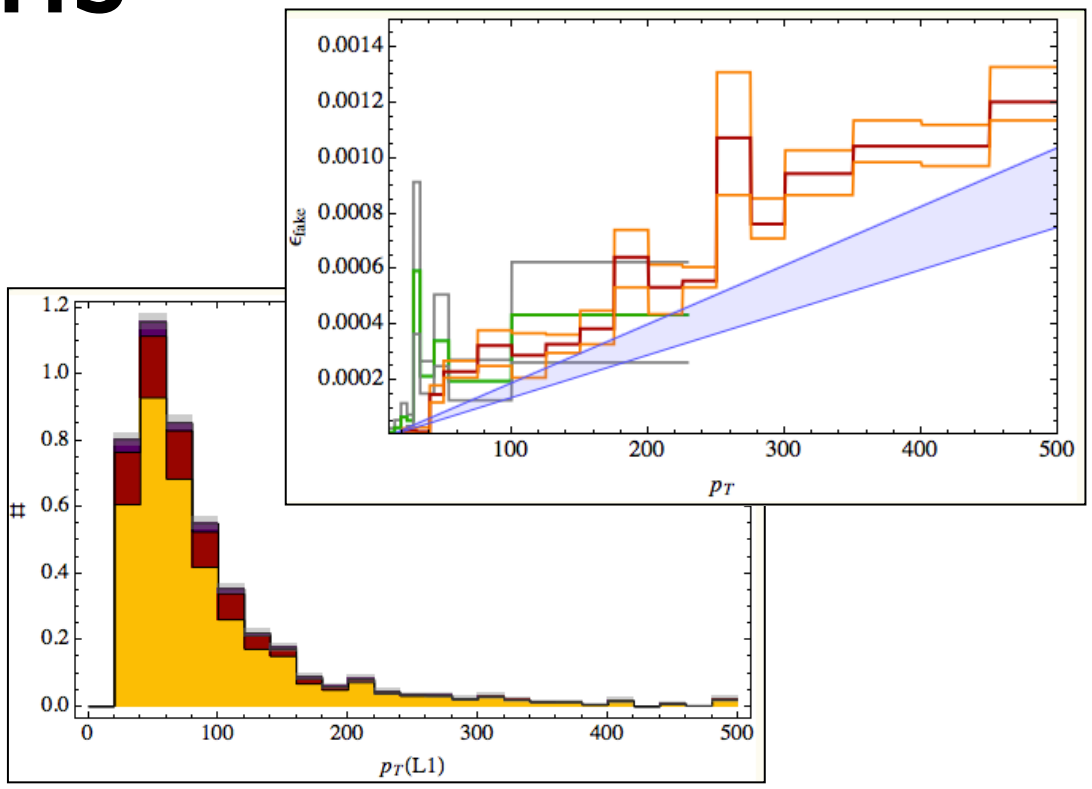
# Efficient Simulation of Fake Leptons




MC4BSM 2014 Workshop
Daejon, South Korea

23 May 2014

David Curtin
Yang Institute for Theoretical Physics
Stony Brook

Based on

**1306.5695** DC, Jamison Galloway, Jay Wacker

# Motivation

# Definition

A "fake" lepton is an object which is reconstructed as an isolated lepton in your detector, but for some reason it is not the kind of lepton you are interested in.

Lepton Fakes are much rarer than e.g. b-mistags, but they can still be the dominant backgrounds for certain "golden channels".

It is this rarity which makes them difficult to study & simulate.

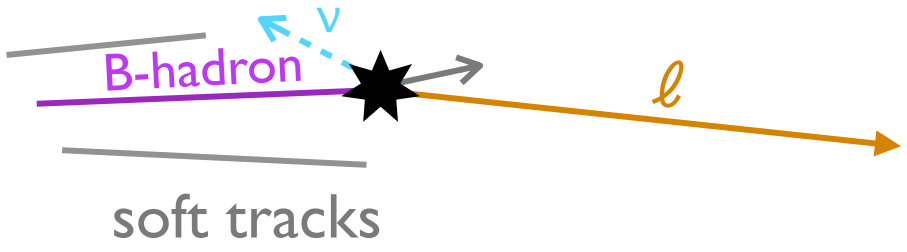
# Where do Fake Leptons come from?

Irreducible (prompt) fakes:

~ all fake $\mu$, half fake e

A real (but "uninteresting") Lepton is produced inside a jet, accidentally passes isolation condition.

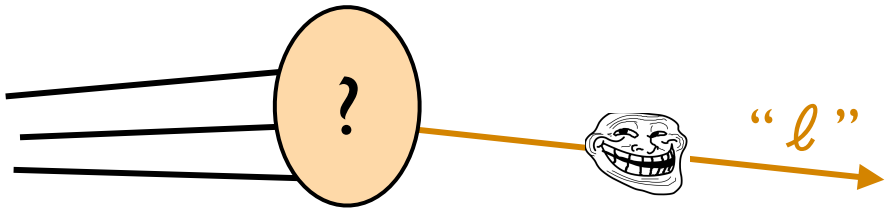
e.g. b-quark accidentally doesn't shower much

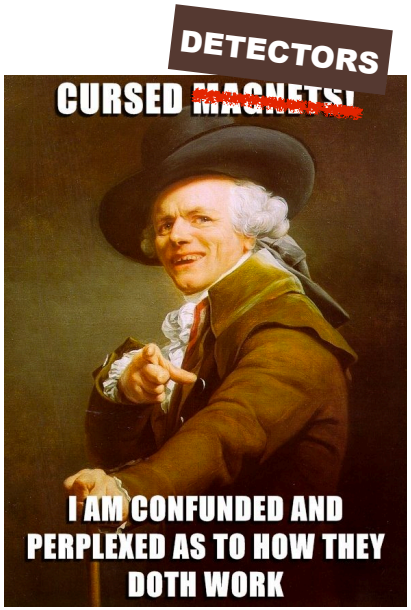



lepton accidentally gets most of pT

"Reducible" (non-prompt) fakes:

Something fakes a lepton in detector: pion, thin jet, detector blip...



# Why study Lepton Fakes?

- Fake Leptons are important backgrounds for otherwise "golden" channels.

  e.g. Same-Sign Dileptons. Very clean, fakes dominate BG.

  *We broke the one-topic-per-paper rule and introduced our Lepton FakeSim in 1306.5695, which explored how to measure tth coupling from SSDL measurements. Let's focus on the FakeSim here....*

- Similarly for photon fakes... and other kinds of rare detector "mis-measurement" (e.g. charge-flips... lepton jets... ??)
- When e.g. SSDL searches become systematics-limited mid-run-II this will become a limiting factor for BSM sensitivity!

# Estimating Fake Lepton BG

If you're a theorist....

- First, identify the 'source processes' which generate your fake lepton backgrounds. Depends on analysis cuts.

  e.g. for SSDL+2b, most lepton fakes come from $t\bar{t}$ & Wbbj

# Estimating Fake Lepton BG

If you're a theorist....

- First, identify the 'source processes' which generate your fake lepton backgrounds. Depends on analysis cuts.

  e.g. for SSDL+2b, most lepton fakes come from $t\bar{t}$ & Wbbj

- Generate samples of your source processes, and...

1. ... multiply your cross section by ~ $10^{-4}$

   efficient, but a little crude

# Estimating Fake Lepton BG

If you're a theorist....

- First, identify the 'source processes' which generate your fake lepton backgrounds. Depends on analysis cuts.

  e.g. for SSDL+2b, most lepton fakes come from $t\bar{t}$ & Wbbj

- Generate samples of your source processes, and...

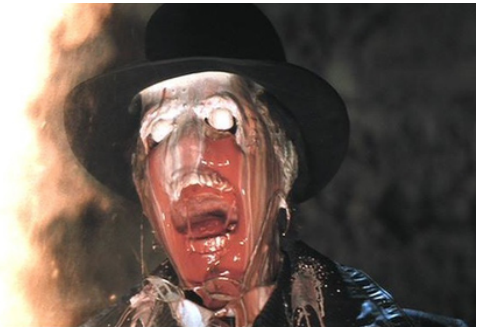
1. ... multiply your cross section by ~ $10^{-4}$

   efficient, but a little crude

2. .. or run > O(10 billion) events through PGS/Delphes

   (a bunch of times if you want to tune)

   "oh god no please make it stop"



# Estimating Fake Lepton BG

If you're an experimentalist....

- Big data-driven input, since fake lepton simulation is very under-developed.

1205.3933

- Look at a particular example: CMS SSDL+2b search

(also used in many other analyses)

# Estimating Fake Lepton BG

CMS SSDL+2b 1205.3933

1. Collect a fake-lepton enriched sample.
   *In this case single loosely-ID'ed lepton + single far-away jet (with W & Z vetoes)*

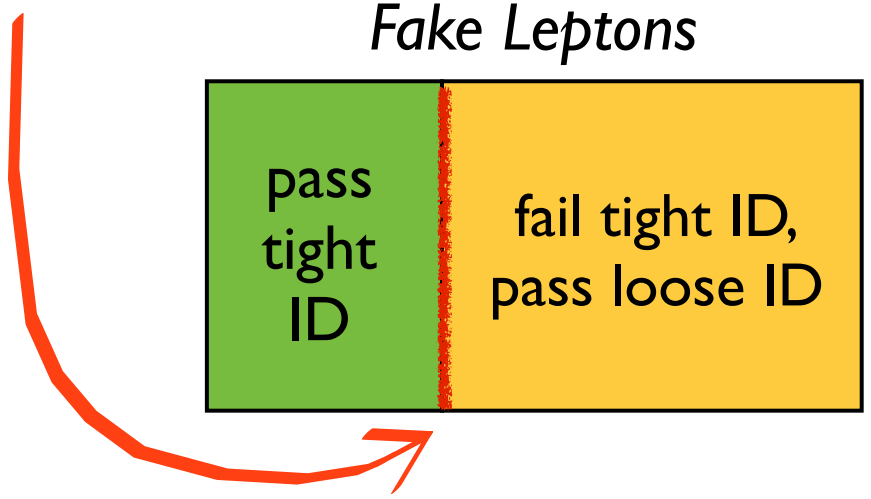
# Estimating Fake Lepton BG

CMS SSDL+2b 1205.3933

1. Collect a fake-lepton enriched sample.
   *In this case single loosely-ID'ed lepton + single far-away jet (with W & Z vetoes)*

2. For this sample, compute the following "T/L ratio"

$$\frac{\text{(\# fake lepton events which pass tight ID)}}{\text{(\# fake lepton events which fail tight ID but pass loose ID)}}$$



# Estimating Fake Lepton BG

CMS SSDL+2b 1205.3933

3. Apply analysis cuts (two SS leptons, 2 b-jets, ...) with tight lepton ID and obtain events in search signal region

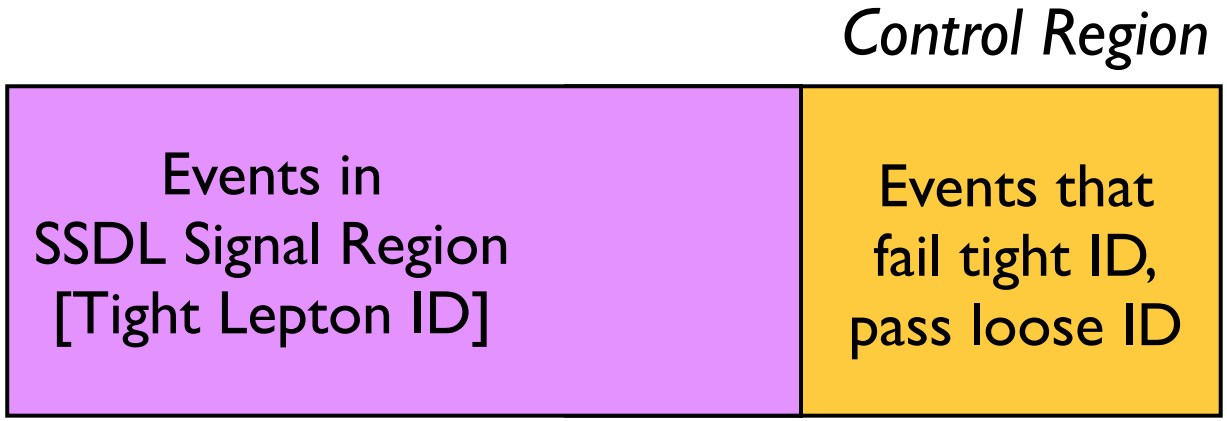
Events in
SSDL Signal Region
[Tight Lepton ID]

# Estimating Fake Lepton BG

CMS SSDL+2b 1205.3933

3. Apply analysis cuts (two SS leptons, 2 b-jets, ...) with tight lepton ID and obtain events in search signal region

4. Define control region with equivalent cuts, but requiring failed TIGHT ID, passed LOOSE ID.



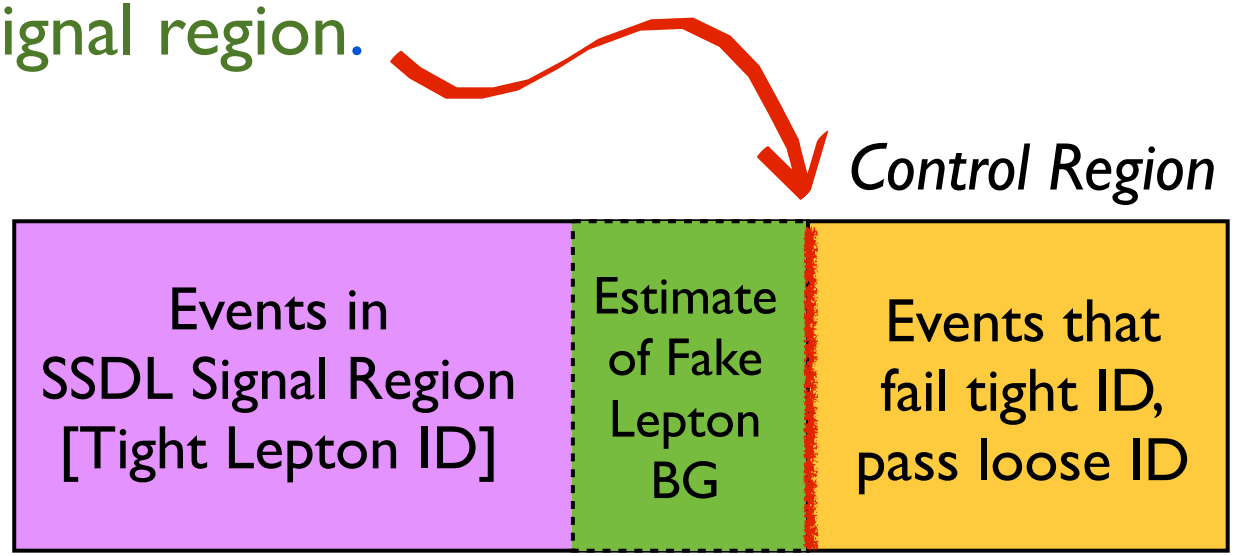
# Estimating Fake Lepton BG

CMS SSDL+2b 1205.3933

3. Apply analysis cuts (two SS leptons, 2 b-jets, ...) with tight lepton ID and obtain events in search signal region

4. Define control region with equivalent cuts, but requiring failed TIGHT ID, passed LOOSE ID.

5. Apply T/L measurement, get estimate of fake lepton contamination in signal region.



# Estimating Fake Lepton BG

If you're an experimentalist....

Problem:

T/L ratio changes between different event samples, and also depends on pT thresholds & b-tags

⇛ 50% systematic error

This can be reduced with extra work, BUT has basic short-coming:

entirely data-driven method ignores **detailed knowledge** we have about fake lepton SOURCE processes!

# Estimating Fake Lepton BG

A better way?

How can we take all the things we DO understand very well from Monte-Carlo & Theory

(kinematics of the fake lepton source processes)

and carry all that into the fake lepton BG estimation, while parameterizing the "last step"

(the way a jet/anything else actually fakes a lepton)

in a data-driven way?

# Building a

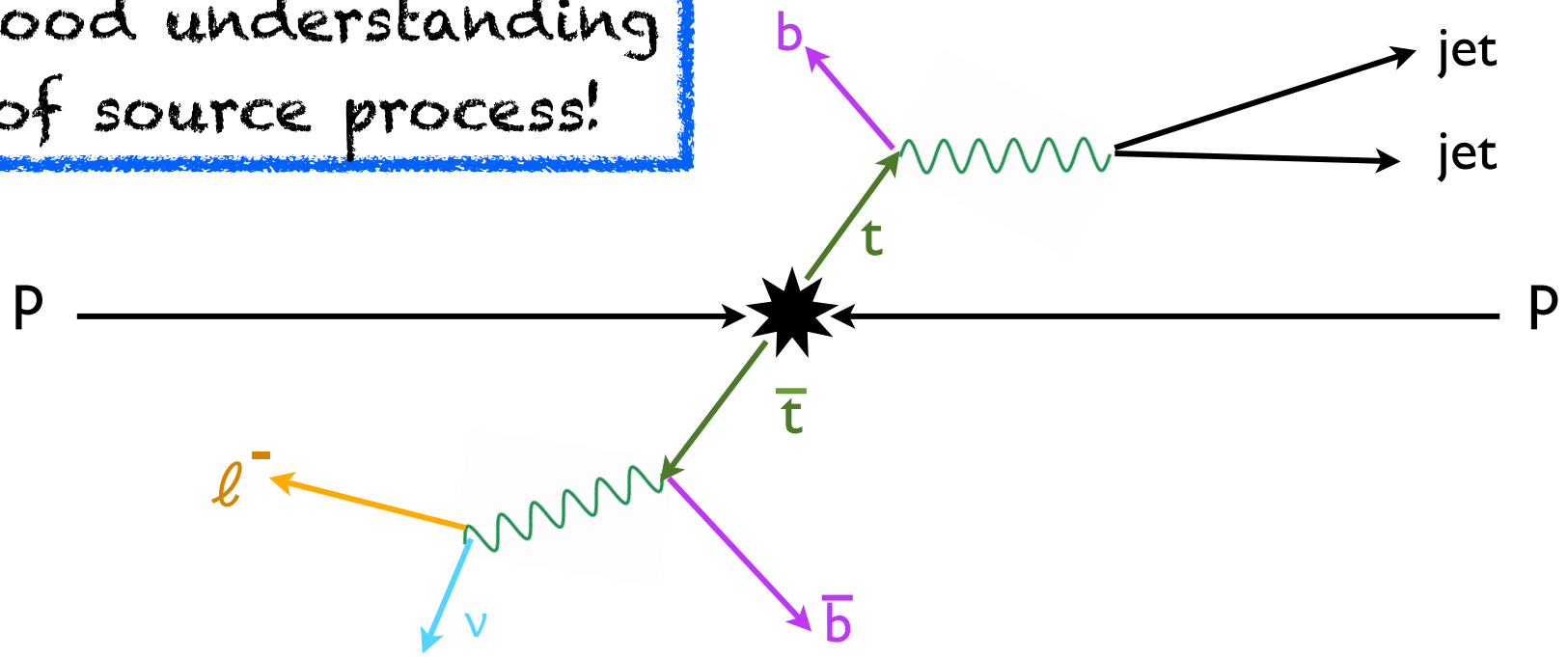
# Fake Lepton Simulator

# Combining MC & Data-Driven

For a given analysis there are some dominant "source processes" which produce most of the fake lepton background.

e.g. semi-leptonic $t\bar{t}$ for SSDL+2b

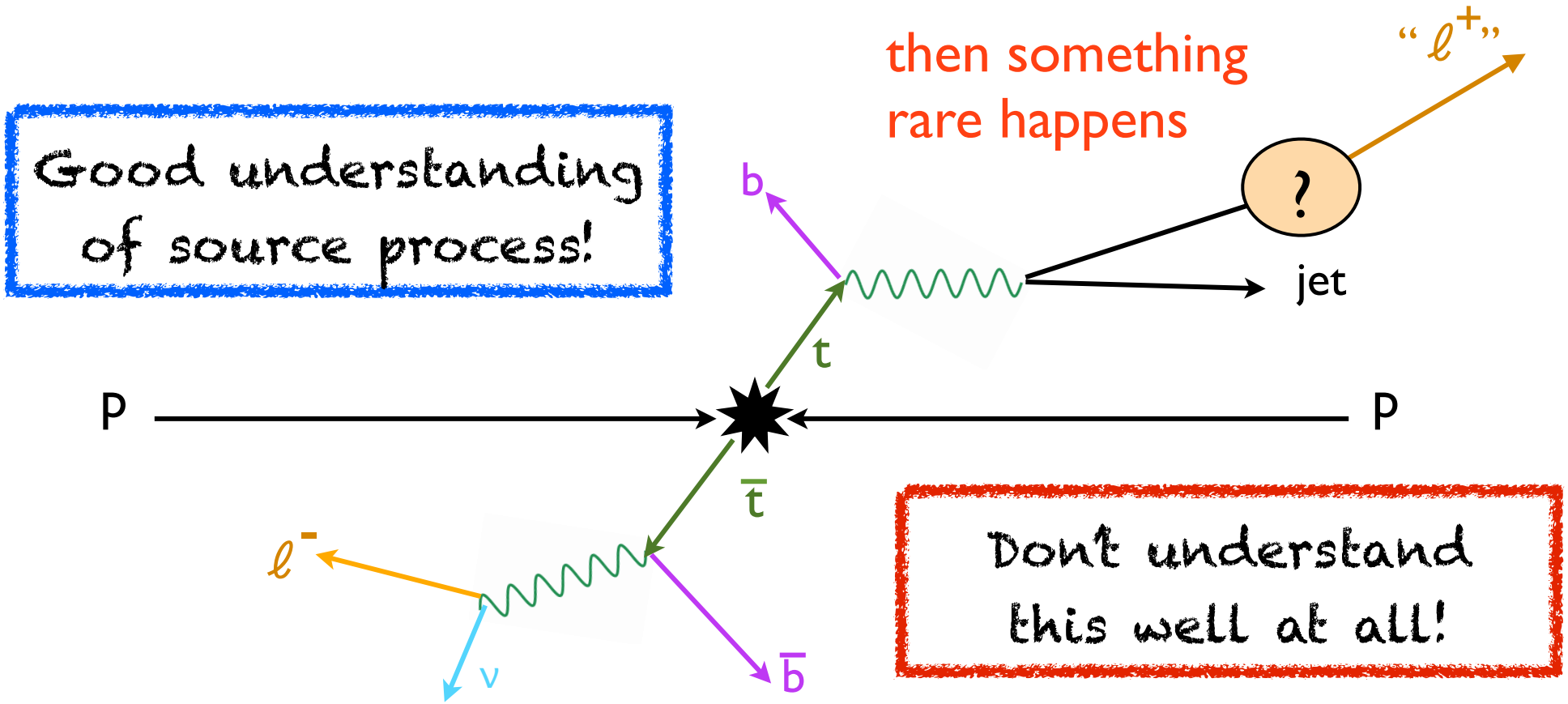
Good understanding of source process!



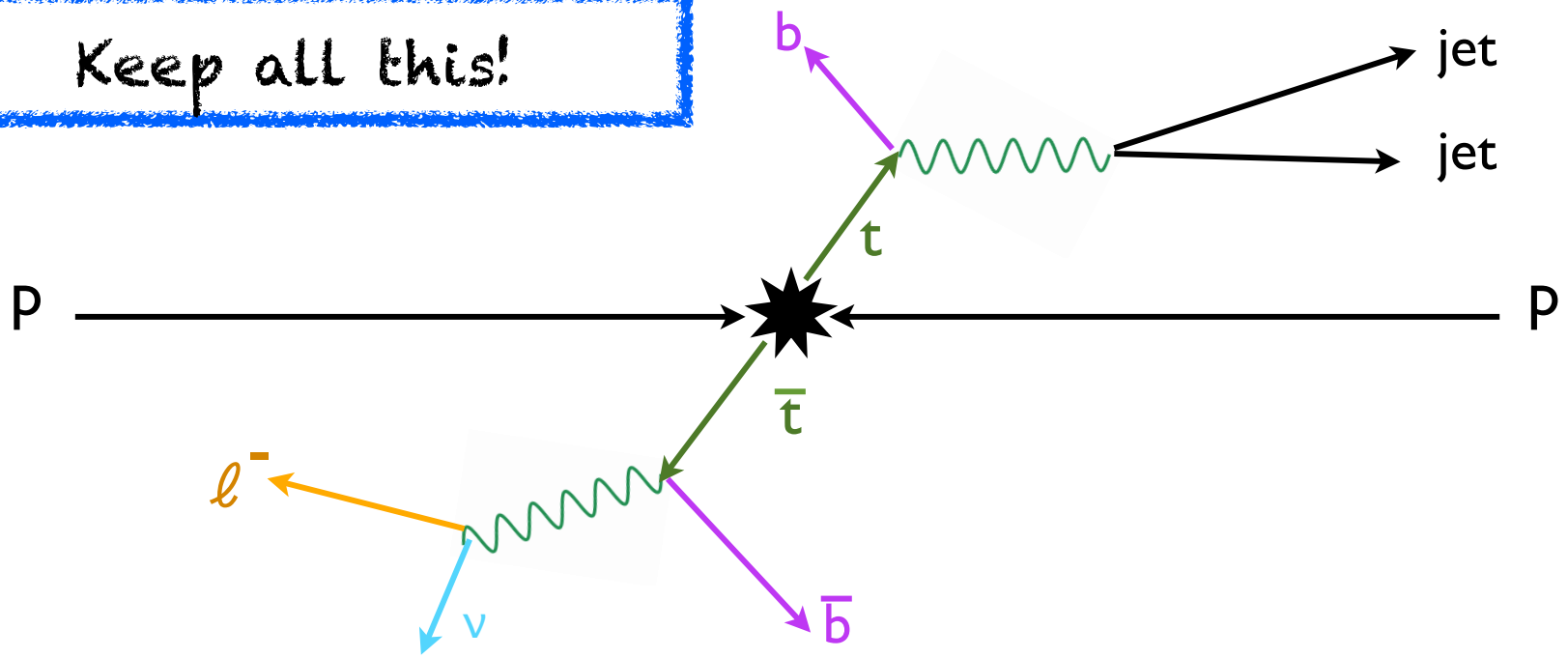
# Combining MC & Data-Driven

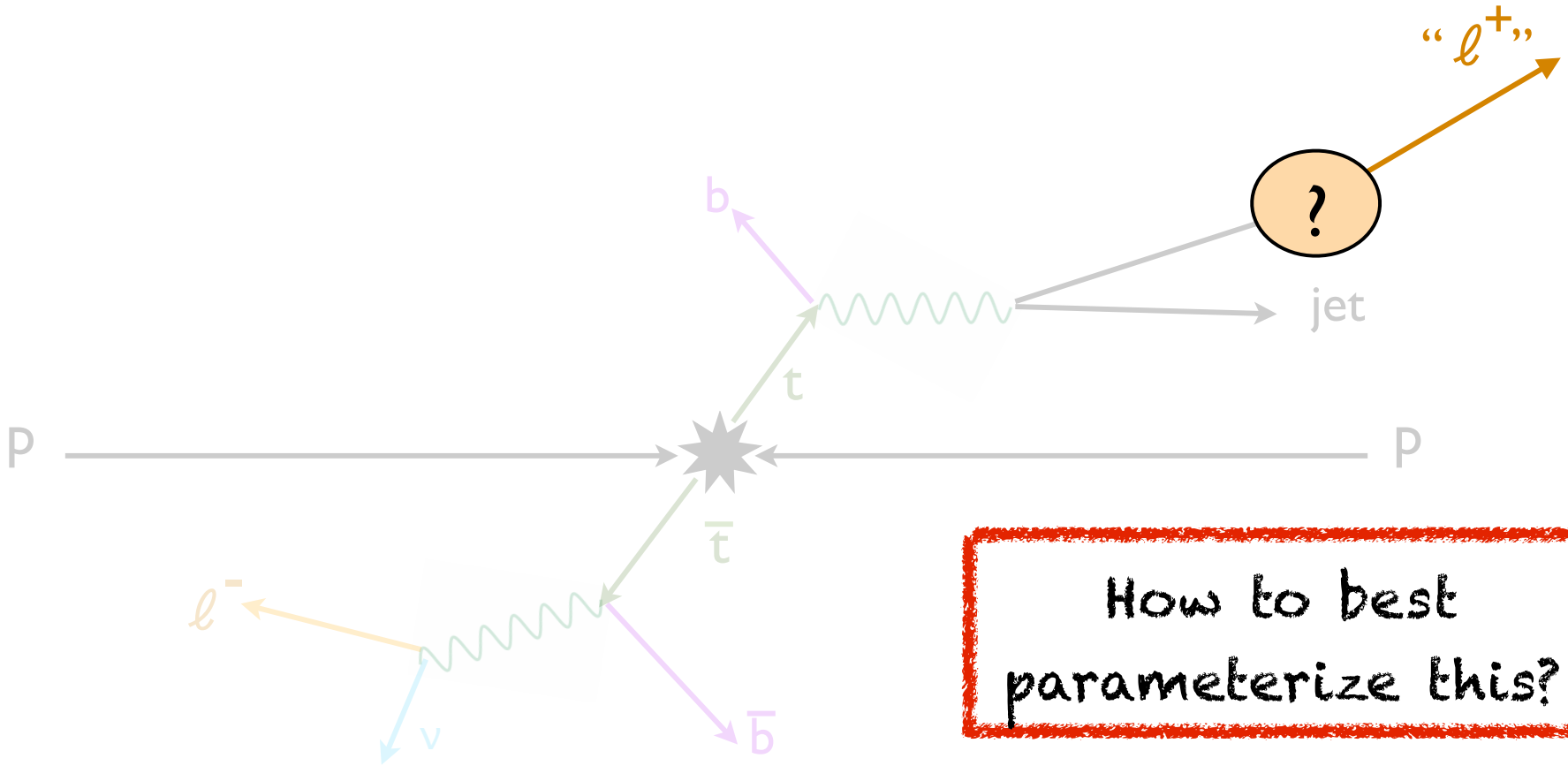
For a given analysis there are some dominant "source processes" which produce most of the fake lepton background.

e.g. semi-leptonic $t\bar{t}$ for SSDL+2b



Keep all this!






How to best parameterize this?

# Parameterizing the "Faking"

We'd like the minimal parameterization for the 'faking' process.

How often does it happen? Define a FAKE RATE:

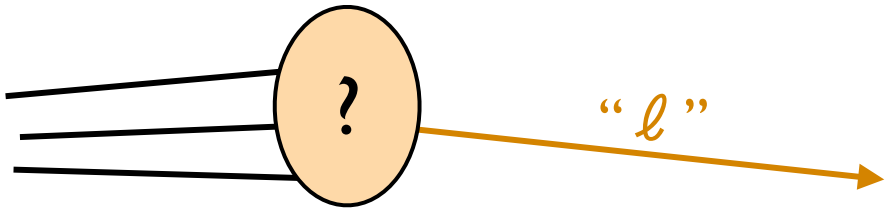
$$\epsilon_{j\to e} = \epsilon_{j\to e}(j_{\text{orig}})$$

Depends on 'source' object properties, e.g. pT

What are the properties of the fake lepton? Define a TRANSFER FUNCTION:

$$\mathcal{T}_{j\to e} = \mathcal{T}_{j\to e}(j_{\text{orig}}; e_{\text{fake}})$$

A PDF for fake lepton properties (e.g. pT), but the PDF can depend on source object.



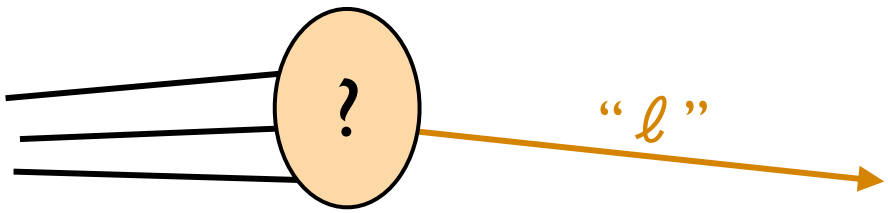

Note that all these depend on lepton & jet reconstruction criteria

# Parameterizing the "Faking"

Simplified Functional Ansatz:

For central fakes, just make this a LINEAR function of pT.

$$\epsilon_{j\to e} = \epsilon_{j\to e}(j_{\text{orig}})$$



# Parameterizing the "Faking"

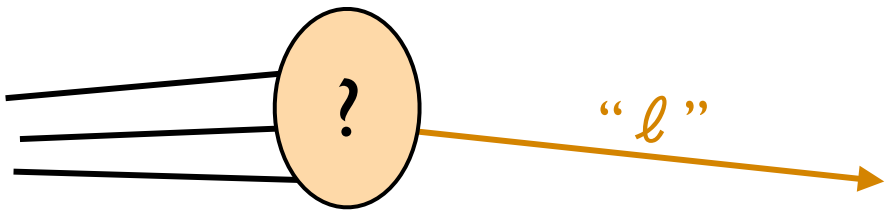
Simplified Functional Ansatz:

For central fakes, just make this a LINEAR function of pT.

$$\epsilon_{j\to e} = \epsilon_{j\to e}(j_{\rm orig})$$

Assume lepton fakes are collinear with "what would have been the original jet". Transfer function is a jet-universal Gaussian for pT-fraction α that is "lost" in converting jet to a fake lepton

$$\mathcal{T}_{j\to e} = \mathcal{T}_{j\to e}(j_{\rm orig}; e_{\rm fake}) = \mathcal{T}_{j\to \ell}(\alpha)$$



# Parameterizing the "Faking"

Simplified Functional Ansatz:

HAS TO BE VERIFIED BY DATA!!

For central fakes, just make this a LINEAR function of pT.

$$\epsilon_{j\to e} = \epsilon_{j\to e}(j_{\text{orig}})$$

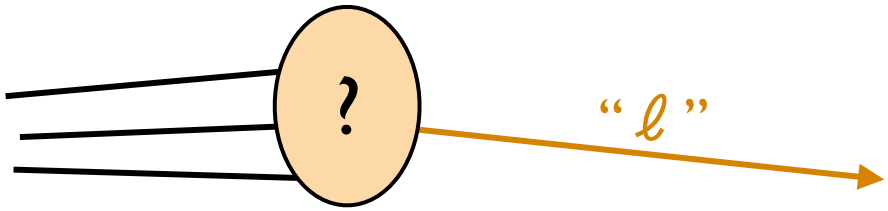
gradient, normalization

Assume lepton fakes are collinear with "what would have been the original jet". Transfer function is a jet-universal Gaussian for pT-fraction α that is "lost" in converting jet to a fake lepton

mean, variance

$$\mathcal{T}_{j\to e} = \mathcal{T}_{j\to e}(j_{\text{orig}}; e_{\text{fake}}) = \mathcal{T}_{j\to \ell}(\alpha)$$

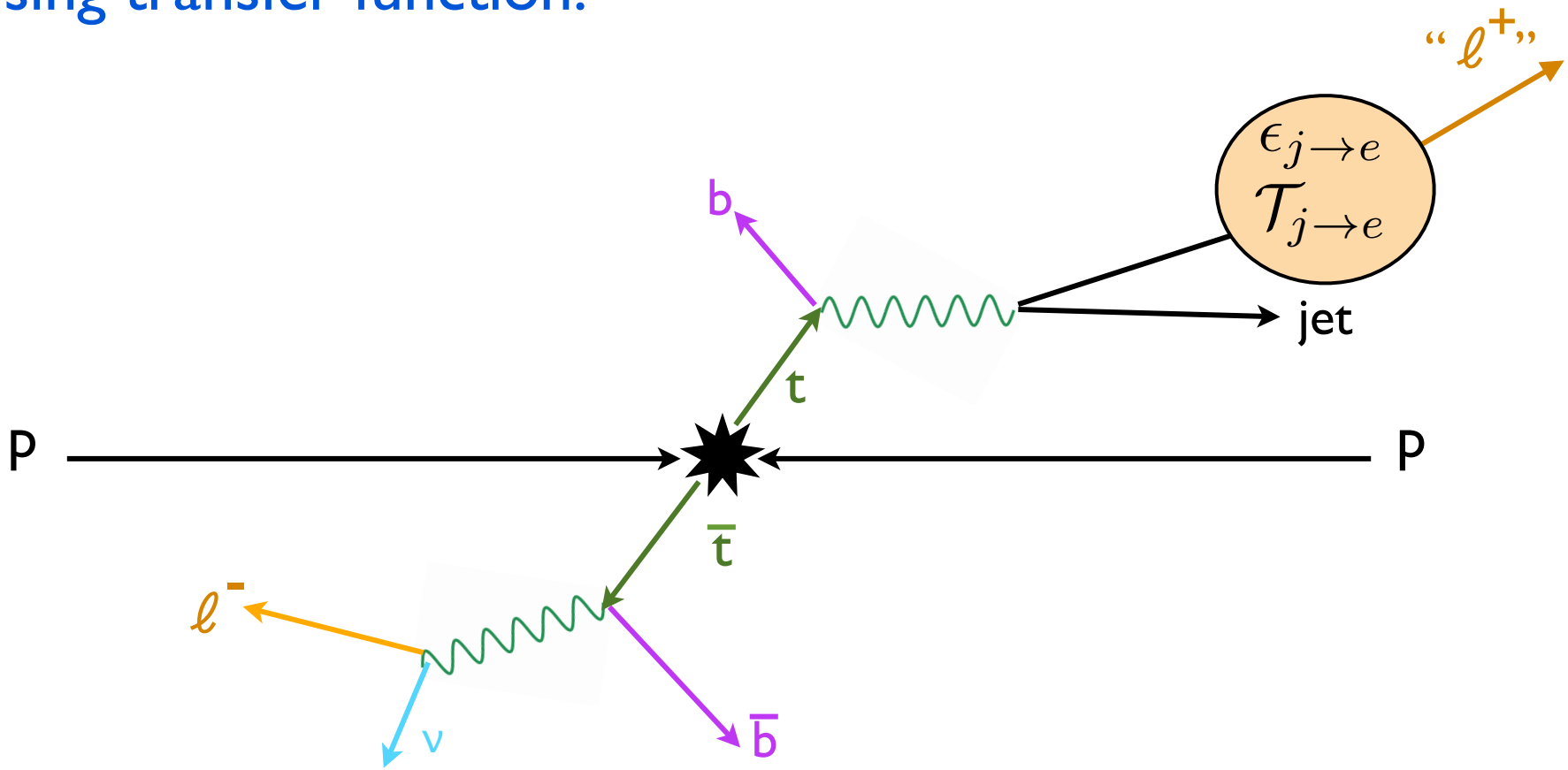
4 parameters that can be tuned to data




(can of course make this more elaborate/realistic, but this captures important physics)

# Implementation of Fast FakeSim

1. Produce source processes at detector-level
2. Reweigh event by fake rate (easy to get large fake lepton samples)
3. Transform jet (or whatever) into lepton using transfer function.



# Fast Fake Lepton Simulator

- This is extremely computationally efficient, generally only need O(10k) events for statistically healthy fake lepton LHC sample.

  *Can increase computational efficiency further by transforming each source event into many weighted fake lepton events to cover whole space of possible fake lepton configurations.*

- Still get correct modification of fake lepton kinematics compared to unmodified source process.

- Fake Rate & Transfer Functions must be tuned to data!

  - EXPERIMENTALISTS: could this be provided? Then theorists could make reasonable fake lepton collider studies.

  - At any rate we can tune to available data & try to extrapolate.

- "Platonic ideal" of FakeSim: Matrix of $\epsilon$, $\mathcal{T}$ for all X faking Y

# Example:

# SSDL + 2b Analysis

# Lepton FakeSim for SSDL+2b

1205.3933, 1212.6194

- Wanted to simulate fake lepton BG for CMS SSDL + 2b analyses @ LHC7, LHC8 to optimize for tth sensitivity.
- For the SSDL+2b final state, we assumed most of the fake leptons come from semileptonic ttbar, leptonic Wbbj, and dileptonic ttbar + j.
- Since final state (SSDL) was not flavor-selected we lumped electrons and muons together → 4 unknown parameters.
- To be able to simulate fake leptons we tuned the FakeSim to reproduce the CMS estimates of fake lepton BG (obtained from T/L ratio method).

# Lepton FakeSim for SSDL+2b

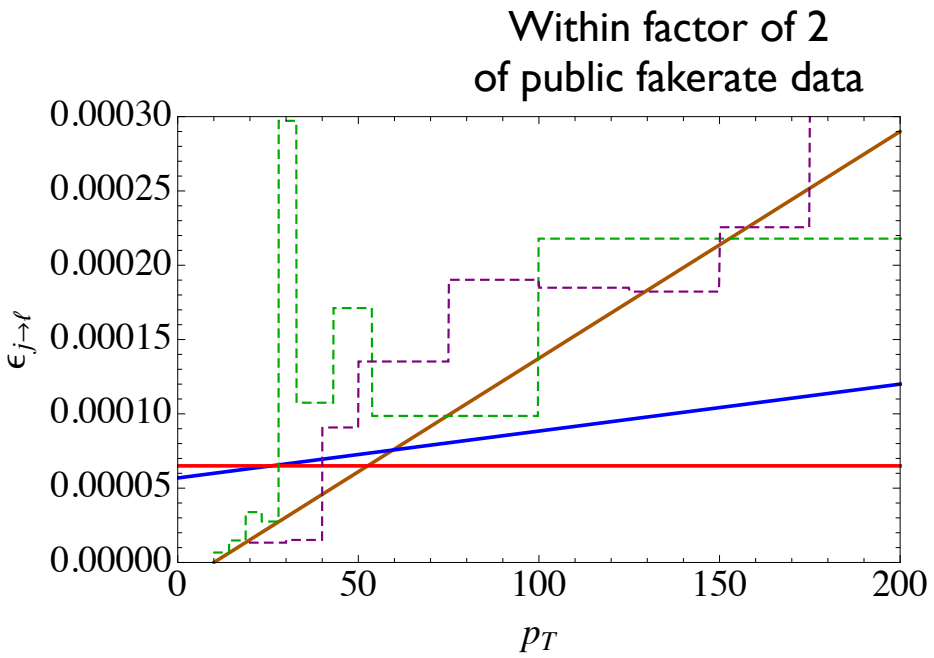
- Interestingly, the CMS fake BG estimates UNDER-constrained the fake rate & transfer functions.

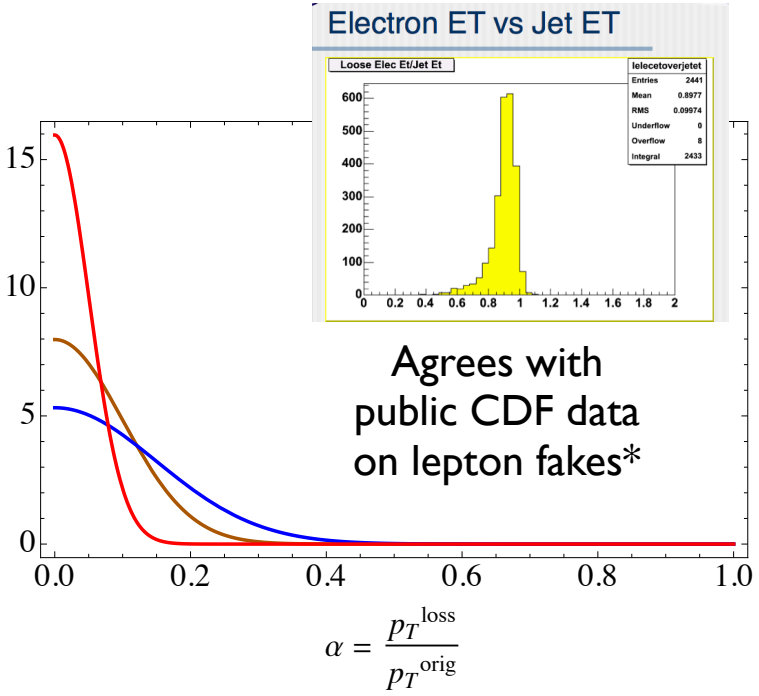
Here are 3 representative best-fit

Efficiency Curves as functions of pT

(like with like)

Transfer Functions (fraction of lost pT)



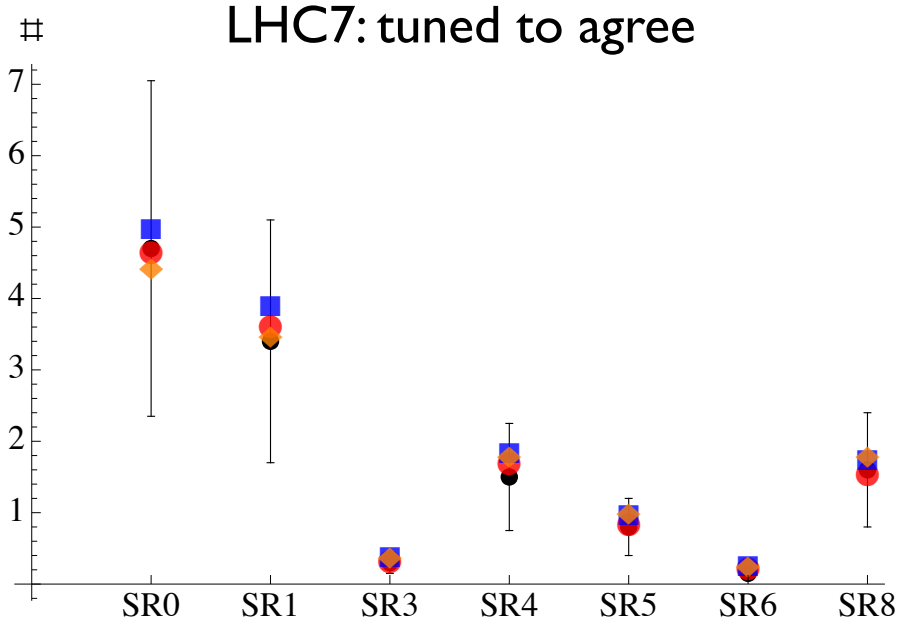

Electron fake rates and trigger efficiency, Mike Flowerdew, U Liverpool

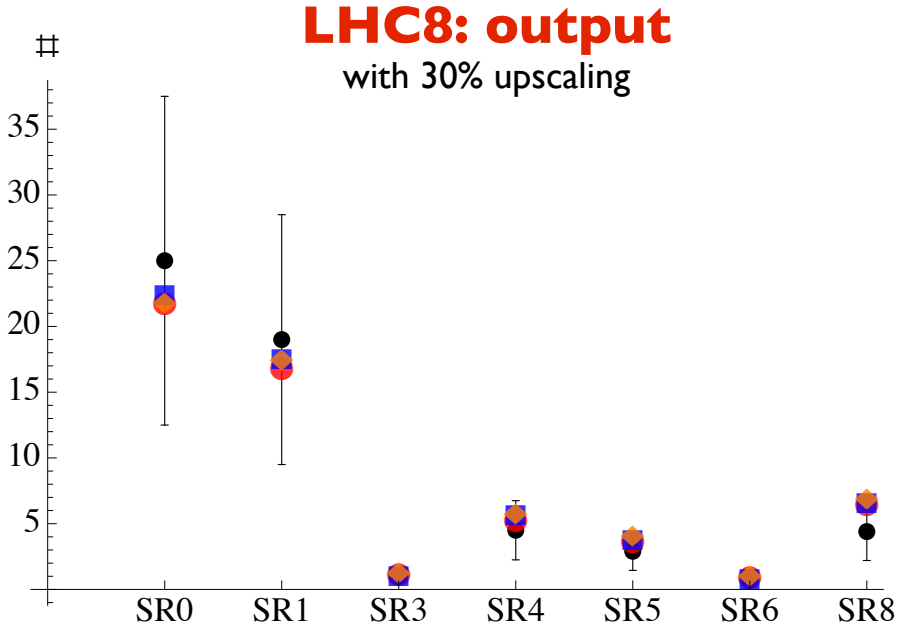
Electron Detection at CMS, Jeffrey Berryhill (FNAL), Aug 3 2009

*Martin Griffiths, Giulia Manca, Beate Heinemann, "Studying the fake lepton rate for the trilepton analy- sis", Seminar given on behalf of the CDF collaboration.

# Lepton FakeSim for SSDL+2b

- Is this underconstrained tune a bad thing? Well...

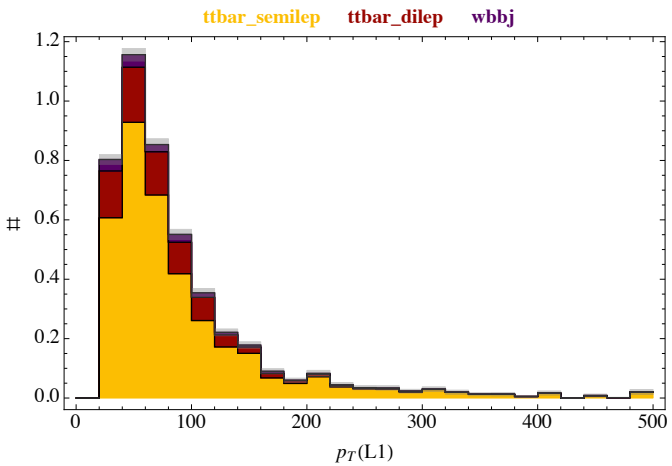


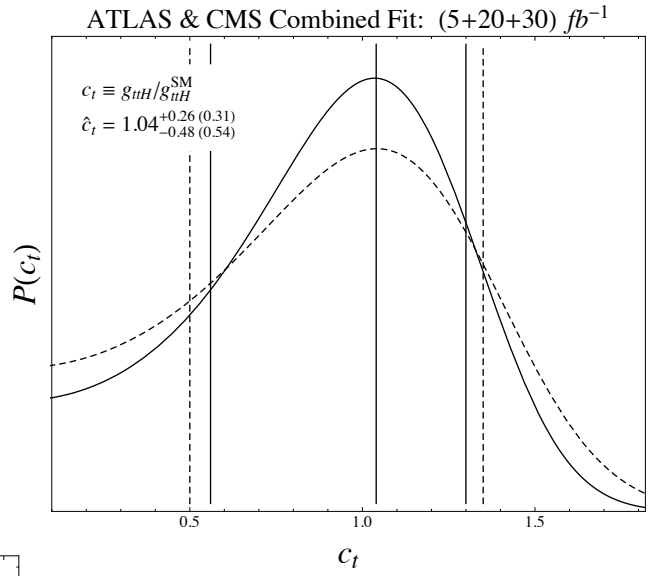



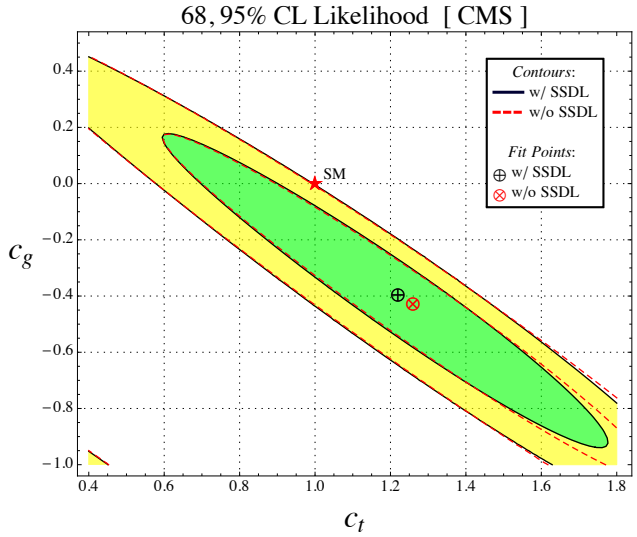

- This means the most important physics is captured by the MC simulation of the source processes, NOT the particulars of the FakeSim parameterization.

# Lepton FakeSim for SSDL+2b

- This tuned Lepton FakeSim then allowed us to optimize the search for tth production and get estimate of coupling sensitivity...




(btw, it's totally worth doing you guys!)

# Applications

# &

# Possible Future Directions

# Applications & Future Directions

- This Fast FakeSim method can be easily adapted by theorists to conduct relatively accurate collider studies with fake BG samples.

  → Cut efficiencies dominated by well-understood MC simulation of source process.

- Generalizes to Fake Photons etc...

- Experimentalists:
  We'd love to see more calibration data for this!

# Experimental Usage?

- This method combines maximal information from well-understood MC with minimal parameterization of the 'faking' process.
- It is possible to study (derive?) the FakeSim parameterization from QCD / hadron physics first principles (shower probabilities, fragmentation functions, hadron decay kinematics).

  Learn more about QCD?
- Experimental application would require *a lot* more work, but the result would be a 'fast fake' simulator with lower systematic error (possible even without further theory input)

  Might become necessary in O(1) years as golden channels become systematics limited!

# Conclusions

# Conclusions

We develop a simple method to simulate "fake" lepton backgrounds for theorist's collider studies. (Easily generalizes to other fakes.)

Our approach is amenable to theoretical study from "first principles", which could teach us more about QCD while improving our understanding of the main "golden channel" background.

With more work & validation, this could be implemented by experimentalists to greatly reduce systematic uncertainties and enhance new physics reach in many searches.