BACKGROUND MODELING FOR DOUBLE HIGGS BOSON PRODUCTION: DENSITY RATIOS AND OPTIMAL TRANSPORT

BY TUDOR MANOLE[1,a], PATRICK BRYANT[2,b], JOHN ALISON[2,c], MIKAEL KUUSELA[3,d]
AND LARRY WASSERMAN[3,e]

[1]*Statistics and Data Science Center, Massachusetts Institute of Technology,* [a]*tmanole@mit.edu*

[2]*Department of Physics and NSF AI Planning Institute for Data-Driven Discovery in Physics, Carnegie Mellon University,* [b]*pbryant2@andrew.cmu.edu,* [c]*johnalison@cmu.edu*

[3]*Department of Statistics and Data Science and NSF AI Planning Institute for Data-Driven Discovery in Physics, Carnegie Mellon University,* [d]*mkuusela@andrew.cmu.edu,* [e]*larry@stat.cmu.edu*

We study the problem of data-driven background estimation, arising in the search of physics signals predicted by the Standard Model at the Large Hadron Collider. Our work is motivated by the search for the production of pairs of Higgs bosons decaying into four bottom quarks. A number of other physical processes, known as background, also share the same final state. The data arising in this problem is, therefore, a mixture of unlabeled background and signal events, and the primary aim of the analysis is to determine whether the proportion of unlabeled signal events is nonzero. A challenging but necessary first step is to estimate the distribution of background events. Past work in this area has determined regions of the space of collider events, where signal is unlikely to appear and where the background distribution is, therefore, identifiable. The background distribution can be estimated in these regions and extrapolated into the region of primary interest using transfer learning with a multivariate classifier. We build upon this existing approach in two ways. First, we revisit this method by developing a customized residual neural network which is tailored to the structure and symmetries of collider data. Second, we develop a new method for background estimation, based on the optimal transport problem, which relies on modeling assumptions distinct from earlier work. These two methods can serve as cross-checks for each other in particle physics analyses, due to the complementarity of their underlying assumptions. We compare their performance on simulated double Higgs boson data.

1. Introduction. The Standard Model (SM) of particle physics is a theory describing the interactions between elementary particles—the building blocks of matter. One key component of the SM is the presumed existence of a quantum field responsible for generating mass in certain elementary particles. This field is known as the Higgs field, originally theorized by Higgs (1964), Englert and Brout (1964). Excitations of the Higgs field produce particles, known as Higgs bosons, which were the subject of an intensive search by experimental particle physicists ever since the mid 1970s. In July 2012, two independent experiments at the Large Hadron Collider (LHC) at CERN (the European Organization for Nuclear Research) announced the observation of a new particle consistent with the SM Higgs boson (ATLAS (2012), CMS (2012)). Having discovered this Higgs-like particle, current work is concerned with detailed studies of its properties in order to confirm or refute those predicted by the SM. One such property is the so-called Higgs boson *self-coupling*, whereby a single excitation of the Higgs field can split into two Higgs bosons without intermediate interactions with other

particles. Observing this phenomenon would provide compelling new information regarding the mechanism of particle mass generation. This paper is concerned with some of the statistical challenges posed by its search.

The LHC is housed in a massive underground tunnel in which two counter-rotating beams of protons are accelerated to nearly the speed of light. When these protons collide, new particles are formed, and their paths within particle detectors are recorded. Individual collisions are referred to as *events*. An event in which two Higgs bosons are generated is called a *double Higgs (or di-Higgs) event*. The Higgs boson is a highly unstable particle; whenever it is produced, it decays into other particles almost immediately, making di-Higgs production impossible to observe directly.

The Higgs boson most commonly decays into a pair of so-called bottom quarks (*b*-quarks). An event in which four *b*-quarks are observed is thus a candidate di-Higgs event but could also have arisen from various other physical processes that produce four *b*-quarks. We say that a di-Higgs event in which the Higgs bosons decay into four *b*-quarks is a *signal event*, while any other event tagged as having four *b*-quarks is called a *background event*. The problem of searching for double Higgs boson production reduces to testing whether the proportion of signal events is nonzero among the observed data. As we describe in Section 3, carrying out this test is a well-understood statistical task when the distributions of both background and signal events are known. While the di-Higgs signal distribution can be approximated to sufficient accuracy with first-principles simulation, simulating the background distribution suffers from large high-order corrections which are computationally intractable (Di Micco et al. (2020)). Instead, the background distribution must be estimated using observed data. This is known as the problem of *data-driven background modeling*, which is the main subject of this paper.

As stated, the background distribution is not a statistically identifiable quantity without further assumptions, due to the potential presence of an unknown proportion of signal in the data. Any analysis strategy must, therefore, make some modeling assumptions to make the background estimation problem tractable. As we discuss below, it is standard to assume that the background distribution is related in some way to the distribution of certain *auxiliary events*, which in turn is identifiable. An example of useful auxiliary events is those consisting of less than four observed *b*-quarks, since they are unlikely to be signal events but are kinematically similar to the background events of interest (Bryant (2018)). Stated differently, the distribution of auxiliary events is an identifiable estimand which has undergone a *distributional shift* relative to the nonidentifiable background distribution of interest. If the analyst has access to a sample of auxiliary events, its empirical distribution provides a first naive approximation of the desired background distribution. To obtain a more precise estimate, one must correct for the distributional shift.

As we discuss in Section 1.2, the most widely-used method for correcting this distributional shift is based on an estimate of the *density ratio* between the background and auxiliary events. This method typically first estimates the density ratio in a signal-free region of the phase space, known as the *Control Region*, and then extrapolates it to the region of primary interest, known as the *Signal Region*. Any deviation of this extrapolated density ratio from unity is used to correct the distributional shift undergone by the auxiliary sample. This extrapolation can be viewed as an instance of transfer learning (Weiss, Khoshgoftaar and Wang (2016)). While a careful choice of the density ratio estimator can greatly improve the accuracy of this extrapolation, it clearly cannot lead to a consistent estimator if the distribution in the Signal Region is unconstrained relative to its counterpart in the Control Region. This procedure thus places an implicit modeling assumption on the underlying distributions, which is challenging to quantify and verify in practice. Nevertheless, variants of this procedure have been used in each of the most recent di-Higgs searches in the four *b*-quark final state (e.g., ATLAS (2018a), ATLAS (2019), ATLAS (2021), CMS (2022), ATLAS (2022)). This raises the important need for cross-checking the modeling assumption made by such an approach.

1.1. *Our contributions.* This paper develops a new methodology for data-driven background modeling in di-Higgs boson searches. Our approach is fully nonparametric and does not involve the extrapolation of density ratios. It hinges upon a characteristic modeling assumption, which is complementary to that of the density ratio method. These two distinct methods can thus serve as cross-checks for each other in di-Higgs searches, an important benefit that will increase the analyst's trust in the obtained background estimates.

Our approach is based on the optimal transport problem (Villani (2003)) between multidimensional distributions of collider events. Optimal transport has already proven to be a powerful tool for transfer learning in classification problems (Courty et al. (2016)), and here we propose to use it rather differently to correct distributional shifts between estimates of the auxiliary and background distributions. Our method involves out-of-sample estimation of optimal transport maps for which we consider two different estimators. While the first is based on smoothing of an in-sample optimal coupling and has previously been proposed in the literature (cf. Section 1.2), our second estimator appears to be new and leverages some strengths of the density ratio approach.

The optimal transport problem requires a cost function on the space of collider events for which we use a variant of the metric proposed by Komiske, Metodiev and Thaler (2019). This metric is itself obtained through the optimal transport problem of matching clusters of energy deposits in collision events. Our approach, therefore, involves a nested use of optimal transport.

As a secondary contribution, we revisit the density ratio approach to background estimation. In particular, we recall how this approach can be reduced to fitting a probabilistic classifier for discriminating auxiliary events from background events, and we develop a powerful new classifier tailored to this application. Our classifier is a customized convolutional neural network with residual layers (He et al. (2016)), whose architecture accounts for the structure and symmetries of collider events with multiple identical final state objects.

We illustrate the empirical performance of these two methodologies on realistic simulated collider data. We observe that both approaches lead to quantitatively similar background estimates, despite the complementarity of their underlying modeling assumptions. In particular, this study illustrates how our methods can be used to cross-check each other in practice.

1.2. *Related work.* Di-Higgs boson production has been the subject of numerous recent searches by the ATLAS and CMS collaborations at the LHC—we refer to the recent survey paper of Di Micco et al. (2020) for an overview. The four b-quark final state is the most common decay channel for di-Higgs events but suffers from a large multijet background. As described previously, each of the most recent searches in this final state performed data-driven background estimation by first estimating a density ratio in a Control Region and extrapolating it to the Signal Region. Certain searches, such as ATLAS (2019), estimate the density ratio using heuristic one-dimensional reweighting schemes, while others, such as CMS (2022), use off-the-shelf multivariate classifiers for this purpose. Part of our work builds upon the latter by designing a new classifier tailored to collider data.

The idea of estimating density ratios using classifiers has a long history in statistics—see, for instance, Fix and Hodges (1951), Silverman and Jones (1989), Qin (1998), Cheng and Chu (2004), Kpotufe (2017)—and appears in a variety of applications in experimental particle physics (e.g., Cranmer, Pavez and Louppe (2015), Brehmer et al. (2020), CMS (2022)). Classification-based estimators have the practical advantage of circumventing the need for high-dimensional density estimation, which can be particularly challenging to perform over the space of collider events. It has been empirically observed that modern classification algorithms, such as deep neural networks, have the ability to transfer well to new distributions (Yosinski et al. (2014)), which further motivates their use for density ratio estimation in our context.

Rather than density ratios, the key object of interest in our new methodology is the notion of optimal transport map. Optimal transport theory has received a surge of recent interest in the statistics and machine learning literature—we refer to Kolouri et al. (2017), Panaretos and Zemel (2019), Panaretos and Zemel (2019b), Peyré and Cuturi (2019) for recent reviews. Closest to our setting are applications of optimal transport to domain adaptation for classification problems; see, for instance, Courty et al. (2016), Redko, Habrard and Sebban (2017), Rakotomamonjy et al. (2022), and references therein. Nested optimal transport formulations, as in our work, have recently been used for other tasks such as multilevel clustering (Ho et al. (2017), Ho et al. (2019), Huynh et al. (2021)) and multimodal distribution alignment (Lee et al. (2019)). Very recently, optimal transport has also been used in high energy physics for calibrating stochastic simulators (Pollard and Windischhofer (2022)), for purposes of exploratory data analysis (Cai et al. (2020), Komiske et al. (2020), Komiske, Metodiev and Thaler (2019)), and for the purpose of defining a geometry on the space of collider events (Komiske, Metodiev and Thaler (2020)). We also note that optimal transport has implicitly been used for one-dimensional template morphing in the early work of Read (1999).

Our methodology relies on estimating optimal transport maps or couplings between distributions of collider events. The question of out-of-sample estimation of optimal transport maps over Euclidean spaces has been the subject of intensive recent study (Hütter and Rigollet (2021), Perrot et al. (2016), Pooladian and Niles-Weed (2021), Forrow et al. (2019), Makkuva et al. (2020), Nath and Jawanpuria (2020), de Lara, González-Sanz and Loubes (2021), Deb, Ghosal and Sen (2021), Manole et al. (2024), Ghosal and Sen (2022), Gunsilius (2022)). While many of these works are tailored to the quadratic cost function, the widely-used nearest-neighbor estimator (Flamary et al. (2021), Manole et al. (2024)) can naturally be defined over general metric spaces and is used in one of our background estimators defined in Section 5.2.2.

Beyond the search of di-Higgs boson production, we emphasize that the question of data-driven background estimation arises in a variety of problems in experimental high-energy physics, where our methodologies could also potentially be applied. We refer to the book Behnke et al. (2013) for a pedagogical introduction to statistical aspects of the subject; see also Appendix 1 of Lyons (1986). Finally, we mention some recent advances on the widely-used sPlot (Barlow (1987), Borisyak and Kazeev (2019), Dembinski et al. (2022), Pivk and Le Diberder (2005)) and ABCD (Alison (2015), ATLAS (2015), Choi and Oh (2021), Kasieczka et al. (2021)) techniques for background estimation, the latter of which can be viewed as a precursor to the methods developed in this paper.

1.3. *Paper outline.* The rest of this paper is organized as follows. Section 2 contains background about the LHC and di-Higgs boson production. Section 3 outlines the statistical procedure used for signal searches in collider experiments at the LHC and mathematically formulates the data-driven background modeling problem. In Section 4 we revisit the density ratio approach to background estimation, based on classifiers for discriminating auxiliary events from background events, and we briefly describe our new classifier architecture for this purpose. In Section 5 we describe our new methodology based on the optimal transport problem. We then compare these methods in a simulated di-Higgs search in Section 6. We close with a discussion in Section 7. In the Supplementary Material (Manole et al. (2024)), Appendix A contains a section-by-section summary of this manuscript in nontechnical language, Appendices B–D contain numerical details deferred from the main text, and Appendix E contains further numerical results.

1.4. *Introduction for the high energy physicist.* This paper is written primarily with the statistics community in mind. This brief section aims to bridge the gap between the language and formalism used by statisticians and that common in high energy physics.

We have a four-tag (4*b*) dataset consisting of background and some a priori unknown amount of signal, parameterized by signal strength μ. The search is done in bins of a discriminating variable, the output of a multivariate classifier trained to separate signal and background. The challenge is to predict the amount of background in each of the classifier output bins. Section 3 introduces the three-tag (3*b*) dataset and defines the Control Region (CR) used to derive the background prediction. The 3*b* dataset, when normalized to the number of 4*b* events, provides a zeroth order estimate of the 4*b* background. The main contribution of our work is in deriving corrections to the 3*b* data to better approximate the true 4*b* background. Section 4 describes a data-driven background estimation method used frequently in HEP (ATLAS (2018a), CMS (2022)), which is a variant of the "ABCD" method. We train a classifier, referred to as the "Four-vs.-Three classifier," to obtain event weights that correct for differences between the 3*b* and 4*b* data in the CR. The predicted background in the Signal Region (SR) is obtained by weighting 3*b* SR events by weights derived in the CR. Section 5 presents a novel method for data-driven background estimation. Instead of extrapolating between the 3*b* and 4*b* samples—assuming the extrapolation is the same in the CR and SR—we propose extrapolating between the CR and SR—assuming the extrapolation is the same in the 3*b* and 4*b* samples. We cannot use a classifier to correct kinematic differences between the samples; a classifier trained on kinematically disjoint samples would achieve perfect separation, and the corresponding weights would be undefined. Instead, we assume that the optimal transport map, which maps events in the CR to the SR, is the same for the 3*b* and 4*b* events, and we describe approaches for estimating such optimal transport maps with collider data.

Readers less interested in the formalism may choose to skip ahead to Section 6, which applies the methods introduced in this work in a simulated di-Higgs search. A more complete description of our work aimed at the high energy physicist can be found in the Supplementary Material, Appendix A.

2. Background.

2.1. *LHC experiments and di-Higgs boson production.* The LHC is the largest particle collider in the world, consisting of a 27 kilometer-long tunnel in which two counter-rotating beams of protons are accelerated to nearly the speed of light. These particles are primarily collided in one of four underground detectors, named ALICE, ATLAS, CMS, and LHCb. ATLAS and CMS are general-purpose detectors used for a wide range of physics analyses, including Higgs boson-related searches, while ALICE and LHCb focus on specific physics phenomena. We focus on the CMS detector in what follows, but similar descriptions can be made for the ATLAS detector.

When two protons collide, their energy is converted into matter in the form of new particles. The goal of the CMS (Compact Muon Solenoid) detector is to measure the momenta, energies and types of such particles. To measure their momenta, CMS is built around a giant superconducting solenoid magnet, depicted in Figure 1, which deforms the trajectories of particles as they move from the center of the detector outward through a silicon tracker. The extent to which the trajectory of a charged particle is bent depends on its momentum and can hence be used to measure the momentum. After the silicon tracker, CMS consists of several layers of calorimeters which measure the energies of the particles. We refer to CMS (2008) for a complete description of the CMS detector.

Proton-proton collisions give rise to highly unstable particles which decay almost instantly into more stable particles. The detector is only able to observe these longer-lived particles. By measuring their energies and momenta, insight can be gained into the physical properties of the unstable particles from which they originate.
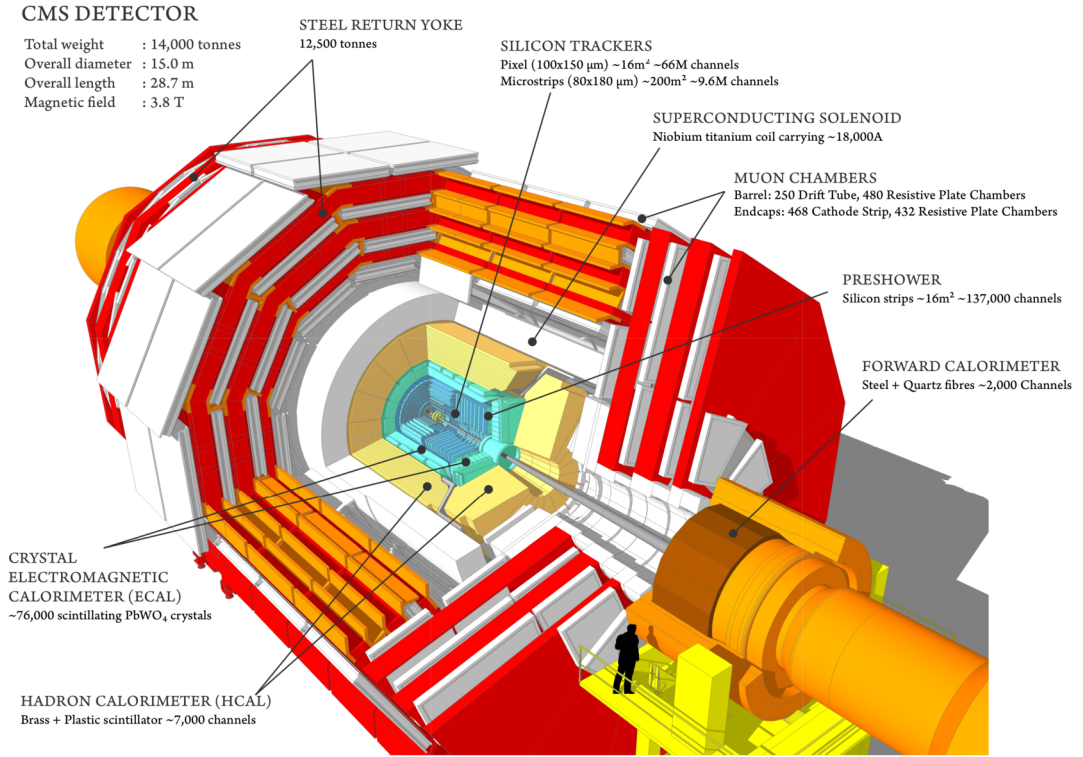
FIG. 1. *Illustration of the CMS detector (*Sakuma and McCauley *(2014)). Counter-rotating beams of protons are made to collide in the center of the detector. The trajectory and mass of each particle emanating from the collision is then recorded.*

The Higgs boson is an example of an unstable particle, which decays within approximately 10^{-22} seconds. The SM predicts that a Higgs boson decays into a pair of bottom quarks (b-quarks) 60% of the time, and this decay channel has indeed been observed experimentally (ATLAS (2018b), CMS (2018a)). Other channels which have been observed experimentally include the decay of a Higgs boson into pairs of photons (ATLAS (2018c), CMS (2018b)), W bosons (ATLAS (2018e), CMS (2019)), Z bosons (ATLAS (2018d), CMS (2018c)), and tau leptons (ATLAS (2019b), CMS (2018d)). The SM further predicts the rare possibility that two Higgs bosons can be produced simultaneously, and this paper is concerned with the statistical challenges arising in the search for this process, which has yet to be observed experimentally. If this process were to occur, the two resulting Higgs bosons would each, in turn, be most likely to decay into two b-quarks, thus making four b-quarks the most common final state of di-Higgs boson events. We focus on this decay channel (abbreviated HH$\to$ 4b) throughout this paper. We note that b-quarks form into bound states with other quarks called b-hadrons, which are themselves unstable and rapidly decay into collimated sprays of stable particles called b-jets, which can be efficiently identified by the CMS detector (CMS (2018e)).

2.2. *Collider events and the CMS coordinate system.* Particles measured by the CMS detector are typically represented in spherical coordinates. Given a particle with momentum vector $p = (x, y, z) \in \mathbb{R}^3$, its azimuthal angle $\phi \in [0, 2\pi)$ is defined as the angle increasing from the positive x-axis to the positive y-axis, while the polar angle $\theta \in [0, \pi)$ is increasing from the positive z-axis to the positive y-axis. The length of its projection onto the (x, y) plane is called the *transverse momentum p_T*. It is common to replace the polar angle θ by the *pseudorapidity* of the particle, given by $\eta = -\log(\tan(\theta/2))$.

In addition to the variables p_T, η and ϕ, the rest mass m of each particle can be obtained from the energy measurements made by the calorimeters in the CMS detector. Altogether, a particle jet is analyzed as a single point in this coordinate system and encoded as a four-dimensional vector (p_T, η, ϕ, m). In our search channel, collisions lead to multiple, say $K \geq 1$, jets measured by the detector, which may be encoded as the $4K$-dimensional vector $(p_{Ti}, \eta_i, \phi_i, m_i : 1 \leq i \leq K)$. We opt for an alternative notation, which will be particularly fruitful for the purpose of defining a metric between collider events in Section 5.3. Specifically, an event will henceforth be represented by the discrete measure

$$(2.1) \qquad g = \sum_{i=1}^{K} p_{Ti} \delta_{(\eta_i, \phi_i, m_i)},$$

where δ_x denotes the Dirac measure placing unit mass at a point $x \in \mathbb{R}^3$. In particular, the representation (2.1) emphasizes the invariance of an event with respect to the ordering of its jets. The transverse momenta p_{Ti} may be viewed as a proxy for the energy of each jet; thus, the total measure of g denotes its total energy, denoted $s_T = \sum_{i=1}^{K} p_{Ti}$. The set of events with K jets of interest is denoted by

$$\mathcal{G}^{(K)} = \left\{ \sum_{j=1}^{K} p_{Tj} \delta_{(\eta_j, \phi_j, m_j)} : p_{Tj}, m_j > 0, \phi_j, \eta_j \in \mathbb{R}, 1 \leq j \leq K \right\},$$

where the definition of ϕ_j is extended from $[0, 2\pi)$ to the entire real line by 2π-periodicity. In the context of double Higgs boson production in the four b-jet final state, the choice $K = 4$ will be most frequently used, and in this case we simply write $\mathcal{G} = \mathcal{G}^{(4)}$.

Finally, we note that events are deemed invariant under the orientation of the x- and z-axes. This fact, together with the periodicity of the angle ϕ, implies that two events $g = \sum_{j=1}^{K} p_{Tj} \delta_{(\eta_j, \phi_j, m_j)} \in \mathcal{G}^{(K)}$ and $g' = \sum_{j=1}^{K} p'_{Tj} \delta_{(\eta'_j, \phi'_j, m'_j)} \in \mathcal{G}^{(K)}$ may be deemed equivalent if they are mirror-symmetric in η, ϕ as well as rotationally symmetric in ϕ, that is, if there exist $\Delta \in 2\pi\mathbb{Z}$ and $\iota_1, \iota_2 \in \{-1, 1\}$ such that

$$(2.2) \qquad \sum_{j=1}^{K} p_{Tj} \delta_{(\iota_1 \eta_j, \Delta + \iota_2 \phi_j, m_j)} = \sum_{j=1}^{K} p'_{Tj} \delta_{(\eta'_j, \phi'_j, m'_j)}.$$

Formally, we define an equivalence relation $\simeq$ between events in $\mathcal{G}^{(K)}$ such that $g \simeq g'$ if and only if there exist Δ, ι_1, ι_2 for which (2.2) holds.

3. Problem formulation.

3.1. *Overview of signal searches at the LHC.*
In order to make inferences about the presence or absence of a signal process in collider data, event counts are commonly analyzed as binned Poisson point processes. While we focus on the setting of double Higgs boson production in the four b-quark final state, the description that follows is representative of a wide range of signal searches for high-energy physics experiments.

Let ν_0 denote a σ-finite Borel measure over the state space $\mathcal{G}$ of collider events, with respect to a fixed choice of Borel σ-algebra on $\mathcal{G}$ denoted $\mathbb{B}(\mathcal{G})$. Let F denote an inhomogeneous Poisson point process (Reiss (2012)) with a nonnegative intensity function $f \in L^2(\mathcal{G})$ on $\mathcal{G}$, that is, F is a random point measure on $\mathcal{G}$ such that:

1. $F(A) \sim \text{Poisson}(\lambda(A))$, where λ is the intensity measure induced by f, defined by $\lambda(A) = \int_A f \, d\nu_0$ for all $A \in \mathbb{B}(\mathcal{G})$;
2. $F(A_1), \ldots, F(A_\ell)$ are independent for all pairwise disjoint sets $A_1, \ldots, A_\ell \in \mathbb{B}(\mathcal{G})$, for all integers $\ell \geq 1$.

Every four b-jet collision event is either a *signal event*, namely, an event arising from two Higgs bosons, or a *background event*, arising from some other physical process. Letting $\mu \geq 0$ denote the rate of signal events, we write the intensity measure λ as

$$\lambda(\cdot) = \beta_4(\cdot) + \mu\sigma(\cdot),$$

where β_4 and σ, respectively, denote nonnegative background and signal intensity measures. σ is typically normalized such that the value $\mu = 1$ corresponds to the theoretical prediction of the signal rate. The measures β_4 and σ typically depend on nuisance parameters related to the calibration of the detector, the uncertain parameters of certain physical processes, such as the parton distribution functions of the proton (Placakyte (2011)), and so on. We suppress the dependence on such nuisance parameters for ease of exposition. The parameter μ is of primary interest, since nonzero values of μ indicate the existence of signal events. A search for the signal process, therefore, reduces to testing the following hypotheses on the basis of observations from the Poisson point process F:

(3.1) $$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu > 0.$$

Given a sequence $G_1, G_2, \ldots$ of observed events, we may write $F = \sum_{i=1}^{M} \delta_{G_i}$, where $M \sim \text{Poisson}(\lambda(\mathcal{G}))$ is independent of the observations, which satisfy

(3.2) $$G_1, G_2, \ldots \overset{\text{iid}}{\sim} \lambda/\lambda(\mathcal{G}) = \epsilon S + (1 - \epsilon)P_4.$$

Here $S = \sigma/\sigma(\mathcal{G})$ and $P_4 = \beta_4/\beta_4(\mathcal{G})$ denote the respective signal and background distributions and $\epsilon = \mu\sigma(\mathcal{G})/\lambda(\mathcal{G})$ the proportion of signal events.

The Poisson point process F is often binned in practice. Let $\xi : \mathcal{G} \to \mathcal{A} \subseteq \mathbb{R}$ denote a dimensionality reduction map, to be discussed below, which will be used to bin the point process using univariate bins. Let $\{I_j\}_{j=1}^{J}$ denote a collection of bins forming a partition of $\mathcal{A}$, and define the event counts

(3.3) $$D_j = F(\xi^{-1}(I_j)) = |\{1 \leq i \leq M : \xi(G_i) \in I_j\}|, \quad j = 1, \ldots, J.$$

The definition of F implies that the random variables D_j are independent and satisfy

(3.4) $$D_j \sim \text{Poisson}(B_j + \mu S_j), \quad j = 1, \ldots, J,$$

where $B_j = \beta_4(\xi^{-1}(I_j))$ and $S_j = \sigma(\xi^{-1}(I_j))$.

The likelihood ratio test with respect to the joint distribution of $D_1, \ldots, D_J$ is typically used to test the hypotheses (3.1) (ATLAS, CMS and Higgs Combination Group (2011)). The binned likelihood function for the parameter μ is given by

(3.5) $$L(\mu) = \prod_{j=1}^{J} \frac{(B_j + \mu S_j)^{D_j}}{D_j!} e^{-(B_j + \mu S_j)}.$$

Di-Higgs events are rare in comparison to background events; thus, the signal-to-background ratio is low. At the time of writing, values of M, which are typically observed at the LHC, may be too small for any test to have power in rejecting the null hypothesis in (3.1) at desired significance levels (Di Micco et al. (2020)). Analyses which fail to reject H_0 instead culminate in an upper confidence bound on μ, also known as an upper limit (ATLAS, CMS and Higgs Combination Group (2011)).

The power of the likelihood ratio test for (3.1) may be increased by choosing a function ξ which maximizes the separation between background and signal event counts across the J bins. Informally, the optimal such choice of ξ is given by

(3.6) $$\xi(g) = \mathbb{P}(G \text{ is a Signal Event} | G = g),$$

which may be estimated using a multivariate classifier, such as a neural network or boosted decision trees, for discriminating background events from signal events.

The signal intensity measure σ is theoretically predicted by the SM and can be approximated well using Monte Carlo event generators (Di Micco et al. (2020)). The background intensity β_4 is, however, intractable due to the strongly interacting nature of quantum chromodynamics (QCD) in which events with the four b-quark final state can be produced via an enormous number of relevant and complex pathways. The intensity measure β_4, or its binned analogue $(B_j)_{j=1}^J$, must, therefore, be estimated from the collider data itself, which we refer to as *data-driven background modeling*. This problem is the primary focus of this paper.

3.2. *Setup for data-driven background modeling.* The aim of this paper is to develop data-driven estimators of the background intensity measure β_4. The primary challenge is the fact that the sample $G_1, \ldots, G_M$ is contaminated with an unknown proportion ϵ of signal events. The background estimation problem is thus statistically unidentifiable as stated, and it will be necessary to impose further modeling assumptions.

In order to formulate these assumptions and our resulting background modeling methods, we assume that the analyst has access to a second Poisson Point Process $T = \sum_{i=1}^N \delta_{H_i}$ consisting of auxiliary events, which were tagged by the CMS detector as having four jets, of which exactly three are b-jets. We refer to such events as "3b events," as opposed to "4b events," which were identified as having four b-jets.[1] We stress that the terms 3b and 4b do not refer to the true number of b-quarks arising from the collision, rather the number of b-jets identified by the detector. As we discuss in Section 6, the majority of 3b events in fact arise from the hadronization of two b-quarks and two charm or light quarks, while a small proportion arise from four b-quarks.[2] For the purpose of a discovery analysis, the 3b sample $H_1, \ldots, H_N$ can, therefore, be treated as having a negligible proportion of signal events (Bryant (2018), CMS (2022)). We treat this proportion as zero for sake of exposition. We henceforth denote the intensity measure of the point process T by β_3, and we denote by $P_3 = \beta_3/\beta_3(\mathcal{G})$ the corresponding probability distribution of the observations $H_1, H_2, \ldots$.

The kinematics of 3b events are similar, but not equal, to those of 4b background events (CMS (2022)). Unlike β_4, however, the intensity measure β_3 is an identifiable estimand due to the lack of signal events in the point process T. Any consistent estimator $\widehat{\beta}_3$ of β_3 can be used to provide a zeroth-order approximation of β_4 (up to a correction for normalization). This approximation is, however, insufficiently accurate to be used as a final estimate of β_4, and our goal is to develop statistical methods for correcting this naive background estimate.

Recall that the four b-jets of any signal event $g \in \mathcal{G}$ are naturally paired, with each pair arising from a Higgs boson. The true pairing of the jets is unknown to the detector; however, it may be approximated, for instance, using an algorithm due to Bryant (2018). We use the same pairing algorithm in our work. Given as input an event g, this deterministic algorithm outputs one among the three distinct unordered pairs of measures $\{g^1, g^2\} \subseteq \mathcal{G}^{(2)}$ which satisfy $g = g^1 + g^2$. We refer to g^1 and g^2 as *dijets*. When g is a signal event, we expect that each dijet arose from a decay of a Higgs boson, whereas when g is a background event, we expect that at least one of the two dijets arose from the decay of a different particle.

The Higgs boson is known to have mass m_H approximately equal to 125 GeV (ATLAS (2012), CMS (2012)). It follows that the two dijets should approximately satisfy $m(g^1) \approx$

---

[1]3b events were used for background estimation in the HH→ 4b channel in the recent analysis of CMS (2022). "2b events" consisting of two, rather then three, b-tagged jets have been used in other recent analyses (e.g., ATLAS (2019), ATLAS (2021)), and our description also applies to such events with only formal changes.

[2]As a result, the expected rate of production of 3b events $\mathbb{E}[N]$ is typically higher than that of 4b events $\mathbb{E}[M]$ by an order of magnitude; cf. Section 6.
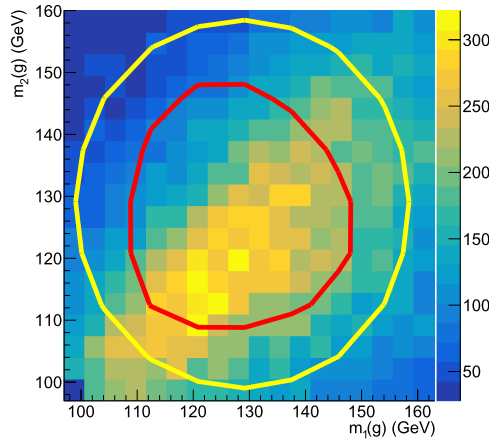
FIG. 2.   *Illustration of the Control and Signal Regions. The two-dimensional histogram represents simulated 4b collider events described in Section 6, plotted in terms of their dijet invariant masses. We emphasize that this is a low-dimensional representation; the events considered in this work are 16-dimensional. The red line indicates the boundary of the Signal Region, while the annulus bounded by the yellow and red lines represents the Control Region. The constants σ_c, r_c, and κ_s used in this figure are stated in Section 6.*

$m(g^2) \approx m_H$, where $m(a)$ denotes the invariant mass[3] of any $a \in \mathcal{G}^{(K)}$, $K \geq 1$. Large deviations of the dijet invariant masses from 125 GeV indicate that g is not a signal event. This provides a heuristic for determining events among $G_1, \ldots, G_M$ which are unlikely to be signal events. To elaborate, we form subsets $\mathcal{G}_c, \mathcal{G}_s \subseteq \mathcal{G}$ such that $\mathcal{G}_c \cap \mathcal{G}_s = \varnothing$, where $\mathcal{G}_s$ is called the *Signal Region*, containing events with dijet masses near m_H, and $\mathcal{G}_c$ is called the *Control Region*, containing all other events which will be used in the analysis. We follow Bryant (2018) and employ the following specific definitions of $\mathcal{G}_c$ and $\mathcal{G}_s$:

$$(3.7) \qquad \mathcal{G}_s = \left\{ g \in \mathcal{G} : \sqrt{\left(1 - \frac{m_H}{m(g^1)}\right)^2 + \left(1 - \frac{m_H}{m(g^2)}\right)^2} \leq \kappa_s \right\},$$

$$(3.8) \qquad \mathcal{G}_c = \left\{ g \in \mathcal{G} : \sqrt{(m(g^1) - \sigma_c m_H)^2 + (m(g^2) - \sigma_C m_H)^2} \leq r_c \right\} \setminus \mathcal{G}_s,$$

for some constants $\sigma_c, r_c, \kappa_s > 0$. These regions are illustrated in Figure 2. We similarly partition the Poisson intensity measures β_3, β_4 by defining, for all $A \in \mathbb{B}(\mathcal{G})$,

$$\beta_j^c(A) = \beta_j(A \cap \mathcal{G}_c), \qquad \beta_j^s(A) = \beta_j(A \cap \mathcal{G}_s), \quad j = 3, 4.$$

These four measures are illustrated in Figure 3. Furthermore, we assume for ease of exposition that these measures are absolutely continuous with respect to the dominating measure ν_0, and we let $b_j^a = d\beta_j^a/d\nu_0$ for all $j = 3, 4$ and $a = c, s$.

Recall that the collider events associated with the intensity measures β_3^c and β_3^s are signal-free by construction, and those from β_4^c are also signal-free under the assumption that negligibly few signal events will fall outside of $\mathcal{G}_s$. These three intensity measures can, therefore, be estimated directly by means of their empirical intensity functions. We have thus reduced the background modeling problem to that of estimating β_4^s, given estimates of β_3^c, β_3^s, and β_4^c.

To this end, we will partition the samples into the sets

$$\{G_1^s, \ldots, G_{m_s}^s\} := \{G_1, \ldots, G_M\} \cap \mathcal{G}_s, \qquad \{H_1^s, \ldots, H_{n_s}^s\} := \{H_1, \ldots, H_N\} \cap \mathcal{G}_s,$$

$$\{G_1^c, \ldots, G_{m_c}^c\} := \{G_1, \ldots, G_M\} \cap \mathcal{G}_c, \qquad \{H_1^c, \ldots, H_{n_c}^c\} := \{H_1, \ldots, H_N\} \cap \mathcal{G}_c,$$

---

[3]If E denotes the sum of the energies of the constituent jets of a and p denotes the magnitude of the sum of their momentum vectors, then the invariant mass of e is defined by $m(a) = \sqrt{E^2 - p^2}$ (Hagedorn (1963)).
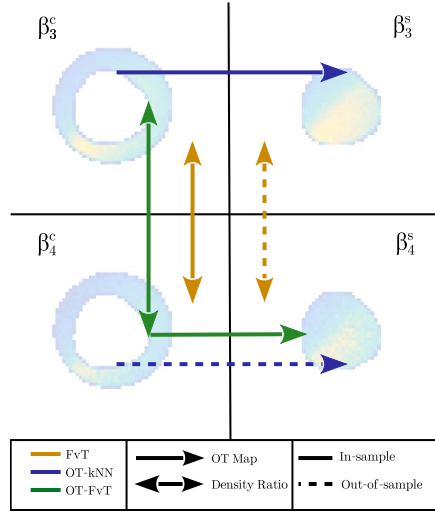
FIG. 3. (*Color online.*) *Illustration of the four Poisson intensity measures* $\beta_3^c, \beta_3^s, \beta_4^c, \beta_4^s$, *among which only the latter is nontrivial to estimate, and summary of the three methods developed in this paper for estimating* β_4^s. *The method FvT (Four vs. Three) estimates the ratio of the two densities in the Control Region using a classifier and then extrapolates it into the Signal Region using out-of-sample evaluations of the classifier. The OT-kNN (Optimal Transport–k Nearest Neighbors) method produces an estimator* $\widehat{T}$ *of the optimal transport map* T *between the 3b Control and Signal Region distributions, and evaluates this estimator out-of-sample on an estimator of the 4b Control Region distribution. The out-of-sample evaluation of* $\widehat{T}$ *is performed using nearest-neighbor extrapolation. The OT-FvT (Optimal Transport–Four vs. Three) method combines these ideas: first, it uses the classifier to produce an estimator of* β_4^c *with the same support as* β_{3,n_c}^c, *and second, it pushes forward this estimator through* $\widehat{T}$, *thereby avoiding out-of-sample evaluations of both the classifier and optimal transport map. The background of the figure consists of bivariate histograms of simulated 3b and 4b samples in the Control and Signal Regions, plotted in terms of their dijet invariant masses, as in Figure* 2.

where $M = m_c + m_s$ and $N = n_c + n_s$. Furthermore, let

$$\beta_{3,n_c}^c = T|_{\mathcal{G}_c} = \sum_{i=1}^{n_c} \delta_{H_i^c}, \qquad \beta_{3,n_s}^s = T|_{\mathcal{G}_s} = \sum_{i=1}^{n_s} \delta_{H_i^s}, \qquad \beta_{4,m_c}^c = F|_{\mathcal{G}_c} = \sum_{i=1}^{m_c} \delta_{G_i^c}$$

denote the empirical estimators of the measures $\beta_3^c, \beta_3^s, \beta_4^c$, illustrated in the background of Figure 3. As previously noted, the measure β_3^s provides a zeroth-order approximation of β_4^s (after a normalization correction); thus, a naive first estimate of β_4^s is given by β_{3,n_s}^s. As we shall see in the simulation study of Section 6, this approximation is insufficiently accurate to be used as a final estimator. Our methodologies improve upon it by modeling the discrepancy between the 3b and 4b distributions in the Control Region via $\beta_{4,m_c}^c, \beta_{3,n_c}^c$ and then using that information in the Signal Region to improve the accuracy of β_{3,n_s}^s as an estimator of β_4^s.

Once we are able to derive an estimator $\widehat{\beta}_4^s$ of β_4^s, based on the signal-free observations $G_1^c, \dots, G_{m_c}^c, H_1^s, \dots, H_{n_s}^s, H_1^c, \dots, H_{n_c}^c$, we may define the fitted histogram $\widehat{B}_j = \widehat{\beta}_4^s(\xi^{-1}(I_j))$, $j = 1, \dots, J$. One may then test the hypotheses (3.1) using the likelihood ratio test, based on the following modification of the likelihood function in equation (3.5):

$$(3.9) \quad \widetilde{L}(\mu) = \prod_{J=1}^{J} \frac{(\widehat{B}_j + \mu S_j)^{D_j^s}}{D_j^s!} e^{-(\widehat{B}_j + \mu S_j)}, \quad \text{where } D_j^s = |\{1 \leq i \leq m_s : \xi(G_i^s) \in I_j\}|.$$

Here $\widetilde{L}$ can be viewed as a restriction of the likelihood L to the Signal Region. Notice that $\widehat{B}_j$ is independent of D_k^s, for any j, k. In practice, it is also necessary to incorporate statistical and systematic uncertainties pertaining to the estimator $\widehat{B}_j$ into the hypothesis testing

procedure (ATLAS, CMS and Higgs Combination Group (2011)). Since formal uncertainty quantification for background modeling is beyond the scope of this work, we omit further details and provide further discussion of this point in Section 7.

The primary difficulty remaining in the testing problem (3.1) is that of deriving estimators of the background intensity measure β_4^s. In what follows, we describe two classes of estimators for β_4^s: one based on density ratio estimation (Section 4), and the second based on optimal transport (Section 5). The former is the most common approach to di-Higgs background modeling and will later be referred to as the FvT method. The latter is new, and we will discuss two different instances of this approach, which will later be referred to as the OT-kNN and OT-FvT methods. These three distinct estimators are summarized in Figure 3.

4. Background modeling via density ratio extrapolation. The discrepancy between $3b$ and $4b$ background distributions may be directly quantified in the Control Region, where no signal events are present. Under a suitable modeling assumption, this discrepancy may be extrapolated into the Signal Region to produce a correction of the $3b$ signal region intensity measure β_3^s, leading to an estimate of β_4^s. This general strategy forms the basis of most background modeling methodologies used in recent di-Higgs searches, as discussed in Section 1. The aim of this section is to recall how this approach may be carried out using a classifier for discriminating $3b$ and $4b$ events. We then propose a classifier specifically tailored to this type of collider data, which will be used in our numerical studies.

Let E denote a random collider event, arising from either the $3b$ or $4b$ distributions, and define the latent binary random variable Z indicating the component membership of E. More specifically, let Z be a Bernoulli random variable with success probability $\mathbb{P}(Z = 1) = \beta_4(\mathcal{G})/(\beta_4(\mathcal{G}) + \beta_3(\mathcal{G}))$, and let E be generated according to the mixture model

$$E|Z = 0 \sim P_3, \qquad E|Z = 1 \sim P_4.$$

Setting $\psi(g) = \mathbb{P}(Z = 1|E = g)$ for all $g \in \mathcal{G}$, it follows from Bayes' Rule that

$$(4.1) \qquad \frac{b_4^c(g)}{b_3^c(g)} = \frac{\psi(g)}{1 - \psi(g)}, \quad g \in \mathcal{G}_c,$$

where we recall that b_j^c denotes the intensity function associated to β_j^c, $j = 3, 4$. Therefore,

$$(4.2) \qquad \beta_4^c(A) = \int_A \frac{\psi(g)}{1 - \psi(g)} \, \mathrm{d}\beta_3^c(g), \quad A \in \mathbb{B}(\mathcal{G}_c).$$

Equations (4.1)–(4.2) are a reformulation for our context of the well-known fact that, up to normalization, a likelihood ratio may be expressed as an odds ratio (Silverman and Jones (1989)). Estimating the ratio of $3b$ to $4b$ intensity functions in the Control Region thus reduces to the classification problem of estimating ψ, say by a classifier $\widehat{\psi}$. This observation has the practical advantage of circumventing the need of performing high-dimensional density estimation. Assuming that the estimator $\widehat{\psi}$ can be evaluated in the Signal Region $\mathcal{G}_s$, disjoint from its training region $\mathcal{G}_c$, we may postulate that the measure

$$(4.3) \qquad A \in \mathbb{B}(\mathcal{G}_s) \longmapsto \int_A \frac{\widehat{\psi}(g)}{1 - \widehat{\psi}(g)} \, \mathrm{d}\beta_3^s(g)$$

provides a reasonable approximation of β_4^s. The quality of such an approximation is driven by the ability of the classifier $\widehat{\psi}$ to generalize between regions of the phase space. To formalize this, we will assume for simplicity that $\widehat{\psi}$ is an empirical risk minimizer taking values in a class $\{\psi_\alpha : \mathcal{G} \to [0, 1] : \alpha \in \Omega\}$, for some parameter space $\Omega \subseteq \mathbb{R}^d$, $d \geq 1$. That is, we assume

$$\widehat{\psi} = \psi_{\widehat{\alpha}}, \quad \text{where } \widehat{\alpha} = \underset{\alpha \in \Omega}{\operatorname{argmin}} \left\{ \frac{1}{n_c} \sum_{i=1}^{n_c} \mathcal{L}(\psi_\alpha(H_i^c), 0) + \frac{1}{m_c} \sum_{j=1}^{m_c} \mathcal{L}(\psi_\alpha(G_j^c), 1) \right\},$$

for some loss function $\mathcal{L} : [0, 1] \times [0, 1] \to \mathbb{R}$. We then make the following assumption.

ASSUMPTION 1. The conditional probability ψ satisfies the following conditions:

(i) (Correct Specification) There exists $\alpha^* \in \Omega$ such that $\psi = \psi_{\alpha^*}$.
(ii) (Generalization) We have

$$\alpha^* = \underset{\alpha \in \Omega}{\operatorname{argmin}} \, \mathbb{E}\big[\mathcal{L}(\psi_\alpha(G), Z)|G \in \mathcal{G}_c\big].$$

Assumption 1 implies that a classifier trained solely in the Control Region can consistently estimate the full conditional probability $\psi(g)$, for events $g \in \mathcal{G}$ lying in both the Control and Signal Regions. Such an assumption guarantees the ability of the classifier $\widehat{\psi}$ to generalize from the Control Region, making the ansatz (4.3) justified. A natural estimator for β_4^s is then obtained by replacing β_3^s in equation (4.3) by its empirical counterpart β_{3,n_s}^s. Doing so leads to the estimator

$$(4.4) \qquad \widehat{\beta}_4^s = \sum_{i=1}^{n_s} \frac{\widehat{\psi}(H_i^s)}{1 - \widehat{\psi}(H_i^s)} \delta_{H_i^s}.$$

$\widehat{\beta}_4^s$ is called the FvT estimator, and we refer to $\widehat{\psi}$ as the FvT (Four vs. Three) classifier.

The validity of Assumption 1 relies crucially upon the choice of the function class $\{\psi_\alpha\}$ or, equivalently, the choice of the classifier $\widehat{\psi}$. Indeed, off-the-shelf classifiers may lack the generalization ability to satisfy Assumption 1(ii). A secondary contribution of our work is to propose a classifier specifically tailored to four-jet collider events, which we now introduce.

The FvT classifier. Our aim is to design a classifier $\widehat{\psi}$ over $\mathcal{G}$, which:

(a) is invariant to the ordering of the constituent jets in an input event g;
(b) is invariant with respect to the equivalence relation $\simeq$ defined in (2.2);
(c) incorporates the dijet substructure of an event $g = g^1 + g^2$.

In Appendix B we describe how these properties can be satisfied using a customized convolutional neural network architecture with residual layers, or ResNet (He et al. (2016)). We refer to the resulting classifier as the FvT classifier and implement the FvT method with this choice throughout our numerical studies in Section 6. Beyond its use for background modeling, we also employ this classifier for the final dimensionality reduction map ξ in equation (3.6). Choosing these two classifiers to have the same architecture is important in practice, since a classifier capable of learning the relevant features for signal extraction should also be capable of learning and then correcting those same features in the background model.

5. Background modeling via optimal transport. The methodology described in the previous section hinged upon the ability of the classifier $\widehat{\psi}$ to accurately extrapolate from the Control Region to the Signal Region, implying that the 3b and 4b intensity functions in the latter region are constrained by their values in the former region. The validity of this assumption is difficult to verify in practice, due to the blinding of the 4b signal region which motivates us to develop a distinct approach with a complementary modeling assumption. In this section, rather than extrapolating the discrepancy between the 3b and 4b intensity functions, we will extrapolate the discrepancy between the Control and Signal Region intensity functions, as illustrated in Figure 3.

We cannot use a density ratio to quantify the discrepancy between the intensity functions in the Control and Signal Regions, because these regions are disjoint. We will instead use the notion of a *transport map*, which will be defined below. In order to employ transport maps, it will be convenient to normalize all intensity functions throughout this section. That is, we will

define an estimator for β_4^s by separately estimating the probability measure $P_4^s = \beta_4^s/\beta_4^s(\mathcal{G}_s)$ and the normalization $\beta_4^s(\mathcal{G}_s)$. More generally, we denote by

$$P_j^c = \beta_j^c/\beta_j^c(\mathcal{G}_c), \qquad P_j^s = \beta_j^s/\beta_j^s(\mathcal{G}_s), \quad j = 3, 4,$$

the four population-level probability measures, with corresponding empirical measures

$$P_{3,n_a}^a = \frac{1}{n_a} \sum_{i=1}^{n_a} \delta_{H_i^a}, \qquad P_{4,m_a}^a = \frac{1}{m_a} \sum_{i=1}^{m_a} \delta_{G_i^a}, \quad a \in \{c, s\}.$$

A transport map (Villani (2003)) between P_3^c and P_3^s is any Borel-measurable function $T : \mathcal{G}_c \to \mathcal{G}_s$ such that, whenever $H \sim P_3^c$, we have $T(H) \sim P_3^s$. When this condition holds, we write $P_3^s = T_\# P_3^c$, and we say T *pushes P_3^c forward* onto P_3^s or that P_3^s is the *pushforward* of P_3^c under T. Equivalently, this condition holds if and only if

$$P_3^s(A) = T_\# P_3^c(A) = P_3^c(T^{-1}(A)), \quad \text{for all } A \in \mathbb{B}(\mathcal{G}_s).$$

We propose to perform background estimation under the following informal modeling assumption, which will be stated more formally in the sequel.

ASSUMPTION 2'. There exists a map $T_0 : \mathcal{G}_c \to \mathcal{G}_s$ such that

(5.1) $$T_{0\#} P_3^c = P_3^s, \quad \text{and} \quad T_{0\#} P_4^c = P_4^s.$$

Assumption 2' requires the $3b$ and $4b$ distributions to be sufficiently similar for there to exist a shared map T_0 which pushes forward their restrictions to the Control Region into their counterparts in the Signal Region. If such a map T_0 were available, it would suggest the following procedure for estimating P_4^s:

(a) Fit an estimator $\widehat{T}$ of T_0 based only on the $3b$ observations;
(b) Given any estimator $\widehat{P}_{4,m_c}^c$ of P_4^c, use the pushforward $\widehat{T}_\# \widehat{P}_{4,m_c}^c$ as an estimator of P_4^s.

For this approach to be practical, we must specify an explicit candidate T_0 satisfying Assumption 2'. We propose to choose T_0 such that its movement of the probability mass from P_3^c into that of P_3^s is minimal. This leads us to consider the classical optimal transport problem, which we now describe.

5.1. *The optimal transport problem.* Assume a metric W on the space $\mathcal{G}$ is given; we provide a candidate for such a metric in Section 5.3. For any transport map T pushing P_3^c forward onto P_3^s, we refer to $W(h, T(h))$ as the cost of moving an event $h \in \mathcal{G}_c$ to an event $T(h) \in \mathcal{G}_s$. The optimal transport problem seeks to find the choice of T which minimizes the expected cost of transporting P_3^c onto P_3^s, which amounts to solving the following optimization problem:

(5.2) $$\underset{T:\mathcal{G}_c \to \mathcal{G}_s}{\arg\min} \int_{\mathcal{G}_c} W(h, T(h)) \, dP_3^c(h), \quad \text{s.t. } T_\# P_3^c = P_3^s.$$

Equation (5.2) is known as the Monge problem (Monge (1781)). When a solution T_0 to the Monge problem exists, it is said to be an *optimal transport map*. We postulate that, when it exists, the optimal transport map from P_3^c to P_3^s is a sensible candidate for the map T_0 appearing in the statement of Assumption 2'.

A shortcoming of this choice is the requirement that there exist a solution to the optimization problem (5.2). It is well-known that the Monge problem over Euclidean space admits a unique solution for absolutely continuous distributions, when the cost function is the squared Euclidean norm (Brenier (1991), Knott and Smith (1984)). While sufficient conditions for

the solvability of the Monge problem in more general spaces are given by Villani (2009, Chapter 9), we do not know whether they are satisfied by the metric space $(\mathcal{G}, W)$ under consideration. Furthermore, the Monge problem may not even be feasible between distributions which are not absolutely continuous, which precludes the possibility of estimating T_0 using the optimal transport map between the empirical measures of P_3^c and P_3^s.

Motivated by these considerations, we introduce a classical relaxation of the Monge problem, known as the Kantorovich optimal transport problem (Kantorovich (1942), Kantorovich (1948)). Let $\Pi(P_3^c, P_3^s)$ denote the set of all joint Borel distributions π over $\mathcal{G}_c \times \mathcal{G}_s$ whose marginals are, respectively, P_c^3 and P_s^3, in the sense that $P_3^c(\cdot) = \pi(\cdot \times \mathcal{G}_s)$ and $P_3^s(\cdot) = \pi(\mathcal{G}_c \times \cdot)$. We refer to such joint distributions as *couplings*. Consider the minimization problem

$$(5.3) \qquad \mathcal{W}(P_3^c, P_3^s) = \inf_{\pi \in \Pi(P_3^c, P_3^s)} \int_{\mathcal{G}_c \times \mathcal{G}_s} W(g, h) \, d\pi(g, h).$$

When the infimum in (5.3) is achieved by a coupling π_0, this last is known as an *optimal coupling*. When an optimal coupling is supported on a set of the form $\{(h, T(h)) : h \in \mathcal{G}_c\}$, for some map $T : \mathcal{G}_c \to \mathcal{G}_s$, it can be seen that T is, in fact, an optimal transport map between P_3^c and P_3^s. The Kantorovich problem (5.3) is, therefore, a relaxation of the Monge problem (5.2). Unlike the latter, however, the minimization problem (5.3) is always feasible since $\Pi(P_3^c, P_3^s)$ is nonempty; indeed, $\Pi(P_3^c, P_3^s)$ always contains the independence coupling $P_3^c \otimes P_3^s$. Moreover, the infimum in the Kantorovich problem is achieved as long as the cost function W is lower semicontinuous, and the measures P_3^c and P_3^s satisfy a mild moment condition (Villani (2009), Theorem 4.1). We also note that the optimal objective value $\mathcal{W}(P_3^c, P_3^s)$ defines a metric between probability measures called the (first-order) Wasserstein distance (Villani (2003)), or Earth Mover's distance (Rubner, Tomasi and Guibas (2000)).

Using the Kantorovich relaxation, we now formalize Assumption 2' into the following condition, which we shall require throughout the remainder of this section.

ASSUMPTION 2. Assume there exists an optimal coupling $\pi_0 \in \Pi(P_3^c, P_3^s)$ between P_3^c and P_3^s. Given a pair of random variables $(H^c, H^s) \sim \pi_0$, let $\pi_0(\cdot|h)$ denote the conditional distribution of H^s, given $H^c = h$, for any $h \in \mathcal{G}_c$. Then, the following implication holds:

$$(5.4) \qquad \begin{array}{c} G^c \sim P_4^c \\ G^s | G^c \sim \pi_0(\cdot|G^c) \end{array} \implies G^s \sim P_4^s.$$

Assumption 2 requires the $3b$ and $4b$ distributions to be sufficiently similar for their restrictions to the Signal and Control Regions to be related by a common conditional distribution. It further postulates that this conditional distribution is induced by the optimal coupling π_0. Heuristically, $\pi_0(\cdot|H)$ plays the role of a multivalued optimal transport map for pushing an event H from the distribution P_3^c onto P_3^s. Assumption 2 requires this map to additionally push the distribution P_4^c onto its counterpart P_4^s in the Signal Region. In the special case where there exists an optimal transport map T_0 from P_3^c to P_3^s, we note that $\pi_0 = (Id, T_0)_\# P_3^c$ is an optimal coupling of P_3^c with P_3^s, where Id denotes the identity map. In this case, equation (5.4) is tantamount to equation (5.1).

5.2. *Background estimation*. We next derive estimators for the background distribution P_4^s under Assumption 2. It follows from equation (5.4) and the law of total probability that

$$P_4^s(\cdot) = \int_{\mathcal{G}_c} \pi_0(\cdot|g) \, dP_4^c(g).$$

Since $\pi_0(\cdot|g)$ is the distribution of H^s, given $H^c = g$, induced by the optimal coupling π_0, it is an identified parameter which can be estimated using only the $3b$ data. Given an estimator $\widehat{\pi}(\cdot|g)$ of this quantity and an estimator $\widehat{P}^c_{4,m_c}$ of P^c_4, it is natural to consider the plugin estimator of the background distribution P^s_4, given by

$$(5.5) \qquad \widehat{P}^s_4(\cdot) := \int_{\mathcal{G}_c} \widehat{\pi}(\cdot|g)\,\mathrm{d}\widehat{P}^c_{4,m_c}(g).$$

In what follows, we begin by defining an estimator $\widehat{\pi}(\cdot|g)$ in Section 5.2.1, followed by two candidates for the estimator $\widehat{P}^c_{4,m_c}$, leading to two distinct background estimation methods described in Sections 5.2.2 and 5.2.3. In Section 5.2.4 we briefly discuss how these constructions also lead to estimators of the unnormalized intensity measure β^s_4. We then provide discussion and comparison of these methodologies in Section 5.2.5.

5.2.1. *The empirical optimal transport coupling.*

A natural plugin estimator for the coupling π_0 is the optimal coupling $\widehat{\pi}$ between the empirical measures P^c_{3,n_c} and P^s_{3,n_s}. In detail, denoting by $\widehat{q} \in \mathbb{R}^{n_c \times n_s}$ the joint probability mass function of $\widehat{\pi}$, the empirical Kantorovich problem takes the following form:

$$(5.6) \qquad \begin{aligned} \widehat{q} = (\widehat{q}_{ij}) \in \underset{(q_{ij})\in\mathbb{R}^{n_c \times n_s}}{\mathrm{argmin}} \sum_{i=1}^{n_c} \sum_{j=1}^{n_s} q_{ij} W(H^c_i, H^s_j), \\[2mm] \text{s.t. } q_{ij} \geq 0, \sum_{i=1}^{n_c} q_{ij} = \frac{1}{n_s}, \sum_{j=1}^{n_s} q_{ij} = \frac{1}{n_c}. \end{aligned}$$

Equation (5.6) is a finite-dimensional linear program for which exact solutions may be computed using network simplex algorithms such as the Hungarian algorithm (Kuhn (1955)). We refer to Peyré and Cuturi (2019) for a survey. We then define the estimator $\widehat{\pi}(\cdot|H^c_i)$, for $i \in [n_c]$, as the discrete distribution over $\{H^s_1, \ldots, H^s_{n_s}\}$ with probability mass function

$$\widehat{q}_{j|i} = \frac{\widehat{q}_{ij}}{\sum_{k=1}^{n_s} \widehat{q}_{ik}} = n_c \cdot \widehat{q}_{ij}, \quad j = 1, \ldots, n_s.$$

We are now in a position to define estimators of the background distribution P^s_4.

5.2.2. *The OT-kNN estimator.*

We first consider the general estimator in equation (5.5) when $\widehat{P}^c_{4,m_c}$ is the empirical measure P^c_{4,m_c}. This choice is perhaps most natural, but it requires us to perform out-of-sample evaluations of the estimator $\widehat{\pi}(\cdot|g)$. Indeed, recall that the latter is defined over $\{H^c_1, \ldots, H^c_{n_c}\}$, whereas P^c_{4,m_c} is supported on $\{G^c_1, \ldots, G^c_{m_c}\}$.

We extend the support of $\widehat{\pi}(\cdot|g)$ to all $g \in \mathcal{G}_c$ using a variant of the nearest neighbors method for nonparametric regression (Biau and Devroye (2015)). A similar procedure has also been used, for instance, by Flamary et al. (2021), Manole et al. (2024). Let $k \geq 1$ be an integer. For all $g \in \mathcal{G}_c$, let $I_k(g)$ denote the indices of the k-nearest neighbors of g with respect to W, among $H^c_1, \ldots, H^c_{n_c}$. Specifically, we set $I(g) = \{j_1, \ldots, j_k\} \subseteq [n_c]$, where

$$W(g, H^c_{j_1}) \leq \cdots \leq W(g, H^c_{j_k}) \leq W(g, H^c_j), \quad \text{for all } j \in [n_c] \setminus \{j_1, \ldots, j_k\}.$$

Furthermore, define the inverse distance weights

$$(5.7) \qquad \omega_i(g) = \frac{1/W(g, H^c_i)}{\sum_{l\in I_k(g)} 1/W(g, H^c_l)}, \quad i \in I_k(g)$$

with the convention $\infty/\infty = 1$. We then define, for all $g \in \mathcal{G}_c$,

$$(5.8) \qquad \widehat{\pi}_{k\mathrm{NN}}(\cdot|g) = \sum_{i\in I_k(g)} \omega_i(g)\widehat{\pi}(\cdot|H^c_i).$$

The estimator $\widehat{\pi}_{kNN}(\cdot|g)$ couples g with all of the events to which its k-nearest neighbors are coupled under $\widehat{\pi}$. The coupling values which correspond to the closest nearest neighbors are assigned higher weights $\omega_i(g)$. Furthermore, we note that, when $g \in \{H_1^c, \ldots, H_{n_c}^c\}$, it holds that $\widehat{\pi}_{kNN}(\cdot|g) = \widehat{\pi}(\cdot|g)$. With these defintions the generic estimator (5.5) takes the form

$$\widehat{P}_{4,kNN}^s(\cdot) := \int_{\mathcal{G}_c} \widehat{\pi}_{kNN}(\cdot|g) \, dP_{4,m_c}^c(g) = \frac{1}{m_c} \sum_{\ell=1}^{m_c} \sum_{i \in I_k(G_\ell^c)} \omega_i(G_\ell^c) \widehat{\pi}(\cdot|H_i^c),$$

or equivalently,

$$\widehat{P}_{4,kNN}^s = \frac{n_c}{m_c} \sum_{j=1}^{n_s} \left(\sum_{\ell=1}^{m_c} \sum_{i \in I_k(G_\ell^c)} \omega_i(G_\ell^c) \widehat{q}_{ij} \right) \delta_{H_j^s}.$$

We refer to $\widehat{P}_{4,kNN}^s$ as the OT-kNN (Optimal Transport–k Nearest Neighbor) estimator of P_4^s.

5.2.3. *The OT-FvT estimator.* The rate of production of $3b$ events typically exceeds that of $4b$ events by one order of magnitude (cf. Section 6). As a result, in the general formulation (5.5) of our optimal transport map estimators, we expect to have access to a smaller sample size m_c for estimating the distribution P_4^c than the sample sizes n_c and n_s for estimating the optimal transport coupling π_0. Motivated by this observation, we next define an estimator $\widehat{P}_{4,m_c}^c$ which can leverage the larger $3b$ sample size n_c.

Let $p_j^c = dP_j^c/d\nu_0$ denote the density of P_j^c for $j = 3, 4$. Recall from Section 4 that, for any event g, $\psi(g)$ denotes the probability that a random event G arose from the $4b$ distribution as opposed to the $3b$ distribution, given that $G = g$. Furthermore, $\widehat{\psi}(g)$ denotes the $[0, 1]$-valued output of the FvT classifier for discriminating $4b$ events from $3b$ events. Recall further that, for any $g \in \mathcal{G}_c$, it holds that $p_4^c(g)/p_3^c(g) = (\beta_3^c(\mathcal{G}_c)/\beta_4^c(\mathcal{G}_c)) \cdot (\psi(g)/(1 - \psi(g))$, or equivalently,

$$P_4^c(A) = \frac{\beta_3^c(\mathcal{G}_c)}{\beta_4^c(\mathcal{G}_c)} \int_A \frac{\psi(h)}{1 - \psi(h)} \, dP_3^c(h), \quad A \in \mathbb{B}(\mathcal{G}_c).$$

We define a plugin estimator of the above quantity via

$$(5.9) \qquad \widehat{P}_{4,m_c}^c(A) = \frac{n_c}{m_c} \int_A \frac{\widehat{\psi}(h)}{1 - \widehat{\psi}(h)} \, dP_{3,n_c}^c(h), \quad A \in \mathbb{B}(\mathcal{G}_c).$$

$\widehat{P}_{4,m_c}^c$ can be viewed as a reweighted version of the empirical measure P_{3,n_c}^c. The weights are chosen to make the $3b$ sample resemble a $4b$ sample by using the FvT classifier to estimate the density ratio p_4^c/p_3^c. Since the $3b$ sample is one order of magnitude larger than the $4b$ sample, we heuristically expect this estimator to have smaller theoretical risk than the empirical measure P_{4,m_c}^c whenever the density ratio p_4^c/p_3^c is smooth.

A second motivation for using the estimator $\widehat{P}_{4,m_c}^c$ is the fact that it is supported on the domain of definition of the in-sample empirical optimal transport coupling $\widehat{\pi}(\cdot|g)$. We, therefore, do not need to extend the domain of this estimator, unlike the previous section. With these choices, the generic estimator in equation (5.5) takes the following form:

$$(5.10) \qquad \widehat{P}_{4,OF}^s := \int_{\mathcal{G}_c} \widehat{\pi}(\cdot|g) \, d\widehat{P}_{4,m_c}^c(g) = \frac{n_c}{m_c} \sum_{j=1}^{n_s} \left(\sum_{i=1}^{n_c} \frac{\widehat{\psi}(H_i^c)}{1 - \widehat{\psi}(H_i^c)} \widehat{q}_{ij} \right) \delta_{H_j^s}.$$

We refer to $\widehat{P}_{4,OF}^s$ as the OT-FvT (Optimal Transport–Four vs. Three) estimator of P_4^s.

5.2.4. *Estimation of the background normalization.* We briefly show how the OT-kNN and OT-FvT estimators can also be used to estimate the unnormalized background intensity function β_4^s. We employ the widely-used ABCD method (Alison (2015), ATLAS (2015), Choi and Oh (2021), Kasieczka et al. (2021)), which requires the following assumption.

ASSUMPTION 3. It holds that $\beta_4^s(\mathcal{G}_s) = \beta_3^s(\mathcal{G}_s)\beta_4^c(\mathcal{G}_c)/\beta_3^c(\mathcal{G}_c)$.

Assumption 3 implies that the ratio of the number of $4b$ to $3b$ events in the Control Region should be the same as that in the Signal Region. Under this assumption a natural estimator for $\beta_4^s(\mathcal{G}_s)$ is simply given by $m_c n_s/n_c$. Therefore, under Assumptions 2–3, the probability measures $\widehat{P}_{4,k\mathrm{NN}}^s$ and $\widehat{P}_{4,\mathrm{OF}}^s$ can be used to define the following two estimators of the unnormalized background intensity measure β_4^s,

$$(5.11) \qquad \widehat{\beta}_{4,k\mathrm{NN}}^s = \frac{m_c n_s}{n_c}\widehat{P}_{4,k\mathrm{NN}}^s, \qquad \widehat{\beta}_{4,\mathrm{OF}}^s = \frac{m_c n_s}{n_c}\widehat{P}_{4,\mathrm{OF}}^s.$$

We, respectively, refer to the above measures as the OT-kNN and OT-FvT estimators of β_4^s, or simply as the OT-kNN and OT-FvT methods.

5.2.5. *Remarks.* We summarize the three background estimation methods, FvT, OT-kNN, and OT-FvT, in Table 1, and make the following remarks:

- Assumption 2 is the primary modeling assumption required by OT-kNN and OT-FvT. We view this condition as being complementary to Assumption 1(ii), required by the FvT method. Indeed, it involves an extrapolation (of an optimal coupling) from the $3b$ to $4b$ distribution rather than an extrapolation (of a density ratio) from the Control Region to the Signal Region.
- The OT-FvT estimator (5.10) can, alternatively, be interpreted through the lens of domain adaptation for the FvT classifier. To make this connection clear, suppose for simplicity that $n_c = m_c$. In this case it can be shown that $\widehat{\pi}$ is, in fact, induced by an optimal transport map, in the sense that there exists a permutation $\widehat{\tau} : [n_c] \to [n_c]$ such that

$$\widehat{q}_{ij} = I(i = \widehat{\tau}(j))/n_c, \quad i, j = 1, \ldots, n_c.$$

The FvT and OT-FvT estimators then take the following form:

$$\widehat{\beta}_{4,\mathrm{FvT}}^s \propto \sum_{j=1}^{n_s} \frac{\widehat{\psi}(H_j^s)}{1 - \widehat{\psi}(H_j^s)}\delta_{H_j^s}, \qquad \widehat{\beta}_{4,\mathrm{OF}}^s \propto \sum_{j=1}^{n_s} \frac{\widehat{\psi}(H_{\widehat{\tau}(j)}^c)}{1 - \widehat{\psi}(H_{\widehat{\tau}(j)}^c)}\delta_{H_j^s}.$$

While the FvT method evaluates the density ratio estimator $\widehat{\psi}/(1 - \widehat{\psi})$ at events H_j^s in the Signal Region, the OT-FvT method evaluates it at the events $H_{\widehat{\tau}(j)}^c$ in the Control Region, to which the events H_j^s are mapped under the empirical optimal coupling $\widehat{\pi}$. The OT-FvT method thus circumvents the evaluation of $\widehat{\psi}$ outside the region where it was trained.

TABLE 1
Summary of the three background estimation methods: FvT, OT-kNN, and OT-FvT. The final estimator for each method takes the form $\widehat{\beta}_4^s \propto \sum_{j=1}^{n_s} v_j \delta_{H_j^s}$, for the values of v_j listed in the table

Estimator (of the form $\propto \sum_{j=1}^{n_s} v_j \delta_{H_j^s}$)	FvT	OT-FvT	OT-kNN
v_j	$\dfrac{\widehat{\psi}(H_j^s)}{1-\widehat{\psi}(H_j^s)}$	$\sum_{i=1}^{n_c}\dfrac{\widehat{\psi}(H_i^c)}{1-\widehat{\psi}(H_i^c)}\widehat{q}_{ij}$	$\sum_{\ell=1}^{m_c}\sum_{i \in I_k(G_\ell^c)}\omega_i(G_\ell^c)\widehat{q}_{ij}$

Optimal transport has similarly been used in past literature as a tool for domain adaptation between train and test data in classification problems (cf. Section 1.2).

- In defining the estimator OT-kNN, we proposed to extend the domain of definition of the empirical optimal transport coupling $\widehat{\pi}(\cdot|g)$ to the entire space $\mathcal{G}_c$ via nearest neighbor extrapolation; cf. equation (5.8). It was shown by Manole et al. (2024) that, for the quadratic optimal transport problem over Euclidean space, such a procedure has statistically minimax optimal risk for estimating the underlying optimal transport map T_0, assuming that it exists and is Lipschitz continuous. Nevertheless, the risk of this estimator suffers severely from the curse of dimensionality and does not generally improve when T_0 enjoys higher regularity. Manole et al. (2024) and Deb, Ghosal and Sen (2021) have instead shown that plugin estimators of T_0, based on density estimates of P_3^c and P_3^s, may achieve improved convergence rates in such settings. In our context it is challenging to perform density estimation over the space of measures $\mathcal{G}$—and particularly over the nonconvex set $\mathcal{G}_c$—thus we did not follow this approach. Our aim was instead to alleviate the curse of dimensionality inherent to the OT-kNN method by introducing the OT-FvT method. Indeed, we view the task of estimating P_4^c as a larger statistical bottleneck than that of estimating π_0, and the estimator $\widehat{P}_{4,m_c}^c$ (used by the OT-FvT method) may potentially achieve smaller risk than the empirical measure P_{4,m_c}^c (used by the OT-kNN method).

- Manole et al. (2024) additionally show that the value $k = 1$ suffices for the estimator $\widehat{\pi}_{k\text{NN}}$ to enjoy optimal theoretical risk. In our work we nevertheless allow for k to be greater than 1 in order to leverage the larger size of the $3b$ sample. For example, when $k = 1$, the estimator $\widehat{\beta}_{4,k\text{NN}}^s$ is supported on at most m_c events, whereas it can be supported on as many as $n_s \gg m_c$ events if k is chosen sufficiently large. In practice, we recommend choosing k to be as small as possible while ensuring that $\widehat{\beta}_{4,k\text{NN}}^s$ has support size on the same order as n_s—this typically amounts to choosing k to be on the order of n_s/m_c. In our simulation study (cf. Section 6), we, therefore, choose the value $k = 10$ but also illustrate the performance of the OT-kNN method for other values of k.

- We have chosen to separately estimate the probability measure P_4^s and the normalization $\beta_4^s(\mathcal{G}_s)$, because the classical optimal transport problem is only well-defined between measures with the same total mass. A possible alternative is to consider the *partial* (Figalli (2010)) or *unbalanced* (Liero, Mielke and Savaré (2018)) optimal transport problems between the unnormalized intensity measures β_3^c and β_3^s. These variants of optimal transport are well-defined between measures that have possibly different mass but have the downside of introducing tuning parameters. As we explain in Section 6, the normalizations $\beta_3^c(\mathcal{G}_c)$ and $\beta_3^s(\mathcal{G}_s)$ are of the same order of magnitude and can, in fact, be made to coincide by tuning the definition of the Control and Signal regions; thus, we have simply focused our attention on the classical (balanced) optimal transport problem in this work. Nevertheless, in the following subsection, we will employ a variant of the partial optimal transport problem to define the metric W.

5.3. *A metric between collider events.* We now describe a candidate for the metric W on $\mathcal{G}$. Recall that the Kantorovich problem in (5.3) gave rise to the Wasserstein distance $\mathcal{W}$ between probability distributions over $\mathcal{G}$. By a recursion of ideas, we will also define W to be a Wasserstein-type metric, arising from the optimal transport problem between constituent jets of events. This approach was introduced by Komiske, Metodiev and Thaler (2019). They propose to metrize $\mathcal{G}$ using a variant of the Wasserstein distance, which is well-defined between measures with nonequal mass (Pele and Werman (2008), Peleg, Werman and Rom (1989)). Given any two collider events $g = \sum_{j=1}^4 p_{T_j}\delta_{(\eta_j,\phi_j,m_j)} \in \mathcal{G}$, $h = \sum_{j=1}^4 p'_{T_j}\delta_{(\eta'_j,\phi'_j,m'_j)} \in \mathcal{G}$,
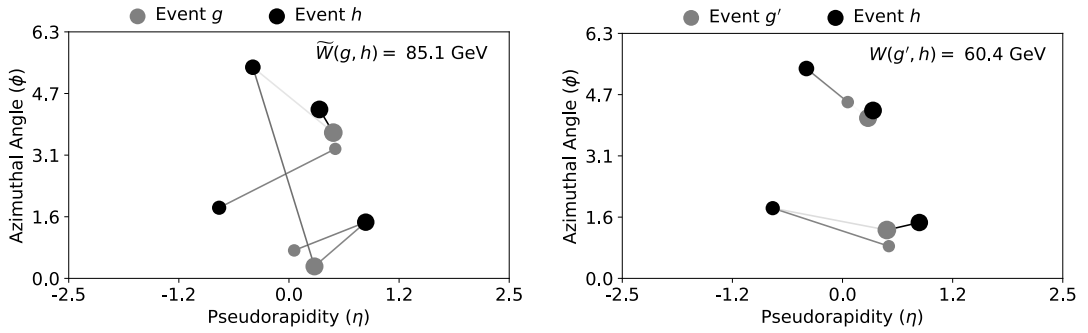
FIG. 4. *Left*: (η, ϕ)-plot of two events $g, h \in \mathcal{G}$. Each point represents a constituent jet, with size proportional to its p_T value. A line connecting the ith jet of event g to the jth jet of event h indicates a nonzero value of the optimal coupling f_{ij}, with line darkness increasing as a function of the magnitude of f_{ij}. *Right*: (η, ϕ)-plot of events $g', h \in \mathcal{G}$, where $g' \simeq g$ is an approximate minimizer in equation (5.13). One has $W(g, h) = W(g', h) < \widetilde{W}(g, h)$.

the metric is defined by

$$
\widetilde{W}(g, h) = \min_{(f_{ij}) \in \mathbb{R}^{4 \times 4}} \frac{1}{R} \sum_{i=1}^{4} \sum_{j=1}^{4} f_{ij} \sqrt{(\eta_i - \eta_j')^2 + (\phi_i - \phi_j')^2} + \left| \sum_{i=1}^{4} (p_{T_i} - p_{T_i}') \right|
$$
(5.12)
$$
\text{s.t. } f_{ij} \geq 0, \sum_j f_{ij} \leq p_{T_i}, \sum_i f_{ij} \leq p_{T_j}', \sum_{i,j} f_{ij} = \min\left(\sum_i p_{T_i}, \sum_j p_{T_j}' \right),
$$

for a tuning parameter $R > 0$. We make several remarks about this definition:

- In the context of particle physics, the coupling f_{ij} is naturally interpreted as a flow of energy (measured in terms of the transverse momentum p_T) from jet i of g to jet j of h, as depicted in Figure 4. $\widetilde{W}(g, h)$ thus measures the smallest possible transport of energy required to rearrange the jets of the event g into those of h.
- We have followed Komiske, Metodiev and Thaler (2019) by omitting the mass variables m_j and m_j' from the definition of $\widetilde{W}$. This choice is further discussed in the context of our simulation study in Section 6.
- The tuning parameter R trades off the influence of the angular variables ϕ_i, η_i and that of the energy variables p_{T_i}. Our choice of R is further discussed in Section 6.

The metric $\widetilde{W}$ does not, however, take into account the equivalence relation $\simeq$ over $\mathcal{G}$ defined in equation (2.2). For example, $\widetilde{W}(g, h)$ could be nonzero, even when g and h are deemed equivalent for our purposes. We, therefore, define our final metric W by

(5.13)
$$
W(g, h) = \inf\{\widetilde{W}(g', h) : g' \simeq g, g' \in \mathcal{G}\}, \quad g, h \in \mathcal{G}.
$$

Strictly speaking, W now becomes a metric over the set of equivalence classes of events induced by $\simeq$. We refer to Figure 4 for an illustration. In practice, we numerically approximate W using a procedure described in Appendix C.

6. Simulation study.

6.1. *Simulation description.* In this section we compare the performance of the three background modeling methods OT-FvT, OT-kNN, and FvT on realistic simulated collider data, generated using the MadGraph particle physics software (Alwall et al. (2011)). Code for reproducing this simulation study is publicly available.[4]

---

[4]https://github.com/tmanole/HH4bsim.

Since b-tagging is imperfect, in practice, we expect the 3b and 4b samples to be composed of a mixture of different multijet scattering processes which do necessarily arise from b-quarks. We perform a study in MadGraph to estimate the relative scale of such processes. Assuming a b-jet tagging efficiency of 75%, a charm jet tagging efficiency of 15% and a light jet tagging efficiency of 1%, we find that the 4b (resp., 3b) sample consists of 90% (10%) events in a final state with four b quarks, 7% (9%) events in a final state with two b quarks and two charm quarks, and 4% (80%) in a final state with two b quarks and two light quarks. In particular, we stress that a fraction of the 3b sample consists of mislabelled 4b events, which could be signal events. This signal contamination is expected to be sufficiently small to be considered negligible for purposes of a signal discovery analysis, as in this paper.

We generate four-quark events in MadGraph according to the percentages listed above. The calorimeters in the CMS detector are not perfect, and the measured jet energies have a finite resolution. The distribution of the observed smeared energy is well-approximated by the normal distribution $N(E, \sigma^2(E))$, where E denotes the true energy of a jet, and $\sigma(E)$ satisfies

$$\left(\frac{\sigma(E)}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2$$

for some constants $S, N, C \geq 0$. We apply this smearing to the quark four-vectors, setting $S = 0.98$, $N = 0$, $C = 0.054$. For simplicity we set the quark masses to zero and omit them from the the metric W. When applying these methods to real data, it may, however, be useful to incorporate the jet masses into the definition of W. We also apply jet-level scale factors to account for the p_T dependence of CMS b-tagging for light, charm, and bottom quark jets,

$$\text{Scale Factor} = \begin{cases} (2.5 p_T e^{-7 p_T} + 0.6)/0.75 & b\text{-quark} \\ (p_T e^{-10 p_T} + 0.2)/0.15 & c\text{-quark} \\ (0.03 p_T + 0.01)/0.01 & u, d, s\text{-quark or gluon,} \end{cases}$$

where p_T is measured in TeV. Events are weighted by the product of the scale factors for the b-tagged jets.

Following this preprocessing of the data, the pairing algorithm described in Section 3.2 is applied to all events, and those falling within the Control and Signal Regions are kept. We define these regions according to equations (3.7)–(3.8), with the parameters $\sigma_c = 1.03$, $\kappa_s = 1.6$, and $r_c = 30$ GeV. The final sample consists of $n_s = 201, 568$ events in the 3b Signal Region, $n_c = 159, 427$ events in the 3b Control Region, $m_s = 28, 980$ events in the 4b Signal Region, and $m_c = 22, 053$ events in the 4b Control Region. The order of magnitude of these sample sizes as well as the proportion of 3b to 4b events is similar to those used in recent di-Higgs analyses at the LHC (ATLAS (2019)). We also simulate a separate 4b sample of size approximately $10(n_s + m_s)$, which we choose not to contain any signal events, and whose distribution we treat as the ground truth for the purpose of validating our background models.

We additionally generate a Monte Carlo sample from the SM di-Higgs signal distribution with which the signal intensity rates $(S_j)_{j=1}^J$, used to form the likelihood function (3.5), can be specified. For the purpose of validating our background models, we train a $[0, 1]$-valued classifier $\widehat{\xi}$ (abbr. Signal vs. Background, or SvB, classifier) to discriminate the 4b data from the Monte Carlo di-Higgs sample. Given that our simulated 4b sample contains no signal events, $\widehat{\xi}$ forms a reasonable proxy for the theoretical binning function ξ in equation (3.6). The SvB classifier has the same architecture as that of the FvT classifier described in Appendix B. In the sequel we refer to $\widehat{\xi}(g)$ as the SvB value corresponding to an event g.

Finally, we discuss our choice of the parameter R arising in the definition of the metric W. In order for the two terms in the definition of $\widetilde{W}$ to be of comparable order, we make the requirement that R lie within the range of the first summand in equation (5.12). We identify

this range as follows. Since b-tagging is only performed for values of η lying in the interval $[-2.5, 2.5]$, we impose $R \leq \sqrt{\pi^2 + 5^2} \approx 5.9$. Furthermore, jet clustering algorithms used by CMS merge particles whose (η, ϕ)-Euclidean distance is within 0.4 (Cacciari, Salam and Soyez (2008), CMS (2017)), thus we impose $R \geq 0.4$. Now, since we expect that the largest discrepancies between the Control and Signal Region distributions arise in the kinematic variables (η, ϕ), we choose the smallest possible value $R = 0.4$ when fitting the empirical optimal transport coupling $\hat{\pi}$. On the other hand, for the nearest-neighbor lookup of the OT-kNN method, we set $R = 2.75$, which is the midpoint of the interval $[0.4, 5.9]$. We make no attempt to tune these values of R, and we leave open the question of choosing them in a data-driven fashion. We compute the metric W in part using the EnergyFlow Python library (Komiske, Metodiev and Thaler (2022)), and we compute optimal couplings between distributions of collider events using the Python Optimal Transport library (Flamary et al. (2021)); see Appendix D for further details.

6.2. *Simulation results.* The fitted intensity measures $\widehat{\beta}_4^s$ produced by the three background methods (FvT, OT-kNN, and OT-FvT) are binned and plotted in Figure 5. Logarithmic scales are used to better visualize signal-rich regions. Plots with additional kinematic variables are given in Appendix E. It can be seen that the three methods yield qualitatively similar estimates of the SvB intensity function. We recall that the SvB variable is of primary interest to model, as it is used as the final discriminant when testing the signal hypothesis (3.1). The m_{HH} variable has also been used as the final discriminant in recent di-Higgs studies (Bryant (2018)). Given an event $g \in \mathcal{G}$ with dijet pairing $\{g^1, g^2\}$, its m_{HH} value is defined as follows,[5] using the notation of Section 3.2,

$$(6.1) \qquad m_{\text{HH}}(g) = m\left(\frac{m_{\text{H}}}{m(g^1)}g^1 + \frac{m_{\text{H}}}{m(g^2)}g^2\right).$$

Once again, we observe that this variable is well-modelled by all three methods. Among the various kinematic variables which we analyzed, the "ΔR_{jj}–Other" variable, appearing in
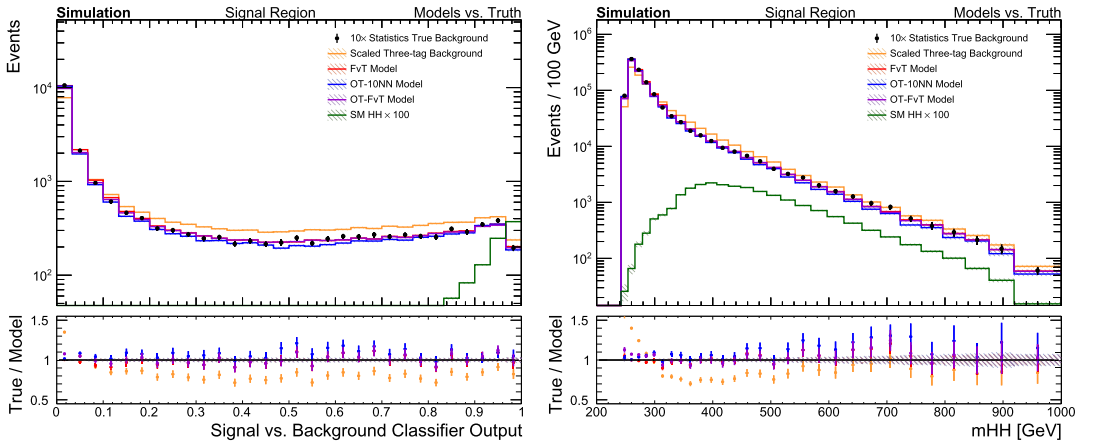

FIG. 5. *Histograms of the the SvB classifier output variable (left) and the m_{HH} variable (right) for the three background models as well as the upsampled 4b data (treated as the ground truth), the 3b data (normalized by the factor $n_s m_c / n_c$), and the di-Higgs signal sample (SM HH). Error bars in the kth bin of any histogram denote $\pm\sqrt{N_k}$, where N_k is the number of events per bin. Error bars in the ratio plot denote $\pm\sqrt{N_k}/N_{0k}$, where N_{0k} is the number of observed 4b events per bin. The dashed lines in the ratio plot denote $\pm\sqrt{1/N_{0k}}$.*

---

[5]Equation (6.1) can be interpreted as the four-body invariant mass after the dijet four-vectors have been corrected to have the Higgs boson mass.

Figure 5 of Appendix E, presents one of the largest qualitative discrepancies between the three methods and appears to be best-modelled by the FvT method. In all cases the three methods provide a significant improvement compared to the uncorrected 3*b* sample.

In order to provide a quantitative comparison of these methods, we develop a heuristic two-sample test for testing equality between the distribution of the fitted background models and of the true upsampled 4*b* data. Specifically, we form a proxy for a two-sample test by training classifiers to discriminate each of the background estimates from the upsampled 4*b* data (similar approaches have previously been used in the high energy physics literature by Krause and Shih (2023a), Krause and Shih (2023b)). For each classifier we record the area under the receiver operating characteristic curve (AUC; Hanley and McNeil (1982)), and any deviation of this quantity from 0.5 is an indication of mismodeling. We again choose our classifiers to be residual neural networks with the architecture described in Appendix B. Although this choice is inherently favourable to the FvT method, and to some extent the OT-FvT method, we use it because it coincides with the SvB classifier architecture and will thus be most powerful at detecting mismodeling in the features which are relevant for the final signal analysis. Another caveat with the use of the AUC as a performance metric is the fact that it quantifies the overall quality of the background estimates over the whole Signal Region and might hence not be sensitive to local deviations in the most signal-rich part of the Signal Region. For this reason it is necessary to also look at the quality of the estimates in terms of metrics that are localized to the most signal-rich part of the phase space, as is done in Figure 5 (left) for the di-Higgs signal. If the background estimates are to be used with any other alternative signal model, similar checks would need to be done with metrics localized for those signals.

The fitted AUC value for each method is reported in Figure 6. Though all AUCs are significantly greater than 0.5, they are substantially lower for our background models than for the benchmark method consisting of the uncorrected 3b sample. The FvT method has the lowest AUC, followed closely by the OT-FvT method and OT-kNN method. While the OT-1NN method has comparable AUC point estimate as the OT-FvT method, we emphasize that its variability interval is wider, which could have been anticipated from the discussion in Section 5.2.5 where we emphasized that the support size of $\beta^s_{4,1NN}$ can be an order of magnitude smaller than n_s. In contrast, the OT-10NN and OT-20NN estimators have narrower variability intervals than OT-1NN but have markedly larger AUC point estimates than the remaining methods. The performance of the OT-kNN method for varying values of k is also illustrated in Figure 7 as a function of the SvB and m_{HH} variables.
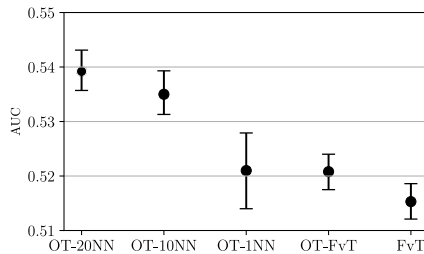

FIG. 6.   *Fitted AUC values obtained by discriminating each background model from the upsampled 4b data using the FvT classifier, in the Signal Region, together with 95% percentile bootstrap variability intervals, obtained by bootstrapping the predicted classifier probabilities. 1000 bootstrap replications are used. Note that this bootstrap procedure does not take into account the variability of the background estimators themselves. For the 3b-tagged data, we obtain the AUC 0.5843, with variability interval [0.5812, 0.5874].*
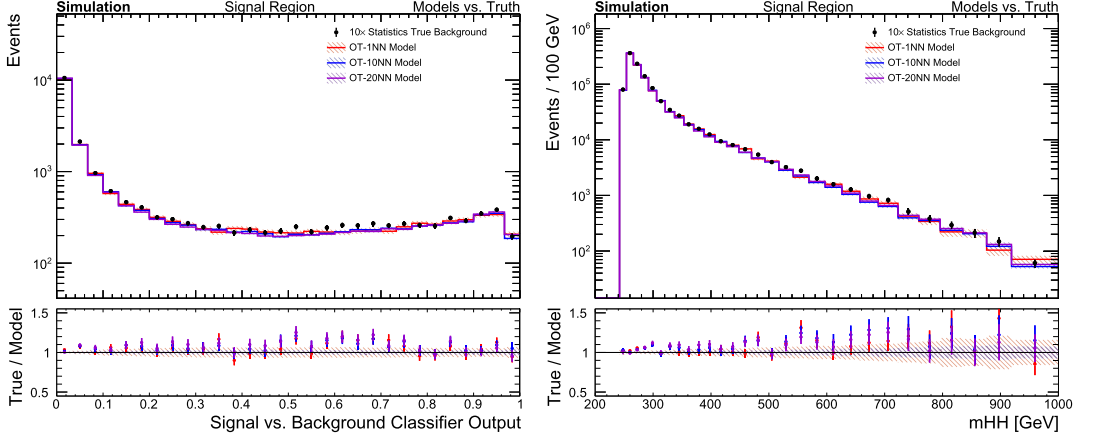
FIG. 7. *Histograms of the SvB classifier output variable (left) and the m_{HH} variable (right) for the OT-kNN estimator, with $k \in \{1, 10, 20\}$.*

We next provide a qualitative comparison of the fitted weights produced by the three background modeling methods. Recall that these methods all take the form

$$\widehat{\beta}_4^s = \sum_{i=1}^{n_s} v_i \delta_{H_i^s}$$

for some nonnegative weights v_i, which are summarized up to normalization in Table 1. In Figure 8, we plot the weights of the two optimal transport methods against those of the FvT method. We observe that the FvT and OT-FvT methods produce weights which are concentrated and symmetric around the identity. This implies that the odds ratio of the FvT classifier at a point H_j^s in the Signal Region behaves similarly to the odds ratio at any point H_i^c in the Control Region to which H_j^s is optimally coupled. This suggests that the transfer learning of the FvT classifier from the Control Region to the Signal Region is, to some extent, well-modelled by the optimal transport coupling $\widehat{\pi}$. This observation heuristically suggests that Assumptions 1–2 both hold in this simulation. In contrast to the method OT-FvT, we observe that the method OT-10NN produces markedly different weights than the FvT method, which can partly be anticipated from the discrete nature of the nearest neighbor extrapolation. We
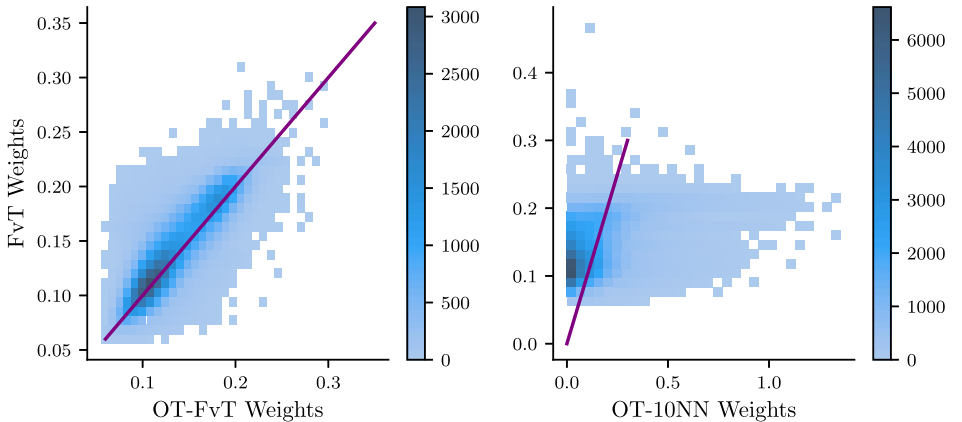

FIG. 8. *Bivariate histogram of the 3b data $H_1^s, \ldots, H_{n_s}^s$ in the Signal Region, plotted in terms of the weights of the OT-FvT method against those of the FvT method (left) and of the OT-10NN method against those of the FvT method (right). The purple lines denote the identity function.*

conjecture that the nearest-neighbor estimator of the optimal transport coupling has poorer theoretical risk than its counterpart in the OT-FvT method.

7. Conclusion and discussion. Our aim has been to study the problem of data-driven background estimation, motivated by the ongoing search for double Higgs boson production in the 4*b* final state. After recalling a widely-used approach to this problem based on transfer learning of a multivariate classifier, our first contribution was to develop the FvT classifier architecture, which is tailored to collider data and which can serve as a powerful tool for implementing this methodology. Our primary contribution was then to propose a distinct background estimation method based on the optimal transport problem. A recurring theme throughout our work has been the complementarity of the modeling assumptions made by these two distinct approaches, which allows them to be used as cross-checks for one another in practice. We substantiate this point with a realistic simulation study, in which these methodologies appear to give consistent results despite their inherently distinct derivations.

Quantifying the uncertainty of our background estimates is a challenging problem left open by our work, which is nonetheless crucial for applying our methods in practice. In the experimental particle physics community, it is commonplace to measure both *statistical* uncertainties—those arising from fluctuations of the data generating process—and *systematic* uncertainties—those arising from potential mismodeling (Heinrich and Lyons (2007)). Both of these forms of uncertainty are challenging to quantify in our context. For instance, a prerequisite for quantifying the statistical uncertainty of the methods OT-kNN or OT-FvT is to perform statistical inference for optimal transport maps. This is a difficult open problem in the statistical optimal transport literature that has only been addressed for some special cases (Ramdas, García Trillos and Cuturi (2017), Rippl, Munk and Sturm (2016)) and for regularized variants of optimal transport maps, which differ from those used in our work (Goldfeld et al. (2024), Gunsilius and Xu (2021), Klatt, Tameling and Munk (2020), González-Sanz, Loubes and Niles-Weed (2022)). The question of quantifying systematic uncertainties is more open-ended and typically involves heuristics for assessing the extent of potential mismodeling by the background estimation methods. Due to the complementarity of assumptions placed by our methods, any lack of closure between them could potentially play a role in quantifying their systematic uncertainties. While further investigation is required to make such a proposal formal, it is our hope that the optimal transport methodology presented in our work can help contribute to the challenging question of systematic uncertainty quantification in the search for di-Higgs boson production or in other searches at the Large Hadron Collider.

SUPPLEMENTARY MATERIAL

Appendices A–E (DOI: 10.1214/24-AOAS1916SUPPA; .pdf). This supplementary material consists of Appendix A, containing a section-by-section summary of this manuscript in nontechnical language, Appendices B–D, containing numerical details deferred from the main text, and Appendix E, containing additional numerical results.

Code for simulation study (DOI: 10.1214/24-AOAS1916SUPPB; .zip). Python code for reproducing the simulation study from Section 6.

REFERENCES

ALISON, J. (2015). The road to discovery: Detector alignment, electron identification, particle misidentification, WW physics, and the discovery of the Higgs boson Ph.D. thesis. Presented 08 Nov 2012.

ALWALL, J., HERQUET, M., MALTONI, F., MATTELAER, O. and STELZER, T. (2011). MadGraph 5: Going beyond. *J. High Energy Phys*. **2011** 128.

ATLAS (2012). Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716** 1–29.

ATLAS (2015). Observation and measurement of Higgs boson decays to WW* with the ATLAS detector. *Phys. Rev. D* **92** 012006.

ATLAS (2018a). Search for pair production of higgsinos in final states with at least three b-tagged jets in 13 TeV pp collisions using the ATLAS detector. *Phys. Rev. D* **98**.

ATLAS (2018b). Observation of $H \to bb$ decays and VH production with the ATLAS detector. *Phys. Lett. B* **786** 59–86.

ATLAS (2018c). Measurements of Higgs boson properties in the diphoton decay channel with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D* **98** 052005.

ATLAS (2018d). Measurements of the Higgs boson production, fiducial and differential cross sections in the 4ℓ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector. ATLAS-CONF-2018-018.

ATLAS (2018e). Measurement of gluon fusion and vector-boson-fusion Higgs boson production cross-sections in the $H \to WW^* \to e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. ATLAS-CONF-2018-004.

ATLAS (2019). Search for pair production of Higgs bosons in the $b\overline{b}b\overline{b}$ final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *J. High Energy Phys*. **2019** 30.

ATLAS (2019b). Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D* **99** 072001.

ATLAS (2021). Search for the HH→ $b\overline{b}b\overline{b}$ process via vector-boson fusion production using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *J. High Energy Phys*. **2021**.

ATLAS (2022). Search for non-resonant pair production of Higgs bosons in the $b\overline{b}b\overline{b}$ final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector Technical Report CERN Geneva. ATLAS-CONF-2022-035.

ATLAS, CMS and HIGGS COMBINATION GROUP (2011). Procedure for the LHC Higgs boson search combination in Summer 2011. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11.

BARLOW, R. (1987). Event classification using weighting methods. *J. Comput. Phys*. **72** 202–219.

BEHNKE, O., KRÖNINGER, K., SCHOTT, G. and SCHÖRNER-SADENIUS, T. (2013). *Data Analysis in High Energy Physics*: *A Practical Guide to Statistical Methods*. Wiley, New York.

BIAU, G. and DEVROYE, L. (2015). *Lectures on the Nearest Neighbor Method. Springer Series in the Data Sciences*. Springer, Cham. MR3445317 https://doi.org/10.1007/978-3-319-25388-6

BORISYAK, M. and KAZEEV, N. (2019). Machine learning on data with sPlot background subtraction. *J. Instrum*. **14** P08020–P08020.

BREHMER, J., LOUPPE, G., PAVEZ, J. and CRANMER, K. (2020). Mining gold from implicit models to improve likelihood-free inference. *Proc. Natl. Acad. Sci. USA* **117** 5242–5249. MR4225023 https://doi.org/10.1073/pnas.1915980117

BRENIER, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math*. **44** 375–417. MR1100809 https://doi.org/10.1002/cpa.3160440402

BRYANT, P. E. (2018). Search for pair production of Higgs bosons in the four bottom quark final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Ph.D. thesis The Univ. Chicago, Chicago, IL.

CACCIARI, M., SALAM, G. P. and SOYEZ, G. (2008). The anti-kt jet clustering algorithm. *J. High Energy Phys*. **2008** 063.

CAI, T., CHENG, J., CRAIG, N. and CRAIG, K. (2020). Linearized optimal transport for collider events. *Phys. Rev. D* **102** 116019.

CHENG, K. F. and CHU, C. K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli* **10** 583–604. MR2076064 https://doi.org/10.3150/bj/1093265631

CHOI, S. and OH, H. (2021). Improved extrapolation methods of data-driven background estimations in high energy physics. *Eur. Phys. J. C* **81** 643.

CMS (2008). The CMS experiment at the CERN LHC. *J. Instrum*. **3** S08004–S08004.

CMS (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B* **716** 30–61.

CMS (2017). Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV.

CMS (2018a). Observation of Higgs boson decay to bottom quarks. *Phys. Rev. Lett*. **121** 121801.

CMS (2018b). Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV. *J. High Energy Phys.* **2018** 185.

CMS (2018c). Measurements of properties of the Higgs boson in the four-lepton final state at $\sqrt{s} = 13$ TeV. CMS PAS HIG-18-001.

CMS (2018d). Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector. *Phys. Lett. B* **779** 283–316.

CMS (2018e). Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. *J. Instrum.* **13** P05011.

CMS (2019). Measurements of properties of the Higgs boson decaying to a W boson pair in pp collisions at $\sqrt{s} = 13$ TeV. *Phys. Lett. B* **791** 96–129.

CMS (2022). Search for Higgs boson pair production in the four b quark final state in proton-proton collisions at $\sqrt{s} = 13$ TeV.

COURTY, N., FLAMARY, R., TUIA, D. and RAKOTOMAMONJY, A. (2016). Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39** 1853–1865.

CRANMER, K., PAVEZ, J. and LOUPPE, G. (2015). Approximating likelihood ratios with calibrated discriminative classifiers. ArXiv Preprint. Available at arXiv:1506.02169.

DE LARA, L., GONZÁLEZ-SANZ, A. and LOUBES, J.-M. (2021). A consistent extension of discrete optimal transport maps for machine learning applications. ArXiv Preprint. Available at arXiv:2102.08644.

DEB, N., GHOSAL, P. and SEN, B. (2021). Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Adv. Neural Inf. Process. Syst.* **34**.

DEMBINSKI, H., KENZIE, M., LANGENBRUCH, C. and SCHMELLING, M. (2022). Custom orthogonal weight functions (COWs) for event classification. *Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip.* 167270.

DI MICCO, B., GOUZEVITCH, M., MAZZITELLI, J. and VERNIERI, C. (2020). Higgs boson potential at colliders: Status and perspectives. *Rev. Phys.* **5** 100045.

ENGLERT, F. and BROUT, R. (1964). Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.* **13** 321–323. MR0174314 https://doi.org/10.1103/PhysRevLett.13.321

FIGALLI, A. (2010). The optimal partial transport problem. *Arch. Ration. Mech. Anal.* **195** 533–560. MR2592287 https://doi.org/10.1007/s00205-008-0212-7

FIX, E. and HODGES, J. L. (1951). Discriminatory analysis. Nonparametric discrimination; consistency properties. Technical Report No. 4, Project No. 21-49-004 USAF School of Aviation Medicine Randolph Field, TX.

FLAMARY, R., COURTY, N., GRAMFORT, A., ALAYA, M. Z., BOISBUNON, A., CHAMBON, S., CHAPEL, L., CORENFLOS, A., FATRAS, K. et al. (2021). POT: Python optimal transport. *J. Mach. Learn. Res.* **22** 1–8.

FORROW, A., HÜTTER, J.-C., NITZAN, M., RIGOLLET, P., SCHIEBINGER, G. and WEED, J. (2019). Statistical optimal transport via factored couplings. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics* 2454–2465.

GHOSAL, P. and SEN, B. (2022). Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *Ann. Statist.* **50** 1012–1037. MR4404927 https://doi.org/10.1214/21-aos2136

GOLDFELD, Z., KATO, K., RIOUX, G. and SADHU, R. (2024). Limit theorems for entropic optimal transport maps and Sinkhorn divergence. *Electron. J. Stat.* **18** 980–1041. MR4718466 https://doi.org/10.1214/24-ejs2217

GONZÁLEZ-SANZ, A., LOUBES, J.-M. and NILES-WEED, J. (2022). Weak limits of entropy regularized optimal transport; potentials, plans and divergences. ArXiv Preprint. Available at arXiv:2207.07427.

GUNSILIUS, F. and XU, Y. (2021). Matching for causal effects via multimarginal optimal transport. ArXiv Preprint. Available at arXiv:2112.04398.

GUNSILIUS, F. F. (2022). On the convergence rate of potentials of Brenier maps. *Econometric Theory* **38** 381–417. MR4407051 https://doi.org/10.1017/S0266466621000037

HAGEDORN, R. (1963). *Relativistic Kinematics: A Guide to the Kinematic Problems of High-Energy Physics*. W. A. Benjamin, New York-Amsterdam. MR0158658

HANLEY, J. A. and MCNEIL, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29–36.

HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.

HEINRICH, J. and LYONS, L. (2007). Systematic errors. *Annu. Rev. Nucl. Part. Sci.* **57** 145–169.

HIGGS, P. W. (1964). Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.* **13** 508–509. MR0175554 https://doi.org/10.1103/PhysRevLett.13.508

HO, N., HUYNH, V., PHUNG, D. and JORDAN, M. (2019). Probabilistic multilevel clustering via composite transportation distance. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics* 3149–3157. PMLR.

HO, N., NGUYEN, X., YUROCHKIN, M., BUI, H. H., HUYNH, V. and PHUNG, D. (2017). Multilevel clustering via Wasserstein means. In *Proceedings of the* 34*th International Conference on Machine Learning* 1501–1509. PMLR.

HÜTTER, J.-C. and RIGOLLET, P. (2021). Minimax estimation of smooth optimal transport maps. *Ann. Statist.* **49** 1166–1194. MR4255123 https://doi.org/10.1214/20-aos1997

HUYNH, V., HO, N., DAM, N., NGUYEN, X., YUROCHKIN, M., BUI, H. and PHUNG, D. (2021). On efficient multilevel clustering via Wasserstein distances. *J. Mach. Learn. Res.* **22** 145. MR4318501

KANTOROVICH, L. V. (1948). On a problem of Monge. In *CR* (*Doklady*) *Acad. Sci. URSS* (*NS*) **3** 225–226.

KANTOROVICH, L. V. (1942). On the translocation of masses. *C. R.* (*Dokl.*) *Acad. Sci. URSS* **37** 199–201. MR0009619

KASIECZKA, G., NACHMAN, B., SCHWARTZ, M. D. and SHIH, D. (2021). Automating the ABCD method with machine learning. *Phys. Rev. D* **103** 035021. MR4225448 https://doi.org/10.1103/physrevd.103.035021

KLATT, M., TAMELING, C. and MUNK, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM J. Math. Data Sci.* **2** 419–443. MR4105566 https://doi.org/10.1137/19M1278788

KNOTT, M. and SMITH, C. S. (1984). On the optimal mapping of distributions. *J. Optim. Theory Appl.* **43** 39–49. MR0745785 https://doi.org/10.1007/BF00934745

KOLOURI, S., PARK, S. R., THORPE, M., SLEPCEV, D. and ROHDE, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Process. Mag.* **34** 43–59.

KOMISKE, P. T., MASTANDREA, R., METODIEV, E. M., NAIK, P. and THALER, J. (2020). Exploring the space of jets with CMS open data. *Phys. Rev. D* **101** 034009.

KOMISKE, P. T., METODIEV, E. M. and THALER, J. (2019). Metric space of collider events. *Phys. Rev. Lett.* **123** 041801. https://doi.org/10.1103/PhysRevLett.123.041801

KOMISKE, P. T., METODIEV, E. M. and THALER, J. (2020). The hidden geometry of particle collisions. *J. High Energy Phys.* **7** 6. MR4138090 https://doi.org/10.1007/jhep07(2020)006

KOMISKE, P. T., METODIEV, E. M. and THALER, J. (2022). EnergyFlow Python package. https://energyflow. network/.

KPOTUFE, S. (2017). Lipschitz density-ratios, structured data, and data-driven tuning. In *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics* 1320–1328. PMLR.

KRAUSE, C. and SHIH, D. (2023a). Fast and accurate simulations of calorimeter showers with normalizing flows. *Phys. Rev. D* **107** 113003.

KRAUSE, C. and SHIH, D. (2023b). Accelerating accurate simulations of calorimeter showers with normalizing flows and probability density distillation. *Phys. Rev. D* **107** 113004.

KUHN, H. W. (1955). The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2** 83–97. MR0075510 https://doi.org/10.1002/nav.3800020109

LEE, J., DABAGIA, M., DYER, E. and ROZELL, C. (2019). Hierarchical optimal transport for multimodal distribution alignment. In *Advances in Neural Information Processing Systems* 13453–13463.

LIERO, M., MIELKE, A. and SAVARÉ, G. (2018). Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Invent. Math.* **211** 969–1117. MR3763404 https://doi.org/10. 1007/s00222-017-0759-8

LYONS, L. (1986). *Statistics for Nuclear and Particle Physicists*. Cambridge Univ. Press, Cambridge.

MAKKUVA, A. V., TAGHVAEI, A., OH, S. and LEE, J. D. (2020). Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning* 6672–6681. PMLR.

MANOLE, T., BALAKRISHNAN, S., NILES-WEED, J. and WASSERMAN, L. (2024). Plugin estimation of smooth optimal transport maps. *Ann. Statist.* **52** 966–998. https://doi.org/10.1214/24-AOS2379

MANOLE, T., BRYANT, P., ALISON, J., KUUSELA, M. and WASSERMAN, L. (2024). Supplement to "Background modeling for double Higgs boson production: Density ratios and optimal transport." https://doi.org/10. 1214/24-AOAS1916SUPPA, https://doi.org/10.1214/24-AOAS1916SUPPB

MONGE, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Hist. Acad. Roy. Sci. Paris*.

NATH, J. S. and JAWANPURIA, P. (2020). Statistical optimal transport posed as learning kernel mean embedding. In *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*, *NIPS '20*, *Vancouver*, *BC*, *Canada*. Curran Associates, Red Hook, NY, USA.

PANARETOS, V. M. and ZEMEL, Y. (2019). Statistical aspects of Wasserstein distances. *Annu. Rev. Stat. Appl.* **6** 405–431. MR3939527 https://doi.org/10.1146/annurev-statistics-030718-104938

PANARETOS, V. M. and ZEMEL, Y. (2019b). *An Invitation to Statistics in Wasserstein Space*. Springer, Berlin.

PELE, O. and WERMAN, M. (2008). A linear time histogram metric for improved SIFT matching. In *European Conference on Computer Vision* 495–508. Springer, Berlin.

PELEG, S., WERMAN, M. and ROM, H. (1989). A unified approach to the change of resolution: Space and gray-level. *IEEE Trans. Pattern Anal. Mach. Intell.* **11** 739–742.

PERROT, M., COURTY, N., FLAMARY, R. and HABRARD, A. (2016). Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems* 29 (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, eds.) 4197–4205. Curran Associates, Red Hook.

PEYRÉ, G. and CUTURI, M. (2019). Computational optimal transport. *Found. Trends Mach. Learn.* **11** 355–607.

PIVK, M. and LE DIBERDER, F. (2005). SPlot: A statistical tool to unfold data distributions. *Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip.* **555** 356–369.

PLACAKYTE, R. (2011). Parton distribution functions. ArXiv Preprint. Available at arXiv:1111.5452.

POLLARD, C. and WINDISCHHOFER, P. (2022). Transport away your problems: Calibrating stochastic simulations with optimal transport. *Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip.* **1027** 166119.

POOLADIAN, A.-A. and NILES-WEED, J. (2021). Entropic estimation of optimal transport maps. ArXiv Preprint. Available at arXiv:2109.12004.

QIN, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* **85** 619–630. MR1665814 https://doi.org/10.1093/biomet/85.3.619

RAKOTOMAMONJY, A., FLAMARY, R., GASSO, G., EL ALAYA, M., BERAR, M. and COURTY, N. (2022). Optimal transport for conditional domain matching and label shift. *Mach. Learn.* **111** 1651–1670. MR4426353 https://doi.org/10.1007/s10994-021-06088-2

RAMDAS, A., GARCÍA TRILLOS, N. and CUTURI, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19** 47. MR3608466 https://doi.org/10.3390/e19020047

READ, A. L. (1999). Linear interpolation of histograms. *Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip.* **425** 357–360.

REDKO, I., HABRARD, A. and SEBBAN, M. (2017). Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 737–753. Springer, Berlin.

REISS, R. D. (2012). *A Course on Point Processes*. Springer, Berlin.

RIPPL, T., MUNK, A. and STURM, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivariate Anal.* **151** 90–109. MR3545279 https://doi.org/10.1016/j.jmva.2016.06.005

RUBNER, Y., TOMASI, C. and GUIBAS, L. J. (2000). The Earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **40** 99–121.

SAKUMA, T. and MCCAULEY, T. (2014). Detector and event visualization with SketchUp at the CMS experiment. In *Journal of Physics: Conference Series* **513** 022032. IOP Publishing, Bristol.

SILVERMAN, B. W. and JONES, M. C. (1989). E. Fix and J.L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on Fix and Hodges (1951). *Int. Stat. Rev.* 233–238.

VILLANI, C. (2003). *Topics in Optimal Transportation. Graduate Studies in Mathematics* **58**. Amer. Math. Soc., Providence, RI. MR1964483 https://doi.org/10.1090/gsm/058

VILLANI, C. (2009). *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *J. Big Data* **3** 1–40.

YOSINSKI, J., CLUNE, J., BENGIO, Y. and LIPSON, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* **27**. Curran Associates, Red Hook.