaven National Laboratory, Upton, New York 11973, USA
[66]Langston University, Langston, Oklahoma 73050, USA
[67]University of Oklahoma, Norman, Oklahoma 73019, USA
[68]Oklahoma State University, Stillwater, Oklahoma 74078, USA
[69]Brown University, Providence, Rhode Island 02912, USA
[70]University of Texas, Arlington, Texas 76019, USA
[71]Southern Methodist University, Dallas, Texas 75275, USA
[72]Rice University, Houston, Texas 77005, USA
[73]University of Virginia, Charlottesville, Virginia 22904, USA
[74]University of Washington, Seattle, Washington 98195, USA

(Dated: December 29, 2013)

The ability to identify jets which originated from b quarks is an important tool of the physics program of the D0 experiment at the Fermilab Tevatron $p\bar{p}$ collider. This article describes a new algorithm designed to select jets originating from b quarks while suppressing the contamination caused by jets from other quark flavors and gluons. Additionally, a new technique, the SystemN method, for determining the misidentification rate directly from data is presented.

PACS numbers: 29.85.+c

I. INTRODUCTION

The identification of heavy flavor jets, in particular those originating from b or c quarks, is an important technique in particle physics and crucial for studies of top quark, the Higgs boson, and other rare processes [1–3]. The b quark is significantly more massive, $m_b \approx 5$ GeV, than the other quarks with the exception of the top quark. This, along with the long lifetimes of b hadrons, is used to create algorithms for identifying jets which originate from b quarks, called b jets. These algorithms are of primary importance for many measurements and searches performed using the full D0 Run II dataset, recorded from April 2002 until September 2011, with an integrated luminosity of 10 fb^{-1}. This paper describes improvements in the D0 b jet identification algorithm beyond those presented in Ref. [1] and a data-driven method for determining the misidentification rates of the algorithms, that utilizes a new template-fitting method to extract the sample composition directly from the data.

*with visitors from [a]Augustana College, Sioux Falls, SD, USA, [b]The University of Liverpool, Liverpool, UK, [c]DESY, Hamburg, Germany, [d]Universidad Michoacana de San Nicolas de Hidalgo, Morelia, Mexico [e]SLAC, Menlo Park, CA, USA, [f]University College London, London, UK, [g]Centro de Investigacion en Computacion - IPN, Mexico City, Mexico, [h]Universidade Estadual Paulista, São Paulo, Brazil, [i]Karlsruher Institut für Technologie (KIT) - Steinbuch Centre for Computing (SCC), D-76128 Karlsrue, Germany, [j]Office of Science, U.S. Department of Energy, Washington, D.C. 20585, USA, [k]American Association for the Advancement of Science, Washington, D.C. 20005, USA and [l]Kiev Institute for Nuclear Research, Kiev, Ukraine

II. THE UPGRADED D0 DETECTOR

The D0 detector is a general purpose hadron collider detector composed of a tracking system, liquid-argon sampling calorimeter, and muon system [4]. The central tracking system consists of a silicon microstrip tracker (SMT) [5] and a central fiber tracker (CFT), both located within a 1.9 T superconducting solenoidal magnet, with designs optimized for tracking and vertexing at pseudorapidities[1] $|\eta| < 3$ and $|\eta| < 2.5$, respectively. The tracking system enables an accurate measurement of a track's impact parameter (IP), i.e. the distance of closest approach of a track to the $p\bar{p}$ interaction vertex.

The calorimetry comprises a liquid-argon and uranium calorimeter, with a central section (CC) covering pseudorapidities $|\eta| \lesssim 1.1$ and two forward sections (EC) extending the coverage to $|\eta| \approx 4.2$ [6]. The muon system, covering $|\eta| < 2$, consists of three layers of tracking detectors and scintillation trigger counters. One layer is located in front of 1.8 T magnetized iron toroids, and two are positioned after the toroids. The luminosity is measured using plastic scintillator arrays located in front of the EC cryostats [7].

[1] D0 uses a right-handed coordinate system with the origin at the nominal collision point in the center of the detector. The direction of the proton beam is the $+z$ axis, and the $+y$ axis points vertically upwards. The polar angle, θ, is defined such that $\theta = 0$ is in the $+z$ direction. Pseudorapidity is defined as $\eta = -\ln(\tan\frac{\theta}{2})$. The azimuthal angle φ is defined relative to the x axis in the plane transverse to the proton beam direction. The momentum of all particles is measured transverse to the beam direction, p_T.

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

III. DATA AND SIMULATED SAMPLES

The Run II data sample is broken into four subsamples based on different beam and detector conditions. All figures and numbers presented within this article will, for conciseness, be from the largest of the four periods, corresponding to the final 4.4 fb^{-1} of integrated luminosity recorded by the D0 detector. The data are selected by triggering on events containing at least two jets.

To simulate these events we use the PYTHIA [8] Monte Carlo (MC) event generator to create a large sample of multijet events. These events contain jets originating from all types of partons. The fragmentation and decay of particles containing b or c quarks is modeled with EVTGEN [9].

For analyzing the simulated events it is important that the generated jet flavor is known [1]. If a jet contains a simulated b hadron, i.e. $\Delta R(\text{jet}, \text{hadron}) = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} < 0.5$, it is flagged as a b jet. If no b hadron is contained within the jet, but a c hadron is contained then it is defined as a c jet. This sequence guards against cases where a b quark decays to a c quark. The remaining jets, which do not contain b or c hadrons, are defined as light jets.

IV. TRACKING AND PRIMARY VERTEX RECONSTRUCTION

Past and current b jet identification algorithms at D0 are based on three main inputs:

- Particle tracks: reconstructed from hits in the CFT and SMT tracking detectors
- Vertices: reconstructed from at least two tracks originating from the same point
- Calorimeter jets: reconstructed from their energy deposition in the calorimeter

After the track finding step we select the primary $p\bar{p}$ interaction vertex, from which we select tracks for use in the identification algorithms (described in Sec. V A). These steps are briefly described below. A more detailed discussion of the various objects can be found in Ref. [1].

A. Track selection

For a track to be reconstructed it must first be detected with at least one hit in the SMT and at least six hits in the CFT for forward tracks and more than seven for central tracks. These tracks are also required to have transverse momentum $p_T^{\text{trk}} > 0.5$ GeV and a distance of closest approach with respect to the the primary interaction vertex (dca) of less than 4 mm along the axis of the beam, z, and 2 mm in the transverse plane with respect to the beam.

B. Primary vertex reconstruction

Knowledge of the $p\bar{p}$ interaction point is needed for the precise reconstruction and measurement of all objects in the calorimeter and provides an important point of reference for measuring lifetime based variables, which are discussed in Sec. VII A. Multiple interactions may occur during a single beam bunch crossing, making it necessary to identify the primary vertex (PV) associated with the interaction of interest. To form a PV candidate [1]:

(i) two tracks must originate less than 2 cm apart in the z direction;

(ii) an initial vertex fitting using a Kalman filter algorithm [10] to obtain a list of candidate vertices;

(iii) a second vertex fitting iteration using an adaptive algorithm to reduce the effect of outlier tracks;

(iv) the PV is selected as the vertex with the lowest probability of originating from a soft underlying event.

C. Jet reconstruction and calibration

Jets are reconstructed from energy deposits in the calorimeter using the iterative midpoint cone algorithm [11] with a cone of radius $R = 0.5$. By design, this algorithm provides reduced sensitivity to the presence of soft or collinear radiation from partons. The energies of jets are corrected for detector response, the presence of noise, multiple $p\bar{p}$ interactions, and for energy deposited outside of the jet reconstruction cone [12].

V. ALGORITHM PREREQUISITES

Jets and their track information have to fulfill certain criteria, described below, before being used as inputs for b jet identification.

A. Taggability

Since b jet identification algorithms are based solely on tracking and vertex information, it is important to require that each jet reconstructed in the calorimeter is associated with tracks in the tracking system. We implement this "taggability" [1] requirement separately from the requirements of the b jet identification algorithm, allowing for the algorithm's performance to be less dependent on possible variations of the tracking system efficiency. For a jet reconstructed in the calorimeter to be considered *taggable* it must be matched to at least two tracks within a cone of radius $R = 0.5$ with the origin set along the jet axis. All identification efficiencies and misidentification rates, which are the rates at which light

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

jets are selected by the algorithm, are measured relative to *taggable* jets. 90% the jets selected for this analysis with $p_T > 20$ GeV will be classified as taggable.

B. V^0 rejection

Neutral hadrons containing strange quarks (V^0) have decay signatures similar to those of b hadrons. In particular, K_S and Λ hadrons have lifetimes of 90 ps and 263 ps, respectively. To suppress this background, we reject secondary vertices with two oppositely charged tracks with the following criteria:

- The z projection of each track must have a $dca < 1$ cm. This requirement suppresses misreconstructed tracks.
- The significance of the dca, $S_d = dca/\sigma_{dca}$, of each track relative to the PV in the transverse plane has $|S_d| > 3$.
- The tracks associated with the V^0 candidate must have $dca < 200$ μm. This guarantees that V^0s from long lived neutral hadrons are rejected, not those which may have originated from b hadron decays.
- The invariant mass of the two tracks must be outside the mass range expected from K_S or Λ, 472 MeV $< m(\pi\pi) <$ 516 MeV and 1108 MeV $< m(p\pi) <$ 1122 MeV.

To reject photon conversions we reject pairs of tracks which have a negligibly small opening angle between an electron and positron in the plane transverse to the beam line. To be rejected the tracks from the electron and positron must be less than 30 μm apart at the point where their trajectories are parallel to each other. In addition their invariant mass must be less than 25 MeV.

VI. b JET IDENTIFICATION ALGORITHMS

For physics analyses prior to the year 2010 D0 used three algorithms based on charged tracks to identify b jets [1].

Counting Signed Impact Parameters (CSIP) - CSIP determines the number of displaced tracks identified to a jet based on the S_d of each track. To be selected by this algorithm a jet must have at least three tracks with $S_d > 2$, or two tracks with $S_d > 3$.

Jet Lifetime Impact Parameter (JLIP) - The JLIP algorithm uses the IP of all tracks associated with a jet to construct a probability that the jet is a light flavor jet. The JLIP probability is constructed such that it is uniformly distributed between 0 and 1 for light flavor jets, while for heavy flavor jets the JLIP probability is close to zero.

TABLE I: Track selection requirements for the five SVT algorithm configurations: Super Loose (SVT1), Medium Loose (SVT2), Loose Extra (SVT3), Loose (SVT4), and Tight (SVT5).

Track cuts		SVT1	SVT2	SVT3	SVT4	SVT5
p_T [GeV]	$>$	0.5	0.5	0.5	1	1
χ^2	$<$	15	15	10	10	3
S_{xy}	$>$	–	1.5	3	3	3.5
S_{dl}	$>$	–	–	5	5	7

Secondary Vertex Tagger (SVT) - The SVT uses tracks that are significantly displaced from the PV to reconstruct secondary vertices. A jet is tagged if it is matched to a secondary vertex (SV), ΔR(jet, SV) < 0.5. This algorithm can be tuned by varying the requirements on the tracks p_T, χ^2 per degree of freedom for the secondary vertex, the transverse impact parameter significance of the tracks with respect to the primary vertex (S_{xy}), and decay length significance of the secondary vertex in the plane transverse to the beam (S_{dl}). These selections are optimized in a set of five SVT algorithms (SVT$1-5$) that provide complementary information about the jet. The track selections for the different configurations are listed in Table I.

In Ref. [1], we described how input variables obtained from these tools were combined using a neural network to construct the D0 NN-algorithm (D0-NN). The D0-NN shows significant performance improvements compared to the first-level algorithms. In the following, we describe how further improvements have been achieved using an extended set of input variables, making use of both decision trees and a neural network. The new algorithm which results from these improvements is called MVA$_{bl}$, standing for a multivariate analysis that discriminates between b quark and light jets.

VII. MVA$_{bl}$ ALGORITHM

To develop the MVA$_{bl}$ algorithm we generate two MC samples: 10^6 di-b jet signal events and 10^6 di-light jet background events. We use variables (discussed below) which separate b jets from light jets to train six random forests (RF) using the ROOT TMVA [13] framework. One RF is trained using the impact parameter properties from the CSIP and JLIP algorithms and one for each set of SVT variables extracted from the five different SVT algorithms configurations.

These six RFs are then combined using a neural network implementation, the TMultiLayerPerceptron

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

(MLP), also within the ROOT [14] framework. This neural network utilizes the non-linear correlations between inputs to produce the MVA_{bl} output. This improves discrimination over the D0-NN by the inclusion of an order of magnitude more variables.

A. Input variables

1. Impact Parameter Variables

To train the RF based on variables derived from the impact parameter properties we combine the following variables:

1. the output of the JLIP algorithm;
2. the output of the CSIP algorithm;
3. the reduced JLIP [1], which is computed by removing the track with the lowest probability of originating from the PV and then recalculating the JLIP;
4. the combined probability [1] associated with the tracks with the highest and second highest probability of coming from the PV;
5. the largest separation in $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ between any two tracks within a jet, $\max[\Delta R(\text{tracks})]$;
6. the sum of the ΔR distances between each track matched to the jet and the center of the calorimeter jet, $\Sigma_{\text{trk}}\Delta R(\text{trk, jet})$;
7. the p_T-weighted ΔR width of the tracks relative to the calorimeter jet defined as
$$\Theta \equiv \frac{\sum\limits_{\text{trk}} p_T^{\text{trk}} \times \Delta R(\text{trk, jet})}{\sum\limits_{\text{trk}} p_T^{\text{trk}}}; \qquad (1)$$
8. the total transverse momentum of all tracks in the jet cone;
9. the total number of tracks matched to the jet.

The resulting RF output distribution is displayed in Fig. 1(a).

2. Secondary Vertex Variables

The SVT algorithms preselect a set of tracks according to their kinematic properties and reconstruction quality. As a consequence, starting from a common set of tracks, the various SVT configurations lead to different secondary vertices with different properties providing a complementary set of variables for each jet. We then train five RFs using variables associated with the secondary vertices.

In total each of the SVT RFs uses 29 input variables:

1. the p_T of the highest p_T track matched to the secondary vertex, p_T^1;
2. the p_T of the second highest p_T track matched to the secondary vertex, p_T^2;
3. the p_T fraction carried by the tracks from the secondary vertex tracks, $p_T^{\text{SVT}}/p_T^{\text{jet}}$;
4. the number of tracks originating from the secondary vertex;
5. the mass of the secondary vertex ($\mathrm{M_{SV}}$), calculated by summing all track four-momentum vectors assuming that all tracks originate from pions;
6. the signed decay length significance of the secondary vertex in the plane transverse to the beam direction;
7. the JLIP probability of the tracks matched to the secondary vertex;
8. the sum of χ^2/n.d.f. of the tracks matched to the secondary vertex;
9. the number of secondary vertices which can be reconstructed from the tracks matched to the jet;
10. the signed IP of the track with the highest momentum measured transverse to the direction of the secondary vertex;
11. the number of tracks matched to the jets;
12. The proper lifetime of the secondary vertex, computed using $\mathrm{M_{SV}}$, in the plane transverse to the beam direction;
13. the decay length of the secondary vertex in the plane transverse to the beam direction;
14. the decay length of the secondary vertex in the beam direction;
15. the p_T of the highest p_T track in the jet divided by the p_T of the secondary vertex (p_T^{SVT}), p_T^1/p_T^{SVT};
16. the p_T of the second highest p_T track normalized to the secondary vertex p_T, p_T^2/p_T^{SVT};
17. the *dca* of the secondary vertex to the PV in the plane transverse to the beam;
18. the *dca* of the secondary vertex to the PV in the beam direction;
19. the p_T of the track which has the highest momentum measured relative to the direction of the secondary vertex;
20. the momentum of the secondary vertex in the plane transverse to the calorimeter jet direction;

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

21. the p_T of the highest p_T track divided by the total jet p_T, $p_T^1/p_T^{\rm jet}$;
22. the p_T of the second highest p_T track divided by to total jet p_T, $p_T^2/p_T^{\rm jet}$;
23. the angle between the tracks emerging from the secondary vertex projected into the plane transverse to the beam direction;
24. the angle between the tracks emerging from the secondary vertex projected in the beam direction;
25. the Θ (as defined above) as measured for tracks matched to the secondary vertex;
26. the max[ΔR(tracks)] of the tracks matched to the secondary vertex;
27. the p_T weighted charge (q) of the jet, measured as $\sum_{\rm trk} p_T^{\rm trk} q^{\rm trk}/p_T^{\rm jet}$;
28. the signed decay length significance of the secondary vertex in the beam direction;
29. the radius of the cone enclosing all the tracks matched to the secondary vertex.

The outputs of the five SVT RFs are shown in Figs. 1(b−f).

B. Optimized MVA$_{bl}$ parameters

The outputs of the six RFs, shown in Fig. 1, are combined using an MLP neural network into a single variable. The training parameters for the six separate RFs and the final MLP are optimized to minimize the misidentification rate for a fixed b jet identification efficiency. The RF parameters are the number of trees in the forest (5) and the number of variables considered at each random split (all). The parameters used for building the final neural network discriminant are the number of nodes (7 input, 1 hidden, and 1 output) and the number of training iterations (50).

C. MVA$_{bl}$ performance in simulation

The performance of the MVA$_{bl}$ algorithm is presented in Fig. 2. A measure of the discriminating power is given by the performance profile, or the identification efficiency of a b jet versus the misidentification rate. The comparison of the performance of the D0-NN and MVA$_{bl}$ algorithms is presented in Fig. 3. At low values of the misidentification rate, the MVA$_{bl}$ preforms significantly better than the D0-NN, while at high values they are similar. We define a set of benchmark points, designated as operating points (OPs) below, and determine the efficiency and misidentification rates of the OPs for use in subsequent analyses. For the MVA$_{bl}$ algorithm, these points are defined in the following way:

L6, MVA$_{bl} > 0.02$;

L5, MVA$_{bl} > 0.025$;

L4, MVA$_{bl} > 0.035$;

L3, MVA$_{bl} > 0.042$;

L2, MVA$_{bl} > 0.05$;

Loose, MVA$_{bl} > 0.075$;

oldLoose, MVA$_{bl} > 0.1$;

Medium, MVA$_{bl} > 0.15$;

Tight, MVA$_{bl} > 0.225$;

VeryTight, MVA$_{bl} > 0.3$;

UltraTight, MVA$_{bl} > 0.4$;

MegaTight, MVA$_{bl} > 0.5$.

These OPs are displayed in Fig. 4 where the identification efficiency for b jets and the misidentification rate for light jets are shown as a function of the MVA$_{bl}$ output for simulated events.

VIII. EFFICIENCY ESTIMATION

Once the algorithm has been defined and its performance is quantified in simulation, we compare the performance measured in data. This is a two step-process where we use the efficiencies in both data and MC to correct the simulation.

A. System8 method

Using the *System8* (S8) formalism, the b jet identification efficiencies can be measured directly from data [1]. A system of eight equations with eight unknowns is constructed so that solution to these nonlinear equations includes the efficiency for selecting b jets.

To determine the efficiency of identifying a b jet we construct a heavy flavor enriched data sample. These events contain two back-to-back jets satisfying $|\Delta\phi({\rm jet}_1, {\rm jet}_2)| > 2.5$, one jet must have $p_T > 15$ GeV and $|\eta| < 2.5$ and be matched to a muon inside a cone of $R = 0.5$ around its centroid (called a muonic jet). The matched muon must have $p_T^\mu > 4$ GeV. These events, now enriched in heavy flavor jets, contain contamination from light jets due to muonic decays of $\pi^\pm$ and $K^\pm$. Since the S8 method only accommodates a single background we combine the c and light jet backgrounds into a single sample referred to as "cl jets".

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

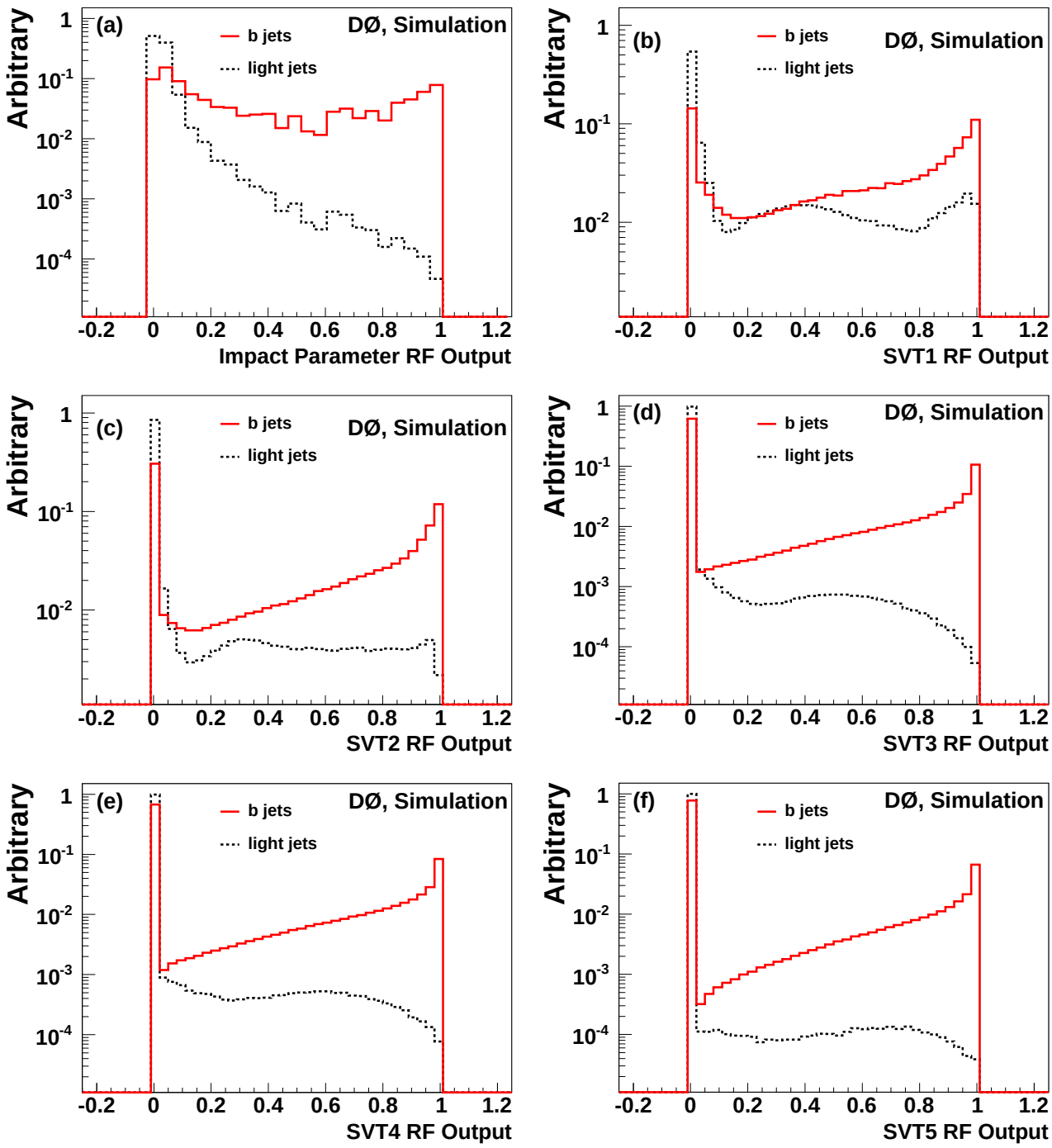

FIG. 1: (color online) Distributions of the six RF outputs for (a) the impact parameter variables and (b−f) the five configurations of the SVT algorithm.

Three additional requirements, or "tags", are individually applied to muonic jets to create subsamples that are further enriched in b jets. The first *tag* selects muonic jet that passes a given MVA_{bl} OP (described in Sec. VII C). The second *tag* is a requirement on p_T^{μ} relative to the direction obtained by adding the muon and jet momenta, known as p_T^{rel}. Requiring that $p_T^{rel} > 0.5$ GeV removes light jets as the large b quark mass leads to large muon p_T^{rel} [15]. The final *tag* is a requirement that the jet which is recoiling from the muonic jet has JLIP < 0.005, this is known as the "away-side tag". The "away-side tag" allows us to select a data sample heavily enriched in pair-produced back-to-back b jets. Using the JLIP to *tag* this *away* jet leads to an enrichment in the overall heavy flavor

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

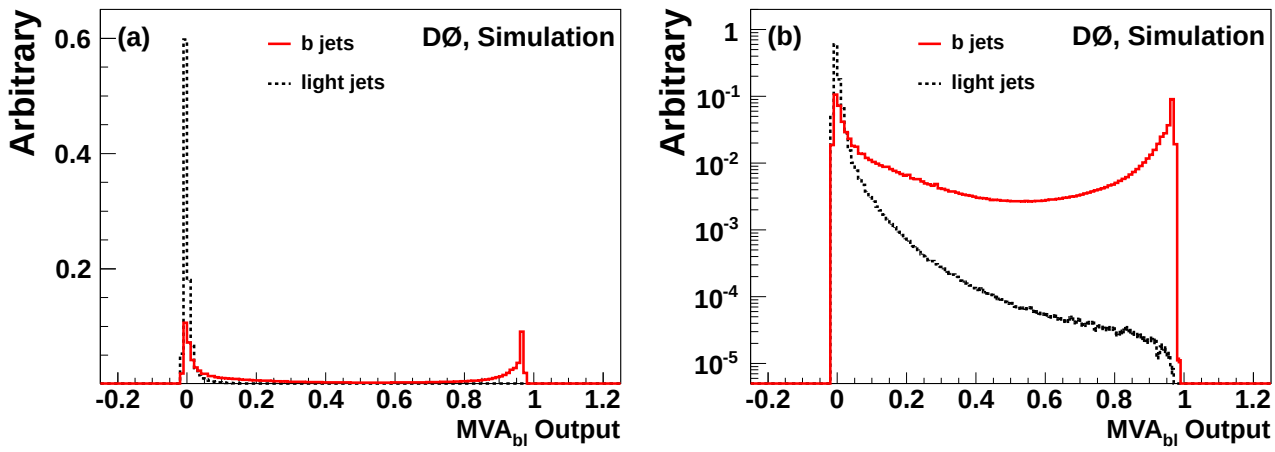

FIG. 2: (color online) The MVA$_{bl}$ output for light flavored and b jets in MC events, with (a) linear and (b) logarithmic scales. Both distributions are normalized to unity.

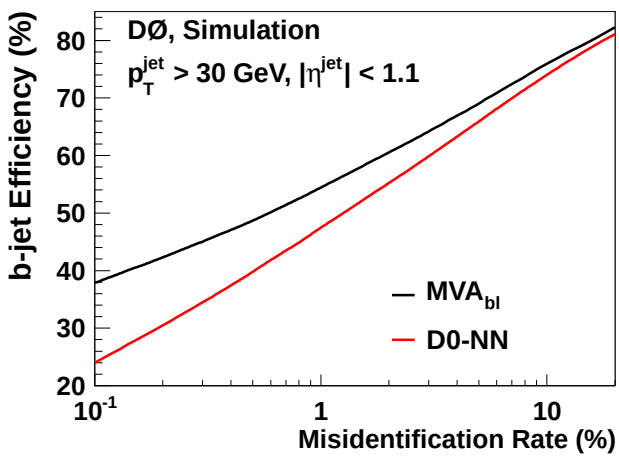

FIG. 3: (color online) The performance of the MVA$_{bl}$ and D0-NN algorithms for jets with $|\eta^{\rm jet}| < 1.1$ and $p_T^{\rm jet} > 30$ GeV.

content without applying any additional requirements on the muonic jet. The following coefficients are introduced into the S8 formulation to account for possible correlations between these *tags*:

β : Correlations between the *away tag* and MVA$_{bl}$ requirements for b jets.

α : Correlations between the *away tag* and MVA$_{bl}$ requirements for cl jets.

κ_b : Correlations between the p_T^{rel} and MVA$_{bl}$ requirements for b jets.

κ_{cl} : Correlations between the p_T^{rel} and MVA$_{bl}$ requirements for cl jets.

The above *tags* are denoted as k, for the MVA$_{bl}$ requirement; m, for the p_T^{rel} requirement; and, b, for the *away* tag. These are applied both individually and concurrently and will appear as superscripts in the following system of S8 equations:

$$\begin{array}{lcll}
f_b & + & f_{cl} & = 1 \\
f_b\varepsilon_b^k & + & f_{cl}\varepsilon_{cl}^k & = Q^k \\
f_b\varepsilon_b^m & + & f_{cl}\varepsilon_{cl}^m & = Q^m \\
f_b\varepsilon_b^n & + & f_{cl}\varepsilon_{cl}^n & = Q^n \\
f_b\kappa_b\varepsilon_b^k\varepsilon_b^m & + & f_{cl}\kappa_{cl}\varepsilon_{cl}^k\varepsilon_{cl}^m & = Q^{k,m} \\
f_b\varepsilon_b^m\varepsilon_b^n & + & f_{cl}\varepsilon_{cl}^m\varepsilon_{cl}^n & = Q^{m,n} \\
f_b\beta\varepsilon_b^n\varepsilon_b^k & + & f_{cl}\alpha\varepsilon_{cl}^n\varepsilon_{cl}^k & = Q^{n,k} \\
f_b\kappa_b\beta\varepsilon_b^k\varepsilon_b^m\varepsilon_b^n & + & f_{cl}\kappa_{cl}\alpha\varepsilon_{cl}^k\varepsilon_{cl}^m\varepsilon_{cl}^n & = Q^{k,m,n},
\end{array} \tag{2}$$

where the subscripts b and cl refer either to b or cl jets, Q refers to the fraction of the total number of selected jets in the sample that pass a given *tag*, f_X denotes the fraction of events of a given flavor X in the initial untagged sample, and ε_X^Y refers to the efficiency of a jet of flavor X passing *tag* Y. Q is determined from the data and α, β, κ_b, and κ_{cl} are determined from simulations [1]. This leaves eight remaining unknowns which form the solution, including the variable we are interested in: ε_b^k, the efficiency of a b jet passing the MVA$_{bl}$ requirement. These equations give two possible solutions for ε_b^Y but this can be resolved by requiring that $\varepsilon_b^Y > \varepsilon_{cl}^Y$.

The b jet identification efficiency obtained with the S8 method is valid for muonic jets. To obtain the efficiency for inclusive b jet decays, a correction factor is determined by using two samples of simulated b jets: muonic and inclusive. The final efficiency is then defined as

$$\varepsilon_b^{\rm data} = \frac{\varepsilon_{b\to\mu X}^{\rm data}}{\varepsilon_{b\to\mu X}^{MC}} \times \varepsilon_b^{MC} = SF \times \varepsilon_b^{MC} \tag{3}$$

where $SF = \varepsilon_{b\to\mu X}^{\rm data}/\varepsilon_{b\to\mu X}^{MC}$ is the data-to-simulation efficiency correction factor, $\varepsilon_{b\to\mu X}^{\rm data}$ is the efficiency for passing all MVA$_{bl}$ OPs as measured by the S8, and $\varepsilon_{b\to\mu X}^{\rm MC}$

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

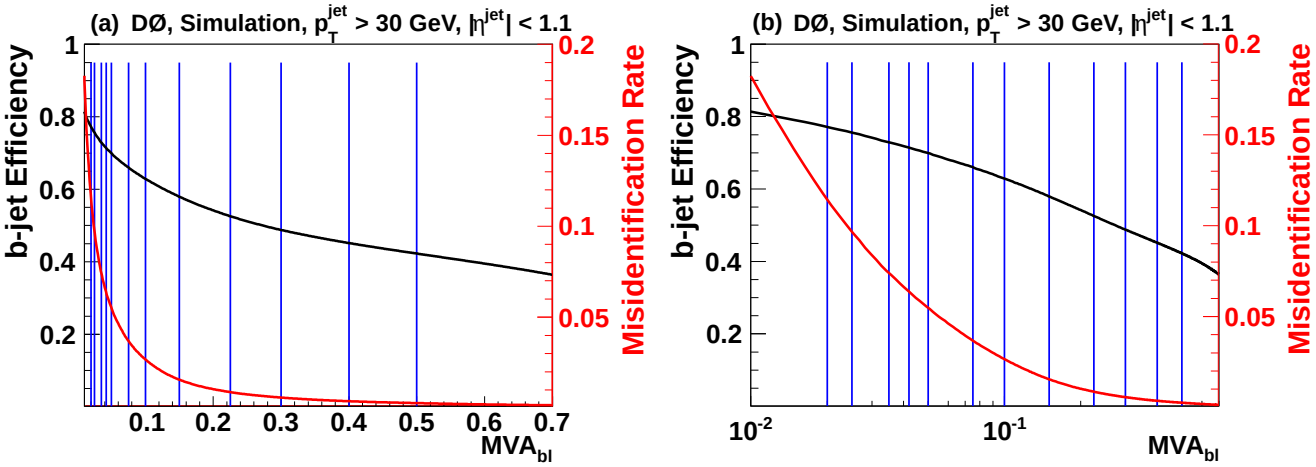

FIG. 4: (color online) The efficiency for selecting a b jet and the light jet misidentification rate as a function of the MVA$_{bl}$ requirement as determined in simulations. The vertical lines correspond to the selected operating points described in Sec. VII C, with (a) linear and (b) logarithmic scales.

is the efficiency measured in simulation. The identification efficiency for c jets is not measured directly from the data. It is assumed that the data-to-simulation scale factor is identical for b and c jets [1]. The c jet identification efficiency is then derived from the simulation as

$$\varepsilon_c^{\text{data}} = SF \times \varepsilon_c^{MC}. \tag{4}$$

B. MVA$_{bl}$ efficiency

Using this methodology we are able to determine ε_b^{data} for the set of OP requirements. We have selected two OPs, Loose and Tight, for demonstration.

In Fig. 5 the efficiency for identifying a muonic b jet, $\varepsilon_{b\to\mu X}$, is shown for data and MC. The ratio of these two efficiencies, SF, is also displayed. Figs. 6 and 7 show the MC and data corrected efficiencies for b and c jets in dijet events, respectively. The data efficiency curves are corrected with the parameterized correction factor derived in Fig. 5. Finally, in Fig. 8, we present the total systematic uncertainty for the S8 method on ε_b^{data}, discussed in Ref. [1], parameterized as a function of jet p_T.

IX. MISIDENTIFICATION RATE DETERMINATION

A precise understanding of the misidentification rates is especially important in searches for rare processes which can be overwhelmed by large backgrounds. Previous methods [1–3] to determine this rate relied heavily on simulation. The method in Ref. [1] for estimating the misidentification rate uses "negatively tagged" (NT) jets, or those with negative IP, with input from simulation. Here we present the *SystemN* (SN) method which extracts misidentification rates directly from data.

A. SystemN method

The SN method uses a series of linear equations to describe the efficiency for light jets to satisfy the various MVA$_{bl}$ OPs. Using a data sample of inclusive dijet events (the *inclusive jet* sample) we separate events as determined by the OP boundaries. If we have n OPs, then there will be n+1 bins, with each bin containing all the jets between the two consecutive OP's MVA$_{bl}$ values. An equation relating the number of jets of each flavor, along with their identification efficiencies, to the total number of retained jets in each bin is formed:

$$N \;=\; \varepsilon_l n_l + \varepsilon_c n_c + \varepsilon_b n_b, \tag{5}$$

where N is the number of selected jets in that bin, ε_X is the efficiency to identify a jet of flavor X, and n_X is the number of jets of flavor X in the total sample. The measured b and c jet efficiencies from the S8 method are used to predict the rate for selecting b and c jets in each bin. For example, the equations describing a selection of five arbitrary OPs is given below (a total of twelve OPs are defined in the real analysis):

$$\begin{array}{llll}
\varepsilon_l^{OP5} n_l & +\varepsilon_c^{OP5} n_c & +\varepsilon_b^{OP5} n_b & = N_{OP5} \\
\varepsilon_l^{OP4-5} n_l & +\varepsilon_c^{OP4-5} n_c & +\varepsilon_b^{OP4-5} n_b & = N_{OP4-5} \\
\varepsilon_l^{OP3-4} n_l & +\varepsilon_c^{OP3-4} n_c & +\varepsilon_b^{OP3-4} n_b & = N_{OP3-4} \\
\varepsilon_l^{OP2-3} n_l & +\varepsilon_c^{OP2-3} n_c & +\varepsilon_b^{OP3-4} n_b & = N_{OP2-3} \\
\varepsilon_l^{OP1-2} n_l & +\varepsilon_c^{OP1-2} n_c & +\varepsilon_b^{OP1-2} n_b & = N_{OP1-2} \\
\varepsilon_l^{aOP1} n_l & +(1-\varepsilon_c^{OP1}) n_c & +(1-\varepsilon_b^{OP1}) n_b & = N_{OP1},
\end{array} \tag{6}$$

where ε_X^{OPi-j} is the efficiency for selecting a jet of flavor X between the the ith and j^{th} OP boundaries. The anti-OP1 point, aOP1, is the set of all jets which fall below the OP1 requirement. The number of jets of a given flavor,

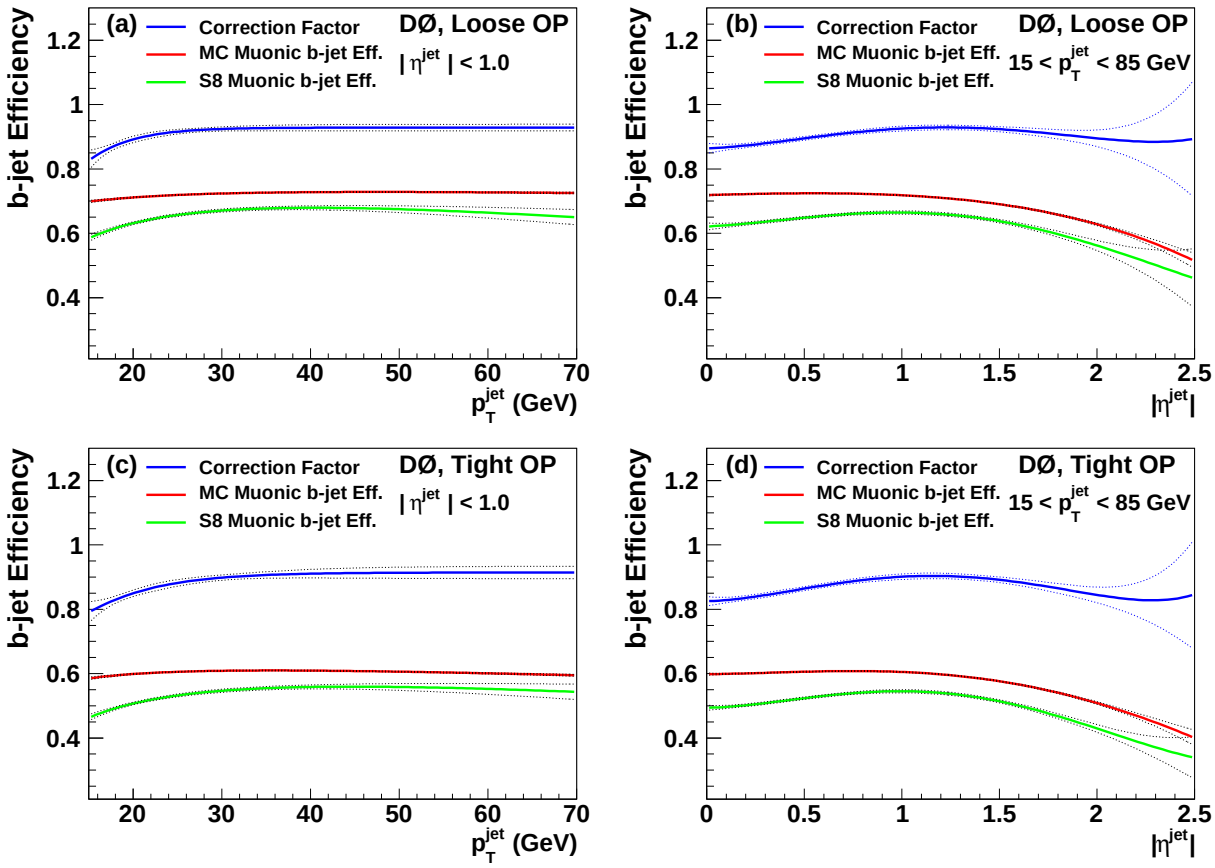

FIG. 5: (color online) The efficiency for selecting a muonic b-jet in MC and data using the S8 method. The correction factor, SF, which is used to model the algorithm's efficiency, is also shown. Two OPs are shown (a,b) the Loose and (c,d) Tight. The efficiencies are parameterized as a function of (a,c) p_T, for central jets and versus (b,d) η. The band which surrounds the lines corresponds to $\pm 1\sigma$ total uncertainties.

n_X, can be extracted from the data using a template fit based on the $\mathrm{M_{SV}}$ distributions corresponding to each jet flavor, as described below.

B. Sample composition

A measurement of the overall flavor composition is obtained by fitting $\mathrm{M_{SV}}$ templates for b, c, and light jets to a data distribution. These fits provide the number of b and c jets after the MVA_{bl} and SVT requirements, $n_b^{M_{SV}}$ and $n_c^{M_{SV}}$. Applying these requirements creates a sample enriched in heavy flavor jets. The sample composition of the *inclusive jet* sample is calculated by extrapolating from this *heavy flavor* sample using b and c jet selection efficiencies measured using the S8 procedure for jets passing MVA_{bl} and SVT requirements. The data sample is divided into several jet p_T and η bins to provide a parameterization of the sample composition.

Data is used to estimate the $\mathrm{M_{SV}}$ template shapes for the different jet flavors. For the b and c jet $\mathrm{M_{SV}}$ templates, a data-to-MC correction factor is estimated by comparing the $\mathrm{M_{SV}}$ distributions in a separate data sample (described in Sec. IX B 1) to the MC templates on a bin-by-bin basis. For light jets, $\mathrm{M_{SV}}$ template shapes are estimated using a data sample enriched in light jets, described in Sec. IX B 2.

1. *Corrections to the heavy flavor templates*

To obtain an estimate of the shape of the heavy flavor jet $\mathrm{M_{SV}}$ distribution from data, a heavy flavor enriched dijet sample is constructed by requiring:

- Two taggable jets with a separation of $|\Delta\phi(\mathrm{jet}_1, \mathrm{jet}_2)| > 2.5$.
- A jet must be selected by passing both an MVA_{bl} and SVT requirement.
- The recoiling jet must be matched to a muon, $p_T^{\mu} >$ 8 GeV, and pass a SVT requirement with $\mathrm{M_{SV}} >$ 1.8 GeV.

The ratio of the data $\mathrm{M_{SV}}$ distribution and the MC predicted $\mathrm{M_{SV}}$ templates for b, c, and light jets are used

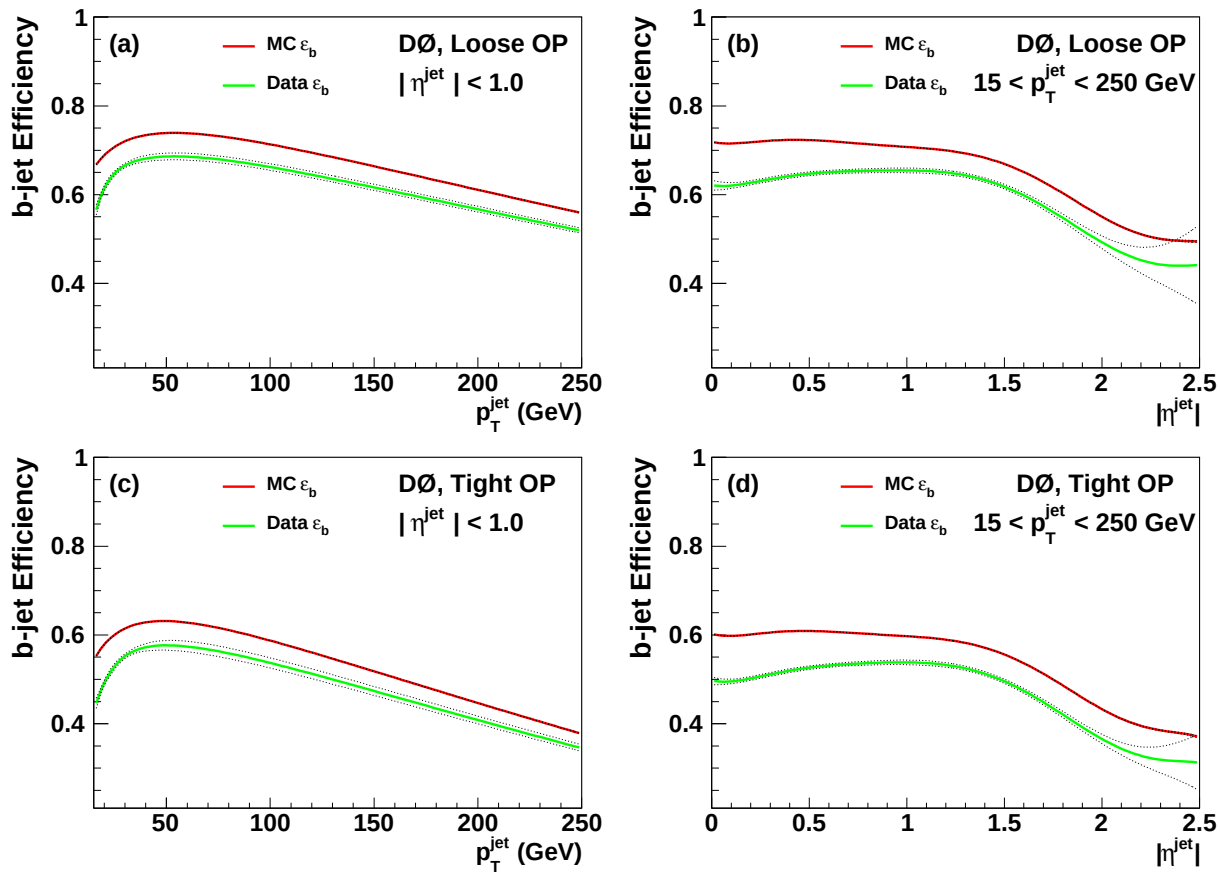

FIG. 6: (color online) The MC b jet identification efficiency, as measured in dijet events along with the data b jet identification efficiency. Two OPs are shown (a,b) the Loose and (c,d) Tight. The efficiencies are parameterized as a function of (a,c) p_T, for central jets and versus (b,d) η.

to create a correction factor. To determine the normalization of the MC templates the predicted sample composition is taken from the MC. This correction factor is then applied to the MC b and c jet $\mathrm{M_{SV}}$ templates to correct their shape to the data in separate jet p_T and η bins, an example of the corrected mass template is shown in Fig. 9.

2. *Data driven light jet templates*

The light jet templates are estimated from $\mathrm{M_{SV}}$ distribution of jets in a NT data sample [1]. This sample comprises jets having a negative IP and passing an SVT selection. The shape of the $\mathrm{M_{SV}}$ distribution corresponding to this sample is affected by contamination due to the presence of heavy flavor jets and as such is not a perfect representation of the light jet $\mathrm{M_{SV}}$ shape in data. The NT template shapes are measured from data in each p_T and η interval. Fig. 10 shows a comparison between the NT $\mathrm{M_{SV}}$ distribution and the MC light jet template. The difference in the shapes is taken as a systematic uncertainty.

3. *Sample composition measurement*

The data driven templates obtained above are used to fit the $\mathrm{M_{SV}}$ distribution in data using a log-likelihood fitter in bins of jet p_T and η. An example of a fit to the $\mathrm{M_{SV}}$ distribution using the b, c, and light jet templates is shown in Fig. 11. This results in a measurement of the fraction of each flavored jet type in that bin. The fits in each of the p_T and η regions are subsequently extrapolated back to the full *inclusive jet* sample using the b and c jet efficiency distributions measured for the MVA_{bl} and SVT algorithms. The number of events of heavy flavor, HF (either b or c), in the *inclusive jet* sample is calculated using the following formula:

$$n_{HF} = N \times f_{HF} = N \times \frac{f_{HF}^{Tag}}{\varepsilon_{HF}^{Tag}} \tag{7}$$

where f_{HF}^{Tag} is the fraction of jets with flavor HF extracted from the heavy flavor enriched sample and ε_{HF}^{Tag} is the S8 efficiency for a MVA_{bl} and SVT requirements, and N is the total number of events in that bin. The efficiency is calculated for the average p_T and η of the jets in

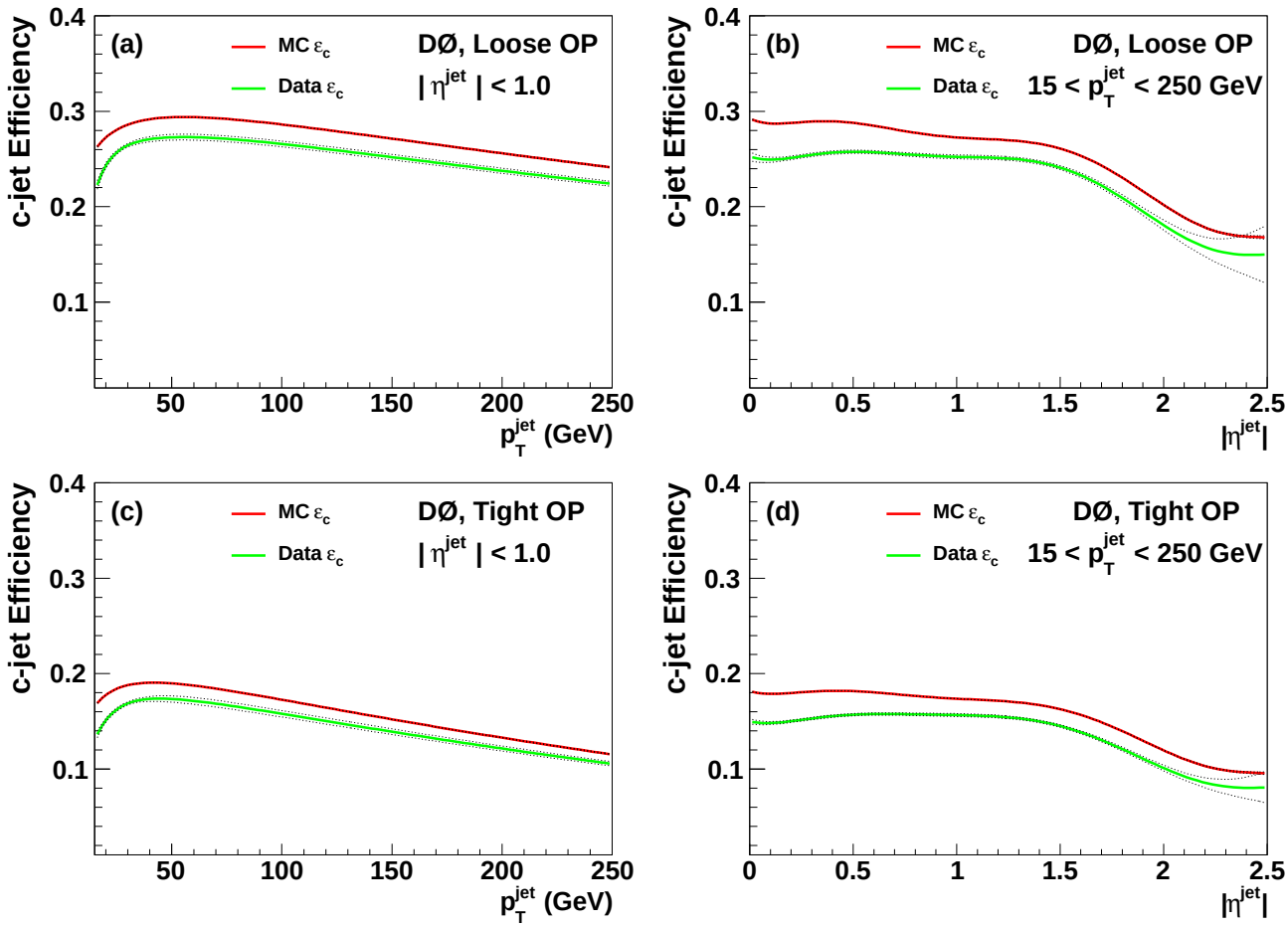

FIG. 7: (color online) The MC c jet identification efficiency, as measured in dijet events along with the data b jet identification efficiency. Two OPs are shown (a,b) the Loose and (c,d) Tight. The efficiencies are parameterized as a function of (a,c) p_T, for central jets and versus (b,d) η.

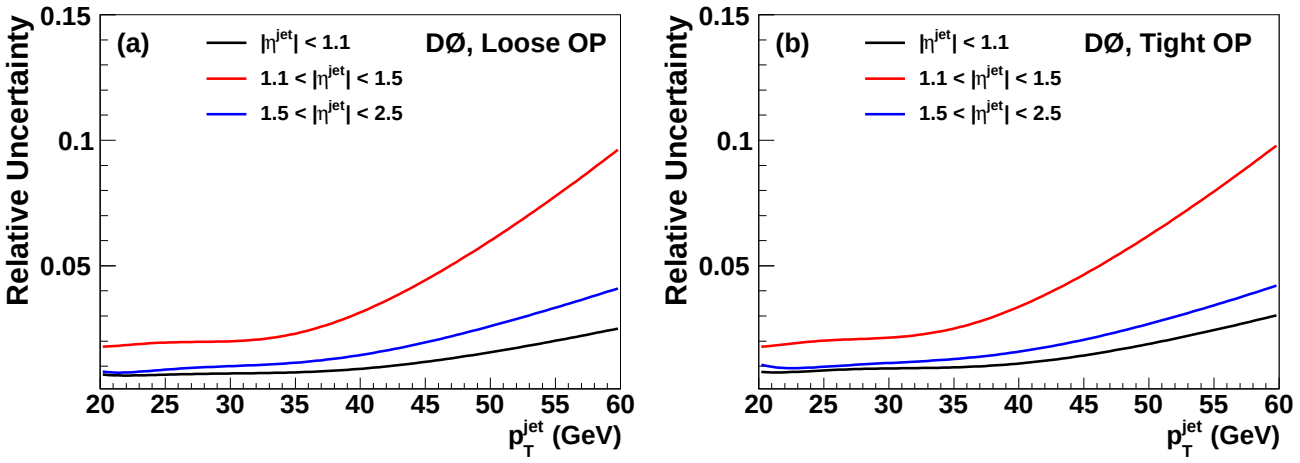

FIG. 8: (color online) The total uncertainty on $\varepsilon_b^{\rm data}$ from the S8 method as a function of p_T for two choices of OPs (a) Loose and (b) Tight.

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

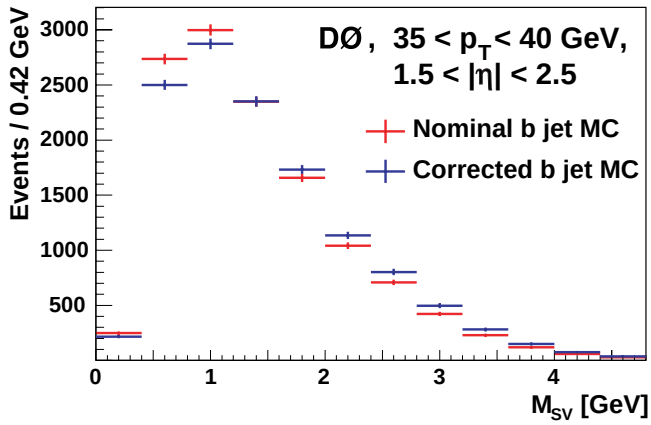

FIG. 9: (color online) Comparison of MC and data corrected b jet M_{SV} template shapes for jets with $1.5 < |\eta| < 2.5$ and $35 < p_T < 45$ GeV. The data corrected M_{SV} template uses a shape reweighting derived in Sec. IX B 1.

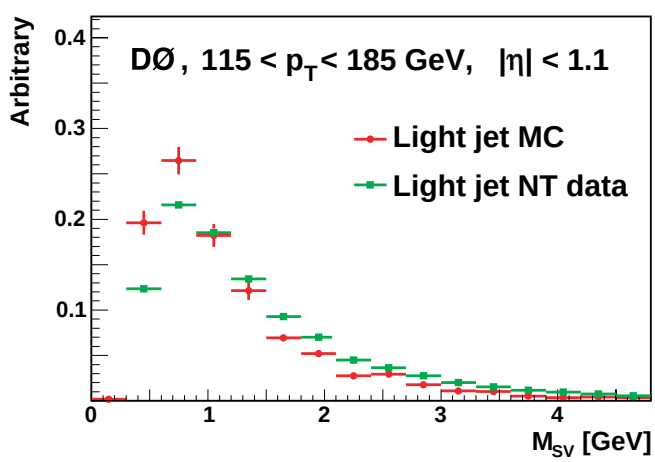

FIG. 10: (color online) Comparison of the MC light jet and NT M_{SV} mass templates for jets with $|\eta| < 1.1$ and $115 < p_T < 185$ GeV. The difference is taken as a measure of the systematic uncertainty due to residual contamination from heavy flavor jets in the NT data.

the region. While f_{HF}^{Tag} can be corrected to the *inclusive jet* sample, the light jet fraction cannot be. The corresponding light jet fraction in the *inclusive jet* sample is then determined from $f_l = 1 - f_b - f_c$.

The parameterization of the *inclusive jet* sample composition is important to obtain the misidentification rate as a function of p_T and to minimize the effect of statistically limited bins at high p_T. However, the choice of parameterization is not straightforward. The optimal parameterizations were determined by considering the χ^2 probability of various functional forms, typically a first order polynomial or a second order logarithmic polynomial.

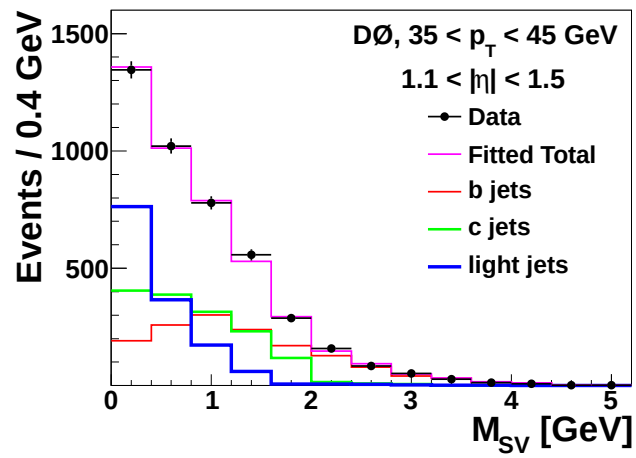

FIG. 11: (color online) An example of the sample composition fit using the M_{SV} for jets which pass MVA_{bl} and SVT requirements and have $35 < p_T < 45$ GeV and $1.1 < |\eta| < 1.5$. The b, c, and light jets are fit to the data resulting in the total fitted distribution.

C. Solutions of the SystemN equations

Instead of solving Eq. 6 analytically, we form a likelihood to improve the stability of the solutions. In this likelihood we take the equations and compare them to what is predicted from simulations. We allow the extracted flavor fractions, f_X, to float within their uncertainties during this fit. To help constrain this likelihood a second set of SN equations is built using a new data sample, the full procedure is repeated and added to the likelihood fit. This new sample is a sub-set of the *inclusive jet* sample which has the additional requirement that the recoiling "away jet" must be matched to a muon. This sample is defined as the "*away* jet sample".

The resulting likelihood is formed by summing over each of the OP bins for both samples:

$$LLH = -2 \sum_{S}^{N_S} \sum_{x=OP}^{N_{OP}} (N_x^S \ln(N_x^{MC}) - N_x^{MC}) \quad (8)$$

where N_x^S is the number of data events in sample S, either *inclusive* or *away jet* sample, in the MVA_{bl} interval x, N_x^{MC} is the predicted number of events in OP bin x. A normalization factor, LLH_{Norm}, is used to ensure that the likelihood values remain well defined:

$$LLH_{Norm} = -2 \sum_{S}^{N_S} \sum_{x=OP}^{N_{OP}} (N_x^S \ln(N_x^S) - N_x^S) \quad (9)$$

which is then subtracted from the likelihood.

We use the b and c jet fractions measured in the previous section to help stabilize the fit through a term which is added to the likelihood:

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

$$d^T E^{-1} d. \tag{10}$$

E is a 2×2 covariance error matrix resulting from the extraction of the b and c jet content from the $\mathrm{M_{SV}}$ fit and d is a vector

$$d = \begin{pmatrix} n_b - n_b^{M_{SV}} \\ n_c - n_c^{M_{SV}} \end{pmatrix}, \tag{11}$$

where $n_x^{M_{SV}}$ is the number of jets, of flavor x, estimated from the $\mathrm{M_{SV}}$ template fits, and n_x are the number of jets, of flavor x, in the *inclusive sample*. The result of this likelihood fit is the extraction of the final data driven light jet efficiency parameterized over jet p_T and η in OP bins. These misidentification rates are shown in Fig. 12.

D. SystemN systematic uncertainties

The three dominant systematic uncertainties on the misidentification rates are:

- The shape of the b and c jet $\mathrm{M_{SV}}$ templates
- The shape of the light jet $\mathrm{M_{SV}}$ template
- The uncertainty on the b and c jet efficiencies from the S8 method

Heavy flavor template shape. The effect of imperfections in the modeling of the b and c jet $\mathrm{M_{SV}}$ templates is estimated by carrying out the sample composition measurement using a set of heavy flavor $\mathrm{M_{SV}}$ templates which are not corrected to data in each of the p_T and η intervals. The full difference between the MC and data corrected sample composition predictions is used as an uncertainty. As described in Sec. IX B 1, the heavy flavor templates are derived using MC inputs. These inputs are then varied and the largest deviation from the nominal shape is used to provide an additional uncertainty.

Light flavor template shape. The uncertainty due to the shape of the light jet $\mathrm{M_{SV}}$ templates is estimated by performing the sample composition fit using both the NT and MC light jet template shapes, taking the difference in the sample composition to assign an uncertainty.

***b* and *c* jet efficiency uncertainty**. When extrapolating the flavor fractions, measured in the heavy flavor enriched sample, to the *inclusive jet* sample the efficiencies from the S8 method are used. To account for the uncertainties inherited in this procedure it is repeated after the efficiencies are varied by $\pm 1\sigma$. This variation will only affect the extrapolation procedure.

The parameterization of the systematic uncertainties is evaluated by carrying out closure tests, where the percentage difference between the number of actually selected jets and the predicted number of jets in various bins in p_T and η regions are compared. The uncertainty is determined from the RMS of the resulting distributions. The total uncertainty on the data-driven misidentification rate attained using the SN method, given by the statistical and systematic uncertainties combined in quadrature, is shown in Fig. 13 for the Loose and Tight OPs of the MVA_{bl} algorithm.

E. Comparison with previous method

A comparison between the misidentification rates of the D0-NN algorithm measured using the SN method and those estimated by the NT method of Ref. [1] is shown in Fig. 14. Both provide comparable uncertainties. For the looser OPs the central value of the new method gives a misidentification rate roughly 20% higher than the central values for the previous method, and for the tighter OPs the difference is closer to 35%. The two methods do agree with each other within uncertainties across the full range of jet p_T, but the misidentification rate for the NT method is systematically lower.

The source of this difference comes from the use of simulation in the NT method. With the removal of the V^0s the main source of misidentified light jets comes from detector resolution and track mis-reconstruction effects. The simulation does not accurately reproduce these effects by modeling ideal detector responses and the resulting misidentification rate as determined by the NT method is systematically underestimated.

F. MVA_{bl} misidentification rates

The final results are the misidentification rate for light jets extracted from our data, as shown in Fig. 12. These are parameterized in terms of p_T for three different η regions. This data-driven measurement of the misidentification rate can be combined with that modeled in simulation and we can derive a MC correction factor, as shown in Fig. 15. These correction factors are applied in the light jet simulations (for jets passing the MVA_{bl} requirements). Table II shows the responses, efficiencies, and misidentification rates, of the MVA_{bl} algorithm as measured in data.

X. SUMMARY AND CONCLUSIONS

The identification of heavy flavor jets is a crucial component of particle physics analyses. Utilizing the unique characteristics of the fragmenting b quark we created algorithms which allow for the identification of b jets with high efficiency and purity. The MVA_{bl} algorithm shows improvements over previous algorithms utilized at D0. For a light jet misidentification rate of 1% we observe an improvement in the efficiency over the D0-NN algorithm

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

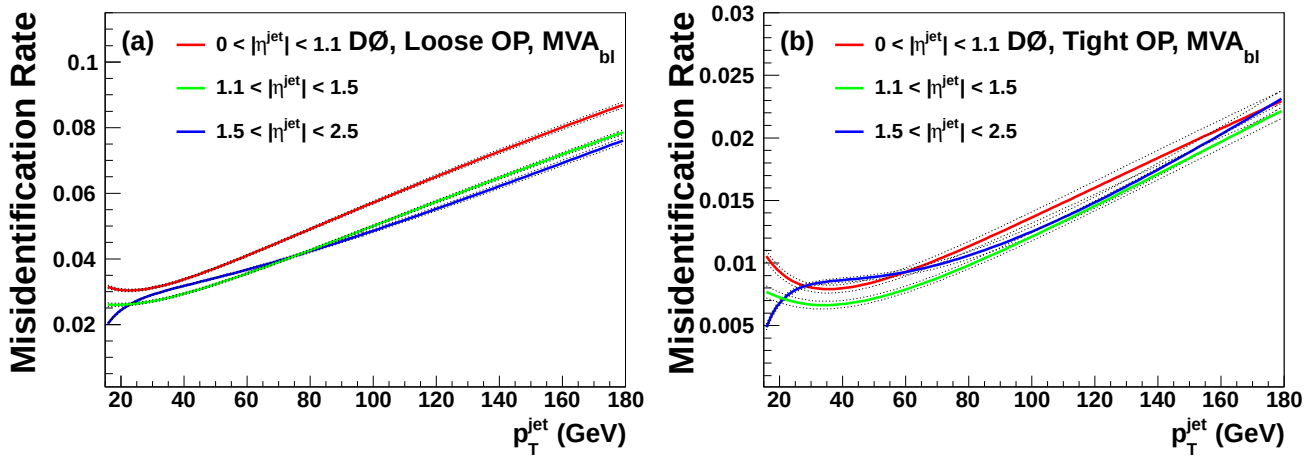

FIG. 12: (color online) The SN data driven misidentification rates for the MVA$_{bl}$ algorithm. Two OPs are shown (a) Loose and (b) Tight. These are further parameterized over jet p_T and for three different jet η intervals: $0 < |\eta| < 1.1$, $1.1 < |\eta| < 1.5$, and $1.5 < |\eta| < 2.5$. The black dotted lines represent the uncertainty on the fit.

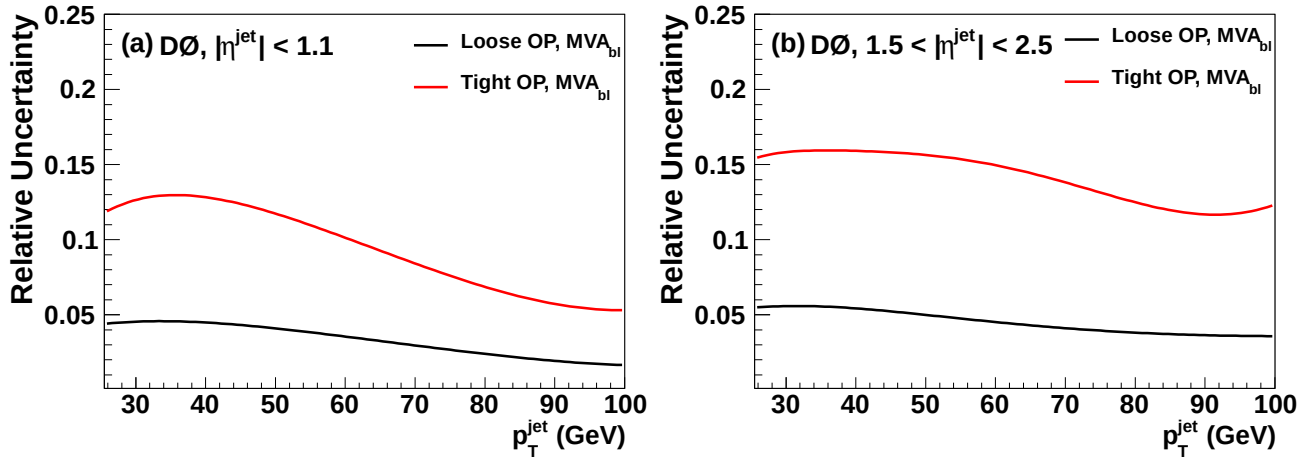

FIG. 13: (color online) The total relative uncertainty on the misidentification rate from the SN method parameterized in terms of jet p_T and for two different η regions: (a) $|\eta| < 1.1$ and (b) $1.5 < |\eta| < 2.5$.

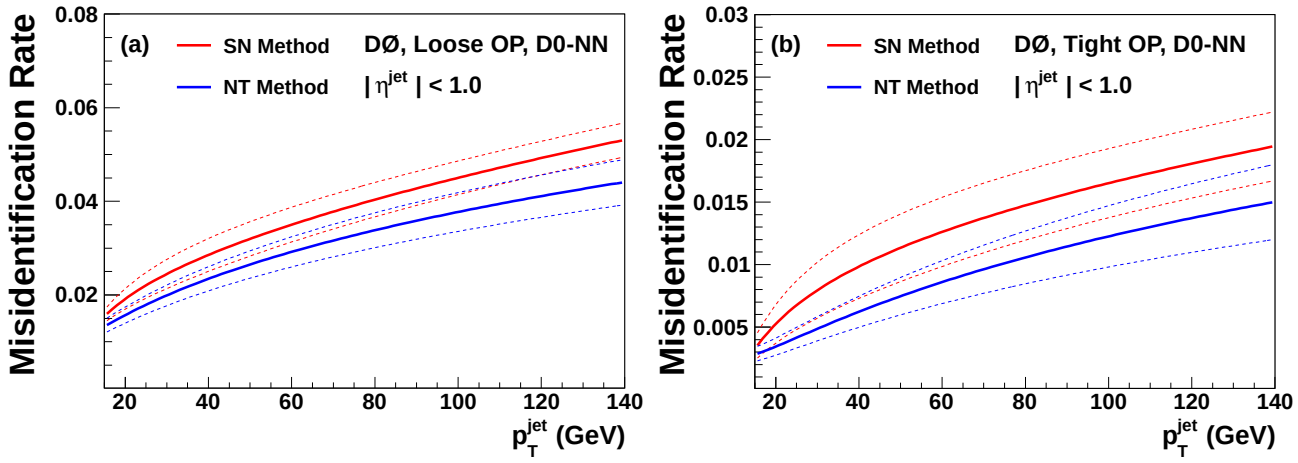

FIG. 14: (color online) Comparison between the misidentification rates of the D0-NN derived for two OPs, (a) Loose and (b) Tight, using the new SN method and the old method described in Ref. [1]. The dashed bands which surround the values correspond to the total uncertainties.

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

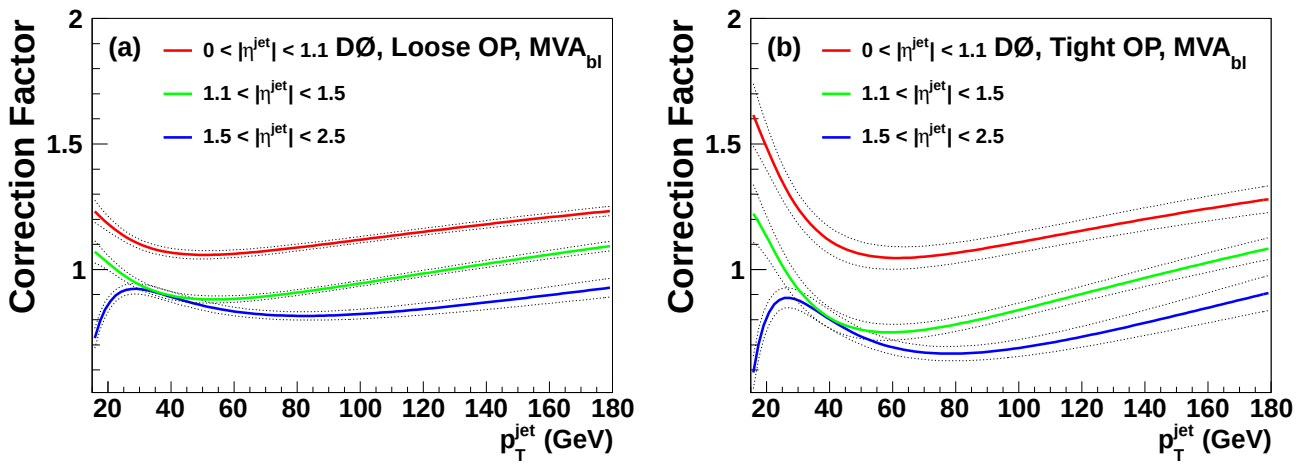

FIG. 15: (color online) The misidentification rate correction factors for the light jet MC which are derived by taking the ratio of the data and MC misidentification rates. Two OPs are shown, (a) Loose and (b) Tight. These are further parameterized over jet p_T and for three different jet η intervals: $0 < |\eta| < 1.1$, $1.1 < |\eta| < 1.5$, and $1.5 < |\eta| < 2.5$. The black dotted lines represent the uncertainty on the fit.

TABLE II: The efficiency of selecting a b, c, or light jet using the MVA$_{bl}$ as determined by using the S8 and SN method directly from data for 12 OPs. The total uncertainties are included along with the OP definitions.

OP Name	Min.	$\lvert\eta^{\rm jet}\rvert < 1.1$			$1.1 < \lvert\eta^{\rm jet}\rvert < 1.5$			$1.5 < \lvert\eta^{\rm jet}\rvert < 2.5$		
	MVA$_{bl}$	$\varepsilon_b^{\rm data}$[%]	$\varepsilon_c^{\rm data}$[%]	$\varepsilon_l^{\rm data}$[%]	$\varepsilon_b^{\rm data}$[%]	$\varepsilon_c^{\rm data}$[%]	$\varepsilon_l^{\rm data}$[%]	$\varepsilon_b^{\rm data}$[%]	$\varepsilon_c^{\rm data}$[%]	$\varepsilon_l^{\rm data}$[%]
L6	0.02	74.8 ± 0.6	39.2 ± 0.3	15.5 ± 0.3	75.3 ± 0.6	38.2 ± 0.3	13.9 ± 0.5	64.8 ± 0.7	31.9 ± 0.4	13.7 ± 0.4
L5	0.025	73.2 ± 0.6	36.8 ± 0.3	13.3 ± 0.3	73.7 ± 0.6	36.0 ± 0.3	11.9 ± 0.5	62.7 ± 0.7	29.6 ± 0.4	11.6 ± 0.2
L4	0.035	70.2 ± 0.6	33.0 ± 0.3	10.5 ± 0.2	70.7 ± 0.7	32.2 ± 0.3	9.4 ± 0.5	59.1 ± 0.8	25.8 ± 0.3	9.1 ± 0.4
L3	0.042	68.9 ± 0.7	31.2 ± 0.3	9.2 ± 0.2	69.3 ± 0.7	30.4 ± 0.3	8.2 ± 0.5	57.5 ± 0.8	24.3 ± 0.3	8.0 ± 0.4
L2	0.05	67.5 ± 0.8	29.6 ± 0.3	8.0 ± 0.2	68.0 ± 0.8	28.9 ± 0.3	7.2 ± 0.5	56.1 ± 0.8	23.0 ± 0.3	7.0 ± 0.4
Loose	0.075	63.8 ± 0.8	25.4 ± 0.3	5.63 ± 0.2	64.3 ± 0.8	24.9 ± 0.3	5.0 ± 0.5	51.9 ± 1.0	19.4 ± 0.3	4.9 ± 0.2
oldLoose	0.1	61.1 ± 0.7	22.8 ± 0.3	4.2 ± 0.2	61.7 ± 0.7	22.3 ± 0.3	3.8 ± 0.5	49.0 ± 0.8	17.2 ± 0.3	3.7 ± 0.1
Medium	0.15	56.7 ± 0.6	19.0 ± 0.2	2.6 ± 0.2	57.4 ± 0.6	18.7 ± 0.2	2.3 ± 0.5	44.5 ± 0.8	14.1 ± 0.2	2.4 ± 0.4
Tight	0.225	51.6 ± 0.7	15.4 ± 0.2	1.4 ± 0.1	52.4 ± 0.7	15.3 ± 0.2	1.3 ± 0.4	39.5 ± 0.7	11.1 ± 0.2	1.3 ± 0.4
VeryTight	0.3	47.4 ± 0.6	12.9 ± 0.2	0.9 ± 0.2	48.3 ± 0.6	12.9 ± 0.2	0.8 ± 0.4	35.4 ± 0.7	9.1 ± 0.2	0.8 ± 0.4
UltraTight	0.4	43.4 ± 0.7	10.9 ± 0.2	0.6 ± 0.2	44.8 ± 0.6	11.0 ± 0.2	0.4 ± 0.4	31.9 ± 0.6	7.4 ± 0.2	0.5 ± 0.3
MegaTight	0.5	40.4 ± 0.6	9.5 ± 0.1	0.4 ± 0.2	41.8 ± 0.6	9.6 ± 0.1	0.2 ± 0.3	29.2 ± 0.7	6.3 ± 0.2	0.4 ± 0.3

for selecting a b jet of 15% per jet. A new method for extracting the misidentification rate directly from data has also been presented. The data-derived misidentification rates of the SystemN method are compatible within uncertainties with previous simulation-based methods, however a systematic difference is observed. This difference is due to the limited ability of the simulation to accurately model resolution and track mis-reconstruction effects. By removing this dependence on simulation the SystemN method provides a more accurate and reliable measurement of the light jet misidentification rates in data.

Acknowledgement

We thank the staffs at Fermilab and collaborating institutions, and acknowledge support from the DOE and NSF (USA); CEA and CNRS/IN2P3 (France); MON, NRC KI and RFBR (Russia); CNPq, FAPERJ, FAPESP and FUNDUNESP (Brazil); DAE and DST (India); Colciencias (Colombia); CONACyT (Mexico); NRF (Korea); FOM (The Netherlands); STFC and the Royal Society (United Kingdom); MSMT and GACR (Czech Republic); BMBF and DFG (Germany); SFI (Ireland); The Swedish Research Council (Sweden); and CAS and CNSF (China).

[1] V. Abazov *et al.*, [D0 Collaboration], Nucl. Instrum. Methods Phys. Res. A **620**, 490 (2010).
[2] J. Freeman *et al.*, Nucl. Instrum. Methods Phys. Res. A **697**, 64 (2013).
[3] S. Chatrchyan *et al.*, [CMS Collaboration], JINST **8**, P04013 (2013).
[4] V. Abazov *et al.*, [D0 Collaboration], Nucl. Instrum. Methods Phys. Res. A **565**, 463 (2006).
[5] S. Ahmed *et al.*, [D0 Collaboration], Nucl. Instrum. Methods Phys. Res. A **634**, 8 (2011).
[6] S. Abachi *et al.*, [D0 Collaboration], Nucl. Instrum. Methods Phys. Res. A **324**, 53 (1993).
[7] B. Casey *et al.*, Nucl. Instrum. Methods **A698**, 208 (2013).
[8] T. Sjostrand *et al.*, Comput. Phys. Commun. **135**, 238 (2001).
[9] D. Lange, Nucl. Instrum. Methods Phys. Res. A **462**, 152 (2001).
[10] R. E. Kalman, Transactions of the ASME–Journal of Basic Engineering **82**, 35 (1960).
[11] G. C. Blazey *et al.*, arXiv:hep-ex/0005012, (2000).
[12] V. M. Abazov *et al.*, [D0 Collaboration], arXiv:hep-ex/13126873, (2013).
[13] A. Hocker *et al.*, PoS **ACAT**, 040 (2007).
[14] R. Brun and F. Rademakers, Nucl. Instrum. Methods Phys. Res. A **389**, 81 (1997).
[15] D. Hedin, T. Kramer, and K. Roberts, C87-11-11 (1987).

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.