A search is presented for the standard model Higgs boson produced in association with top quarks using the full Run II proton-antiproton collision data set, corresponding to $9.45 \text{ fb}^{-1}$, collected by the Collider Detector at Fermilab. No significant excess over the expected background is observed, and 95% credibility-level upper bounds are placed on the cross section $\sigma(t\bar{t}H \rightarrow \text{lepton} + \text{jets})$. For a Higgs boson mass of 125 $\text{GeV}/c^2$, we expect to set a limit of 12.6 and observe a limit of 20.5 times the standard model rate. This represents the most sensitive search for a standard model Higgs boson in this channel to date.

DOI: 10.1103/PhysRevLett.109.181802 PACS numbers: 14.80.Bn, 13.85.Rm

The mechanism of electroweak symmetry breaking [1] in the standard model (SM) [2] predicts the existence of a massive particle called the Higgs boson. The CDF and D0 Collaborations have reported evidence for a particle consistent with the SM Higgs boson with a mass between 120 and 135 $\text{GeV}/c^2$ produced in association with a $W$ or $Z$ boson with decays to two $b$ quarks [3]. The CMS and ATLAS Collaborations have reported the observation of a particle consistent with the SM Higgs boson with a mass of approximately 125 $\text{GeV}/c^2$, which decays to two photons, two $W$ bosons, or two $Z$ bosons [4]. Many other predicted couplings of the SM Higgs boson are currently neither observed nor excluded. In the SM, the fermion masses are generated by Yukawa couplings between the Higgs and the fermion fields with coupling strength proportional to the fermion masses. As the most massive known fermion, the top quark is predicted to couple most strongly to the Higgs boson. Higgs bosons may then be produced with a top-quark pair ($t\bar{t}H$) via radiation or top-quark fusion [5,6]. Samples of top-quark pair events with a few percent-level contamination from other processes can be selected at CDF, offering smaller background uncertainties than in searches for the SM Higgs boson. Hence, the top-quark pair associated production channel provides an important contribution to SM Higgs boson physics. Furthermore, proposed extensions to the SM could significantly enhance the rate of $t\bar{t}H$ production [8]. This enhancement might allow the observation of a non-SM Higgs boson in this search before reaching sensitivity to a SM Higgs boson and could help to distinguish a candidate Higgs boson in other searches from the SM Higgs boson.

This Letter reports a search for the SM Higgs boson produced in association with top quarks. We utilize the full data set recorded with the CDF II detector. The data set consists of proton-antiproton collisions at a center-of-mass energy of $\sqrt{s} = 1.96 \text{ TeV}$ and corresponds to an integrated

181802-3
momentum are ET of a reconstructed lepton (electron or muon), missing transverse energy (ET) [16], and four or more calorimeter energy clusters (jets). The details of the online selection, lepton identification, and jet identification are identical to those described in Ref. [7]. At least two of the jets in each event are required to be consistent with the fragmentation of a b quark (b-tagged) [7]. Because a low-mass (mH \leq 135 \text{ GeV}/c^2) SM Higgs boson is expected to decay mostly to pairs of b quarks, or pairs of W bosons, that will decay predominantly to pairs of u, d, s, or c quarks, large b-tag and jet multiplicities are requested by the selection. Approximately 90\% of the selected search sample is composed of top-quark pairs, with the remainder consisting of W or Z bosons accompanied by jets (W/Z + jets), single-top quarks, dibosons, and strong force mediated (QCD) multijets. Table I shows the expected composition of the data sample.

To select events during data taking, we require the presence of a charged lepton (electron e or muon \( \mu \)) candidate with transverse momentum \( p_T \geq 18 \text{ GeV}/c \).

We further require that the lepton candidate satisfies identification quality requirements, as in Ref. [17]. We require that \( E_T \) be greater than 10, 20, or 25 GeV in events containing a muon candidate, an electron candidate satisfying \(|\eta| \leq 1.1\), and an electron candidate satisfying \(|\eta| > 1.1\), respectively. These \( E_T \) requirements are chosen to optimize the signal selection efficiency and the rejection of instrumental backgrounds, which differ in the three samples. Jets are reconstructed using a cone-based clustering algorithm, with a cone radius \( |\Delta \phi| = \sqrt{\Delta \eta^2 + (\Delta \eta)^2} \) of 0.4 [18]. Jet energies are corrected for instrumental effects [19], and the corrected jets are required to have \( E_T > 20 \text{ GeV} \) and \(|\eta| < 2.0\). We also reconstruct lower energy clusters (12 < \( E_T < 20 \text{ GeV} \)) but do not define them as jets. We use two different algorithms to tag b jets, as in Ref. [7]. One algorithm relies on the reconstruction of secondary decay vertices from long-lived hadrons within the jet cone [20], while the other estimates the likelihood that not all tracks in the jet cone intersect the beam line [21]. Jets identified by either algorithm are considered as tagged, offering higher tagging efficiency than obtained by the use of one algorithm alone.

We model the various backgrounds using a combination of Monte Carlo simulation and data. We simulate the \( t \bar{t} \), diboson, W/Z + jets, and single-top backgrounds using the POWHEG [22], PYTHIA [23], ALPGEN [24], and MADEVENT [25] generators, respectively. We model the QCD multijet background using a data-driven model [17]. For backgrounds involving top quarks, we have

<table>
<thead>
<tr>
<th>Process</th>
<th>4 jets</th>
<th>5 jets</th>
<th>( \geq 6 ) jets</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t \bar{t} ) + jets</td>
<td>962 \pm 89</td>
<td>294 \pm 27</td>
<td>77 \pm 7.1</td>
</tr>
<tr>
<td>( t \bar{t} + b \bar{b} )</td>
<td>32 \pm 27</td>
<td>17 \pm 14</td>
<td>8.2 \pm 6.9</td>
</tr>
<tr>
<td>W/Z + jets</td>
<td>105 \pm 32</td>
<td>26 \pm 8.0</td>
<td>7.1 \pm 2.2</td>
</tr>
<tr>
<td>Multijet</td>
<td>31 \pm 16</td>
<td>0.0 \pm 1.0</td>
<td>0.0 \pm 1.0</td>
</tr>
<tr>
<td>Single top</td>
<td>19 \pm 2.2</td>
<td>3.7 \pm 0.43</td>
<td>0.61 \pm 0.070</td>
</tr>
<tr>
<td>Diboson</td>
<td>5.2 \pm 0.44</td>
<td>1.2 \pm 0.11</td>
<td>0.25 \pm 0.025</td>
</tr>
<tr>
<td>Total background</td>
<td>1150 \pm 106</td>
<td>340 \pm 33</td>
<td>93 \pm 11</td>
</tr>
<tr>
<td>Observed</td>
<td>1133</td>
<td>368</td>
<td>114</td>
</tr>
<tr>
<td>( t \bar{t}H )</td>
<td>0.65 \pm 0.075</td>
<td>1.1 \pm 0.13</td>
<td>1.2 \pm 0.14</td>
</tr>
<tr>
<td>WH</td>
<td>0.52 \pm 0.061</td>
<td>0.07 \pm 0.008</td>
<td>Negligible</td>
</tr>
<tr>
<td>ZH</td>
<td>0.09 \pm 0.011</td>
<td>0.02 \pm 0.002</td>
<td>Negligible</td>
</tr>
</tbody>
</table>
used $m_t = 172.5$ GeV/$c^2$. Signal models are generated by PYTHIA, with Higgs boson masses in $5$ GeV/$c^2$ increments in the range $100 \leq m_H \leq 150$ GeV/$c^2$. The signal samples are normalized to their next-to-leading-order cross section, as described in Ref. [5]. The CTEQ5L parton-distribution functions [26] and a detailed simulation of the response of the CDF II detector using GEANT3 [27] are employed in all Monte Carlo samples.

The search sample is subdivided into independent categories of different expected signal-to-background ratio and background composition to maximize the search sensitivity [28]. Under the selection requirements described above, the reconstructed jet-multiplicity spectrum in $t\bar{t}H$ events peaks at five jets, while the reconstructed jet-multiplicity spectrum for $t\bar{t}$ peaks at four jets. Hence, we separate events with four, five, or six or more jets. The jet-multiplicity samples are then separated by $b$-tag multiplicity. The events with six or more jets, at least three of which are $b$-tagged, feature the largest expected signal-to-background ratio and provide the most sensitivity for a low-mass Higgs boson.

After defining our search sample, we enhance the isolation of a SM Higgs signal using artificial neural networks (NNs) [29]. Each neural network is trained to separate simulated Higgs signal events from background, with individual networks optimized for each Higgs boson mass hypothesis in each of the previously described event categories. Each network uses 18 input variables to discriminate the Higgs boson signal from the backgrounds. These variables are missing transverse energy, maximum jet $E_T$, second-largest jet $E_T$, third-largest jet $E_T$, maximum $E_T$ among $b$-tagged jets, mean jet $E_T$, invariant mass of the combination of all objects (jets, leptons, $p_T$), vector sum of the transverse energies of all objects, scalar sum of the transverse energies of all objects, scalar sum of the transverse energies of all objects with $E_T$ between 12 and 20 GeV, minimum separation in $\eta$-$\phi$ space between $b$-tagged jets, separation in azimuth between the lepton and the missing transverse energy, transverse mass of the lepton and missing transverse energy [30], mass of the vector sum of the lepton and nearest jet in $\eta$-$\phi$ space, minimum mass of the vector sum of any pair of jets, mass of the vector sum of the two non-$b$-tagged jets with the largest $E_T$, and mass of the vector sum of the two $b$-tagged jets with the largest $E_T$. The modeling of the input distributions has been validated in the subset of the data with only four jets and only two $b$ tags, which is expected to contain a negligible number of signal events relative to the background yield. This region is used to constrain the various systematic uncertainties in situ. Two of these distributions can be seen in Figs. 1 and 2, and the output of the discriminant trained to identify a Higgs boson of mass 125 GeV/$c^2$ is shown in Fig. 3. A more detailed presentation of the modeling of these distributions is contained in Ref [28].

We consider several sources of systematic uncertainty that affect the rate of the involved processes and the shape of the discriminant distributions. Because of the high jet and $b$-tag multiplicities considered, the dominant systematic uncertainties are associated with estimates of the $b$-tag efficiency and the jet-energy scale. These affect both the rates and the discriminant shapes, and we estimate the effects by independently varying the estimated $b$-tag
efficiency and the jet-energy scale within 1 standard deviation. These variations in jet-energy scale and tagging efficiency alter the expected acceptance for signal and background by between 1 and 20%, depending on the selection category. In addition, to account for uncertainties on the theoretical cross sections of background processes, we assume the following systematic uncertainties on the normalization of simulated backgrounds: 6% for diboson production, 6% for single-top-quark production, 10% for inclusive $t\bar{t}$ production, and 40% for $W/Z + jets$ [31–34]. Smaller uncertainties include those on the amount of initial- and final-state radiation, parton-distribution function choice, the probability to $b$-tag light-quark jets, and a 6% uncertainty on the measurement of the integrated luminosity [28,35]. The total theoretical uncertainty on the signal samples is estimated to be 10%, as derived from that computed in Ref. [5], but accounting for a reduced uncertainty due to the measurement of the top-quark mass [17].

No measurement is available of the cross section for top-quark production with additional $b$ quarks generated from QCD radiation. The next-to-leading-order corrections to leading-order calculations of the production rate of top-quark pairs with additional $b$ quarks have been estimated to be on the order of a factor of 2 in some regions of phase space [36]. To account for this unknown and potentially large systematic uncertainty, inclusive $t\bar{t}$ simulated events were separated into subsamples with additional $b$ quarks generated from QCD radiation ($t\bar{t} + b\bar{b}$) and without ($t\bar{t} + jets$). We assume an uncertainty of 10% on the normalization of the $t\bar{t} + jets$ component and assume an uncertainty of 100% on the normalization of the $t\bar{t} + b\bar{b}$ component. We estimate the effect of individual systematic uncertainties by calculating the expected exclusion sensitivity considering all uncertainties and then comparing this value to that derived by considering all but one uncertainty. The uncertainty due to the jet-energy scale, $b$-tag efficiency, inclusive top pair cross section, and potential next-to-leading-order effects for $t\bar{t} + b\bar{b}$ individually degrade the expected exclusion sensitivity of the analysis by 7.8%, 5.4%, 6.9%, and 9.0%, respectively.

We compare the distribution of discriminant output observed in data to that of the expected background model. Observing no evidence for Higgs boson production in the discriminant distributions, we calculate a Bayesian 95% credibility-level (C.L.) limit for each mass hypothesis using the combined binned likelihood of the NN output distributions. Each of the three jet-multiplicity categories is subdivided into five independent tagging categories. A posterior density is obtained by multiplying this likelihood by Gaussian prior densities for the background normalizations and systematic uncertainties, leaving the cross section $\sigma(t\bar{t}H \rightarrow \ell + E_T + jets)$ with a uniform prior density, with priors truncated to prevent negative predictions. A 95% C.L. limit is determined such that 95% of the posterior density for the cross section accumulates below the limit [37]. The expected limits with one and two standard deviation uncertainty bands and the observed limits are shown as a function of assumed Higgs boson mass in Fig. 4. Because none of the discriminant function input variables act as an estimator for the reconstructed Higgs boson mass, the upper limits at different candidate Higgs boson masses are strongly correlated. An excess in the data produces an observed limit that exceeds the expected limit at all masses, at a level of approximately 1 standard deviation compared to the background-only hypotheses.

In conclusion, we have presented a search for a SM Higgs boson produced in association with a pair of top quarks.
quarks, in a final state involving a lepton, missing transverse energy, jets, and $b$-tagged jets. For a Higgs boson mass of 125 GeV/$c^2$, we expect a limit of 12.6 and observe a limit of 20.5 times the SM rate, which represents agreement with the background-only prediction at the level of approximately 1 standard deviation. The introduction of neural networks and other improvements to the techniques employed in this analysis produce a factor of 17 improvement in sensitivity over the previous search in this channel

We thank the Fermilab staff and the technical staffs of the participating institutions for their vital contributions. This work was supported by the U.S. Department of Energy and National Science Foundation; the Italian Istituto Nazionale di Fisica Nucleare; the Ministry of Education, Culture, Sports, Science, and Technology of Japan; the Natural Sciences and Engineering Research Council of Canada; the National Science Council of the Republic of China; the Swiss National Science Foundation; the A.P. Sloan Foundation; the Bundesministerium für Bildung und Forschung, Germany; the Korean World Class University Program and the National Research Foundation of Korea; the Science and Technology Facilities Council and the Royal Society, U.K.; the Russian Foundation for Basic Research; the Ministerio de Ciencia e Innovación and the Programa Consolider-Ingenio 2010, Spain; the Slovak R&D Agency; the Academy of Finland; and the Australian Research Council (ARC).


The calorimeter missing $E_T$ [$\vec{E}_T(\text{cal})$] is defined by the sum over calorimeter towers, $\vec{E}_T(\text{cal}) = -\sum_i E_i \hat{n}_i$, where $i$ is the calorimeter tower number with $|\eta| < 3.6$ and $\hat{n}_i$ is a unit vector perpendicular to the beam axis and pointing at the $i$th calorimeter tower. The reconstructed missing transverse energy, $\vec{E}_T$, is derived by subtracting from $\vec{E}_T(\text{cal})$ components of the event not registered by the calorimeter, such as muons and jet-energy adjustments. $E_T$ is the scalar magnitude of $\vec{E}_T$. 


