galaxies

MDPI

Article

Search for High-Confidence Blazar Candidates and Their MWL Counterparts in the *Fermi*-LAT Catalog Using Machine Learning

Sabrina Einecke

Institute of Physics, Technische Universität Dortmund, Dortmund, D-44221, Germany, sabrina.einecke@tu-dortmund.de; Tel.: +49-231-755-8501

Academic Editors: Jose L. Gómez, Alan P. Marscher and Svetlana G. Jorstad
Received: 15 July 2016; Accepted: 23 August 2016; Published: 26 August 2016

Abstract: A large fraction of the gamma-ray sources presented in the Third *Fermi*-LAT source catalog (3FGL) is affiliated with counterparts and source types, but 1010 sources remain unassociated and 573 sources are associated with active galaxies of uncertain type. The purpose of this study is to assign blazar classes to these unassociated and uncertain sources, and to link counterparts to the unassociated. A machine learning algorithm is used for the classification, based on properties extracted from the 3FGL, an infrared and an X-ray catalog. To estimate the reliability of the classification, performance measures are considered through validation techniques. The classification yielded purity values around 90% with efficiency values of roughly 50%. The prediction of high-confidence blazar candidates has been conducted successfully, and the possibility to link counterparts in the same procedure has been proven. These findings confirm the relevance of this novel multiwavelength approach.

Keywords: Blazars; *Fermi*-LAT; 3FGL; *Swift*-XRT; 1SXPS; WISE; ALLWISE; Machine Learning

1. Introduction

The Large Area Telescope (LAT) on board the *Fermi* satellite conducted the deepest all-sky survey in gamma-rays so far. Despite outstanding achievements in assigning source types, 1010 sources in the Third *Fermi*-LAT Source Catalog (3FGL) remain without plausible associations, and 573 sources are associated with active galaxies of uncertain type [1]. Assigning blazar classes to unassociated and uncertain sources—and linking counterparts to the unassociated ones—will improve our knowledge of the population of gamma-ray -emitting objects tremendously.

The application of machine learning algorithms has become an integral part of exploring astrophysical data. Previous machine learning strategies to assign source types were based solely on properties extracted from gamma-ray observations. Ackermann et al. [2] performed a binary classification between Active Galactic Nuclei (AGN) and pulsars to the 1FGL catalog, using logistic regression and classification trees as classification algorithms. For the same source types, Mirabal et al. [3] applied random forests to the 2FGL catalog. Hassan et al. [4] , meanwhile, used random forests and support vector machines to discriminate BL Lacs (BLLs) and Flat Spectrum Radio Quasars (FSRQs) in the 2FGL catalog, and Doert and Errando [5] classified AGN and non-AGN with a combination of random forests and neural networks.

The extension to multiwavelength information—especially the relation between properties extracted from different parts of the energy spectrum—provides additional source type-specific characteristics for better classification. This has been shown in several studies by Massaro et al. (e.g., [6] and [7]). At the same time, it offers the possibility to determine the most likely corresponding counterpart. The source localization accuracy of *Fermi* measurements (given by the 95% confidence

region) is in the order of several arcminutes. Typically, several hundred possible counterparts are located within this region, making the association ambiguous. To figure out the most likely counterpart, the associated sample is used to generate machine learning classification models. For any particular 3FGL source, all possible combinations with measurements of one additional energy range are considered. In this work, the Wide-Field Infrared Survey Explorer (WISE) source catalog and the *Swift* X-ray Point Source (1SXPS) catalog are taken into account.

2. Data Samples

The 3FGL catalog comprises 3033 gamma-ray sources in the energy regime between 100 MeV and 300 GeV, detected by *Fermi*-LAT in the first four years [1]. The most numerous sources are associated with blazars—namely 660 BLLs and 484 FSRQs —and a classification of those source types is most promising. The *Swift*-XRT Point Source Catalog (1SXPS) incorporates 151 524 X-ray sources, detected by the X-Ray Telescope (XRT) of the *Swift* satellite in 8 years of operation [8]. The telescope covers an energy range of 0.3 to 10 keV. A mid-infrared survey at the wavelengths 3.4 , 4.5 , 12 and 22 µm was conducted by the Wide-field Infrared Survey Explorer (WISE) [9]. The corresponding ALLWISE Source Catalog contains 747 634 026 sources (http://wise2.ipac.caltech.edu/docs/release/allwise/). In addition to the names and positions of all sources, the introduced catalogs provide features like information about fluxes, variabilites, and spectral properties, among others.

While the source localization accuracy (given by the 95% confidence region) of 3FGL sources is in the order of several arcminutes, the accuracies of ALLWISE and 1SXPS sources are more precise and in the order of some arcseconds (depending on the flux of the source). This implies, for instance, several hundred possible ALLWISE associations for one particular 3FGL source, and illustrates the difficulty of an association procedure and the necessity of a sophisticated method. The idea is to consider all possible combinations between a certain 3FGL source and possible counterparts (from 1SXPS or ALLWISE) located within the confidence region of the 3FGL source.

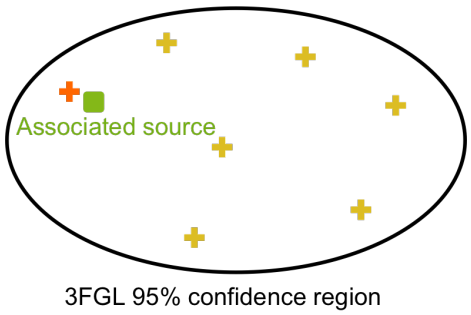

Figure 1. Sketch of the training set creation. Crosses illustrate the possible counterparts (from 1SXPS or ALLWISE).

To create a data set to train the classification procedure (see Section 3), only the associated 3FGL sources (with known position of the associated counterpart) plus counterparts (from 1SXPS or ALLWISE) within the confidence region are used. The closest counterpart (from 1SXPS or ALLWISE) to the associated counterpart of the 3FGL source (cf. Figure 1), which is associated with a BLL or an FSRQ, is assigned to the classes BLL or FSRQ, respectively. The remaining counterparts are assigned to the non-blazar class. The resulting classification task to be performed is a three-class problem.

Subsequent to the construction of the classification model with the training set, two data sets are classified. One set comprises the unassociated 3FGL sources, and the other one the AGNs of uncertain type.

3. Classification Procedure

A typical classification procedure comprises, among other things, a feature generation and selection, a choice of an appropriate classification algorithm, a tuning of the parameters thereof, and a validation of the procedure to estimate the performance. This procedure is performed within the data mining framework RapidMiner (http://rapidminer.com), offering a variety of classification algorithms. Within the feature generation, additional features to the ones extracted directly from the catalogs are generated; e.g., by combining features from different wavelengths. This extension improves upon previous studies and opens up new opportunities, allowing the determination of the most likely counterpart in the same process as the source type assignment. The number of features used in the classification algorithm has to be optimized, as the generation of the classification model with a large number of features is time-consuming, and many features do not create any added value concerning the performance. A feature selection based on the minimum-redundance and maximum-relevance algorithm [10] reduces the set to features with a maximum relevance regarding the classes and a minimum redundancy among the selected ones. The data set with 3FGL and ALLWISE features is reduced to 30 features, and the set with 3FGL and 1SXPS features to 40. The random forest is chosen as a supervised classification algorithm, as it is known to be very robust [11]. The number of decision trees in the random forest is set to 100, and the number of considered features per knot in every tree to 5. For each event (e.g., counterpart combination) entering the generated classification model, a confidence value between 0 and 1 is calculated, related to the probability that the event is assigned to a certain class (e.g., `confidence(BLL) = 1` indicates a high probability that the event is affiliated with the class BLL). For the affiliation of an event with a particular class, it is necessary to define confidence limits. If the confidence of an event is above this limit, it is affiliated with this class.

To estimate the performance of the model, the training set is split (e.g., with a ratio of 9:1). One set is used to create the model, and the other one to assign classes to the events and compare them to the true classes. This comparison is quantified by performance values for each class separately in dependence of the confidence level. Purity (the ratio of correctly and falsely classified events) and efficiency (the ratio of correctly classified events and the total number in this class) are chosen here. By iterating the split set, the performance value is calculated multiple times, and by averaging these values, an uncertainty in the form of the standard deviation is determined. This kind of validation is called cross-validation. Here, data sets are iterated 10 times.

According to the aim of a study and its criteria regarding purity and efficiency, different confidence limits are chosen. As this work is a proof of principle, only exemplary limits are selected. These limits are applied for the classification of unassociated sources and AGNs of uncertain type.

4. Results

In Figures 2–4, the confidence distributions for the classes BLL and FSRQ are displayed for the data samples extracted from the 3FGL and the ALLWISE catalog. For confidence limits greater than approximately 0.5, the true class dominates the remaining classes in the confidence distribution (cf. Figure 2). A confidence limit of 0.6 leads to a purity of (92 ± 5) % and an efficiency of (60 ± 8) % for BLLs, and (84 ± 10) % and (49 ± 8) %, respectively, for FSRQs. Applying this limit for the classification of AGNs of uncertain type, 107 BLLs and 20 FSRQs are predicted. The classification of the unassociated sources yields 23 BLLs and 6 FSRQs.

The confidence distributions for the classes BLL and FSRQ for the data samples extracted from the 3FGL and the 1SXPS catalog are depicted in Figures 5–7. Again, the true class clearly dominates the remaining classes (cf. Figure 5). A purity of (90 ± 8) % and an efficiency of (59 ± 9) % for BLLs and (79^{+21}_{-32}) % and (18 ± 13) %, respectively, for FSRQs are obtained for a confidence limit of 0.5. Using this limit for the classification of AGNs of uncertain type, 46 BLLs and 4 FSRQs are predicted. The classification of the unassociated sources yields only 11 BLLs and no FSRQs.

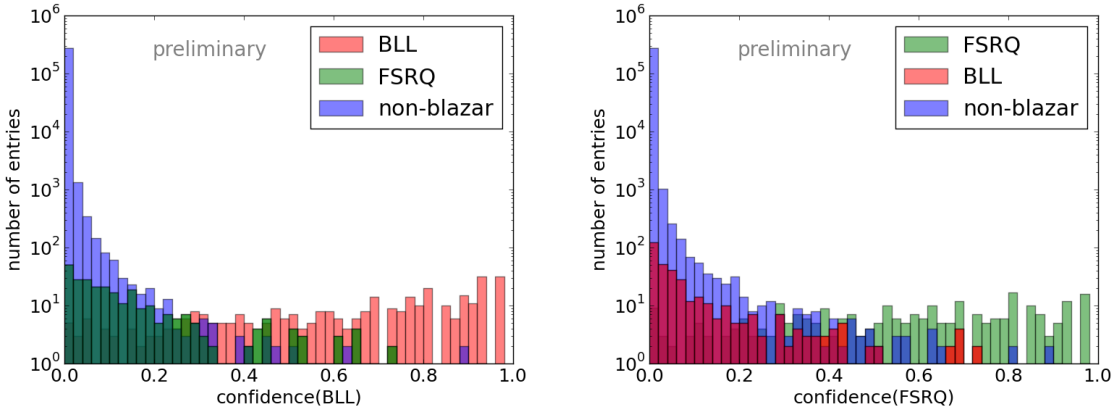

Figure 2. Confidence distribution for the BLL (*left*) and the FSRQ class (*right*). The classification model is applied to the training set extracted from the 3FGL and the ALLWISE catalog.

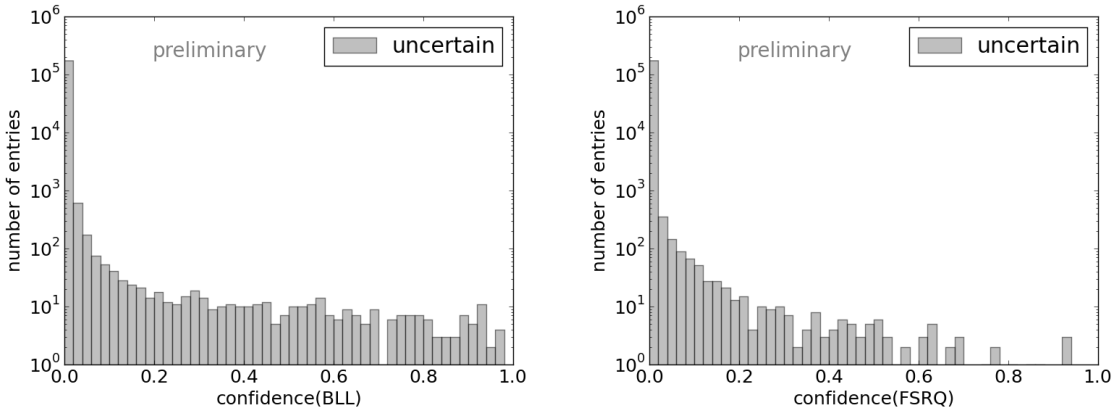

Figure 3. Confidence distribution for the BLL (*left*) and the FSRQ class (*right*). The classification model is applied to the data set of AGNs of uncertain type extracted from the 3FGL and the ALLWISE catalogs.

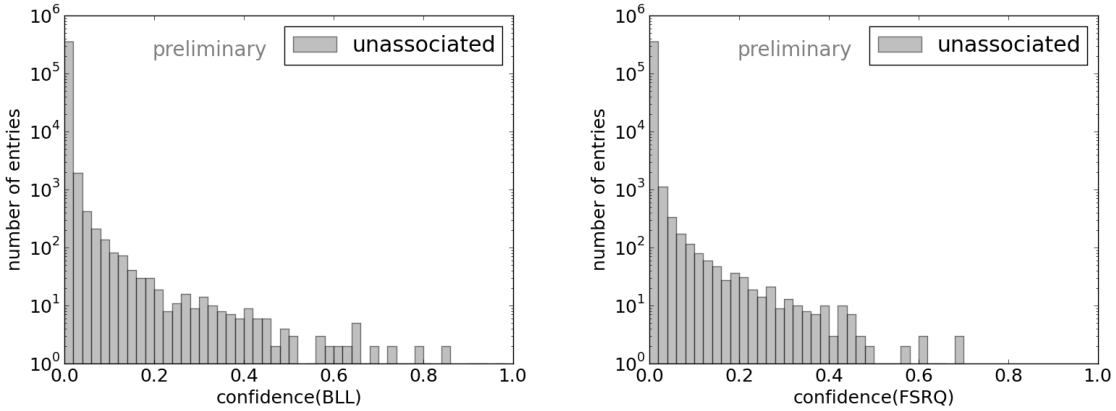

Figure 4. Confidence distribution for the BLL (*left*) and the FSRQ class (*right*). The classification model is applied to the data set of unassociated sources extracted from the 3FGL and the ALLWISE catalogs.

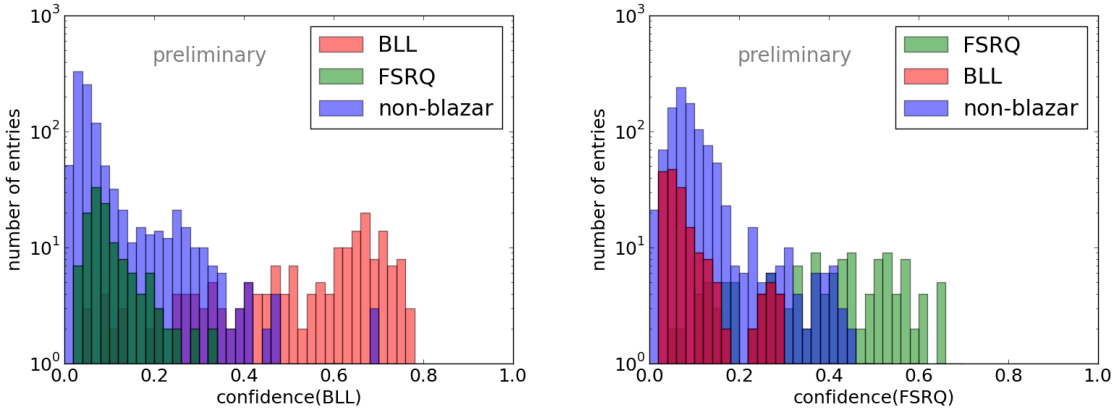

Figure 5. Confidence distribution for the BLL (*left*) and the FSRQ class (*right*). The classification model is applied to the training set extracted from the 3FGL and the 1SXPS catalogs.

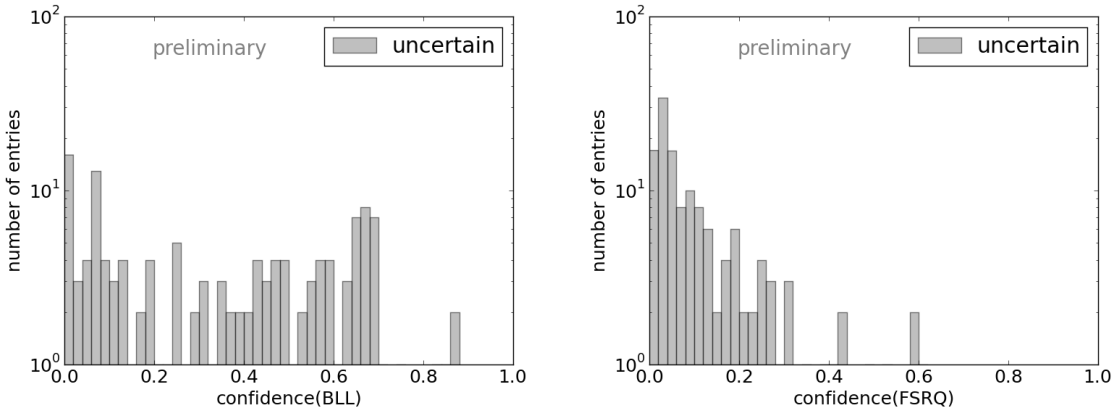

Figure 6. Confidence distribution for the BLL (*left*) and the FSRQ class (*right*). The classification model is applied to the data set of AGNs of uncertain type extracted from the 3FGL and the 1SXPS catalogs.

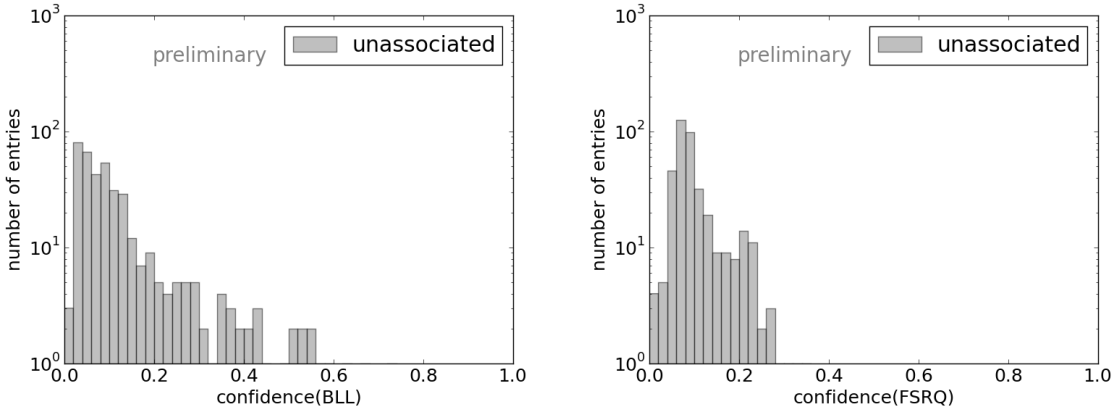

Figure 7. Confidence distribution for the BLL (*left*) and the FSRQ class (*right*). The classification model is applied to the data set of unassociated sources extracted from the 3FGL and the 1SXPS catalogs.

5. Discussion

In this work, random forest models have been used to assign BLL and FSRQ classes to unassociated and uncertain 3FGL sources and to link X-ray and infrared counterparts to the unassociated ones. The results show that the approach of combining features extracted from catalogs of different wavelength is very promising, and they confirm the relevance of multiwavelength studies. High-confidence blazar candidates have been successfully predicted for both unassociated and uncertain sources. The unassociated sources tend to be weaker than the associated ones; the training sample and therefore the model do not include many of those weak sources, leading to a smaller number of candidates for unasscociated sources than for uncertain. Probably, more blazars are still hidden, and more sensitive instruments and observation time is required to reveal them. In general, the classification performs better for BLLs than for FSRQs, probably due to the similarity between FSRQs and low-peaked BLLs, and a high separation power for intermediate- and high-peaked BLLs from FSRQs. Even though the counterpart linking performs better with infrared than X-rays, which might be related to the emission processes and the connection between the different wavelengths, typical for a specific source type, both approaches (mid-infrared and X-ray) perform extremely well.

The next logical step is to merge the most probable candidates gathered from different catalogs, exploiting the power of multiwavelength strategies, which will lead to conclusions with even higher confidence concerning blazar counterpart candidates.

Acknowledgments: Part of this work is supported by Deutsche Forschungsgemeinschaft (DFG). This work made use of data supplied by the UK Swift Science Data Centre at the University of Leicester. This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, and NEOWISE, which is a project of the Jet Propulsion Laboratory/California Institute of Technology. WISE and NEOWISE are funded by the National Aeronautics and Space Administration. This research has made use of the NASA/IPAC Infrared Science Archive, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Acero, F.; Ackermann, M.; Ajello, M.; Albert, A.; Atwood, W.B.; Axelsson, M.; Baldini, L.; Ballet1, J.; Barbiellini, G.; Bastieri, D.; et al. *Fermi* Large Area Telescope Third Source Catalog. *Astrophys. J. Suppl. Ser.* **2015**, *218*, 23.
2. Ackermann, M.; Ajello, M.; Allafort, A.; Antolini, E.; Baldini, L.; Ballet, J.; Barbiellini, G.; Bastieri, D.; Bellazzini, R.; Berenji, B.; et al. A Statistical Approach to Recognizing Source Classes for Unassociated Sources in the First *Fermi*-LAT Catalog. *Astrophys. J.* **2012**, *753*, 83.
3. Mirabal, N.; Frias-Martinez, V.; Hassan, T.; Frias-Martinez, E. *Fermi*'s `SIBYL`: Mining the Gamma-Ray Sky for Dark Matter Subhalos. *Mon. Not. R. Astron. Soc.* **2012**, *424*, L64–L68.
4. Hassan, T.; Mirabal, N.; Contreras, J.L.; Oya, I. Gamma-Ray Active Galactic Nucleus Type through Machine-learning algorithms. *Mon. Not. R. Astron. Soc.* **2013**, *428*, 220–225.
5. Doert, M.; Errando, M. Search for Gamma-Ray-Emitting Acitve Galactic Nuclei in the *Fermi*-LAT unassociated Sample using Machine Learning. *Astrophys. J.* **2014**, *782*, 41.
6. Massaro, F.; D'Abrusco, R.; Tosti, G.; Ajello, M.; Gasparrini, D.; Grindlay, J.E.; Smith, H.A. The *WISE* Gamma-Ray Strip Parametrization: The Nature of the Gamma-Ray Active Nuclei of Uncertain Type. *Astrophys. J.* **2012**, *750*, 138.
7. Paggi, A.; Massaro, F.; D'Abrusco, R.; Smith, H.A.; Masetti, N.; Giroletti, M.; Tosti, G.; Funk, S. Unveiling the Nature of the Unidentified Gamma-Ray Sources. IV. The *Swift* Catalog of Potential X-Ray Counterparts. *Astrophys. J. Suppl. Ser.* **2013**, *209*, 9.
8. Evans, P.A.; Osborne, J.P.; Beardmore, A.P.; Page, K.L.; Willingale, R.; Mountford, C.J.; Pagani, C.; Burrows, D.N.; Kennea, J.A.; Perri, M.; et al. 1SXPS: A deep *Swift* X-Ray Telescope Point Source Catalog with Light Curves and Spectra. *Astrophys. J. Suppl. Ser.* **2014**, *210*, 8.

9. Wright, E.L.; Eisenhardt, P.R.M.; Mainzer, A.; Ressler, M.E.; Cutri, R.M.; Jarrett, T.; Kirkpatrick, J.D.; Padgett, D.; McMillan, R.S.; Skrutskie, M.; et al. The Wide-Field Infrared Survey Explorer (WISE): Mission Description and Initial On-Orbit Performance. *Astron. J.* **2010**, *140*, 1868.
10. Ding, C.; Peng, H.; Long, F. Feature Selection based on Mutual Information: Criteria of Max-dependency, Max-relevance, and Min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238.
11. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.