

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Advancing Particle Physics with Sophisticated Computational Frameworks

Permalink

<https://escholarship.org/uc/item/4cs3f5fp>

Author

Howard, Jessica Nicole

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Advancing Particle Physics with Sophisticated Computational Frameworks

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Physics

by

Jessica Nicole Howard

Dissertation Committee:
Professor Timothy M.P. Tait, Co-Chair
Professor Daniel Whiteson, Co-Chair
Professor Michael Ratz

2022

Chapter 4 and Appendix C © 2022 Scientific Reports
Portion of Chapter 5 © 2022 Journal of High Energy Physics (JHEP)
All other materials © 2022 Jessica Nicole Howard

TABLE OF CONTENTS

	Page
LIST OF FIGURES	iv
LIST OF TABLES	ix
ACKNOWLEDGMENTS	x
VITA	xii
ABSTRACT OF THE DISSERTATION	xiii
1 Conventions	1
2 Introduction	3
2.1 The Standard Cosmological History of the Universe	3
2.2 Dark Matter	6
2.3 Large Hadron Collider (LHC) Simulations	13
2.4 Motivation and Outline of This Thesis	16
3 Background of Techniques Used	19
3.1 Optimal Transport Theory	19
3.1.1 Wasserstein Distance	21
3.2 Machine Learning	27
3.2.1 Unsupervised Generative Machine Learning	29
3.3 Practical Tools in Quantum Field Theory	34
3.3.1 Effective Field Theory	35
3.3.2 QCD Confinement and Chiral Symmetry Breaking	37
4 Foundations of a Fast, Data-Driven, Machine-Learned Simulator	43
4.1 Introduction	43
4.2 Theoretical Background	47
4.3 Objective and Related Work	48
4.4 Proposed Solution	52
4.4.1 Our Approach: OTUS	52
4.4.2 OTUS in Practice	56
4.5 Results	59
4.5.1 Demonstration in $Z \rightarrow e^+e^-$ decays	59

4.5.2	Demonstration in semileptonic top-quark decays	62
4.6	Methods	66
4.6.1	Data Generation	66
4.6.2	Model	70
4.6.3	Training	74
4.6.4	Evaluation	79
4.7	Conclusion	80
4.8	Data Availability	83
4.9	Code Availability	83
4.10	Author Contributions Statement	84
5	Dark Matter Freeze-out during $SU(2)_L$ Confinement	85
5.1	Introduction	85
5.2	Weak Confinement and Dark Matter	88
5.3	Dark Matter freeze-out	98
5.3.1	Annihilation Cross Section	98
5.3.2	Freeze-out	101
5.3.3	Deconfinement	102
5.3.4	Numerical Results	103
5.4	Three Generations of Standard Model doublets and Dark Matter	105
5.5	Methods	107
5.5.1	Outline of Code	107
5.5.2	Statistical Handling of Parameter Scan Results	109
5.6	Discussion	111
6	Conclusion and Outlook	114
6.1	Thoughts on Applying Machine Learning to Problems in Particle Theory	116
	Bibliography	119
	Appendix A Chiral transformation invariance of QCD lagrangian	138
	Appendix B OTUS Statistical Matching	141
	Appendix C OTUS Ablation Study	143
	Appendix D Matrices for One Generation	150
	Appendix E Statistics Supplement	152
	Appendix F Small Angle Approximation	156
	Appendix G Handling Derivative Field Interactions	158
	Appendix H Neglecting Gauge Interactions	160


LIST OF FIGURES

	Page
<p>2.1 Schematic diagram depicting the timeline of the Universe assuming a standard cosmological history. During the earliest times of the Universe (first few minutes) particles were the dominant objects. Our earliest direct experimental probes only goes as far back as Big Bang Nucleosynthesis (BBN). Under the standard cosmological history, we can use our understanding from particle experiments to give us clues about what earlier times might have been like.</p>	5
<p>2.2 Schematic diagram of the four stages in current LHC simulation methods: (1) parton interactions, (2) showering, (3) detection, and (4) reconstruction. See the text for further description. We list the typical number of parameters in each stage. Notice that the initial and final stages are low dimensional ($\mathcal{O}(10)$ parameters) compared to the intermediate stages ($\mathcal{O}(100)$ to $\mathcal{O}(10^6)$ parameters). The arrows indicate the names of typical software packages which perform the calculations in these stages.</p>	15
<p>3.1 Schematic diagram showing a failing of VAEs. The latent space loss function of a VAE matches the encoding distribution, $p_E(z x)$, to the latent space prior, $p(z)$. This encourages every data sample, x_i, to be mapped to the entire latent space prior $p(z)$, which inadvertently results in an information collapse between the data space, \mathcal{X}, and latent space, \mathcal{Z} and spoils the conditionality of the encoder (decoder) mapping on x (z). On the other hand, the latent space loss function of a WAE fixes this problem by encouraging the matching of the marginalized encoding distribution, $p_E(z)$, to the latent space prior, $p(z)$, instead. Every data sample, x_i, is matched to a sub-distribution, $p_E(z x_i)$ which conspire to build the learned latent distribution, $p_E(z) \approx p(z)$.</p>	33

4.1	Schematic of the problem and the solution. Current simulations map from a physical latent space, \mathcal{Z} , to data space, \mathcal{X} , attempting to mimic the real physical processes at every step. This results in a computationally intensive simulation. Previous Machine Learning (ML) solutions can reproduce the distributions in \mathcal{X} but are not conditioned on the information in \mathcal{Z} ; instead they map from unphysical noise to \mathcal{X} , which limits their scope. We introduce a new method which provides the best of both worlds. OTUS provides a simulation $\mathcal{Z} \rightarrow \mathcal{X}$ (Decoder) which is conditioned on \mathcal{Z} yet is computationally efficient. Advantageously, it also inadvertently provides an equivalently fast unfolding mapping from $\mathcal{X} \rightarrow \mathcal{Z}$ (Encoder).	49
4.2	Schematic diagram of how OTUS can be used in an abstract analysis. The gray surface represents \mathcal{Z} . Different theoretical models, θ_i , will produce different signatures $\{z_i \theta_i\}$ which lie in \mathcal{Z} . The goal of OTUS is to learn a general mapping from $\mathcal{Z} \rightarrow \mathcal{X}$ which is independent of the underlying theory, θ , and only depends on the information contained in $\{z \in \mathcal{Z}\}$. One trains OTUS using control region data which span \mathcal{Z} and have known outcomes in \mathcal{X} . These allow us to pair distributions in \mathcal{Z} with distributions in \mathcal{X} . From these examples, OTUS interpolates to the rest of \mathcal{Z} and can then be used to generate $\{x_i\}$ from samples $\{z_i \theta_i\}$ from regions not used during training, including the blinded signal region. This can then be used to search for new particles.	57
4.3	Performance of OTUS for $Z \rightarrow e^+e^-$ decays. a Matching of the positron's p_x , p_y , and E distributions in \mathcal{Z} . It shows distributions of samples from the theoretical prior, $\{z \sim p(z)\}$ (solid black), as well as the output of the encoder, $\{\tilde{z}\}$; the encoder transforms samples of testing data in experimental space, \mathcal{X} , to the latent space, \mathcal{Z} , and is shown as $x \rightarrow \tilde{z}$ (dashed cyan). b Matching of the positron's p_x , p_y , and E distributions in \mathcal{X} . It shows the testing sample $\{x \sim p(x)\}$ (solid black) in the experimental space, \mathcal{X} , as well as output from the decoder applied to samples drawn from $p(z)$, labeled as $z \rightarrow \tilde{x}'$ (dashed purple). Also shown are samples passed through both the decoder and encoder chain, $x \rightarrow \tilde{z} \rightarrow \tilde{x}$ (dotted green). Dotted green and solid black distributions are matched explicitly during training. Enhanced differences between dashed purple and solid black indicate the encoder's output needs improvement, as $p_E(z)$ does not fully match $p(z)$. If performance were ideal, the distributions in every plot would match up to statistical fluctuations. Residual plots show bin-by-bin ratios with statistical uncertainties propagated accordingly (see Section 4.6.4).	61
4.4	Visualization of the transformation from $\mathcal{Z} \rightarrow \mathcal{X}$ in the $Z \rightarrow e^+e^-$ study for positron energy. a The learned transformation of the decoder, $p_D(x z)$. b The true transformation from the simulated sample, for comparison, though the true (z, x) pairs are not typically available and were not used in training. Colors in the \mathcal{X} projection indicate the source bin in \mathcal{Z} for a given sample.	62

4.5	Performance of OTUS for $Z \rightarrow e^+e^-$ decays in a physically important derived quantity, the invariant mass of the electron-positron pair, M_Z.	a Matching of the M_Z distribution in \mathcal{Z} . It shows distributions of samples from the theoretical prior, $\{z \sim p(z)\}$ (solid black), as well as the output of the encoder, $\{\tilde{z}\}$; the encoder transforms samples of testing data in experimental space, \mathcal{X} , to the latent space, \mathcal{Z} , and is shown as $x \rightarrow \tilde{z}$ (dashed cyan). b Matching of the M_Z distribution in \mathcal{X} . It shows the testing sample $\{x \sim p(x)\}$ (solid black) in the experimental space, \mathcal{X} , as well as output from the decoder applied to samples drawn from $p(z)$, labeled as $z \rightarrow \tilde{x}'$ (dashed purple). Also shown are samples passed through both the decoder and encoder chain, $x \rightarrow \tilde{z} \rightarrow \tilde{x}$ (dotted green). Dotted green and solid black distributions are matched explicitly during training. Enhanced differences between dashed purple and solid black indicate the encoder's output needs improvement, as $p_E(z)$ does not fully match $p(z)$. If performance were ideal, the distributions in every plot would match up to statistical fluctuations. Note that this projection was not explicitly used during training, but was inferred by the networks. Residual plots show bin-by-bin ratios with statistical uncertainties propagated accordingly (see Section 4.6.4).	63
4.6	Performance of OTUS for semileptonic $t\bar{t}$ decays.	a Matching of the b quark's p_x , p_y , and E distributions in \mathcal{Z} . It shows distributions of samples from the theoretical prior, $\{z \sim p(z)\}$ (solid black), as well as the output of the encoder, $\{\tilde{z}\}$; the encoder transforms samples of the testing data in experimental space, \mathcal{X} , to the latent space, \mathcal{Z} , and is shown as $x \rightarrow \tilde{z}$ (dashed cyan). b Matching of the leading jet's p_x , p_y , and E distributions in \mathcal{X} . It shows the testing sample $\{x \sim p(x)\}$ (solid black) in the experimental space, \mathcal{X} , as well as output from the decoder applied to samples drawn from the prior $p(z)$, labeled as $z \rightarrow \tilde{x}'$ (dashed purple). Also shown are samples passed through both the decoder and encoder chain, $x \rightarrow \tilde{z} \rightarrow \tilde{x}$ (dotted green). Dotted green and solid black distributions are matched explicitly during training. Enhanced differences between dashed purple and solid black indicate the encoder's output needs improvement, as $p_E(z)$ does not fully match $p(z)$. If performance were ideal, the distributions in every plot would match up to statistical fluctuations. Residual plots show bin-by-bin ratios with statistical uncertainties propagated accordingly (see Section 4.6.4).	67
4.7	Visualization of the transformation from $\mathcal{Z} \rightarrow \mathcal{X}$ in the $t\bar{t}$ study for the energy of the b quark in \mathcal{Z} to energy of the leading jet in \mathcal{X}.	a The learned transformation of the decoder, $p_D(x z)$. b The true transformation from the simulated sample, for comparison, though the true (z, x) pairs are not typically available and were not used in training. Note that the b quark will not always correspond to the leading jet, see the text for details. Colors in the \mathcal{X} projection indicate the source bin in \mathcal{Z} for a given sample.	68

4.8	Performance of OTUS for semileptonic $t\bar{t}$ decays in physically important derived quantities in \mathcal{X}. a Matching of the invariant mass of the combined $t\bar{t}$ pair. b Matching of the invariant mass of the hadronically decaying W -boson, M_W . c Matching of the invariant mass of the top-quark, M_t , reconstructed using information from the leptonically decaying W -boson. d Matching of the invariant mass of the top-quark, M_t , reconstructed using information from the hadronically decaying W -boson. These show the testing sample $\{x \sim p(x)\}$ (solid black) in the experimental space, \mathcal{X} , as well as output from the decoder applied to samples drawn from $p(z)$, labeled as $z \rightarrow \tilde{x}'$ (dashed purple). Also shown are samples passed through both the decoder and encoder chain, $x \rightarrow \tilde{z} \rightarrow \tilde{x}$ (dotted green). Dotted green and solid black distributions are matched explicitly during training. Enhanced differences between dashed purple and solid black indicate the encoder's output needs improvement, as $p_E(z)$ does not fully match $p(z)$. Residual plots show bin-by-bin ratios with statistical uncertainties propagated accordingly (see Section 4.6.4).	69
4.9	Schematic diagrams of the network and loss structures used in this study for the base training strategy. a Diagram showing the full OTUS model where gray indicates information used in the calculation of losses only. b Diagram showing the internal structure present in both the encoder and decoder models. c Diagram showing the setup used for the post processing decoder network loss. See the text for more details.	76
5.1	Schematic diagram of the weak confinement and dark pion freeze-out. Upper panel: A sketch of the cosmological history of the Universe where we assume a period of weak confinement begins at Λ_W , at which point the DM (χ_1, χ_2) and SM (q, ℓ) doublets are bound into weak pions. During this epoch, the freeze-out of dark pions takes place at T_{fo} , followed by deconfinement at T_{dc} . Lower panel: The evolution of the dark pion abundance for a representative value of the freeze-out temperature $x_{\text{fo}} = m_1/T_{\text{fo}} \simeq 30$, corresponding to a temperature of $0.2m_{\text{DM}}$. In our notation, m_1 and m_{DM} denote the lightest dark pion and the constituent dark matter masses respectively, see Section 5.3 for details.	87
5.2	Pion masses as a function of m_{DM}, assuming $C_G = C_A = C_Z = -1$, $C_W = 1$ and $\kappa = 1$, for two benchmark points: BP1 where $g_s, e_Q \simeq g'$ and $s_Q \simeq g'/\sqrt{3g_s^2 + g'^2}$ are found by running g_s and g' to $\Lambda_W = 4\pi f \simeq 800$ TeV; and BP2 where we take $g_s = 0.1$ and $e_Q = 0.01$. $M_{13} = 0$ is not shown. . . .	95
5.3	Four-point pion interaction diagrams contributing to the process $\Pi_a \Pi_b \rightarrow \Pi_c \Pi_d$. The dashed lines denote fields on which derivatives act, contributing a factor of the corresponding momentum. An incoming (outgoing) field contributes a negative (positive) momentum factor to the matrix element.	99

5.4	<p>The region of interest for the constituent dark matter mass, m_{DM}, and the weak confinement scale, f, for one generation (left) and three generation (right) cases. The solid and dashed lines show where the DM relic density is consistent with observations at 1 and 2 σ respectively. We show the velocity-averaged effective cross section during freeze-out given in Eq. (5.31). The grey shaded area is inconsistent with unitarity constraints. Note that for both cases we start our scan at $m_{\text{DM}} = 500$ GeV and that the highest points for our scans are $m_{\text{DM}} = 8.5$ TeV and 10.5 TeV for one generation and three generation case respectively. For the benchmark shown above, BP1, $g_s = 0.8$, $e_Q = 0.5$ and $s_Q^2 = 0.12$.</p>	104
5.5	<p>Schematic outline of the code showing the dependencies of the various script files. There are two main script files <code>preScan.py</code> and <code>omegaH2.py</code> which call functions in several helper script files located in the directory <code>utilityFunctions/</code>. The arrows indicate the order in which functions are called in the helper script files. Relevant file outputs of <code>preScan.py</code> and <code>omegaH2.py</code> are shown in the orange sections. The code is publicly available  [134].</p>	108

LIST OF TABLES

	Page
5.1 Masses of the pions (for the one SM generation case) in the small mixing limit, along with their $U(1)_Q \times SU(2)_C$ charges and constituent $SU(2)_L$ doublet content.	93
5.2 Table of mass squared values corresponding to mass basis states along with the relevant $SU(2)_C \times U(1)_Q$ charges. Three SM generations with χ_1 and χ_2 are included.	106

ACKNOWLEDGMENTS

Firstly, I would like to thank my advisors, Tim Tait and Daniel Whiteson, for their continued support and guidance through my PhD. I am particularly grateful for the freedom you have given me to explore my (very) wide array of interests. The advice you have given me during my time at UCI continues to be invaluable, and I strive to emulate your mentorship.

I would like to express my unending gratitude the UCI particle theory group as a whole; the environment that you have cultivated is truly special, and I feel lucky to have experienced it. In particular, I would like to thank Arvind Rajaraman for guiding my first projects in particle theory and giving invaluable feedback on my NSF GRF application. I would also like to thank both Arvind Rajaraman and Yuri Shirman for encouraging my indecisive self to take a leap and pursue particle theory during my PhD.

I would like to thank the many amazing instructors that I had while at UCI. In particular, the lectures by Michael Ratz, Arvind Rajaraman, and Tim Tait all helped solidify my understanding and love of particle physics and quantum field theory. Thank you all for fostering a wonderful learning environment.

I would like to thank the many collaborators I have had during my stay at UCI: Pierre Baldi, Dillon Berger, Lukas Blecher, Anja Butter, Julian Collado, Djuna Croon, Taylor Faucett, Seyda Ipek, Gregor Kasieczka, Fabian Keilbach, Stephan Mandt, Tilman Plehn, Arvind Rajaraman, Rebecca Riley, Tim Tait, Tony Tong, Jessica Turner, Daniel Whiteson, and Yibo Yang for working with me tirelessly on a wide variety of interesting topics.

I would also like to thank my amazing mentors outside of research for helping me to stay balanced during my PhD. In particular, I would like to thank Aomawa Shields. It has been a privilege to be a part of SCECIE and Rising Stargirls over these years; I treasure all that I have learned.

I would like to thank the incredible network of friends in my life, in particular the ones I have made while at UCI, for always being there for me. To my family, thank you for the encouragement and love you have shown me throughout my life. You all have kept me grounded, and I could not have done it without each and every one of you.

A final note to Haytham: Thank you for your unending love and support, for being there to talk through the good and the bad of grad school, for making me laugh, and always reminding me to breathe. I am a better person for having you by my side. You have my gratitude and my love.

Thank you to the UCI machine learning and physical sciences (MAPS) program for research and collaboration opportunities through your honorary fellowship. This work received support in part by the National Science Foundation under grants DGE-1633631 and DGE-1839285 and the U.S. Department of Energy, Office of Science under the grant DE-SC0009920.

Chapter 4 of this dissertation is a reprint of the material as it appears in [1], used with permission from Scientific Reports. The co-authors listed in this publication are Stephan Mandt, Daniel Whiteson, and Yibo Yang. Chapter 5 (with the exception of Section 5.5) of this dissertation is a reprint of the material as it appears in [2], used with permission from the Journal of High Energy Physics (JHEP). The co-authors listed in this publication are Seyda Ipek, Tim M.P. Tait, and Jessica Turner.

VITA

Jessica Nicole Howard

EDUCATION

Doctor of Philosophy in Physics and Astronomy	2022
University of California, Irvine	<i>Irvine, CA</i>
Bachelor of Science in Physics (Minor in Mathematics)	2017
University of California, Davis	<i>Davis, CA</i>

RESEARCH EXPERIENCE

Graduate Research Assistant	2018–2022
University of California, Irvine	<i>Irvine, CA</i>

TEACHING EXPERIENCE

Teaching Assistant	2017 – 2018
University of California, Irvine	<i>Irvine, CA</i>

SELECTED RESEARCH AWARDS AND HONORS

Graduate Research Fellow	2019 – 2022
National Science Foundation	
Honorary Fellow	2018 – 2021
UC Irvine’s Machine Learning and the Physical Sciences (MAPS) Program	

ABSTRACT OF THE DISSERTATION

Advancing Particle Physics with Sophisticated Computational Frameworks

By

Jessica Nicole Howard

Doctor of Philosophy in Physics

University of California, Irvine, 2022

Professor Timothy M.P. Tait, Co-Chair

Professor Daniel Whiteson, Co-Chair

The Standard Model (SM) of particle physics is one of the most complete mathematical models of physical phenomena to date. Even so, it cannot explain experimental results like the existence of particle dark matter and the fact that neutrino masses are non-zero. Explaining such results will necessitate developing a beyond the SM (BSM) theoretical description of particle physics. What form this BSM physics will take has become increasingly unclear; many elegant theories which were expected to appear in recent experiments have not emerged. Thus, we find ourselves at a cross-roads, in need of new perspectives and new computational frameworks to push our theoretical description of physics forward.

New perspectives will come from challenging previously-held assumptions in the pursuit of fundamentally new descriptions, but challenging such assumptions often presents practical computational challenges. Therefore, these new perspectives must also be accompanied by new computational frameworks. Computational frameworks can come in many forms, from the purely mathematical to the largely numerical. In particular, in recent years machine learning (ML) has become an increasingly accessible and powerful computational tool for scientific applications. Crafting novel BSM theories will require us to investigate and embrace the full spectrum of computational frameworks. Additionally, one of the best ways to spark

new insights is to closely collaborate with and draw inspiration from other fields, such as mathematics and computer science.

In this thesis, we present two examples of how advanced computational frameworks can be used to aid in investigating new physics perspectives. In one example, we see how the purely mathematical framework of optimal transport (OT) theory can be used in tandem with advanced ML methods to enable a new perspective on particle physics simulations. The result is a novel strategy which lays the foundations for a completely data-driven, end-to-end simulation of particle collisions at the Large Hadron Collider. In a second example, we see work which considers a new perspective on what the history of our universe might have looked like. In particular, we consider how the abundance of a WIMP dark matter candidate could be altered by considering a phase of electroweak force confinement early in the universe. Considering this model while making relatively few assumptions was aided by the application of advanced numerical computational tools.

We begin with both high-level and technical background on the topics relevant to these works. We conclude by discussing future directions for these works, as well as briefly giving general thoughts on strategies for applying the computational framework of ML to problems in theoretical particle physics more broadly.

Chapter 1

Conventions

This chapter briefly summarizes the conventions and notations used in this thesis:

- Any abbreviations will be defined at their first appearance in each chapter.
- For definitions in mathematical expressions we use the symbol “:=” or “=:” to define the term on the left-hand-side or right-hand-side, respectively. For example, $x := y$ defines x in terms of y whereas $x =: y$ defines y in terms of x . This is done to be more explicit than the more physics-typical symbol “ \equiv ”.
- Throughout, instead of calling \mathcal{L} the Lagrangian density, we will simply refer to it as the Lagrangian. This is a common abuse of notation. If we do, in fact, discuss the Lagrangian itself, $L := \int d^4x \mathcal{L}$, we will use the more precise Lagrangian density terminology to refer to \mathcal{L} in that discussion.
- Below we list some handy QFT notation reminders:
 - As is standard in particle physics, we use the $(+, -, -, -)$ convention for the Minkowski metric, e.g. $m^2c^4 = p^\mu p_\mu = \eta_{\mu\nu} p^\mu p^\nu = +p^0p^0 - p^1p^1 - p^2p^2 - p^3p^3 = E^2 - \vec{p}^2c^2$.

- $\partial_\mu := \frac{\partial}{\partial x^\mu}$.
 - The Feynman slash notation is a shorthand for contraction with γ^μ , e.g. $\not{\partial} := \gamma^\mu \partial_\mu = \gamma_\mu \partial^\mu$.
 - The "bar" notation is also a shorthand, e.g. $\bar{\psi} = \psi^\dagger \gamma^0$.
- We assume throughout that $\hbar = c = 1$ (and where relevant $k_B = 1$), which means all quantities in QFT have units of energy to some power:
 - **Mass:** $[m] = \text{E}$.
 - **Momentum:** $[p^\mu] = \text{E}$.
 - **Length:** $[x^\mu] = \text{E}^{-1}$.
 - **Spatial Derivatives:** $[\partial_\mu] = \text{E}$ (because x^μ is "in the denominator").
 - **Action:** $[S] = \text{E}^0 = 1$.
 - **Lagrangian Density:** $[\mathcal{L}] = \text{E}^D$ where D is the number of space-time dimensions (because $S := \int d^D x \mathcal{L}$).
 - **Vector Field:** $[A_\mu] = \text{E}^{(D-2)/2}$, for $D = 4$ then $[A_\mu] = \text{E}$.
 - **Fermion Field:** $[\Psi_\mu] = \text{E}^{(D-1)/2}$, for $D = 4$ then $[\Psi_\mu] = \text{E}^{3/2}$.
 - **Scalar Field:** $[\Phi_\mu] = \text{E}^{(D-2)/2}$, for $D = 4$ then $[\Phi_\mu] = \text{E}$.

Chapter 2

Introduction

In this chapter we briefly review some relevant topics found in this thesis: the standard cosmological history of the Universe, dark matter (DM), and simulations of particle collisions at the Large Hadron Collider (LHC). These high-level reviews are meant to concisely summarize important features while pointing to valuable resources for more detailed information. Finally, the last section of this chapter describes the motivation for the work in this thesis and provides an outline of the rest of the chapters.

2.1 The Standard Cosmological History of the Universe

Cosmology is the scientific study of the origins and the development of the Universe into what we observe today. The current accepted model of how the Universe began, the hot big bang model, establishes that the Universe's beginning was both hot and dense. As time progressed the Universe expanded, the density dropped, and the Universe cooled; therefore, the further back we go in the history of the Universe, the higher the average temperature will be at that time. Hotter average temperatures imply that there was more energy available to the

occupants of the Universe, most notably particles. From controlled, laboratory experiments, such as the Large Hadron Collider (LHC), we know that the behavior and interactions of particles can change drastically with energy. Thus, studying the interactions of fundamental particles at high energies will help us discover what the Universe looked like shortly after the big bang singularity [3].

The estimated age of the Universe is the time from the big bang singularity until now — 13.787 ± 0.020 billion years ago [4].¹ While we have experimental probes of the vast majority of this history, probes of very early times, when the majority of interesting particle physics was occurring, are still out of reach. The earliest direct experimental probe is from the era of Big Bang Nucleosynthesis (BBN), corresponding to an average temperature of 10 MeV to 0.1 MeV or about 10^{-2} seconds to 3 minutes after the big bang singularity [3]. During this time, the light nuclei, which are more complicated than hydrogen nuclei (just a single proton), were formed. In particular, the nuclei of deuterium (D), helium-3 (^3He), helium-4 (^4He), and lithium-7 (^7Li) were formed. Since few subsequent astrophysical processes affect the abundances of these nuclei,² the observed abundances of these elements today provides a precise constraint on our understanding of particle physics at this time in history [3].

The early-universe quantities which these measurements can constrain are g_* , which counts the total number of effectively massless degrees of freedom at a given temperature (i.e. $m \ll T$), the neutron half-life, $\tau_{1/2}(n)$, and the ratio of normal matter (baryons) to radiation (photons), $\eta := n_N/n_\gamma$; η is also a measure of entropy, when $\eta \ll 1$ the Universe is “hot” and entropy is high. The abundance of ^4He provides a constraint on the ratio of neutrons to protons at the time when reactions converting neutrons into protons froze-out. This, in turn, places a constraint on g_* as well as $\tau_{1/2}(n)$. While different values of η will also affect the abundance of ^4He , the effect is much less pronounced. Instead, the abundances

¹This age was estimated from 2018 Planck data and includes Baryon Acoustic Oscillations (BAO) data.

²Note that this is less true for ^4He which can also come from stellar processes, however the primordial contribution is still significant and can be estimated by studying metal-poor objects [3].

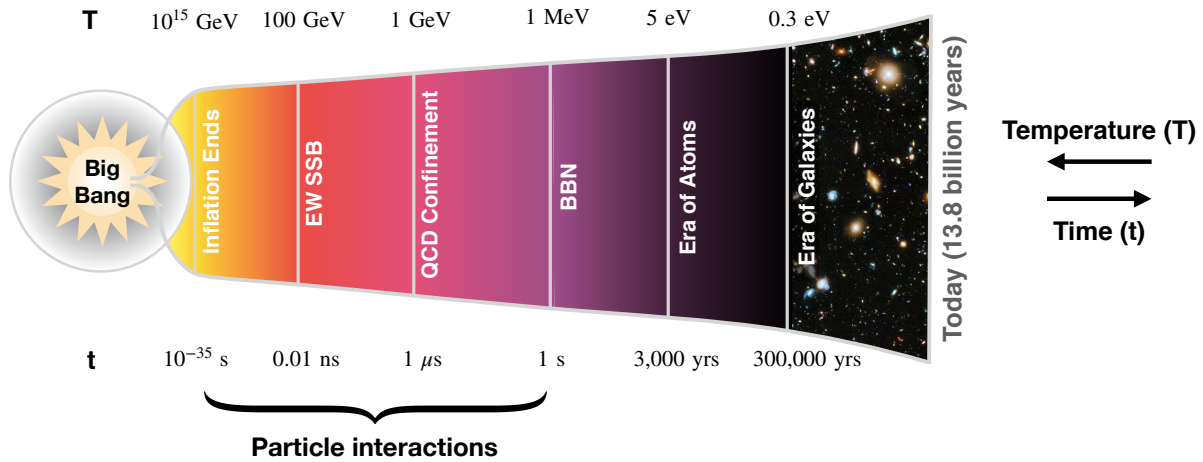


Figure 2.1: **Schematic diagram depicting the timeline of the Universe assuming a standard cosmological history.** During the earliest times of the Universe (first few minutes) particles were the dominant objects. Our earliest direct experimental probes only goes as far back as Big Bang Nucleosynthesis (BBN). Under the standard cosmological history, we can use our understanding from particle experiments to give us clues about what earlier times might have been like.

of D, ^3He , and ^7Li place far stronger constraints on the value of η . These measured values are consistent with how these objects would interact assuming the Standard Model (SM) of particle physics. We can therefore say that the SM of particle physics can describe the historic interactions between particles up to $T = 10$ MeV ($t = 10^{-2}$ s) [3].

From a modern experimental particle physics perspective, $T = 10$ MeV is rather low-energy. For example, the second run of the LHC was colliding protons at 13 TeV energies, which is larger by 6 orders of magnitude. If we assume that particle physics in the lab corresponds to particle physics in the early universe, then we can say quite a bit more about what the early universe looked like (see Fig. 2.1 for details). This correspondence is the stance taken by the standard cosmological history — the SM of particle physics at high energies corresponds to what the Universe looked like at early times.

However, we know that the SM is incomplete; missing pieces like the observed matter/anti-matter asymmetry, the nature of DM, and non-zero neutrino masses will certainly affect an early-universe description of particle interactions [3, 5]. So while we can use the SM as a guide in the early universe, extrapolating it back to earlier times (higher temperatures) should not be a strict requirement. Until experimental probes of the first instances of the Universe are available, it is worth entertaining a wide variety of possible early-universe models to fully explore all possibilities.

Recent work which relaxed the assumption of a standard cosmological history has had success in being able to (partially) explain many BSM observations, such as baryogenesis and DM [2, 6–12]. In Chapter 5 of this thesis, we will see an example of this related to DM [2]. This work considers an alternate cosmological history with a phase of electroweak force confinement contemporary with DM freeze-out. This alteration expands regions of DM mass parameter space which were previously considered unviable, and shows that they could produce the observed DM relic abundance while having evaded experimental detection.

2.2 Dark Matter

One of the biggest unsolved mysteries in modern physics is the observation that there is more mass (matter) in the Universe than we expect based on what we can see. More precisely, the visible matter that we can observe (i.e. matter which interacts typically with light) is not enough to describe the gravitational dynamics of large-scale structures in the Universe (e.g. galaxies). We are thus left with two options: our description of gravity needs to be revised or there is matter which we cannot observe via our traditional, light-based astronomical means [5].

Since the discovery of this discrepancy many explanations have been proposed and subse-

quent experiments devised. Explanations which modify gravity are less favorable as they must contend with the overwhelming experimental success of the theory general relativity and also fail to explain many experimental observations [5]. Similarly, it is unlikely that this missing mass is arising solely from familiar “dark” objects such as planets [3].

Instead, observations to-date strongly support the notion that DM is a new kind of massive particle. This is also motivated by the fact that we *know* there are other issues with our theoretical description of particle physics (e.g. non-zero neutrino masses), so it is reasonable to hope that the solutions may be related. However, besides the fact that DM behaves roughly like a massive particle, we do not know much about its properties. For example, does it only interact gravitationally with normal matter? Could it interact feebly with normal matter via another force? Does it interact with itself? Is it fundamental or composite? While there is experimental evidence that can attempt to answer these questions, the overall story is muddled, and often very dependent on the specific theoretical model being used. However, in order to make practical headway, the vast majority of DM models assume that DM interacts with normal matter non-gravitationally, at least to some extent; this can allow DM to better fit into a BSM physics model and also gives us a chance of seeing it in earth-based experiments [3, 5].

A historically popular (and economic) choice for DM is that it is a Weakly Interacting Massive Particle (WIMP). The idea is that DM interacts only via the Weak Force (i.e. it is charged under the SM’s $SU(2)_L$ gauge symmetry, but not under any of the other SM symmetries). Because DM interacts Weakly, this means that, at some point in the Universe’s cosmological history, DM was in thermal equilibrium with all other SM particles (i.e. reactions converting DM particles \leftrightarrow SM particles were happening at equal rates). If DM stayed in thermal equilibrium throughout the Universe’s history its abundance today would be negligible³

³The abundance of a massive particle species in thermal equilibrium decreases with a factor $\exp(-m/T)$. Given the Universe’s present temperature $T = 2.7260 \text{ K} = 2.3491 \times 10^{-4} \text{ eV}$ [4], this exponential factor would make the abundance nearly negligible.

and not reflect what we observe. When a particle species departs thermal equilibrium, its reactions (i.e. $\text{DM} \rightarrow \text{SM}$) are said to have "frozen-out", fixing its abundance to what we observe today. In particular, we measure the contribution of DM to the energy density of the Universe. Assuming that the DM present today comprises a single particle species, the energy density of DM today will be the mass of the DM particle times its number density, $\rho_{\text{DM}} := m_{\text{DM}} n_{\text{DM}}$. This is commonly reported as a normalized density parameter, $\Omega_{\text{DM}} := \rho_{\text{DM}}/\rho_{\text{crit}}$, where ρ_{crit} is the critical density such that the Universe would be flat i.e. have zero curvature. Because of this normalization, the sum of the density parameters of all sources is one, $\sum \Omega_i = 1$

If DM's interactions to the SM are through the Weak Force, then this lets us know when DM froze-out and thus allows us to calculate the number density of DM today. This leaves DM's mass as the only free parameter. Therefore, matching the measured value of Ω_{DM} will uniquely determine m_{DM} . This proposed model is attractive because it does not involve extra particle physics mechanisms, only a new massive particle which interacts Weakly; moreover, WIMP candidates for DM exist in many historically popular theories (e.g. SUSY), so this scenario was historically seen as a two-for-one deal [5]. To-date, however, experiments which have searched for WIMPs have largely ruled out the mass values favored in this scenario. While there are WIMP models that have not been ruled out, they usually are not as elegantly simple as the original scenario described above [13].

However, an implicit assumption in the above discussion is that the early universe followed the standard cosmological history. An alternate cosmological history, in particular one which affects the strength of the Weak Force during DM freeze-out, would greatly alter this picture. In Chapter 5 we will see exactly how such an alternate cosmological history might arise and how this expands regions of DM mass parameter space which were previously considered unviable [2].

The remainder of this section briefly describes how one calculates the relic number density

of a particle species which froze-out at some point in the early universe. This largely follows the discussion found in Ref. [3].

We are interested in finding out how the phase-space distribution function of a particle species evolves as a function of time. This evolution is simple both when this species is in thermal equilibrium with all other particle species (the “plasma”) and when it has completely frozen-out. In the former case, we say that the particle species is coupled to the plasma, and in the latter, we say that it has decoupled. Whether a species is coupled to the plasma or not depends on how the per-particle interaction rate Γ (for reactions which would keep the particle species in thermal equilibrium with the plasma) relates to the expansion rate of the Universe, H . If $\Gamma \gg H$, then interactions are occurring sufficiently rapidly to maintain equilibrium (coupled); on the other hand if $\Gamma \ll H$, then reactions are not rapid enough to maintain equilibrium (decoupled). It is in the intermediate case, $\Gamma \approx H$, where the description gets tricky.

In general, the phase-space distribution function $f(p^\mu, x^\mu)$ evolves according to the Boltzmann equation⁴

$$\hat{\mathbf{L}}[f] = \mathbf{C}[f], \tag{2.1}$$

where \mathbf{C} is the collision operator and $\hat{\mathbf{L}}$ is the Liouville operator, which is generally given by

$$\hat{\mathbf{L}} = p^\alpha \frac{\partial}{\partial x^\alpha} - \Gamma_{\beta,\gamma}^\alpha p^\beta p^\gamma \frac{\partial}{\partial p^\alpha}, \tag{2.2}$$

where the affine connection incorporates gravitational effects. Schematically, the RHS of 2.1 details changes to $f(p^\mu, x^\mu)$ arising from interactions while the LHS details changes arising from space-time considerations (e.g. the expansion of the Universe).

⁴An interesting aside worth mentioning here is that historically the field of optimal transport theory, which will be seen later in this thesis, has been applied to the study of simple Boltzmann equations [14].

Under the FRW model, the phase-space density will be spatially homogeneous and isotropic, i.e. $f(p^\mu, x^\mu) \rightarrow f(p := |\vec{p}|, t)$ (or equivalently $f(E, t)$), and the Liouville operator can be rewritten as

$$\hat{\mathbf{L}} = E \frac{\partial}{\partial t} - H p^2 \frac{\partial}{\partial E}. \quad (2.3)$$

Defining number density in terms of phase-space density as

$$n(t) = \frac{g}{(2\pi)^3} \int d^3p f(E, t), \quad (2.4)$$

we can rewrite Eq. (2.1). We begin by multiplying both sides by

$$\frac{g}{(2\pi)^3} \int \frac{d^3p}{E}. \quad (2.5)$$

The LHS will then give us,

$$\frac{g}{(2\pi)^3} \int \frac{d^3p}{E} \hat{\mathbf{L}}[f] = \frac{g}{(2\pi)^3} \int \frac{d^3p}{E} \left(E \frac{\partial f}{\partial t} - H p^2 \frac{\partial f}{\partial E} \right) \quad (2.6)$$

$$= \frac{g}{(2\pi)^3} \int d^3p \frac{\partial f}{\partial t} - H \frac{g}{(2\pi)^3} \int d^3p \frac{p^2}{E} \frac{\partial f}{\partial E} \quad (2.7)$$

$$= \frac{\partial}{\partial t} \left(\frac{g}{(2\pi)^3} \int d^3p f \right) - H \frac{g}{(2\pi)^3} \int (dp p^2) \frac{p^2}{p} \frac{\partial f}{\partial p} \quad (2.8)$$

$$= \frac{\partial}{\partial t} n(t) - H \frac{g}{(2\pi)^3} \int dp p^3 \frac{\partial f}{\partial p} \quad (2.9)$$

$$= \frac{dn}{dt} + H \frac{g}{(2\pi)^3} \int dp \frac{\partial(p^3)}{\partial p} f \quad (2.10)$$

$$= \frac{dn}{dt} + 3H \frac{g}{(2\pi)^3} \int (dp p^2) f \quad (2.11)$$

$$= \frac{dn}{dt} + 3Hn, \quad (2.12)$$

where we used 1) the fact that $E^2 = m^2 + p^2$ implies $p dp = E dE$, 2) the fact that $d^3p = p^2 dp$ and 3) the relation $-\int dp p^3 \frac{\partial f}{\partial p} = \int dp \frac{\partial(p^3)}{\partial p} f$ obtained via integration by parts, with the

surface term neglected.

We can therefore rewrite the Boltzmann equation Eq. (2.1) in the following form

$$\frac{dn}{dt} + 3Hn = \frac{g}{(2\pi)^3} \int \frac{d^3p}{E} \mathbf{C}[f(E, t)]. \quad (2.13)$$

Thus, the Boltzmann equations are a coupled set of integral-partial differential equations for the phase-space distributions of all particle species in the system. Under some reasonable assumptions this can be simplified. In particular, we can focus on the phase-space distribution (and thus number density) of one⁵ particle species, ψ , and group the rest into an effective plasma phase-space distribution, leaving us instead with one integral-partial differential equation for the species of interest, ψ .

With some additional assumptions (see Ref. [3] for more details) we can express the collision term entirely in terms of the number density of ψ , n , and the thermally averaged cross section times the relative velocity, $\langle\sigma v\rangle$, for reactions which would encourage ψ to go into equilibrium with the plasma. In particular,

$$\frac{dn}{dt} + 3Hn = -\langle\sigma v\rangle (n^2 - n^{\text{eq}2}) \quad (2.14)$$

where n^{eq} is the number density when ψ is in equilibrium. If ψ is a non-relativistic particle species

$$n^{\text{eq}} = g \left(\frac{mT}{2\pi} \right)^{3/2} e^{-x} \quad (2.15)$$

where $x := m/T$ and m is the mass of ψ . For non-relativistic interactions the relative velocity, v , is small so we can expand $\langle\sigma v\rangle = a + b\langle v^2\rangle + \mathcal{O}(v^4)$. Then solving Eq. (2.14) for the value of n at late times (low T) will give us the relic abundance of ψ .

⁵Technically, there is also a way to simplify a case of N species with some additional caveats, more on this below.

If we wanted to instead consider N different species, in principle we would have to solve N such equations. This can get quite complicated, especially when interactions couple the equations together. However, in the case where the N particle species of interest, ψ_i for $i = 1, \dots, N$, have approximately degenerate masses (i.e. $\Delta m_{ij} \sim T_F$, where T_F is the freeze-out temperature) and are interrelated such that all ψ_i for $i > 1$ will decay quickly into the lightest ψ_1 (with mass m_1), we can write a single effective equation in terms of, n , the number density of ψ_1 . Namely,

$$\frac{dn}{dt} + 3Hn = -\langle \sigma_{\text{eff}} v \rangle (n^2 - n_{\text{eq}}^2), \quad (2.16)$$

where

$$\sigma_{\text{eff}} = \sum_{ij}^N \sigma_{ij} \frac{g_i g_j}{g_{\text{eff}}^2} (1 + \Delta_i)^{3/2} (1 + \Delta_j)^{3/2} e^{-x(\Delta_i + \Delta_j)}, \quad (2.17)$$

with

$$g_{\text{eff}} = \sum_{i=1}^4 g_i (1 + \Delta_i)^{3/2} e^{-x\Delta_i}, \quad (2.18)$$

where g_i is the number of degrees of freedom of ψ_i , $x = m_1/T$, σ_{ij} is the cross section for the reaction $\psi_i \psi_j \rightarrow \text{SM}$ (summed over all kinematically accessible SM particles in the final state), and $\Delta_i := (m_i - m_1)/m_1$ is the mass difference between the heavier $\psi_{i \neq 1}$ and ψ_1 . This scenario is called coannihilation (see Ref. [15] and Ref. [16] for more details). The Boltzmann equation can be solved either with analytical approximations or numerical techniques.

2.3 Large Hadron Collider (LHC) Simulations

In an ideal world, a scientist would be able to propose a theoretical model for how some physical phenomenon works, devise an observable and distinguishable feature of this model, and measure this feature directly in the lab to see whether it matches the model's prediction. If it does, then they have a working theoretical description for that physical phenomenon. If it does not, then they must incorporate this information to improve their theoretical description. And repeat. This is the essence of how scientific fields in the physical sciences would ideally operate.

So where does reality diverge from this idealistic scenario? The answer is sometimes you get lucky and reality matches this scenario; but for many physical sciences, particularly particle physics, the wrench in the works is that direct measurements of your distinguishable features are practically impossible. To be more precise, in particle physics theoretical predictions are usually restricted to a space where we cannot make direct measurements.

For example, a theory may predict that a new particle is created, but it is too short-lived to be seen by our detectors. Instead of measuring the properties of this new particle directly, we must infer these properties from measuring its longer-lived decay products, which we *can* detect. These indirect measurements must be reconstructed into a theoretically meaningful form before we can draw conclusions about the validity of a theory. And not only are these measurements indirect, but their reconstruction is also imperfect, often muddled by noise and biases coming from the detectors.

To use an analogy, this task is like finding the exact age of a fossil. We cannot travel back in time to know exactly when the fossil was formed (direct measurement) and instead are limited to using techniques such as carbon dating (indirect measurement) to ascertain the approximate age of the fossil.

In statistical terms, the inaccessible direct data belong to a latent, theoretical model space. Different theories, with different parameters, will make distinguishable predictions in this latent space. Statistical inference then requires understanding how points in this unobserved latent space are transformed into experimental data. This transformation is usually non-trivial, involving complex physical interactions that cannot be described analytically. To bridge this gap, particle physicists typically employ computationally expensive numerical simulations, so-called simulation based inference [17]. Theoretical calculations create predictions in this latent space, \mathcal{Z} , which depend on model parameters, Θ , and simulations map from \mathcal{Z} into the experimental data space, \mathcal{X} . The simulation, in effect, samples from a conditional probability distribution, $p_{\text{sim}}(x | z)$, so that theoretical predictions can be compared in \mathcal{X} .

Over the years, many different simulation strategies have been employed (e.g. matrix element method [18, 19]), but the most widely used methods (e.g. Pythia+GEANT4 [20, 21]) attempt to use our limited knowledge of particle-matter interactions to numerically map from \mathcal{Z} to \mathcal{X} , using Monte-Carlo methods and data from independent experimental analyses to patch where our theoretical understanding fails. The final numerical mapping typically spans many different numerical software packages but can be segmented into four stages which parallel what (we think) happens in reality (see Fig. 2.2). (1) The first stage concerns fundamental particle interactions. The computations in this stage are relatively quick and also theoretically well-framed within quantum field theory (QFT). These are usually computed with a numerical software called MadGraph [22]. (2) The next stage concerns the propagation of the final states from the previous stage to the particles which our detector will eventually see. For quarks and gluons, this involves the modeling of hadronic showers. Our theoretical knowledge of this process is limited and so it is approximated with heuristic methods and Monte-Carlo techniques. These computations are usually performed with a numerical software called Pythia [20]. (3) This stage attempts to recreate how the detector will respond, including biases and inefficiencies, to the particles which it encounters. (4)

Four Stages of Current LHC Simulations

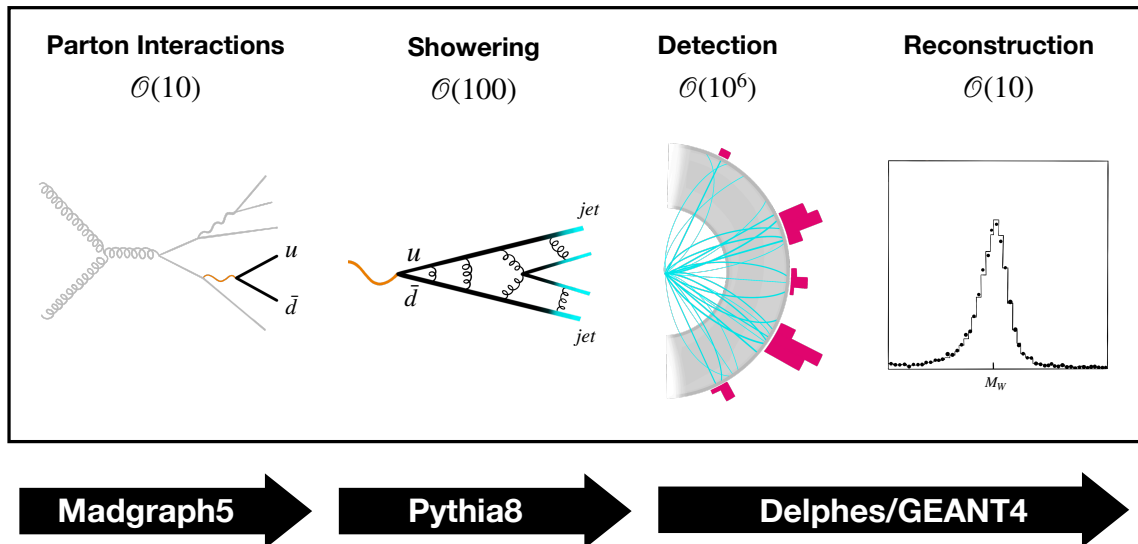


Figure 2.2: **Schematic diagram of the four stages in current LHC simulation methods: (1) parton interactions, (2) showering, (3) detection, and (4) reconstruction.** See the text for further description. We list the typical number of parameters in each stage. Notice that the initial and final stages are low dimensional ($\mathcal{O}(10)$ parameters) compared to the intermediate stages ($\mathcal{O}(100)$ to $\mathcal{O}(10^6)$ parameters). The arrows indicate the names of typical software packages which perform the calculations in these stages.

Finally, this raw detector data is reconstructed into a physicist-interpretable form for use in analyses. These last two stages are usually contained within a single numerical software, typically GEANT4 [21] or Delphes [23].

The idea behind this segmented simulation strategy is to try to leverage as much knowledge as we have about the intermediate processes, with the hope that the final end-to-end mapping will mimic what happens in reality. While we would like for this to happen out-of-the-box, it unfortunately does not. These simulations must undergo rigorous tuning to separate experimental measurements to help account for the gaps in our knowledge (e.g. about the mechanism of hadronization) [1].

Another downside is that these simulations are painfully slow. While Nature can produce

detector events nearly instantaneously, it takes about 10 minutes to simulate 1 event through this intensive chain; experimental analyses typically need millions to billions of events. Moreover, whenever the simulator must be re-tuned, these events must be re-simulated which further slows the process. The slow simulation crisis is the cause of the current dominant uncertainty (statistical), and this problem is only likely to worsen in future experiments (e.g. the high luminosity LHC) [1, 24, 25].

Therefore, many physicists have set to the task of making faster simulators. Often using techniques from machine learning to speed up the intermediate stages. However, Chapter 4 of this thesis explores an alternate perspective, which is to use machine learning to learn an end-to-end mapping from \mathcal{Z} to \mathcal{X} . Other works [26, 27] have also tried this end-to-end strategy but must incorporate simulated samples during training to learn a conditional mapping, thus inheriting some of the above issues. On the other hand, Chapter 4 explores how to learn this mapping directly from reconstructed data, without requiring simulated data samples, using an optimal-transport based machine learning method.

2.4 Motivation and Outline of This Thesis

The current state of particle physics is one of many possibilities and few clear directions in which to push forward. We have numerous reasons to believe that the SM is incomplete, most notably the existence of DM and non-zero neutrino masses [5]. Explaining such experimental results necessitates a beyond the SM (BSM) theoretical description of particle physics. Additionally, many of the deviations from the SM have come in places we were not expecting, and thus are fundamentally challenging our previous guesses as to how the Universe works. To meet this challenge and solve this puzzle, we will need both new perspectives and more clues.

More clues will undoubtedly come from the results of new experiments; but unfortunately this has an associated time-cost of years to decades. While efforts to design and build these future experiments are underway, it is worth asking if we might be able to obtain new clues in other, more immediate, ways — possibly by reexamining old clues (prior experimental data) to see if they lead to new insights. Moreover, we must also ask ourselves how we can best prepare to interpret the results of these future experiments when they do arrive. Both of these pursuits will require new and improved tools. On the other hand, developing new perspectives will require us to challenge previously-held assumptions in the pursuit of fundamentally new descriptions. And all of this will require a healthy dose of creativity.

When at such a crossroad, questioning long-held assumptions and adding creative new tools are not only prudent but a necessity. This time provides us with an exciting opportunity to reflect and revise our practices. While we wait on the next era of particle experiments, which seek to explore some of the most apparent experimental deviations from the SM, we have the freedom to chart new possibilities and examine past results and practices with new perspectives. Charting the possibilities, no matter how crazy they might initially seem, ultimately builds the framework for discovering the truth.

To hearten us on this journey, we may take comfort in the past, where we have seen similar stories play out in physics before. For example, at the end of the 19th century an unexpected observation in the spectrum of black-body radiation led Max Planck to propose the crazy idea that energy could be quantized. This heuristic law (Planck's law) was able to explain the behavior of this system, and eventually paved the way for the theory of quantum mechanics as we know it today. Often times, the success of such unfettered exploration of possibilities relies on the consideration of new practical tools. These tools are often found in, or inspired by, adjacent fields like mathematics and computer science. Therefore, close collaborations between fields will enhance these pursuits.

This thesis attempts to take this philosophy of questioning, creativity, and collaboration to

heart, providing two small examples of the wide range of possible forms that such research can take.

Chapter 4 contains research which develops the foundations of a fundamentally new strategy for simulation tools used in the analysis of data from experiments at the LHC. This new strategy has the potential to save orders of magnitude of computational time, thus meeting the demands of current and future experiments. This method is unique in its ability to learn a physically realistic transformation from theoretical model information to observed data in a completely data-driven way. This proof-of-principle work lays the foundations of this method by drawing on techniques from the fields of machine learning and optimal transport theory [1].

Chapter 5 contains research which questions assumptions about events in our cosmological history and explores the implications on the nature of DM today. In particular, this work considers how a phase of electroweak force ($SU(2)_L$) confinement which was contemporary with DM freeze-out may help alleviate current experimental constraints on WIMP DM. This work is an example of how questioning assumptions, which are not yet experimentally founded, might expand the space of possible descriptions. In general, such shifts may drastically alter the extent to which certain theoretical descriptions are ruled out — in turn affecting the analysis of current and future experimental data and even the design of future experiments. Therefore, it is crucial to question assumptions in order to fully chart the possibilities.

The remaining chapters of this thesis are organized as follows. Chapter 3 provides a background review of the main technical topics used in both of these works. Chapter 6 provides a high-level summary of these works and specific follow-up directions. It also contains broader thoughts concerning future research directions at the interface of machine learning, optimal transport theory, and theoretical particle physics.

Chapter 3

Background of Techniques Used

This chapter covers the background of the main technical topics found in this thesis. Section 3.1 and Section 3.2 cover background material relevant for Chapter 4 while Section 3.3 covers background material relevant for Chapter 5.

3.1 Optimal Transport Theory

Optimal transport (OT) theory is a branch of mathematics with a sporadic yet fascinating history. Despite being formulated in the 18th century, the mathematical study of this topic has been relatively infrequent when compared to other branches of mathematics¹ and has only begun to flourish in earnest in recent years. Over much of its history, only a handful of individuals were concerned with studying OT theory. As a result, breakthroughs were separated by decades. The problem was first formulated in 1781 by French engineer Gaspard Monge [29] but did not see any progress towards a mathematical solution until the 1930s. During World War II major advances were made by Leonid Kantorovich resulting in one of

¹For example, Group Theory, which is another branch of mathematics with wide applications in physics and beyond, was actually developed after OT (in the 19th century [28]) but has seen far more study since.

the most lasting formulations of an OT problem (Monge-Kantorovich formulation). After that, advances remained relatively few until the late 1980s when work by John Mather, Yann Brenier and Mike Cullen found wide-reaching applications of OT to problems in geometry, partial differential equations, and meteorology (not to mention statistics) [14]. Since then, interest in this area has only increased, and is continuing to grow. This is especially true as scientific applications of OT are explored; for example, in the last decade, OT theory has provided fundamental improvements to many machine learning methods [30, 31] and is even being explored as a natural description for the products of particle collisions [32–34]. While scientific applications of OT are relatively nascent, the potential of OT to advance scientific fields is vast and exciting and warrants continued study.

OT theory concerns itself with finding the optimal way to transform one probability distribution into another. This can be intuitively visualized by imagining a probability distribution as a pile of dirt. The task is then to move this pile of dirt, one shovel-full at a time, some distance while reshaping it to have a different form in the process. After defining the cost of transporting a given shovel-full of dirt, you can then ask the question, “What is the transportation plan that allows me to perform this task optimally (i.e. with the least overall cost)?” This scenario of transporting piles of dirt is actually the inspiration for the first formulation of the OT problem by Monge who was interested in whether there was a mathematically optimal way to construct a soil embankment [14].

Discussing the mathematics of OT problems necessitates a notion of distance between probability distributions — the more similar the probability distributions, the easier it would be to optimally transport one into the other, and thus the smaller the distance. The Wasserstein distance (or Kantorovich–Rubinstein metric) was developed to meet this need.^{2,3} [35].

²This metric was originally proposed in 1939 by Russian engineer Leonid Kantorovich in the context of optimal transport of goods. But in 1969, Russian mathematician Leonid Vaseršteĭn (or Wasserstein in German spelling) used this distance in work on automatas. The use of these results in subsequent work led to the coining of the name “Wasserstein” distance, despite the metric originating with Kantorovich.

³Note that, especially when describing the distance between one-dimensional probability distributions, this is also sometimes called the Earth Mover’s metric in reference to the original OT problem.

The Wasserstein distance has become crucial in the study of OT problems, becoming nearly synonymous with OT in many cases.

In Section 3.1.1 we give an overview of the mathematics of the Wasserstein distance. We also discuss a strategy for its explicit computation which is used to achieve the results in Chapter 4.

3.1.1 Wasserstein Distance

The Wasserstein distance is a robust way to describe the degree of similarity between probability distributions (or more accurately, between probability measures). It is formally a mathematical metric, giving it an edge over other common distances such as the KL-divergence, which has pathological failings such as diverging when probability distributions do not overlap in the same space [36, 37]. However, in dimensions greater than one, the explicit computation of the Wasserstein distance becomes computationally intractable leading to the development of many approximations (e.g. [33, 37]). In this section, we summarize the mathematical description of the Wasserstein distance and discuss one strategy for approximating its computation for the case of multi-dimensional probability distributions.

We begin by first noting some definitions which will be useful in the following description:

Probability space: A probability space, Ω , is the set of events to which you can assign a probability [38].

Probability measure: A probability measure determines how you assign probability to an event. It is a real-valued function⁴ defined over a set of events in a probability space.

A probability measure μ must map from a set of events, $E = \{E_i \mid E_i \in \Omega\}$, to the unit

⁴Note that there has been interesting work done generalizing to complex probability measures [39, 40], but in this thesis we restrict ourselves to real measures, as is more standard.

interval, $\mu : E \rightarrow [0, 1]$. If E is the empty set, $\mu(E) = 0$ and if E contains all possible events $\mu(E) = 1$. μ also has the countable additivity property which states that for all countable collections of pairwise disjoint event sets, $\{E_i\}$, $\mu(\cup_i E_i) = \sum_i \mu(E_i)$ [41].

Borel measure: A Borel measure is a measure which maps from the Borel σ -algebra of some topological space to the real number line [42].

Borel probability measure: If the Borel measure is also a probability measure, then it is called a Borel probability measure [42].

Lebesgue measure: The Lebesgue measure extends the typical idea of length, area, and volume to more complicated sets. A measure μ is continuous with respect to the Lebesgue measure, λ , if, for every measurable set A , $\lambda(A) = 0 \Rightarrow \mu(A) = 0$ [42].

Pushforward (measure): A pushforward (or pushforward measure)⁵ is a measurable function which is used to transfer a measure from one measurable space to another — analogous to changing variables. Given two measurable spaces Ω_1 and Ω_2 and a measurable mapping $f : \Omega_1 \rightarrow \Omega_2$ the pushforward of probability measure $\mu : E_1 \rightarrow [0, 1]$ is $f_{\#}(\mu) : E_2 \rightarrow [0, 1]$ such that $f_{\#}(\mu)(B) = \mu(f^{-1}(B))$ for $B \in E_2$ [42].

Probability density function: A probability density function, $I_{\mu}(x)$, determines the probabilities you assign to a given set of events with respect to a given probability measure, μ . Specifically, it is used to specify the probability that the random variable, x , will fall within a particular range of values, $[x_L, x_U]$: $P(x \in [x_L, x_U]) = \int_{x_L}^{x_U} I_{\mu}(x) dx = \int_{\mu(x_L)}^{\mu(x_U)} d\mu(x)$ [37].

Cumulative distribution function: The cumulative distribution function (CDF), $F_X(x)$, gives the probability that a real-valued random variable X will take on a value that is less than or equal to x . For continuous random variables, the CDF is usually denoted as $F(x)$ and is closely related to the probability density function, $I(x)$, via the fundamental theorem

⁵Note that this is sometimes also called “image measure” in the literature.

of calculus: $I(x) = \frac{dF}{dx}$, provided $I(x)$ is differentiable. Inverting this relation gives an expression for $F(x) = \int_{-\infty}^x I(t)dt$ [37].

Equipped with these definitions, we are now ready to define the Wasserstein distance and discuss its practical computation. This discussion largely summarizes work from the following sources [14, 36, 37, 43, 44].

Let (Ω, d) be a d -dimensional probability space. Let $P_p(\Omega)$ be the set of Borel probability measures defined on (Ω, d) with finite p^{th} -moment. Let $\mu \in P_p(X)$ and $\nu \in P_p(Y)$ for $X, Y \subseteq \Omega$. Let the corresponding probability density functions be $I_\mu(x)$ and $I_\nu(y)$. Recall that this means that $d\mu = I_\mu(x) dx$ and $d\nu = I_\nu(y) dy$.

Then the p -Wasserstein (typically abbreviated to Wasserstein) distance, for $p \in [1, \infty)$, is defined as

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{X \times Y} c(x, y) d\gamma(x, y) \right)^{\frac{1}{p}} \Leftrightarrow (\inf \mathbf{E}[c(x, y)])^{\frac{1}{p}} \quad (3.1)$$

where $c(\cdot, \cdot)$ is the cost function. The right-hand side shows the equivalent, and more practically useful, definition in terms of the expectation value, $\mathbf{E}[\cdot]$, of the cost function. The cost function is commonly chosen to be $\|x - y\|^p$ when Ω is a compact subset of \mathbb{R}^d . Unless stated otherwise, we will assume $c(x, y) = \|x - y\|^p$ in the rest of this chapter. $\Gamma(\mu, \nu)$ is the set of transportation plans to take μ to ν which, for $\gamma \in \Gamma(\mu, \nu)$, satisfy

$$\gamma(A \times Y) = \mu(A) \Leftrightarrow \int \gamma(x, y) dy = \mu(x) \quad (3.2)$$

$$\gamma(X \times B) = \nu(B) \Leftrightarrow \int \gamma(x, y) dx = \nu(y), \quad (3.3)$$

for Borel subsets $A \subseteq X$ and $B \subseteq Y$.

Intuitively, the first of these relations is saying that you cannot remove more dirt out of x

than was there to begin with. And the second is saying you must only place enough dirt at any point y to get the desired distribution shape. The total amount of dirt moved out of (into) an infinitesimal region in the initial (final) pile is $\mu(x) dx$ ($\nu(y) dy$). Said another way, the joint transport plan, $\gamma(x, y)$, must marginalize to μ or ν , respectively.

When μ and ν are absolutely continuous with respect to the Lebesgue measure this definition can be rewritten as

$$W_p(\mu, \nu) = \left(\inf_{f \in MP(\mu, \nu)} \int_X c(x, f(x)) d\mu(x) \right)^{\frac{1}{p}} \quad (3.4)$$

where $MP(\mu, \nu) = \{f : X \rightarrow Y \mid f_{\#}\mu = \nu\}$. Where $f_{\#}$ is a pushforward of measure μ .

Note that calculating the Wasserstein distance in Eq. (3.1) involves optimizing over all transportation plans (or equivalently over all $MP(\mu, \nu)$ in Eq. (3.4)). The number of possibilities will thus grow with the dimensionality of the probability measures, making finding the optimal mapping increasingly intractable.

However, for one-dimensional, continuous probability measures there is a unique, monotonically increasing transport (pushforward) map from μ to ν which minimizes the cost. Knowing this mapping a priori removes the need to optimize. This map is $f(x) = F_{\nu}^{-1}(F_{\mu}(x))$ where F_{μ} and F_{ν} are the cumulative distribution functions (CDF) of I_{μ} and I_{ν} respectively. In this case, the p-Wasserstein distance can now be written as

$$W_p(\mu, \nu) = \left(\int_0^1 c(F_{\mu}^{-1}(z), F_{\nu}^{-1}(z)) dz \right)^{\frac{1}{p}} \quad (3.5)$$

where $z := F_{\mu}(x)$ (and thus $x = F_{\mu}^{-1}(z)$).

Oftentimes, there is no known analytic form for the CDFs (or for the probability density functions). In such cases where only samples from I_{μ} and I_{ν} are available, one can approximate the inverse CDF $F_{\mu}^{-1}(z)$ ($F_{\nu}^{-1}(z)$) by sorting the samples of I_{μ} (I_{ν}) in ascending order.

Let these sets of sorted samples be $\{x_i \mid x_i \leq x_{i+1}\}$ and $\{y_i \mid y_i \leq y_{i+1}\}$ respectively. The cost function is then calculated on each pair of samples $\{x_i, y_i\}$ and the result is averaged over the number of samples to approximate Eq. (3.5).

The OT problem solution in Eq. (3.5) allows us to easily calculate the Wasserstein distance for one-dimensional probability density functions, but in higher dimensions the computation is still intractable. One approximation strategy, tries to leverage the special one-dimensional case to solve the problem more generally.

The idea behind the Sliced Wasserstein (SW) distance is to convert the problem of calculating the Wasserstein distance between higher-dimensional probability density functions into many calculations of the simple one-dimensional Wasserstein distance. The high-dimensional probability density functions are "sliced" along many one-dimensional axes (i.e. projected onto that dimension). The one-dimensional Wasserstein distance is then calculated along this dimension. This is done for many slices and the results are averaged. This averaged result can be shown to be a good approximation to the Wasserstein distance [36, 37] and is computationally tractable.

We define a projection mapping $P_\theta : \mathbb{R}^d \rightarrow \mathbb{R}$ where $\theta \in \mathbb{S}^{d-1}$ is the unit vector passing through a point on the d-dimensional unit sphere, \mathbb{S}^{d-1} . This mapping is a pushforward of a measure, ρ , defined on \mathbb{R}^d to a measure, ρ' , defined on \mathbb{R} i.e. $P_{\theta\#}(\rho) = \rho'$. The 2-Sliced Wasserstein (Sliced Wasserstein) distance is then defined as

$$SW_2(\mu, \nu) = \left(\int_{\mathbb{S}^{d-1}} W_2^2(P_{\theta\#}(\mu), P_{\theta\#}(\nu)) d\theta \right)^{\frac{1}{2}} \quad (3.6)$$

where the integral is with respect to the surface measure on \mathbb{S}^{d-1} .

Part of the reason that the Sliced Wasserstein distance is so useful, is that it, like the Wasserstein distance, is also a true distance metric. Namely, it satisfies (1) the triangle

inequality, (2) symmetry, and (3) the identity of indiscernibles (i.e. $SW_2(\mu, \nu) = 0 \Leftrightarrow \mu = \nu$). The first two are inherited from the Wasserstein distance (and the fact that we are assuming $c(x, y) = \|x - y\|^2$). The “ \Rightarrow ” direction of the last requirement is a bit more difficult to show.

In order to show that $SW_2(\mu, \nu) = 0 \Rightarrow \mu = \nu$, we first note that $SW_2(\mu, \nu) = 0$ implies that $W_2(P_{\theta\#}(\mu), P_{\theta\#}(\nu)) = 0$ for all θ i.e. all slices will be identical). Next, we use the fact that W_2 is a true distance metric, and thus is zero if and only if its arguments are equal i.e. $P_{\theta\#}(\mu) = P_{\theta\#}(\nu)$. Therefore, we have $SW_2(\mu, \nu) = 0 \Rightarrow P_{\theta\#}(\mu) = P_{\theta\#}(\nu)$. Intuitively, one can imagine that if the projections of μ and ν are equal for every angle θ , then μ and ν must be equal. But to prove this rigorously, we use the fact that $P_{\theta\#}(\cdot)$ is a Radon Transform and thus is a bijective mapping (i.e. invertible), which immediately implies that $\mu = \nu$. \square

Moreover, what makes the Sliced Wasserstein distance a valid approximation of the Wasserstein distance is the fact that it obeys the following relation: There exists a constant $\alpha > 0$ such that for $\beta = (2(d + 1))^{-1}$

$$SW_2(\mu, \nu) \leq W_2(\mu, \nu) \leq \alpha [SW_2(\mu, \nu)]^\beta \quad (3.7)$$

therefore, minimizing $SW_2(\mu, \nu)$ will also minimize $W_2(\mu, \nu)$. Note also that, when restricting the domain of our measures to compact sets of \mathbb{R}^d , these distances induce the same topology [36, 43, 45].

Practically, calculating the Sliced Wasserstein distance (replacing all integrals with a finite-sample average) is far more computationally efficient than computing the Wasserstein distance directly. The Sliced Wasserstein distance swaps its computational dependence on the dimensionality of the problem, d , for a dependence on the number of Monte-Carlo samples from the distributions, M , and the number of slices used, N . This can be as fast as $N \times \mathcal{O}(M)$ but is at worst $N \times \mathcal{O}(M \log M)$. Theoretically, increasing both the number of

slices and the number of samples will improve the accuracy of the approximation. However, practically, increasing M (for sufficiently large, fixed N) improves the accuracy far more than an analogous increase in N for sufficiently large, fixed M . This empirical observation is supported by several works looking to improve the information given by slices [37] or use mild assumptions to speed up calculations [46].

3.2 Machine Learning

Machine learning (ML) is a topic which has become ubiquitous in nearly every scientific field. A main reason for this is the sheer variety of problems that ML methods can solve. Moreover, if a ML method does not yet exist for a scientific problem, chances are it will soon be invented. A complete overview of the various kinds of ML is out of the scope of this thesis, and will likely be outdated in a matter of months. However, here are a few resources which the interested reader should investigate [47–49].

At its core, ML is the practice of systematically training a machine to solve a problem for which we cannot program a solution directly, but know when we have succeeded (i.e. we have a measure of success). For example, say you were tasked with sorting millions of images into two piles: cat images and dog images⁶. Sorting these images by hand would take forever, so you decide to write an algorithm to distinguish cat images from dog images. You sit down at the computer to write the algorithm... but where do you start? How do you convey to the computer that it should identify shapes within the pixels of the images? And beyond that, how do you define a given shape as “cat-like” or “dog-like”? Suddenly, the thought of sorting these images by hand seems like it would be faster!

For a human, it is easy to tell whether an image is of a cat or a dog, but it is difficult to process images quickly (which is necessary to complete the task in a reasonable amount of

⁶In fact, this very task is a common introductory ML problem.

time). The strategy of how to implement ML in this scenario is this: use your slow but precise human-intuition to correctly label a limited set of images as “cat” or “dog”, then train a machine to do this faster. From this set of examples, you can systematically train a machine to predict the label for a given image. This training is done by measuring the machine’s success rate (i.e. what percentage of the time does the predicted label match the true label) and passing this score to the machine so that it can make better predictions next time. Iterating this process eventually results in a fast, nearly perfect classifier capable of sorting millions of images at lightning speeds.

While the range of methods which fall under the umbrella of ML is vast, much of the recent success of ML methods lies in the use of artificial neural networks⁷. This is largely due to the fact that networks are extremely flexible function approximators capable of mapping a wide variety of types of inputs to a wide variety of types of outputs [48]. In the example above, the inputs are images and the output is a label, 0 for “cat” and 1 for “dog”.

In the rest of this thesis, all ML methods that we discuss will consist of different types of network architectures (inputs, outputs, and how inputs are transformed into outputs within the network) which are stitched together and trained in various creative ways. In this section, we focus specifically on generative networks, where the output is a continuous quantity (rather than a label as in the above example), which are trained using unsupervised learning (no known input/output pairs⁸). In Chapter 4, we show how this kind of method could be used for making a fast, data-driven simulator LHC events to use in simulation-based inference [17].

⁷Sometimes you will see this abbreviated as ANNs to distinguish them from physical neural networks in biological systems. But in this thesis we sacrifice specificity for fewer acronyms and will simply call them “networks”.

⁸To emphasize just how tricky this is, we are trying to get the network to learn a function, $f(x)$, when we do not know what the output, $y = f(x)$, should be for a given specific input, x .

3.2.1 Unsupervised Generative Machine Learning

In supervised ML, the goal is to learn to approximate a function, f , by comparing the output of the network, $\tilde{f}(z)$, and $f(z)$ for a given input z . This is analogous to performing a functional fit, except our network is not parameterized a priori and thus is more flexible [50].

Put simply, unsupervised ML is the process of trying to learn a function, f , of some input, z , without any known examples of pairs, $\{z, f(z)\}$. In other words, for a specific input, we do not know what the output should be, but we want to try to learn $f(z)$ anyways. This sounds like it very well could be a hopeless cause, but we just need to pick a different kind of objective.

While we might not know the specific value $f(z)$ for each given z we might still have information on a *distribution* level. Namely, for a set of samples $\{z\}$ we know approximately how the set of samples $f(\{z\}) := \{f(z_i) \mid z_i \in \{z\}\}$ is distributed. Our objective is thus to try to learn \tilde{f} such that the distributions match, i.e. $\tilde{f}(\{z\}) \approx f(\{z\})$. Then, using a new (but statistically identical) set of samples $\{z'\}$, we can *generate* samples of the desired final distribution i.e. $\tilde{f}(\{z'\})$.

Formally, this problem is ill-posed, as there could be many mappings of $\{z\}$ that produce the same final distribution, $f(\{z\})$, so there is no guarantee that, for a finite set of examples $\{z\}$, we will learn the exact function f . However, even this ill-posed strategy might be good enough to solve our desired task. This is especially true if there is knowledge about the general behavior of f which we can give to the network — so-called inductive biases [50].

In the world of unsupervised, generative ML there are two main strategies from which many interesting derivatives and adaptations arise.

The first is Generative Adversarial Networks (GANs) [51]. The objective of GANs is to pit two networks against each other. The first network (often called the generator) is tasked

with transforming Gaussian distributed noise, $\{\epsilon\}$, into a distribution, $\tilde{f}(\{\epsilon\})$, which matches $f(\{z\})$ for a fixed underlying set of $\{z\}$. The second network is tasked with distinguishing between the real distribution, $f(\{z\})$, and the fake distribution, $\tilde{f}(\{z\})$. The first network is rewarded when it “fools” the second. And the second network is rewarded when it “catches” the fake distribution.⁹ However, note that the GAN has not learned a function of z but rather a function of ϵ which mimics the distribution $f(\{z\})$; in other words, it is not *conditioned* on z , so for this to work, the underlying distribution of $\{z\}$ must stay the same (up to statistical fluctuations).

The other strategy is probabilistic autoencoders [31, 52]. But first, we must describe autoencoders (AEs) [53]. AEs also utilize two networks, but instead of pitting them against one another they are instead encouraged to work collaboratively towards a solution. The first network is called the encoder, $E(\cdot)$; it maps a data sample $x \sim f(\{z\})$ to a *latent* space, creating an “encoded” representation of it. The second network is called the decoder, $D(\cdot)$; it maps the encoded sample back to the data-space (“decodes” it) and tries to match it to the initial input, $x \approx D(E(x))$. When the dimension of the latent space matches that of the data space, the ideal mapping $D(E(\cdot))$ is the identity mapping. But commonly, the latent space has a smaller dimension than the data space and thus is a means of compressing the data, x .

Now say you want to encode a new set of data drawn from the same distribution, $\{x' \mid x' \sim f(\{z\})\}$, with what accuracy will $x' \approx D(E(x'))$? One would hope that the accuracy would be high, since $\{x'\} \approx \{x\}$ up to statistical fluctuations. However, this is not necessarily the case when the latent distribution produced by the encoder is unconstrained.¹⁰ Constraining the latent space to match a certain distribution can provide incentive for the AEs solution

⁹Unsurprisingly, this adversarial game is often described using a counterfeiter/cop analogy.

¹⁰As a pathological example, a perfect AE could map each individual sample x to its own latent coordinate — thus ensuring that it can perfectly encode and decode each sample $x \sim \{x\}$. As long as a sample x' belongs to the training set, $\{x\}$, the reconstruction will be perfect. But there is no guarantee that a sample x' which is *close* to a sample x will be mapped similarly. In other words, there is no incentive for the encoding and decoding mappings to generalize.

to generalize (i.e. be able to handle new, but related, data). Additionally, now that we know what distribution the latent space *should* have, after training the AE, we can pass different samples from this distribution through the decoder to generate new data samples. This is the main goal of probabilistic autoencoders. Note that a typical choice for the latent distribution is a Gaussian, in which case the decoder network is, in essence, the same as the generative network in the GAN case — mapping from Gaussian distributed noise to generate new data samples. The difference lies in how these two generative networks are trained.

So for both GANs and probabilistic autoencoders, we have the ability to mimic our data distribution, $\{x\} = f(\{z\})$, for a fixed underlying distribution of $\{z\}$. But we still have yet to solve our initial task, which is to learn a function, $\tilde{f}(z) \approx f(z)$, which is conditioned on, z . For GANs, there are two options. The first is called a conditional GAN [54] but this sacrifices its unsupervised setup to achieve this conditional goal.¹¹ The second stays unsupervised by simply changing the initial Gaussian distribution to the distribution of the desired inputs $\{z\}$. And while this would fit our requirements, practically speaking, training GANs can be difficult and finicky due to the training’s adversarial nature [55].

Therefore, in this thesis we focus on achieving this goal with probabilistic autoencoders. This turns out to be both practically easier and have several other attractive features. We give more details in the next section.

3.2.1.1 Probabilistic Autoencoders using Wasserstein Distance

As a reminder our goal is to learn $\tilde{f}(z) \approx f(z)$. Ideally, this function maps a sample z to the target output $x = f(z)$. In an autoencoder setup, we want one of our networks to learn to approximate this map from z to x . The natural choice is trying to force our decoder to

¹¹Namely, a conditional GAN (cGAN), in addition to the Gaussian distributed noise, also feeds in $\{z\}$, thus making the learned function dependent on the inputs as desired. However, training a cGAN uses known pairs, $\{z, f(z)\}$, to ensure that learned function is correctly conditioned on the input, z .

be this map, $\tilde{f}(\cdot) := D(\cdot) \approx f(\cdot)$. This immediately signals what we need to choose as our latent space — we want this latent space to align with our desired input. In particular, the latent space is the space of possible values of z and the distribution we draw from is the distribution $\{z\}$.

Now the question is, what do we choose as our objective in order to encourage the distribution of the output of the encoder $\{\tilde{z}\} := \{\tilde{z}_i \mid \tilde{z}_i = E(x_i)\}$ to match the desired distribution $\{z\}$? Namely, how do you minimize the difference between two (possibly high-dimensional) probability distributions?

One popular strategy is called a Variational Autoencoder (VAE) [52], which uses the KL-divergence to constrain the latent space. This strategy comes along with several restrictions. For example, the desired distribution $\{z\}$ must have an analytic, parameterized form for its probability density distribution, $p(z)$, in order to practically calculate the KL-divergence during training. As alluded to in the last section, the KL-divergence also has some pathological failings, which also make this unattractive (see references [36] and [37] for more details). Lastly, and arguably most importantly, in a VAE the distributions being matched effectively spoils the conditionality between the latent and data spaces [1, 31]. This is very problematic as conditionality was the whole goal in the first place!

To see how this conditionality is spoiled, we need to look at the latent-space matching term in the VAE objective. It attempts to minimize the KL-divergence between the Encoding distribution, $p_E(z \mid x)$,¹² and the latent prior, $p(z)$. The problem is that this is encouraging *every* sample x to be mapped to the *whole* latent distribution — this effectively spoils the conditionality of z on x in the encoder mapping (and by extension the conditionality of x on z in the decoder mapping). Instead, what we really want is to match the *marginalized* encoding distribution $p_E(z)$ to the desired latent distribution $p(z)$. See Fig. 3.1 for a visual description.

¹²Note that $E(x)$ is essentially a sample drawn from $p_E(z \mid x)$.

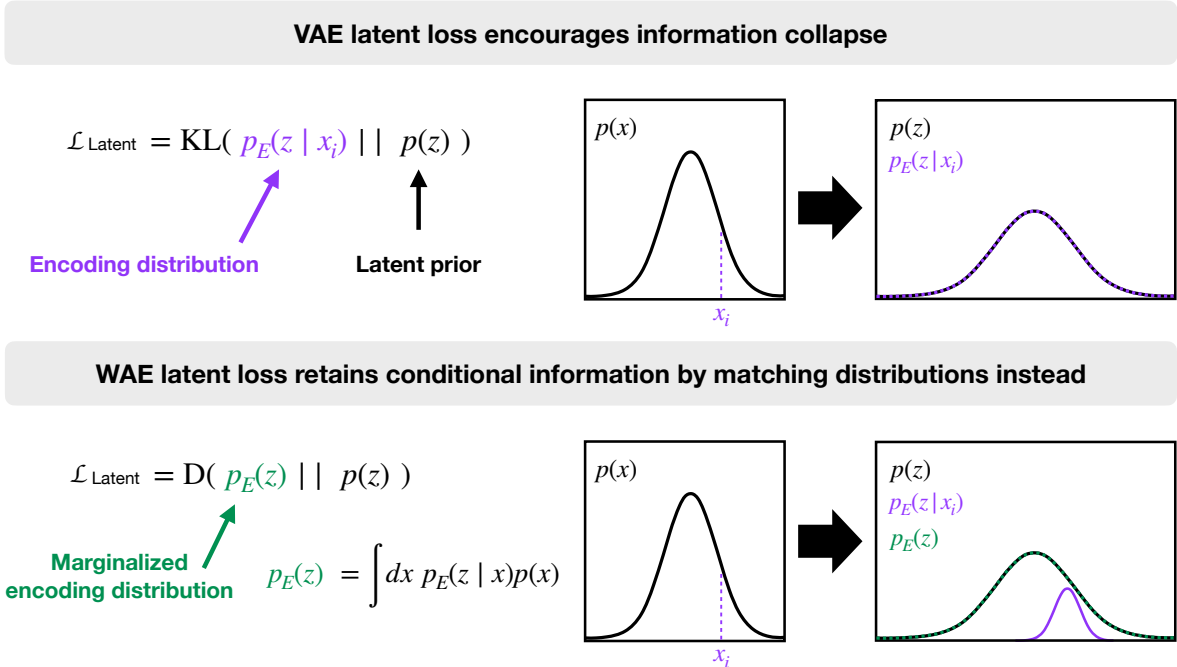


Figure 3.1: **Schematic diagram showing a failing of VAEs.** The latent space loss function of a VAE matches the encoding distribution, $p_E(z | x)$, to the latent space prior, $p(z)$. This encourages every data sample, x_i , to be mapped to the entire latent space prior $p(z)$, which inadvertently results in an information collapse between the data space, \mathcal{X} , and latent space, \mathcal{Z} and spoils the conditionality of the encoder (decoder) mapping on x (z). On the other hand, the latent space loss function of a WAE fixes this problem by encouraging the matching of the marginalized encoding distribution, $p_E(z)$, to the latent space prior, $p(z)$, instead. Every data sample, x_i , is matched to a sub-distribution, $p_E(z | x_i)$ which conspire to build the learned latent distribution, $p_E(z) \approx p(z)$.

Fortunately, there is another strategy which tries to minimize the distance between $p_E(z)$ and $p(z)$ while also avoiding the other issues mentioned above. This strategy is Wasserstein Autoencoders (WAEs) [31], and for cases where there is no analytic formulation of $p(z)$ its derivative Sliced Wasserstein Autoencoders (SWAEs) [36].¹³ WAEs reformulate the autoencoding objective using OT theory — ultimately aiming to minimize the Wasserstein distance between the marginalized decoded distribution, $p_D(x)$, and the data distribution, $p(x)$, while also forcing matching between $p_E(z)$ and $p(z)$. Several possible strategies to measure the difference between $p_E(z)$ and $p(z)$ are suggested, but the formulation, in principle, allows for

¹³Note that there is another derivative strategy called Sinkhorn Autoencoders (SAEs) [56] which can also be applied in cases where there is no analytic form for $p(z)$. These strategies are largely comparable, with some trade-offs between accuracy and computational efficiency [1]. In this thesis, we focus on SWAEs.

the use of any type of difference between $p_E(z)$ and $p(z)$ (under minor assumptions) [31].

SWAEs took this a step further, arguing that the difference between $p_E(z)$ and $p(z)$ should also be OT-based i.e. some approximation of the Wasserstein distance between $p_E(z)$ and $p(z)$. In particular, SWAEs chose to use the Sliced Wasserstein distance approximation of the Wasserstein distance, because it does not require a known analytic form for $p(z)$. This choice allows this method to be applied to a wide variety of kinds of problems. In Chapter 4 we illustrate one such application where $p(z)$ represents the physically-meaningful theoretical model space for fundamental particle collisions.

3.3 Practical Tools in Quantum Field Theory

Quantum field theory (QFT) is currently the main mathematical tool used to describe fundamental particle interactions. It arose out of a desire to reconcile quantum mechanics and relativity. In quantum mechanics, time, t , is treated as a coordinate and space, x , as an operator. But relativity tells us that space and time are not quite so different and, in fact, are related in order to keep the physics in all inertial frames the same (Lorentz covariance). Therefore, treating them differently (one as a coordinate and the other as an operator) makes it impossible to write a theory that is consistent with both quantum mechanics and relativity. We are therefore left with two options: promote t to an operator, or demote x to a coordinate. QFT is the latter choice,¹⁴ where particle states are now functions called fields, which take a different value at each spacetime, (t, x) , coordinate.¹⁵

A full review of QFT is out of the scope of this thesis work,¹⁶ so instead we aim for the much more manageable goal of reviewing a few practical QFT techniques which are used to obtain

¹⁴For a discussion on why the former choice is problematic see Ref. [57].

¹⁵At risk of being overly pedantic, note that the name “field” is chosen in the calculus sense (i.e. scalar or vector field) and not in the algebraic sense.

¹⁶The interested reader is encouraged to consult the following resources [57, 58].

the results in Chapter 5. We first schematically review the advantages of using Effective Field Theories (EFTs) in practical calculations. Next we give a historical and pedagogical demonstration of how an EFT (along with Spontaneous Symmetry Breaking) can accurately predict the spectrum and interactions of composite particles (Chiral Perturbation Theory). This demonstration serves as a simple framework to give context to the calculations done in Chapter 5.

3.3.1 Effective Field Theory

The idea of Effective Field Theory (EFT) comes from the observation that, for many physical systems, having separate theories to describe the physics at separate scales works remarkably well. For example, one does not need quantum mechanics to describe a ball rolling down a hill, but would need it to describe the behavior of an atom in that ball. In this example, the relevant scale is distance. For sufficiently small distances we need one theory (quantum mechanics), but as we consider larger and larger distances, at some point we swap our quantum theory for a classical one because it more easily describes the behavior of the system. The key to this strategy working lies in the observation that the quantum behavior has negligible effects on the macroscopic (classical) behavior, allowing us to use whatever theory corresponds to the distance scale under consideration. The hope is that theories of particle interactions follow this same general schematic; at high energies (short distances) you have one theoretical description, termed the UV theory, and at low energies (larger distances) you have another description, termed the IR theory.

As a brief aside, recent work [59, 60] is actually questioning whether the assumption that UV and IR particle theories can be generally separated, or if there are cases where there exists information which connects the two (so called, UV/IR mixing).¹⁷ This is not completely

¹⁷Questioning this assumption is being done to help explain an aesthetic problem in many particle theories called *naturalness*.

unprecedented either; turbulent flow systems are an example of how small-scale differences do not average out and can greatly affect macroscopic properties [61]. However, there is no questioning the historical success of using EFTs to make sense of experimental results and formulate more predictive theoretical descriptions. In fact, one could argue that constructing the current Standard Model of particle physics was made possible by leveraging this paradigm.

The attractive feature of an EFT is that it allows you to have a predictive, analytic theory as long as you stay within a valid (energy) scale window. Functionally, the only alternative is to have a UV theory which you use to explicitly calculate IR results. However, this can come with theoretical difficulties which necessitate approximate solutions that require prohibitively expensive computational tools. The best example of this is the high-energy theory quantum chromodynamics (QCD). At sufficiently low energies calculations become non-perturbative¹⁸ meaning analytical calculations are intractable and instead must be performed numerically using e.g. lattice QCD.¹⁹ The EFT alternative to this is Chiral Perturbation Theory which we describe in more detail in the next section. This low-energy description is formulated in terms of hadrons which behave as asymptotically free, interacting particles at these energies. Because of this, as long as we are in the low-energy regime, analytic QFT calculations can be performed.

In short, EFTs are a game of segmenting your theory into a spectrum of theories at different energy levels. Each level has a different cast of particles whose interactions can be calculated analytically using QFT. Chiral Perturbation Theory is one very successful example of how an EFT can be used to analytically calculate particle interactions at low-energies.

¹⁸Intuitively, to perform QFT calculations, one must assume that particle states are asymptotically free. In QCD this is only true of the quarks and gluons at high energies. At low enough energies, the quarks and gluons are no longer free particles and are instead confined into hadrons.

¹⁹Lattice QCD, and in general Lattice QFT, is a computationally intensive task which discretizes spacetime in order to explicitly perform QFT calculations.

3.3.2 QCD Confinement and Chiral Symmetry Breaking

A quintessential example of the necessity of having drastically different theories to describe particle interactions at different energy scales can be found in quantum chromodynamics (QCD). At high energies, QCD describes the interactions between quarks and gluons (partons). At these high energies these particles interact freely. However, at low energies there are no free quarks and gluons; they have been replaced by another cast of particles called hadrons, e.g. the protons and neutrons in the atom.

Since these two theories have drastically different degrees of freedom (partons vs hadrons), we will need two different theories to describe them. This is exactly the situation where an EFT description really shines. More precisely, QCD becomes non-perturbative at low energies, so if we want to do analytic QFT calculations, we need a low-energy EFT which can describe the interactions between hadrons. However, we would also like to understand how this low-energy theory relates to the high-energy QCD.

Historically, this understanding came from using another technique called Spontaneous Symmetry Breaking (SSB) to help understand how (the lightest) hadrons are (to a first-order approximation) constructed from quarks that have been confined — bound together by gluons. To leading order, the kind of quarks which are bound together will describe which type of hadron you will get — uud and you get a proton, udd and you get a neutron, and so on. Chiral Perturbation Theory (ChPT) is the low-energy EFT which was used to understand the IR/UV connection for the lightest hadrons.

ChPT is one of the great successes of an EFT description. Historically, it helped to make sense of a messy spectrum of hadrons by predicting their spectrum and properties in a straightforward way from QCD. This strategy has been applied to consider the confinement due to other forces (i.e. the SSB of a different gauge symmetry) [2, 11] which might have happened early in the Universe's history and thus have had downstream cosmological effects.

Chapter 5 is one example of this; it investigates the effects of having a phase of Electroweak force confinement in the early universe on the relic abundance of a WIMP Dark Matter candidate.

In the rest of this section, we review a simplified version of ChPT to describe the lightest hadrons (pions), closely following existing resources on the subject [57, 61–63]. The framework of this strategy is schematically equivalent to what is done in Chapter 5.

Our goal is to describe how the lightest hadrons (pions) arose out of QCD via the spontaneous breaking of the chiral symmetry in the QCD Lagrangian. This spontaneous breaking confines the two lightest quarks (u and d), in various combinations, into the three pions (π^+ , π^- , π^0).

For simplicity, we will only consider the u and d flavors of quarks. We then begin by noting that if we neglect the pion masses,²⁰ the QCD Lagrangian, \mathcal{L}_{QCD} , is invariant under the chiral transformation $SU(2)_L \times SU(2)_R$.²¹ Namely,

$$\mathcal{L}_{QCD} = -\frac{1}{4}F_{\mu\nu}^a F^{a\mu\nu} + i\bar{u}\not{D}u + i\bar{d}\not{D}d = -\frac{1}{4}F_{\mu\nu}^a F^{a\mu\nu} + i\bar{Q}\not{D}Q \quad (3.8)$$

where $F_{\mu\nu}^a$ is the gluon field strength tensor and $Q := (u, d)^T$ (and $\bar{Q} := (\bar{u}, \bar{d})$). We have also neglected the very small electromagnetic interaction under which the u quark has a different charge than the d quark (i.e. the covariant derivative $D^\mu \approx \partial^\mu - ig_s T^a A_a^\mu$). Pedagogical details of how to see that this is invariant are in Appendix A .

Under these assumptions, this $SU(2)_L \times SU(2)_R$ is an exact symmetry of the high-energy theory (\mathcal{L}_{QCD}). The question now is what happens to this symmetry at low energies. If this symmetry remains exact in the low-energy theory, we would expect to see them (ap-

²⁰We will discuss what happens when we include them later.

²¹Note that in fact it is invariant under the symmetry $U(2)_L \times U(2)_R$ which can be decomposed as follows, $U(2)_L \times U(2)_R = SU(2)_L \times SU(2)_R \times U(1)_L \times U(1)_R$. Note that the $U(1)_L \times U(1)_R$ piece is typically denoted as $U(1)_V \times U(1)_A$ for “vector” and “axial” respectively. It turns out that the $U(1)_A$ symmetry is anomalous — present at tree-level but broken at loop-level — and thus, in some sense, is not a valid symmetry of the problem. On the other hand, $U(1)_V$ is a valid symmetry and can be identified with baryon number. In what follows, we will be focusing on just the $SU(2)_L \times SU(2)_R$ piece.

proximately) manifest in the properties of hadrons. If it is broken spontaneously, we would expect to see massless Goldstone Bosons.

What we find is that the hadrons do not retain this chiral symmetry, and we do have a set of particles which seem to be (approximately) Goldstone Bosons. Specifically, the pions are not exactly massless (as Goldstone Bosons should be) but they are much lighter than the other hadrons which are composed of u and d quarks.²² The fact that the pions are not massless is not all that surprising since the u and d quarks are not really massless. In other words, the chiral symmetry of \mathcal{L}_{QCD} is only approximate, it is explicitly broken by the masses of the u and d quarks.²³ However, this approximate symmetry can still be spontaneously broken, and therefore must give rise to pseudo Goldstone Bosons (the pions). If the chiral symmetry were exact the pions would be massless Goldstone Bosons but since it is not exact the pions get a mass correction.

So what is the residual symmetry in the low-energy theory? Looking at the mass spectrum of the hadrons, it appears that the $SU(2)_L \times SU(2)_R$ has been broken down to $SU(2)$. The different groups of hadron states form different representations of this residual symmetry. For example, the pions form a triplet and the proton and neutron form a doublet. Therefore, we have $SU(2)_L \times SU(2)_R \rightarrow SU(2)$ via spontaneous symmetry breaking.

In order for there to be spontaneous symmetry breaking, we need some combination of high-energy states to get a non-zero vacuum expectation value (vev). This vev must be invariant under the residual symmetry of the theory, $SU(2)$. It must also be a color and electromagnetic force singlet to reproduce the properties of hadrons that are observed (i.e. no bare color charges and electromagnetic charge conservation). Finally, as is always the case, the vev must be a Lorentz scalar in order to not break Lorentz invariance.

²²For reference, the mass of the pions is around 135 MeV (the charged pions are a few MeV heavier) whereas the mass of the proton and neutron is around 1 GeV.

²³Specifically it is broken to the extent that $M := \text{diag}(m_u, m_d) \neq LMR^\dagger$, which breaks the invariance of \mathcal{L}_{QCD} under $SU(2)_L \times SU(2)_R$ transformations.

The winning candidate winds up being $\bar{Q}Q_L = \bar{Q}_i(Q_j)_L = \bar{Q}_i P_L Q_j$.²⁴ Let us choose the vev such that

$$\langle \bar{Q}Q_L \rangle = \langle \bar{Q}_i P_L Q_j \rangle = \delta_{ij} v^3 = \mathbb{1} v^3 \quad (3.9)$$

We can easily see that this is invariant under an $SU(2)$ transformation

$$\mathbb{1} v^3 \xrightarrow{SU(2)} U^\dagger (\mathbb{1} v^3) U = U^\dagger U v^3 = \mathbb{1} v^3 \quad (3.10)$$

$SU(2)$ is said to be a vector subgroup of $SU(2)_L \times SU(2)_R$; physically it corresponds to the nuclear iso-spin symmetry obeyed by hadrons. Note that we cannot exactly predict what value v will take (since the symmetry breaking is non-perturbative) but we expect $v \sim \Lambda_{QCD}$ (the EFT cutoff), since that is the relevant mass scale in the problem.

We now want to use this knowledge to construct a perturbative theory of pions. The method we will use is called a non-linear sigma model (NLSM). The main idea is that we want to construct our effective \mathcal{L} such that it is obvious how it transforms under the UV symmetry ($SU(2)_L \times SU(2)_R$). Doing this allows us to easily ensure that we have accounted for all broken generators

We therefore choose

$$\tilde{\Sigma}(x) = \exp \left[i \frac{\Pi^a(x)}{f_\Pi} T^a \right] \begin{pmatrix} v + \sigma(x) & 0 \\ 0 & v + \sigma(x) \end{pmatrix} \quad (3.11)$$

where f_Π is called the pion decay constant; it has units of [mass] and is introduced so the $\Pi^a(x)$ are canonically normalized. Note that this also includes the field $\sigma(x)$ which parameterizes small perturbations away from the potential minimum after SSB. In general

²⁴Note that some resources use $\bar{Q}Q$ instead. These descriptions are equivalent up to an unphysical phase. If we restrict to a real vev, then the descriptions are exactly equivalent.

this field is quite heavy (near to or greater than the cutoff of our EFT) and thus can be safely ignored in this EFT description. In general, a formulation which keeps only the Goldstone bosons (Π^a) and not the heavy dof (σ) is called an NL σ M.

We therefore define

$$\Sigma(x) := \frac{\tilde{\Sigma}(x)}{v} \Big|_{m_\sigma \rightarrow \infty} = \exp \left[i \frac{\Pi^a}{f_\Pi} T^a \right] \quad (3.12)$$

This may look rather mysterious but the thinking is that if $m_\sigma \rightarrow \infty$ the effect of the field $\sigma(x)$ (on any Feynman diagram) would be suppressed by $\sim (1/m_\sigma^2)$ due to the propagator. Therefore as $m_\sigma \rightarrow \infty$ the effect will $\rightarrow 0$. Therefore, we can safely skip including the field explicitly.

Under the UV symmetry, $SU(2)_L \times SU(2)_R$, $\Sigma(x)$ has a straightforward transformation $\Sigma(x) \rightarrow L\Sigma(x)R^\dagger$. We can now write our effective (IR) lagrangian in terms of $\Sigma(x)$. Which terms we include will depend somewhat on the observed behavior of the hadrons which we are trying to describe, but at the very least, we must include the kinetic term. The kinetic term ignoring EM interactions (i.e. $D^\mu = \partial^\mu$) will be

$$\mathcal{L}_{\text{kinetic}} = f_\Pi^2 \text{Tr} [(\partial_\mu \Sigma(x))^\dagger (\partial^\mu \Sigma(x))]. \quad (3.13)$$

If we expand out the exponential $\Sigma(x) \approx \mathbb{1} + \frac{i}{f_\Pi} \Pi^a T^a + \dots$ and plug this in we will find

$$\mathcal{L}_{\text{kinetic}} = \frac{1}{2} (\partial_\mu \Pi^a) (\partial^\mu \Pi^a) + \mathcal{O}(\Pi^4). \quad (3.14)$$

Note that this contains the usual kinetic terms, but it also contains an infinite number of interactions with well-defined coefficients and characterized by inverse powers of f_Π .

In particular, if we concentrate on the $\mathcal{O}(\Pi^4)$ we will find that it describes interactions like $\Pi\Pi \rightarrow \Pi\Pi$ and comes with a pre-factor $1/(f_\Pi^2)$. By studying pion interactions we can actually

measure this value and find that $f_{\Pi} \approx 92$ MeV. Even though this is non-renormalizable (because of the mass⁻² dimension of the pre-factor), if we are safely in our EFT range, the effects of higher order processes on the value of $1/f_{\Pi}$ can be systematically controlled.

We can estimate the value of the EFT cutoff scale by considering a loop diagram of pion interactions. Comparing it to a tree level diagram, we know that there will be two extra factors of p/f_{Π} corresponding to each "side" of the loop, where p is the momentum in the loop. Schematically, the result of the loop integral (over p) should look like:

$$(\text{tree level contribution}) \times (\text{energy scale of process})^2 \times \left(\frac{1}{f_{\Pi}^2}\right) \times \left(\frac{1}{16\pi^2}\right) \quad (3.15)$$

where the last piece is the usual loop factor.

We would like the loop factor to be smaller than the tree-level results, which means

$$(\text{energy scale of process})^2 \times \left(\frac{1}{f_{\Pi}^2}\right) \times \left(\frac{1}{16\pi^2}\right) < 1 \quad (3.16)$$

$$\Rightarrow (\text{energy scale of process}) < 4\pi f_{\Pi}. \quad (3.17)$$

Our EFT is only well-defined (perturbative) when we have a valid loop expansion (each successive term in the expansion is smaller than the last). Therefore, we can say our EFT is valid at energies, $E < 4\pi f_{\Pi}$, so $4\pi f_{\Pi}$ is our EFT cutoff.

Chapter 4

Foundations of a Fast, Data-Driven, Machine-Learned Simulator

This chapter is heavily based on work previously published in collaboration with Stephan Mandt, Daniel Whiteson, and Yibo Yang [1].

4.1 Introduction

From measuring masses of particles to deducing the likelihood of life elsewhere in the Universe, a common goal in analyzing scientific data is statistical inference — drawing conclusions about values of a theoretical model’s parameters, θ , given observed data, x . The likelihood model of observed data, $p(x|\theta)$, is a central ingredient in both frequentist and Bayesian approaches to statistical inference; however, it is typically intractable, due to the complexity of a full probabilistic description of the data generation process. One way to circumvent this difficulty is to *simulate* experimental data for a given value of the theoretical parameters, θ , from which a probability model of the likelihood, $p(x|\theta)$, can be constructed

and used for downstream statistical inference regarding θ . This is known as simulation-based inference, and has found application across scientific disciplines ranging from particle physics to cosmology [17].

However, traditional approaches to simulation, which attempt to faithfully model complex physical phenomena, can be computationally expensive — a limitation we aim to overcome in this work. In simulation-based inference, experimental data arising from a physical system typically depend on an initial configuration of the system, z , that is unobserved, or belonging to a *latent* space, while the parameters θ govern the underlying mechanistic model. In many cases, the transformation from the latent state to experimental data is non-trivial, involving complex physical interactions that cannot be described analytically, but can be *simulated* numerically by Monte-Carlo algorithms. In particle physics, for example, the parameters θ govern theoretical models that describe fundamental particle interactions. These fundamental interactions produce secondary particles, z , which are not directly observable and often transform in flight before passing through layers of detectors whose indirect measurements, x , can help reconstruct their identities and momenta. The transformation from the unobserved latent space, particles produced in the initial interaction, to the experimental data, is stochastic, governed by quantum mechanical randomness, and has no analytical description.

Instead, Monte-Carlo-based numerical simulations of in-flight and detection processes generate samples of possible experimental data for a given latent space configuration [20, 21, 23–25]. This approach is computationally expensive [24, 25] because it requires the propagation and simulation of every individual particle, each creating subsequent showers of thousands of derivative particles. Additionally, these simulations contain hundreds of parameters which must be extemporaneously tuned to give reasonable results in control regions of the data where the latent space has been well-established by results from previous experiments.

In particle physics, like many other fields in the physical sciences [64–67], the computational cost of numerical simulations has become a central bottleneck. A fast, interpretable, flex-

ible, data-driven generative model which can transform between the latent space and the experimental data would be significant for these fields. Recent advances in the flexibility and capability of machine learning (ML) models have allowed for their application as computationally inexpensive simulators [26, 27, 68–74]. Applications of these techniques have made progress towards this goal but fall short in crucial ways. For example, approaches leveraging Generative Adversarial Networks (GANs) are able to mimic experimental data for fixed distributions in the latent space [26, 68, 69], but are unable to generate predictions for new values of latent variables, a crucial requirement for a simulator. Other efforts condition on latent variables [70] but require training with labeled pairs generated by slow Monte-Carlo generators, incurring some of the computational cost they seek to avoid.

We lay the foundations and provide a proof-of-principle demonstration for Optimal-Transport-based Unfolding and Simulation (OTUS). We use unsupervised learning to build a flexible description of the transformation from latent space, \mathcal{Z} , to experimental data space, \mathcal{X} , relying on theoretical priors, $p(z)$, where $z \in \mathcal{Z}$ and a set of samples of experimental data $\{x \in \mathcal{X}\}$ but, crucially, no labeled pairs, (z, x) . Our model applies a type of probabilistic autoencoder [31, 36], which learns two mappings: encoder (data \rightarrow latent, $p_E(z | x)$) and decoder (latent \rightarrow data, $p_D(x | z)$). Typical probabilistic autoencoders (i.e. variational autoencoders (VAEs) [52, 69]) use a simple, unphysical latent space, \mathcal{Y} , for computational tractability during learning. However, this causes VAEs to suffer from the same weakness as GANs: doomed to mimic the data distribution, $p(x)$, for a fixed physical latent space, $p(z)$, unless the model compromises to requiring expensive simulated pairs (e.g. a conditional VAE approach [75]). OTUS’s innovation is to align the probabilistic autoencoder’s latent space, \mathcal{Y} , with that of our inference task, \mathcal{Z} . With this change, our decoder becomes a computationally inexpensive, conditional simulator mapping $\mathcal{Z} \rightarrow \mathcal{X}$ as well as a tractable transfer function, $p_D(x | z)$. See Fig. 4.1 for a visual description.

For VAEs, identifying \mathcal{Y} with \mathcal{Z} is difficult because the training objective requires the ability

to explicitly compute the latent space prior, $p(y)$ for $y \in \mathcal{Y}$. In particle physics, such explicit computations are intractable. We therefore turn to a new form of probabilistic autoencoder: the Sliced Wasserstein Autoencoder (SWAE) [31, 36], which alleviates this, and other, issues by reformulating the objective using the Sliced Wasserstein distance and other ideas from optimal transport theory. This reformulation lets us identify \mathcal{Y} with \mathcal{Z} and also allows the encoder and decoder network mappings to be inherently stochastic.

We suggest that an SWAE [36] can be used to achieve the broad goal of simulators: learning the mapping from the physical latent space to experimental data directly from samples of experimental data $\{x \sim p(x)\}$ and theoretical priors $\{z \sim p(z)\}$ in control regions. The resulting decoder ($\mathcal{Z} \rightarrow \mathcal{X}$) can be applied as a simulator, generating samples of experimental data from latent variables in a fraction of the time, and probed and visualized to ensure a physically meaningful transformation. Additionally, the decoder’s numerically tractable detector response function, $p_D(x | z)$, would be useful in other applications, such as direct calculation of likelihood ratios via integration [76]. The encoder network’s $\mathcal{X} \rightarrow \mathcal{Z}$ mapping can also be used in unfolding studies [77, 78]. Lastly, the mathematical attributes of the SW distance allow for the inclusion of informed constraints on the mappings.

In this work, we first present background on the problem, the objective, and discuss related work. In Section 4.4, we present the foundations for the OTUS method and discuss steps toward scaling OTUS to a full simulation capable of replacing current Monte-Carlo methods in particle physics analyses. In Section 4.5, we give initial proof-of-principle demonstrations on Z -boson and semileptonic top-quark decays. In Section 4.6, we discuss the details of our methods. We then conclude by discussing directions for future work and also briefly discuss how OTUS might be applied to problems in other scientific fields.

4.2 Theoretical Background

The primary statistical task in particle physics, as in many areas of science, is inferring the value of a model parameter, θ , based on a set of experimental data, $\{x\}$. For example, physicists inferred the mass of the Higgs boson from Large Hadron Collider data [79, 80]. Inference about θ requires a statistical model, $p(x | \theta)$, which can be used to calculate the probability to make an observation, x , given a parameter value, θ . Unfortunately, such analytical expressions are unavailable due to the indirect nature of observations and the complexity of detectors. Previous solutions to this problem have relied on numerical Monte-Carlo-based simulations [20, 21, 23].

Fundamental particle interactions, like the decay of a Higgs boson, produce a set of particles which define an unobserved latent space, \mathcal{Z} . The statistical model $p(z | \theta)$ is usually well-understood and can often be expressed analytically or approximated numerically. However, experimenters only have access to samples of experimental data, $\{x\}$. Therefore, calculating $p(x | \theta)$ requires integrating over the unobserved $\{z \sim p(z | \theta)\}$; namely, $p(x | \theta) = \int dz p(x | z) p(z | \theta)$.

The transfer function, $p(x | z)$, represents the multi-staged transformation from the unobserved latent space, \mathcal{Z} , to the experimental data space, \mathcal{X} . As latent space particles travel they may decay, interact, or radiate to produce subsequent showers of hundreds of secondary particles. These particles then pass through the detector, comprising many layers and millions of sensors resulting in a high-dimensional response of order $\mathcal{O}(10^8)$. Finally, the full set of detector measurements are used to reconstruct an estimate of the identities and momenta of the original unobserved particles in the latent space. For the vast majority of analyses this final, experimental data space, \mathcal{X} , has a similar dimensionality¹ to that of \mathcal{Z} , usually $\mathcal{O}(10^1)$. However, the complex, stochastic, and high-dimensional nature of the transforma-

¹The dimensionality is not necessarily equal due to the imperfect nature of the detection process. For example, \mathcal{Z} may represent four quarks but \mathcal{X} may only contain three jets.

tion makes it practically impossible to construct a closed-form expression for the transfer function $p(x | z)$. Instead, particle physicists use simulations as a proxy for the true transfer function.

To arrive at $p(x | \theta)$, samples of $\{z \sim p(z | \theta)\}$ are transformed via simulations into effective samples of $\{x \sim p(x | \theta)\}$, approximating the integral above. Current state-of-the-art simulations strive to faithfully model the details of particle propagation and decay via Monte-Carlo techniques. This approach is computationally expensive and limited by our poor understanding of the processes involved. Ad-hoc parameterizations often fill gaps in our knowledge but introduce arbitrary parameters which must be tuned to give realistic results using data from control regions, where the underlying $p(z | \theta)$ is well-established from previous experiments, freeing $p(x | \theta)$ of surprises. Examples of control regions include decays of heavy bosons (e.g. Z) or the top quark (t).

The computational cost of current simulations is the dominant source of systematic uncertainties and the largest bottleneck in testing new models of particle physics [81]. A computationally-inexpensive, flexible simulator which can map from \mathcal{Z} to \mathcal{X} such that it effectively approximates $p(x | z)$ would be a breakthrough.

4.3 Objective and Related Work

The development of OTUS was guided by the goals of the simulation task and the information available for training. Specifically, the simulator has access to samples from model priors, $p(z | \theta)_{\text{control}}$, and experimental data samples, $\{x_{\text{control}}\}$. Critically, $\{x_{\text{control}}\}$ samples come from experiments, where the true $\{z_{\text{control}}\}$ are unknown, such that no $(z_{\text{control}}, x_{\text{control}})$ pairs exist. Instead, the distribution of $\{z_{\text{control}}\}$ are known to follow $p(z | \theta)_{\text{control}}$ and the distribution of $\{x_{\text{control}}\}$ is observed.

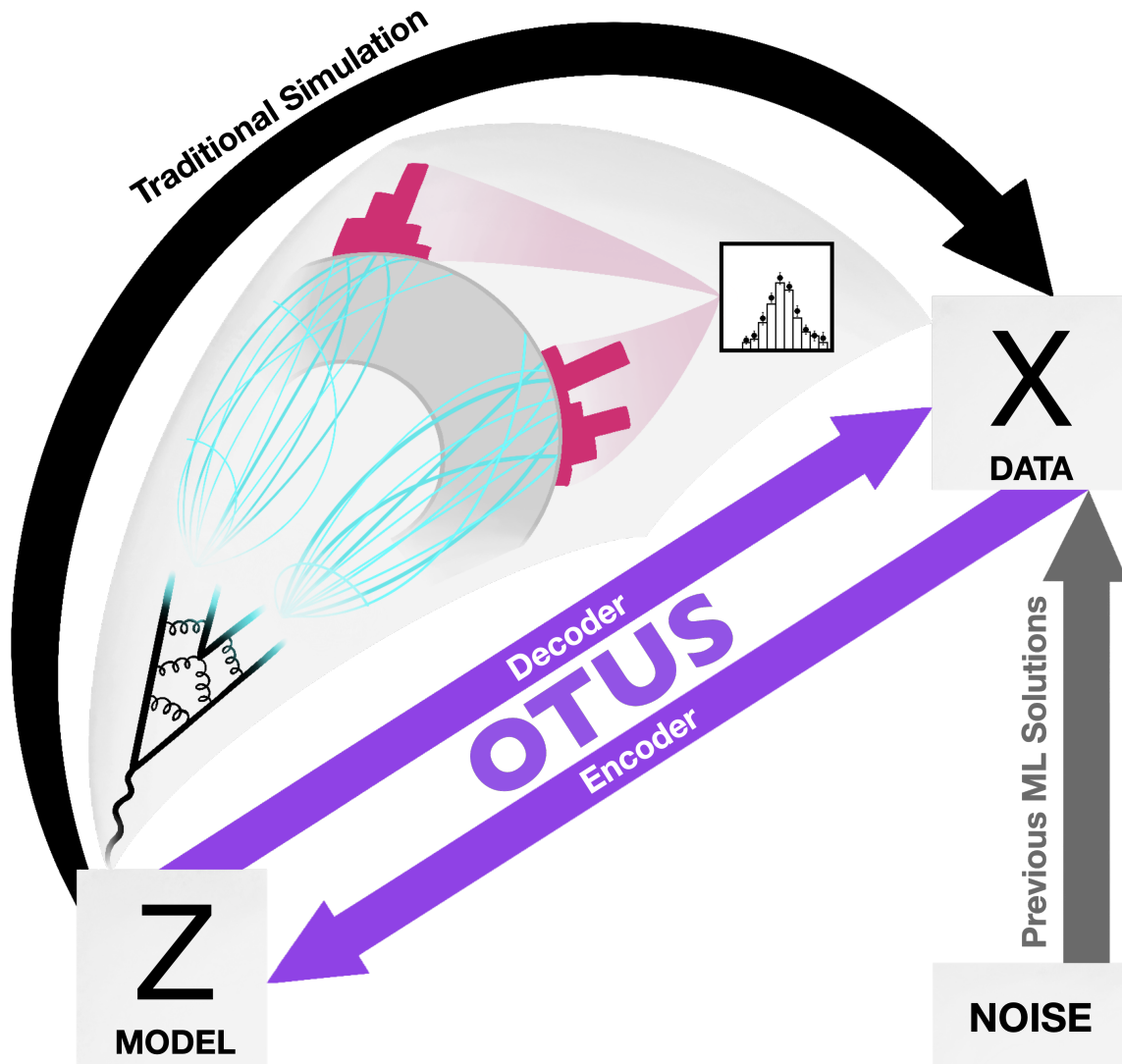


Figure 4.1: **Schematic of the problem and the solution.** Current simulations map from a physical latent space, \mathcal{Z} , to data space, \mathcal{X} , attempting to mimic the real physical processes at every step. This results in a computationally intensive simulation. Previous Machine Learning (ML) solutions can reproduce the distributions in \mathcal{X} but are not conditioned on the information in \mathcal{Z} ; instead they map from unphysical noise to \mathcal{X} , which limits their scope. We introduce a new method which provides the best of both worlds. OTUS provides a simulation $\mathcal{Z} \rightarrow \mathcal{X}$ (Decoder) which is conditioned on \mathcal{Z} yet is computationally efficient. Advantageously, it also inadvertently provides an equivalently fast unfolding mapping from $\mathcal{X} \rightarrow \mathcal{Z}$ (Encoder).

The simulator should learn a stochastic transformation $\mathcal{Z} \rightarrow \mathcal{X}$ such that samples $\{z\}$ drawn from $p(z | \theta)_{\text{control}}$ can be transformed into samples $\{x\}$ whose distribution matches that of the experimental data $\{x_{\text{control}}\}$. Additionally, these control regions should be robust so that the simulator can approximate $p(x | \theta)$ for different, but related, values of θ . Traditional Monte-Carlo simulators such as GEANT4 [21] face related challenges.

The flexibility of ML models at learning difficult functions across a wide array of contexts suggests that these tools could be used to develop a fast simulator. The objectives described above translate to four constraints on the class of ML model and methods of learning. Generating samples of $\{x \in \mathcal{X}\}$ requires a (1) generative ML method. For $z \in \mathcal{Z}$, the simulator maps $z \rightarrow x$ such that the output x depends on the input z , meaning the mapping is (2) conditional. The problem’s inherent and unknown randomness prevents us from assuming any particular density model, suggesting that our simulator should preferably be (3) inherently stochastic. The lack of (z, x) pairs mandates an (4) unsupervised training scheme. Additionally, the chosen method should produce a simulation mapping ($\mathcal{Z} \rightarrow \mathcal{X}$) which is inspectable and physically interpretable.

Generative ML models can produce realistic samples of data in many settings, including natural images. Generative Adversarial Networks (GANs) transform noise into artificial data samples and have been adapted to particle physics simulation tasks for both high-level and raw detector data, which can resemble images [26, 27, 68–70]. However, while GANs have successfully mimicked existing datasets, $\{x\}$, for a fixed set of $\{z\}$, they have not learned the general transformation $z \rightarrow x$ prescribed by $p(x | z)$, and so cannot generate fresh samples $\{x'\}$ for a new set of $\{z'\}$, thus failing condition (2). Other GAN-based approaches [70] condition the generation of $\{x\}$ on values of $\{z\}$, but in the process use labeled pairs (x, z) , which are only obtained from other simulators, rather than from experiments, thus failing condition (4). Relying on simulated (x, z) pairs incurs the computational cost we seek to avoid, and limits the role of these fast simulators to supplementing traditional simulators,

rather than replacing them.

An alternative class of unsupervised, generative ML models are variational autoencoders (VAEs). While GANs leverage an adversarial training scheme, VAEs instead optimize a variational bound on the data’s likelihood by constructing an intermediate latent space, \mathcal{Y} , which is distributed according to a prior, $p(y)$ [82]. An encoder ($\mathcal{X} \rightarrow \mathcal{Y}$) network transforms $x \rightarrow \tilde{y}$, where the $\tilde{}$ distinguishes a mapped sample from those drawn from $p(y)$. Similarly, a decoder ($\mathcal{Y} \rightarrow \mathcal{X}$) network transforms a sample produced by the encoder back to the data space, $\tilde{y} \rightarrow \tilde{x}$. The autoencoder structure is the combined encoder-decoder chain, $x \rightarrow \tilde{y} \rightarrow \tilde{x}$. During training, the distribution of the encoder output, $p_E(y | x)$, is constrained to match the latent space prior, $p(y)$, via a latent loss term which measures the distance between the distributions. At the same time, the output of the autoencoder, \tilde{x} , is constrained to match the input, x , which are compared pairwise. New samples from \mathcal{X} following the distribution of the data, $p(x)$, can then be produced by decoding samples, $\{y\}$, drawn from $p(y)$, via $y \rightarrow \tilde{x}'$.

The form of $p(y)$ is usually independent of the nature of the problem’s underlying theoretical model, and is often chosen to be a multi-dimensional Gaussian for simplicity. This choice provides sufficient expressive power even for complex datasets (i.e. natural images). However, in the particle physics community, optimizing the encoding mapping to match this latent space is seen as an extra, unnecessary hurdle in training [26]. Therefore, GANs have been largely favored over VAEs in the pursuit of a fast particle physics simulator. Some studies investigated VAEs in this context, but retained the unphysical form of $p(y)$ (i.e. multi-dimensional Gaussian) [69, 71, 72], preventing them from being conditional generators, failing requirement (2).

4.4 Proposed Solution

4.4.1 Our Approach: OTUS

In this work, we aim to align the probabilistic autoencoder’s latent space, \mathcal{Y} , with that of our inference task, \mathcal{Z} . This will allow us to learn a conditional simulation mapping from our theoretical model latent space to our data space, $\mathcal{Z} \rightarrow \mathcal{X}$. Therefore, we construct a probabilistic autoencoder where the latent space prior, $p(y)$, is identical to the physical latent space, $p(y) \equiv p(z) = p(z | \theta)$, for the choice of particular parameters, θ . The decoder then learns $p_D(x | z)$ providing precisely the desired conditional transformation, $z \rightarrow x$. Additionally, $p_D(x | z)$ can act as a tractable transfer function in approaches which estimate $p(x | \theta)$ via direct integration [76]. The encoder’s learned $p_E(z | x)$ is of similar interest in unfolding applications [77, 78].

This is not possible with VAEs because optimizing the variational objective requires explicit computation of the densities $p(y)$, $p_E(y | x)$, and $p_D(x | y)$. Therefore, $p(y)$ is often assumed to be a standard isotropic Gaussian for its simplicity and potential for uncovering independent latent factors of the data generation process. However, in particle physics the true prior, $p(z)$, which is governed by quantum field theory, is highly non-Gaussian and computing its density explicitly requires an expensive numerical procedure. Similarly, as we have little knowledge about the true underlying stochastic transforms, assuming any particular parametric density model for $p_E(y | x)$ or $p_D(x | y)$, like a multivariate Gaussian, would be inappropriate and overly restrictive. These concerns led us to use inherently stochastic (i.e. implicit) models for $p(z)$, $p_E(z | x)$, and $p_D(x | z)$ that are fully sample-driven.

Additionally, the VAE objective’s use of KL-divergence introduces technical disadvantages. The KL-divergence, $D_{\text{KL}}(\cdot || \cdot)$, is not a true distance metric, and will diverge for non-overlapping distributions often leading to unusable gradients during training [30, 36]. Moreover, the spe-

cific use of $D_{\text{KL}}(p_E(z | x) || p(z))$ within the VAE loss forces $p_E(z | x)$ to match $p(z)$ for every value of $x \sim p(x)$ [31]. This term must be carefully tuned (e.g. with a β -VAE approach [69, 83]) to avoid the undesirable effect of the encoder mapping different parts of \mathcal{X} to the same overlapping region in \mathcal{Z} , which can be particularly problematic if \mathcal{Z} represents a physically meaningful latent space.

We resolve these issues by applying an emerging class of probabilistic autoencoders, based instead on the Wasserstein distance, which is a well-behaved distance metric between arbitrary probability distributions rooted in concepts from optimal transport theory [31, 36].

The original Wasserstein Autoencoder (WAE) [31] loss function is

$$\mathcal{L}_{\text{WAE}}(p(x), p_D(x | z), p_E(z | x)) = \mathbb{E}_{x \sim p(x)} \mathbb{E}_{p_E(z|x)} \mathbb{E}_{\tilde{x} \sim p_D(x|z)} [c(x, \tilde{x})] + \lambda d_z(p_E(z), p(z)), \quad (4.1)$$

4.1A
4.1B

where \mathbb{E} denotes the expectation operator and $c(\cdot, \cdot)$ is a cost metric. For the optimal $p_E(z | x)$, \mathcal{L}_{WAE} becomes an upper bound on the Wasserstein distance between the true data distribution, $p(x)$, and the decoder’s learned distribution, $p_D(x) = \int dz p_D(x | z) p(z)$; the bound is tight for deterministic decoders.

Term A of Equation (4.1) constrains the output of the encoder-decoder mapping, \tilde{x} , to match the input, x , while term B of Equation (4.1) constrains the encoder mapping. The hyperparameter λ provides a relative weighting between the two terms. The difference between the marginal encoding distribution, $p_E(z) = \int dx p_E(z | x) p(x)$, and the latent prior, $p(z)$, is measured by $d_z(\cdot, \cdot)$.² Unfortunately, the originally proposed options for $d_z(\cdot, \cdot)$ [31] had undesirable features which made them ill-suited for this particle physics problem (see Section 4.6.2.1).

²Comparing $p_E(z)$ and $p(z)$ rather than $p_E(z | x)$ and $p(z)$ is the crucial innovation which allows different parts of \mathcal{Z} to remain disjoint.

The more recent Sliced Wasserstein Autoencoder (SWAE) [36] uses the Sliced Wasserstein (SW) distance as the $d_z(\cdot, \cdot)$ metric. The SW distance, $d_{\text{SW}}(\cdot, \cdot)$, is a rigorous approximation to the Wasserstein distance, $d_W(\cdot, \cdot)$. The SWAE completely grounds the loss function in optimal transport theory as each term and the total loss can be identified as approximating the Wasserstein distances between various distributions and allows $p(y)$ to be any sampleable distribution, including the physical, $p(z)$. Additionally, the (S)WAE method allows the encoder and decoder to be implicit probability models, while avoiding an adversarial training strategy which can lead to problems like mode collapse [55].

Both d_W and d_{SW} are true distance metrics [36]. The KL-divergence and adversarial schemes lack this property resulting in divergences and meaningless loss values which lead to problems during training and make it difficult to include additional, physically-motivated constraints. The Wasserstein distance is the cost to transport probability mass from one probability distribution to another according to a cost metric, $c(\cdot, \cdot)$, following the optimal transportation map. However, it is difficult to calculate for multivariate probability distributions when pairs from the optimal transportation map are unknown. However, for univariate probability distributions, there is a closed-form solution involving the difference between the inverse Cumulative Distribution Functions (CDF⁻¹s) of the two probability distributions. The SW distance approximates the Wasserstein distance by averaging the one-dimensional Wasserstein distance over many randomly selected slices — one-dimensional projections of the full probability distribution [36] (see Section 4.6.3).

The SWAE loss takes the general form of the WAE loss

$$\mathcal{L}_{\text{SWAE}}(p(x), p_D(x | z), p_E(z | x)) = \underbrace{\mathbb{E}_{x \sim p(x)} \mathbb{E}_{p_E(z|x)} \mathbb{E}_{\tilde{x} \sim p_D(x|z)} [c(x, \tilde{x})]}_{4.2A} + \lambda \underbrace{d_{\text{SW}}(p_E(z), p(z))}_{4.2B}. \quad (4.2)$$

Term A of Equation (4.2) compares pairs (x, \tilde{x}) , where \tilde{x} is the output of the encoder-decoder

mapping. In term B of Equation (4.2), matched pairs are not available so we instead use the SW distance approximation. Both loss terms use the cost metric $c(u, v) = \|u - v\|^2$ [36].

The SWAE allows us to train a probabilistic autoencoder that transforms between \mathcal{X} and \mathcal{Z} with a physical prior $p(z)$. However, since we are in an unsupervised setting, the true $p(x | z)$ is unknown. It is therefore crucial to ensure that the learned transformation is plausible and represents a series of physical interactions. To encourage this, we can easily impose supplemental physically-meaningful constraints on the SWAE model. These constraints can be relations between \mathcal{Z} and \mathcal{X} spaces or constraints on the internal properties of these respective spaces. In this work, we use one constraint from each category.

From the first category, we add a term comparing the unit vector parallel to the momentum of an easily identifiable particle in the latent and experimental spaces. This can be thought of as analogous to choosing a consistent basis and can be helpful for problems containing simple inversion symmetries. An example of such an inversion symmetry exists in the $Z \rightarrow e^+e^-$ study below. In particle experiments, misidentification of lepton charge in the process of data reconstruction is known to be extremely rare. This means a learned mapping which frequently maps electron/positron (e^\mp) information in \mathcal{Z} to positron/electron (e^\pm) information in \mathcal{X} , and vice versa, would be unphysical. For a generative mapping $G : \mathcal{U} \rightarrow \mathcal{V}$, this anchor term takes the general form

$$\mathcal{L}_A(p(u), p_G(v | u)) = \mathbb{E}_{u \sim p(u)} \mathbb{E}_{v \sim p_G(v|u)} [c_A(u, v)]. \quad (4.3)$$

We chose $c_A(u, v) = 1 - \hat{\mathbf{p}}_u \cdot \hat{\mathbf{p}}_v$, where $\hat{\mathbf{p}}$ is the unit vector of the electron's momentum. We add the anchor loss in \mathcal{Z} space, $\mathcal{L}_A(p(x), p_E(z | x))$, and in \mathcal{X} space, $\mathcal{L}_A(p(z), p_D(x | z))$, to the SWAE loss with hyperparameter weightings β_E and β_D respectively.

From the second category, we enforce the Minkowski metric constraint internally for \mathcal{Z} and \mathcal{X} spaces respectively. A particle's nature, excluding discrete properties such as charge

and spin, is described by four quantities related by the Minkowski metric. Arranging these quantities into a 4-vector defined as $p^\mu = (\mathbf{p}, E)$ where E is a particle’s energy and \mathbf{p} is a vector of its momentum in the $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$ direction respectively, the constraint becomes

$$p^\mu p_\mu = E^2 - \mathbf{p}^2 = m^2, \tag{4.4}$$

where m is the particle’s mass. We directly enforce this relationship in the model for all particles.³

Adding more physically-motivated constraints would be straightforward, however, in this work we only assume this minimal set and recommend that more robust data structures be considered first, as such constraints may become unnecessary (see Section 4.7).

4.4.2 OTUS in Practice

In this section we briefly outline how OTUS might eventually be applied to problems in particle physics such as searches for new particles. However, we emphasize that this work only demonstrates a proof-of-principle version of OTUS. Follow-up work will be necessary to overcome some technical hurdles before OTUS could be applied to such a problem (see Section 4.7).

A main goal of particle physics is to discover the complete set of fundamental units of matter: particles. Therefore, searches for exotic particles are common practice in this field. These searches typically proceed by looking for anomalies in data which are better described by simulations which assume the existence of a new particle. It is therefore phrased as a hypothesis test between two theoretical models, θ_{SM} , which assumes only the particles in the

³We note that initial experiments lacked this constraint yet the networks automatically learned this relationship from the data. However, directly including this constraint in the model architecture improved performance overall.

Abstract example of OTUS application

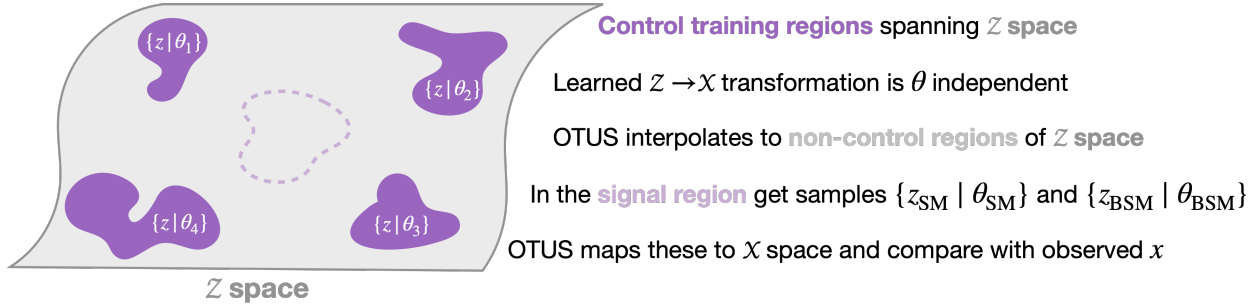


Figure 4.2: **Schematic diagram of how OTUS can be used in an abstract analysis.** The gray surface represents \mathcal{Z} . Different theoretical models, θ_i , will produce different signatures $\{z_i | \theta_i\}$ which lie in \mathcal{Z} . The goal of OTUS is to learn a general mapping from $\mathcal{Z} \rightarrow \mathcal{X}$ which is independent of the underlying theory, θ , and only depends on the information contained in $\{z \in \mathcal{Z}\}$. One trains OTUS using control region data which span \mathcal{Z} and have known outcomes in \mathcal{X} . These allow us to pair distributions in \mathcal{Z} with distributions in \mathcal{X} . From these examples, OTUS interpolates to the rest of \mathcal{Z} and can then be used to generate $\{x_i\}$ from samples $\{z_i | \theta_i\}$ from regions not used during training, including the blinded signal region. This can then be used to search for new particles.

Standard Model (SM), and θ_{BSM} , which assumes the existence of one or more new particles that lie Beyond the Standard Model (BSM). These distinct models will generate distinct latent signatures, $\{z_{SM} | \theta\}$ and $\{z_{BSM} | \theta\}$, which lie in \mathcal{Z} . As particle physics experiments do not observe the latent $\{z\}$ directly, the hypothesis test is performed in the observed space \mathcal{X} , see Section 4.1 and Section 4.3 for more details.

The goal for OTUS is to learn a simulation mapping from $\mathcal{Z} \rightarrow \mathcal{X}$ which is independent of the underlying model, θ , and can be applied to any z . This is achieved by carefully selecting control regions, $\{z_i | \theta_i\}$, which span \mathcal{Z} and for which observed data, $\{x\}$, is available for training. See Fig. 4.2 for a visual description. These control regions have known distributions of outcomes in \mathcal{X} , which allows us to properly match distributions in \mathcal{Z} to distributions in \mathcal{X} for training OTUS. Since these control regions are chosen to span \mathcal{Z} , OTUS will then be able to interpolate to unseen signal regions. Neural networks in general are known to perform well at interpolation tasks [84], and recent work has shown that autoencoders in particular are proficient at learning manifold interpolation [85]. Still more work has suggested there

might be a deeper connection to the structure of this manifold and optimal transport [32]. Therefore, it is reasonable to expect that OTUS will be able to interpolate well in this space. However, these claims should be thoroughly investigated in future work.

A signal region is a region in \mathcal{Z} space where signatures of new particles might occur. SM predictions, $\{z_{\text{SM}} \mid \theta_{\text{SM}}\}$, and BSM predictions, $\{z_{\text{BSM}} \mid \theta_{\text{BSM}}\}$, would then be passed to OTUS to produce two simulated data samples $\{x_{\text{SM}}\}$ and $\{x_{\text{BSM}}\}$ which would be compared with observed data, $\{x\}$, via a hypothesis test to calculate the relative likelihood of the SM and BSM theories. This technique, simulation-based inference, is standard practice in particle physics and is applied to existing simulation methods.

As a concrete example, let our BSM theory be the SM with the addition of a new particle, Z' , with a mass of $0.030 \text{ [TeVc}^{-2}\text{]}$, which decays into a pair of leptons, a flagship search for the Large Hadron Collider [86]. The latent space \mathcal{Z} would include the two leptons produced by the decay of the Z' , and the observed space \mathcal{X} would include the leptons identified and measured by the detector. For OTUS to be able to predict the observed signatures from this latent space, it would need to interpolate between control regions which have similar relationships. Decays of existing particles to leptons, such as the $0.091 \text{ [TeVc}^{-2}\text{]} Z$ and the $0.002 \text{ [TeVc}^{-2}\text{]} J/\psi$ would allow OTUS to learn the mapping from latent leptons to observed leptons. Our theoretical Z' has a mass which lies between those of the particles in our control regions. OTUS would need to interpolate along this axis; control regions at various masses provided by the Z and J/ψ decays are therefore essential to describe and determine the nature of the interpolation. To verify the interpolation, one might compare the prediction of OTUS to observed data in the intermediate range between the Z' and the Z .

Alternatively, the Z' could have a heavier mass, e.g. $1 \text{ [TeVc}^{-2}\text{]}$. In this scenario, OTUS would be required to extrapolate along the mass axis. Naively, this sounds problematic as extrapolation is generally much less sound than interpolation, however this task is also required of current simulations for this scenario. Simulations succeed in such tasks when they

have inductive biases which control their behavior even outside of training (tuning) regions. These inductive biases are based on physics principles and scale to the signal regions of interest. For neural networks, it has been shown that architectures with inductive bias constraints succeed at such extrapolation tasks [50]. Since a mature version of OTUS will manifestly include such inductive biases (see Section 4.7) it is reasonable to assume it can achieve this task as well as current simulation methods can.

4.5 Results

4.5.1 Demonstration in $Z \rightarrow e^+e^-$ decays

We first test OTUS on an important control region: leptonic decays of the Z -boson to electron-positron pairs, $Z \rightarrow e^+e^-$. The theoretical prior is well-known, and its parameters $\{\theta\}$, like the Z -boson’s mass and its interaction strengths, are tightly constrained by precision experiments. We identify \mathcal{Z} with the Z -boson’s decay products: the electron, e^- , and positron, e^+ , whose four-momenta span the space. We compose these into an eight-dimensional vector

$$z := \{z_{e^-}, z_{e^+}\} = \{\mathbf{p}^{e^-}, E^{e^-}, \mathbf{p}^{e^+}, E^{e^+}\}. \quad (4.5)$$

This simplistic vector description excludes categorical properties such as charge.

The model prior $p(z)$ can be simply expressed with quantum field theory and sampled. The subsequent step, where the electron and positron travel through the layers of detectors, depositing energy and causing particle showers, cannot be described analytically; a model will be learned by OTUS from data in control regions. Here we use simulated data samples, but specific (z, x) pairs are not used to mimic the information available when training

from real data. The complex intermediate state with many low-energy particles and high-dimensional detector readouts is reduced and reconstructed yielding estimates of the electron and positron four-momenta. Therefore, \mathcal{X} has the same structure and dimensionality as \mathcal{Z} , though the distribution $p(x)$ reflects the impact of the finite resolution of detector systems (see Section 4.6.1).

Fig. 4.3 shows distributions of testing data, unpaired samples from \mathcal{X} and \mathcal{Z} in several projections, and the results of applying the trained encoder and decoder to transform between the two spaces. Visual evaluation indicates qualitatively good performance, and quantitative metrics are provided. Measuring overall performance, the SW distances are as follows: $d_{\text{sw}}(p(z), p_E(\tilde{z})) = 0.984$ [GeV²], $d_{\text{sw}}(p(x), p_D(\tilde{x})) = 1.33$ [GeV²], $d_{\text{sw}}(p(x), p_D(\tilde{x}')) = 3.03$ [GeV²]. Additionally, several common metrics are reported for each projection in Tables 1 and 2 of Appendix B. Details of the calculations are provided in Section 4.6.4.

To ensure that the learned decoder reflects the physical processes being modeled, we inspect the transformation from $\mathcal{Z} \rightarrow \mathcal{X}$ in Fig. 4.4. The learned transfer function, $p_D(x | z)$, shows reasonable behavior, mapping samples from \mathcal{Z} to nearby values of \mathcal{X} . This reflects the imperfect resolution of the detector while avoiding unphysical transformations such as mapping information on the far-end distribution tails in \mathcal{Z} to the distribution peaks in \mathcal{X} .

Finally, we examine the distribution of a physically important derived quantity, the invariant mass of the Z -boson, see Fig. 4.5. This quantity was not used as an element of the loss function, and so provides an alternative measure of performance. The results indicate a high-quality description of the transformation from \mathcal{Z} to \mathcal{X} . The performance of the transformation from \mathcal{X} to \mathcal{Z} is less well-described, likely because this relation is more strict in \mathcal{Z} causing a sharper peak in the distribution. Such strict rules are difficult for networks to learn when not penalized directly or hard-coded as inductive biases, again signaling that a robust data representation will be crucial to improving performance (see Section 4.7).

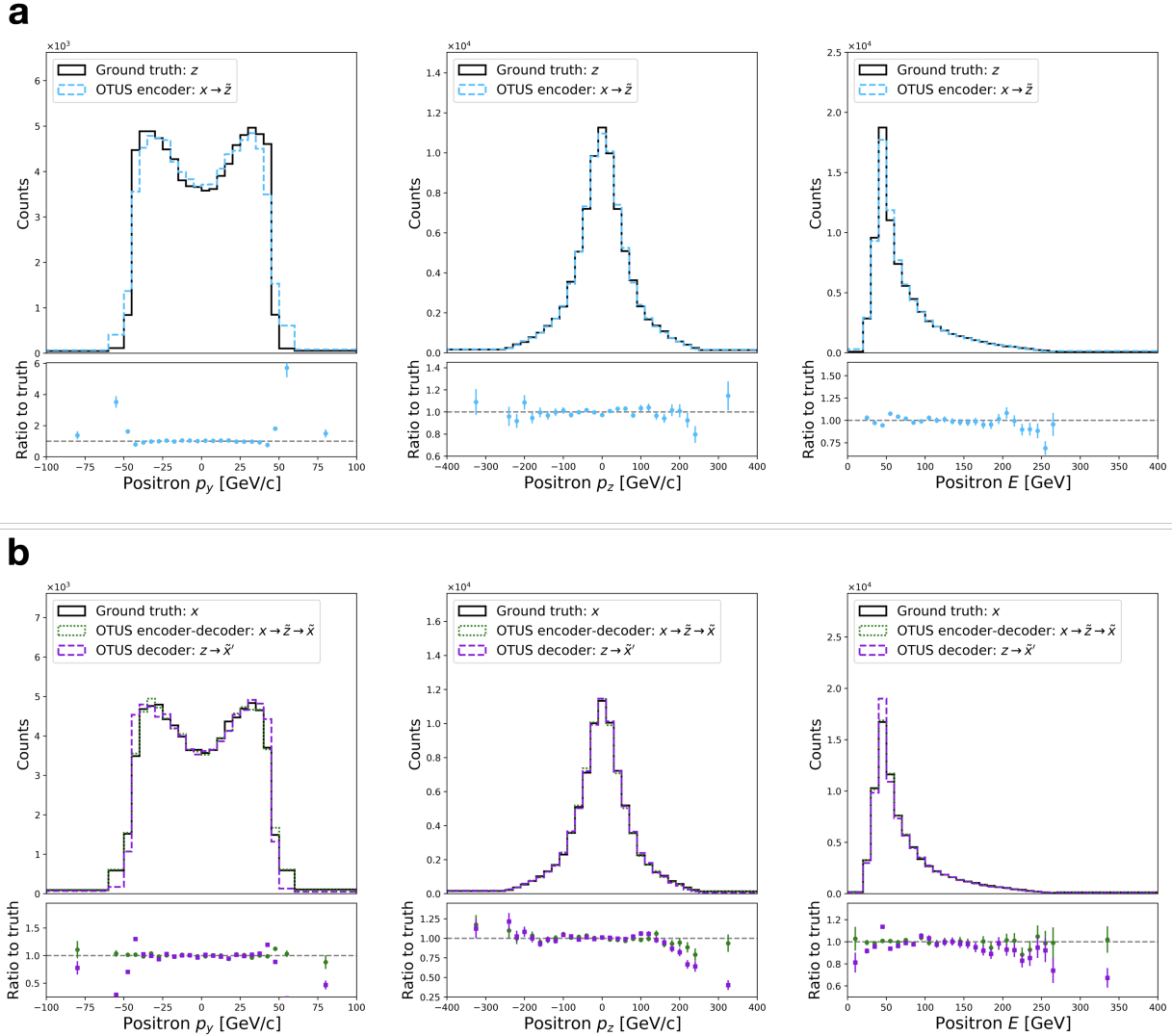


Figure 4.3: **Performance of OTUS for $Z \rightarrow e^+e^-$ decays.** **a** Matching of the positron’s p_x , p_y , and E distributions in \mathcal{Z} . It shows distributions of samples from the theoretical prior, $\{z \sim p(z)\}$ (solid black), as well as the output of the encoder, $\{\tilde{z}\}$; the encoder transforms samples of testing data in experimental space, \mathcal{X} , to the latent space, \mathcal{Z} , and is shown as $x \rightarrow \tilde{z}$ (dashed cyan). **b** Matching of the positron’s p_x , p_y , and E distributions in \mathcal{X} . It shows the testing sample $\{x \sim p(x)\}$ (solid black) in the experimental space, \mathcal{X} , as well as output from the decoder applied to samples drawn from $p(z)$, labeled as $z \rightarrow \tilde{x}'$ (dashed purple). Also shown are samples passed through both the decoder and encoder chain, $x \rightarrow \tilde{z} \rightarrow \tilde{x}$ (dotted green). Dotted green and solid black distributions are matched explicitly during training. Enhanced differences between dashed purple and solid black indicate the encoder’s output needs improvement, as $p_E(z)$ does not fully match $p(z)$. If performance were ideal, the distributions in every plot would match up to statistical fluctuations. Residual plots show bin-by-bin ratios with statistical uncertainties propagated accordingly (see Section 4.6.4).

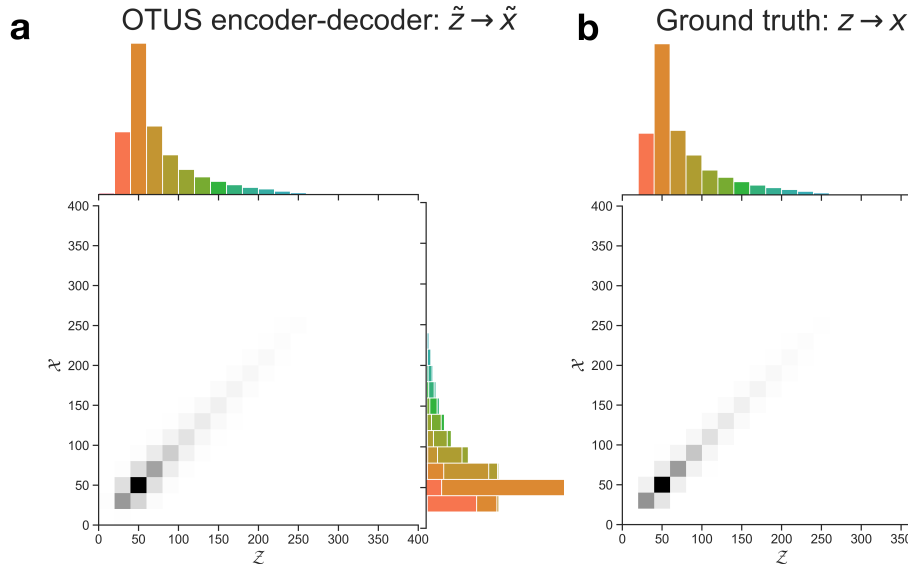


Figure 4.4: **Visualization of the transformation from $\mathcal{Z} \rightarrow \mathcal{X}$ in the $Z \rightarrow e^+e^-$ study for positron energy.** **a** The learned transformation of the decoder, $p_D(x | z)$. **b** The true transformation from the simulated sample, for comparison, though the true (z, x) pairs are not typically available and were not used in training. Colors in the \mathcal{X} projection indicate the source bin in \mathcal{Z} for a given sample.

4.5.2 Demonstration in semileptonic top-quark decays

The Z -boson control region is valuable for calibrating simulations of leptons such as electrons or muons, which tend to be stable and well-measured. We next test OTUS on the challenging task of modeling the decay and detection of top-quark pairs featuring more complex detector signatures. This control region has more observed particles and introduces additional complexities: unstable particles decaying in flight, significantly degraded resolution relative to leptons, undetected particles, and a stochastically variable number of observed particles.

The initial creation of top-quark pairs, their leading-order decay $t \bar{t} \rightarrow W^+ b W^- \bar{b}$, and the subsequent W -boson decays are well-described using quantum field theory, so $p(z | \theta)$ can be sampled. We select the modes $W^- \rightarrow e^- \bar{\nu}_e$ and $W^+ \rightarrow u \bar{d}$ as examples and assign our

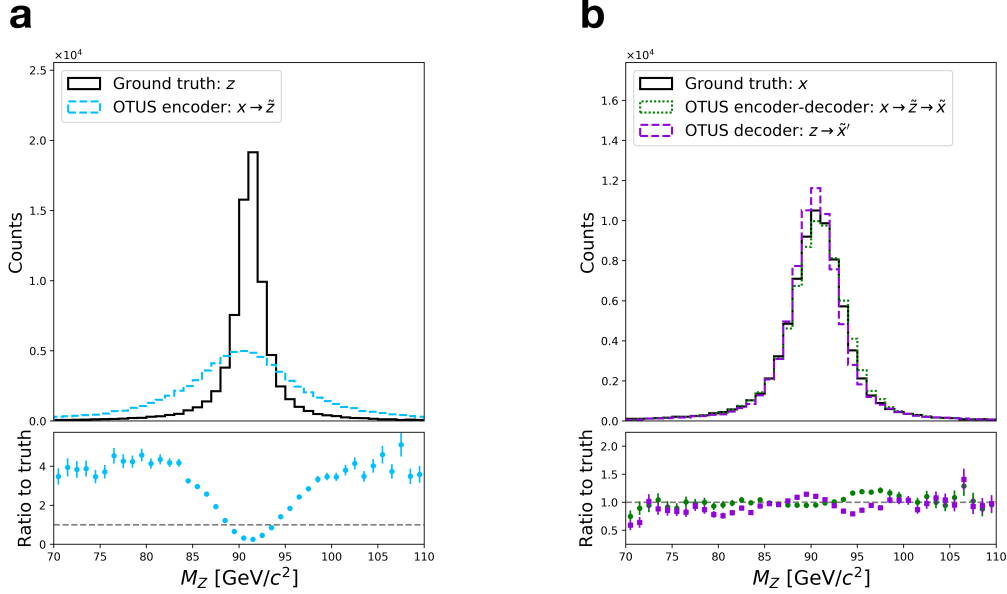


Figure 4.5: **Performance of OTUS for $Z \rightarrow e^+e^-$ decays in a physically important derived quantity, the invariant mass of the electron-positron pair, M_Z .** **a** Matching of the M_Z distribution in \mathcal{Z} . It shows distributions of samples from the theoretical prior, $\{z \sim p(z)\}$ (solid black), as well as the output of the encoder, $\{\tilde{z}\}$; the encoder transforms samples of testing data in experimental space, \mathcal{X} , to the latent space, \mathcal{Z} , and is shown as $x \rightarrow \tilde{z}$ (dashed cyan). **b** Matching of the M_Z distribution in \mathcal{X} . It shows the testing sample $\{x \sim p(x)\}$ (solid black) in the experimental space, \mathcal{X} , as well as output from the decoder applied to samples drawn from $p(z)$, labeled as $z \rightarrow \tilde{x}'$ (dashed purple). Also shown are samples passed through both the decoder and encoder chain, $x \rightarrow \tilde{z} \rightarrow \tilde{x}$ (dotted green). Dotted green and solid black distributions are matched explicitly during training. Enhanced differences between dashed purple and solid black indicate the encoder's output needs improvement, as $p_E(z)$ does not fully match $p(z)$. If performance were ideal, the distributions in every plot would match up to statistical fluctuations. Note that this projection was not explicitly used during training, but was inferred by the networks. Residual plots show bin-by-bin ratios with statistical uncertainties propagated accordingly (see Section 4.6.4).

latent space to describe the four-momenta of these six of particles:

$$z := \{z_{e^-}, z_{\bar{\nu}_e}, z_b, z_{\bar{b}}, z_u, z_{\bar{d}}\} = \{\mathbf{p}^{e^-}, E^{e^-}, \mathbf{p}^{\bar{\nu}_e}, E^{\bar{\nu}_e}, \mathbf{p}^b, E^b, \mathbf{p}^{\bar{b}}, E^{\bar{b}}, \mathbf{p}^u, E^u, \mathbf{p}^{\bar{d}}, E^{\bar{d}}\} \quad (4.6)$$

with a total of twenty-four dimensions.

Unlike in the $Z \rightarrow e^+e^-$ study, the \mathcal{X} space's structure is considerably different from that of the \mathcal{Z} space. While the electron e^- is stable and readily identifiable, the other particles are more challenging. The neutrino, $\bar{\nu}_e$, is stable, yet invisible to our detectors, providing no estimate of its direction or momentum; instead its presence is inferred using momentum conservation $\mathbf{p}^\nu = -\sum \mathbf{p}^{\text{observed}}$. Unfortunately, soft initial state radiation and detector inefficiencies also contribute to missing momentum. The aggregate quantity is labeled \mathbf{p}^{miss} . The four quarks \bar{b} , u , \bar{d} and b are strongly-interacting particles each producing complex showers of particles that are clustered together into jets to estimate the original quark momenta and directions. Unfortunately, despite significant recent progress [87–89], we cannot assume a perfect identification of the source particle in \mathcal{Z} for a given jet observed in \mathcal{X} , causing significant ambiguity.

Additionally, a complete description of the $\mathcal{Z} \rightarrow \mathcal{X}$ transformation should include the possibilities for the number of jets in \mathcal{X} to exceed the number of quarks, due to radiation and splitting, or to fail to match the number of quarks, due to jet overlap or detector inefficiency. We leave this complexity for future work and restrict our \mathcal{X} space to contain exactly four jets.

The final complexity introduced in this study is the presence of a sharp lower threshold in transverse momentum, p_T . Experimental limitations require that jets with $p_T < 20$ [GeVc⁻¹] be discarded and therefore are not represented in the training dataset, as they would be unavailable in control region data. Mimicking this experimental effect, we directly impose

this threshold on the decoder’s output instead of the network learning it. Paralleling reality, such events are discarded before computing losses. This strategy requires modifications to both the model and training strategy (see Section 4.6).

Our experimental data is the vector

$$x := \{x_{e^-}, x_{\text{miss}}, x_{\text{jet1}}, x_{\text{jet2}}, x_{\text{jet3}}, x_{\text{jet4}}\} \quad (4.7)$$

$$= \{\mathbf{p}^{e^-}, E^{e^-}, \mathbf{p}^{\text{miss}}, E^{\text{miss}}, \mathbf{p}^{\text{jet1}}, E^{\text{jet1}}, \mathbf{p}^{\text{jet2}}, E^{\text{jet2}}, \mathbf{p}^{\text{jet3}}, E^{\text{jet3}}, \mathbf{p}^{\text{jet4}}, E^{\text{jet4}}\}, \quad (4.8)$$

with a total of twenty-four dimensions. If quark-jet assignment were possible, it would be natural to align the order of the observed jets with the order of their originating quarks in \mathcal{Z} space. Lacking this information, it is typical to order jets by descending $|\mathbf{p}_T| = \sqrt{p_x^2 + p_y^2}$, where jet 1 has the largest $|\mathbf{p}_T|$.

Fig. 4.6 shows distributions of testing data, unpaired samples from \mathcal{X} and \mathcal{Z} in several projections, and the results of applying the trained encoder and decoder to transform between the two spaces. Visual evaluation indicates qualitatively good performance, and quantitative metrics are also provided. Measuring overall performance the SW distances are as follows: $d_{\text{sw}}(p(z), p_E(\tilde{z})) = 22.3$ [GeV²], $d_{\text{sw}}(p(x), p_D(\tilde{x})) = 232$ [GeV²], $d_{\text{sw}}(p(x), p_D(\tilde{x}')) = 120$ [GeV²]. Additionally, several common metrics are reported for each projection in Table 3 and 4 of Appendix B. Details of the calculations are provided in Section 4.6.4.

To probe the $\mathcal{Z} \rightarrow \mathcal{X}$ transformation, we inspect the learned transfer function, $p_D(x | z)$ in Fig. 4.7. While the overall performance is worse in this more complex case, it still shows reasonable behavior, mapping samples from \mathcal{Z} to nearby values of \mathcal{X} and avoiding unphysical transformations such as mapping information on the far-end distribution tails in \mathcal{Z} to the distribution peaks in \mathcal{X} . Additionally, cross-referencing with the true simulation’s mapping shows the similar nature of the mappings.

Finally, we examine the distribution of physically important derived quantities, the invariant masses of the top-quarks and W -bosons estimated by combining information from pairs and triplets of objects, see Fig. 4.8. No exact assignments are possible due to the ambiguity of the jet assignment and the lack of transverse information for the neutrino, but a comparison can be made between the experimental sample in \mathcal{X} and the mapped samples $\mathcal{Z} \rightarrow \mathcal{X}$. As in the $Z \rightarrow e^+e^-$ case, we see imperfect but reasonable matching on such derived quantities which the network was not explicitly instructed to learn.

4.6 Methods

This section provides details on the methods used to produce the results in the previous section. We first describe the data generation process. We then describe the machine learning models used and strategies for how they were trained. Finally, we give details on the qualitative and quantitative evaluation methods used in the visualizations of the results.

4.6.1 Data Generation

The data for this work was generated with the programs Madgraph5 v.2.6.3.2 [22], Pythia v.8.240 [20], and Delphes v.3.4.1 [23]. ROOT v.6.08/00 [90] was used to interface with the resulting Delphes output files. We used the default run cards for Pythia, Delphes, and Madgraph. Where relevant, jets were clustered using the anti-kt algorithm [91] with a jet radius of 0.5. The card files can be found with the code for this analysis (see Section 4.9).

Samples of the physical latent space, \mathcal{Z} , were extracted from the Madgraph LHE files to form the 4-momenta of the particles. Samples of the data space, \mathcal{X} , were extracted from Delphes' output ROOT files. We selected for the appropriate final state: e^+ , e^- in the $Z \rightarrow e^+e^-$ study and e^- , missing 4-momentum (i.e. $\text{MET} = (\mathbf{p}^{\text{miss}}, E^{\text{miss}})$), and 4 jets in the

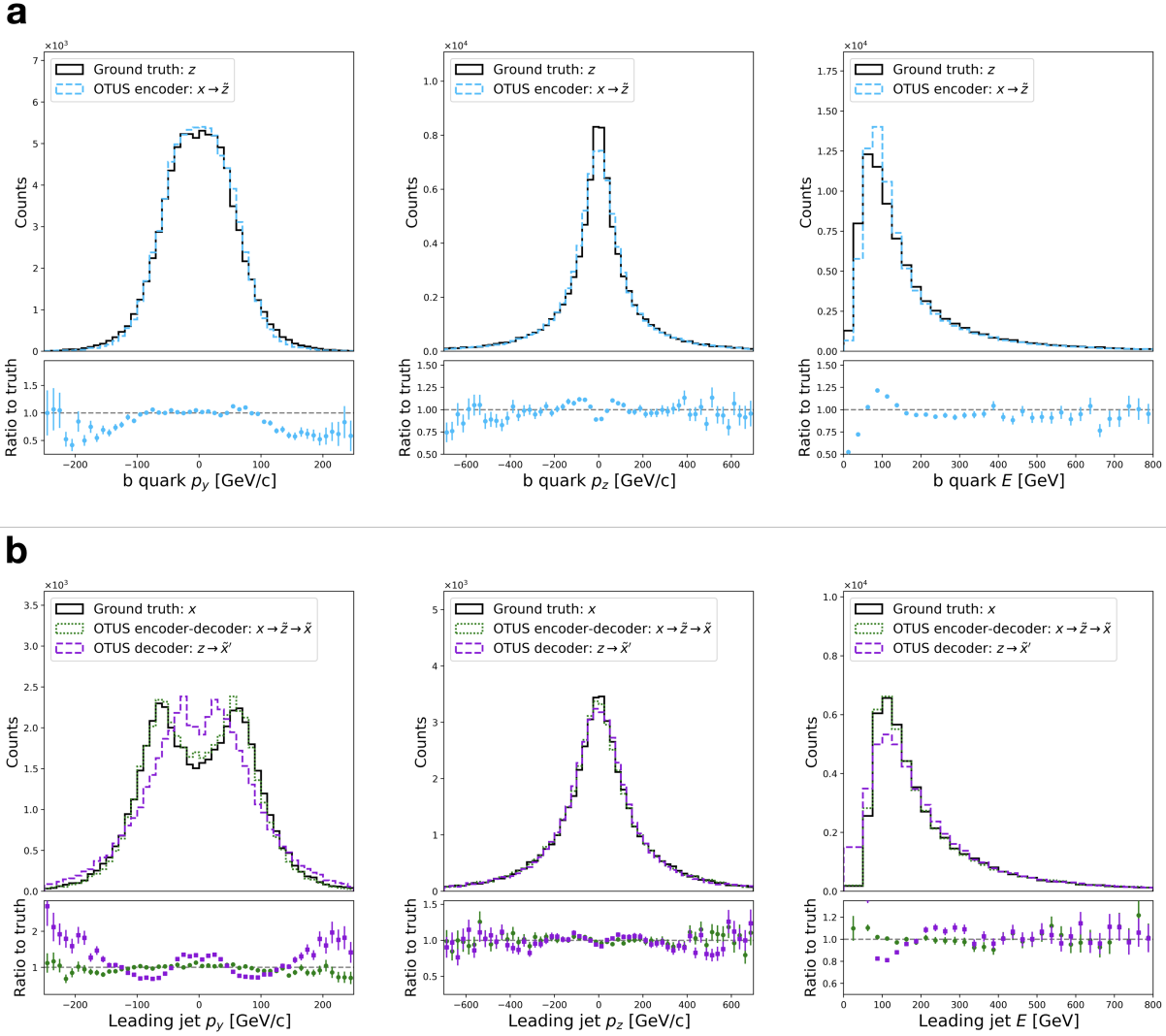


Figure 4.6: **Performance of OTUS for semileptonic $t\bar{t}$ decays.** **a** Matching of the b quark's p_x , p_y , and E distributions in \mathcal{Z} . It shows distributions of samples from the theoretical prior, $\{z \sim p(z)\}$ (solid black), as well as the output of the encoder, $\{\tilde{z}\}$; the encoder transforms samples of the testing data in experimental space, \mathcal{X} , to the latent space, \mathcal{Z} , and is shown as $x \rightarrow \tilde{z}$ (dashed cyan). **b** Matching of the leading jet's p_x , p_y , and E distributions in \mathcal{X} . It shows the testing sample $\{x \sim p(x)\}$ (solid black) in the experimental space, \mathcal{X} , as well as output from the decoder applied to samples drawn from the prior $p(z)$, labeled as $z \rightarrow \tilde{x}'$ (dashed purple). Also shown are samples passed through both the decoder and encoder chain, $x \rightarrow \tilde{z} \rightarrow \tilde{x}$ (dotted green). Dotted green and solid black distributions are matched explicitly during training. Enhanced differences between dashed purple and solid black indicate the encoder's output needs improvement, as $p_E(z)$ does not fully match $p(z)$. If performance were ideal, the distributions in every plot would match up to statistical fluctuations. Residual plots show bin-by-bin ratios with statistical uncertainties propagated accordingly (see Section 4.6.4).

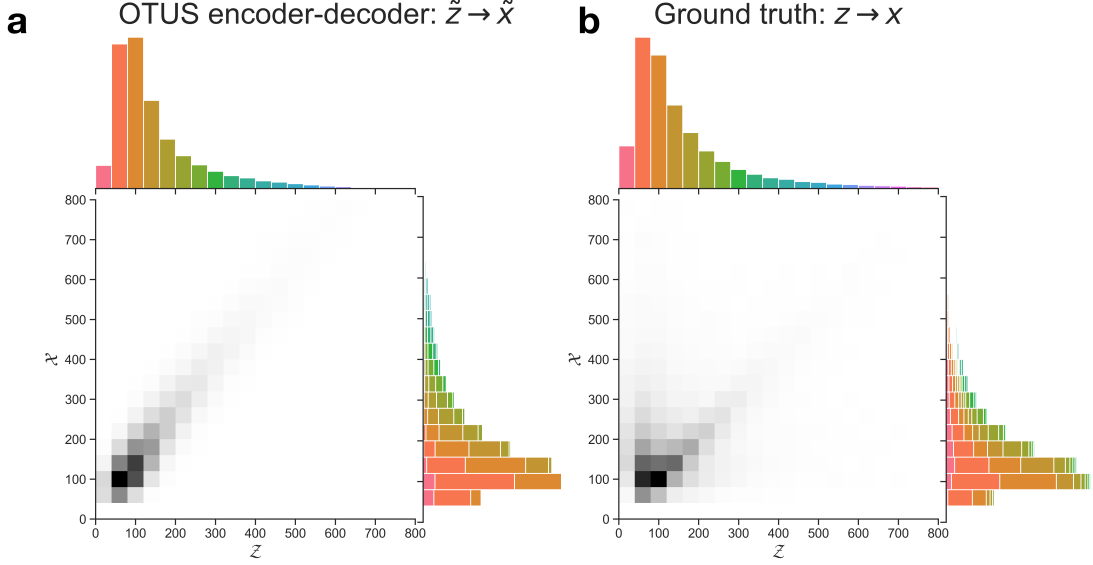


Figure 4.7: **Visualization of the transformation from $\mathcal{Z} \rightarrow \mathcal{X}$ in the $t\bar{t}$ study for the energy of the b quark in \mathcal{Z} to energy of the leading jet in \mathcal{X} .** **a** The learned transformation of the decoder, $p_D(x | z)$. **b** The true transformation from the simulated sample, for comparison, though the true (z, x) pairs are not typically available and were not used in training. Note that the b quark will not always correspond to the leading jet, see the text for details. Colors in the \mathcal{X} projection indicate the source bin in \mathcal{Z} for a given sample.

semileptonic $t\bar{t}$ study. If an event failed this selection, the corresponding \mathcal{Z} event was also removed. Reconstructed data in \mathcal{X} was extracted by default as (p_T, η, ϕ) of the object and converted into (\mathbf{p}, E) via the following relations

$$\mathbf{p} := (p_x, p_y, p_z) = (p_T \cos(\phi), p_T \sin(\phi), p_T \sinh(\eta)) \quad (4.9)$$

$$E = \sqrt{(p_T \cosh(\eta))^2 + m^2}, \quad (4.10)$$

where m is the particle's definite mass and is zero for massless particles. Note that we are assuming natural units where the speed of light, c , is equal to unity. This equates the units of energy, E , momentum, \mathbf{p} , and mass, m . In our case, $m_{e^+} = m_{e^-} = 0$ [GeV c^{-2}] is a standard assumption given that the true value is very small compared to the considered energy scales. We additionally set $m = 0$ [GeV c^{-2}] for the 4 jets and MET since these objects have atypical definitions of mass.

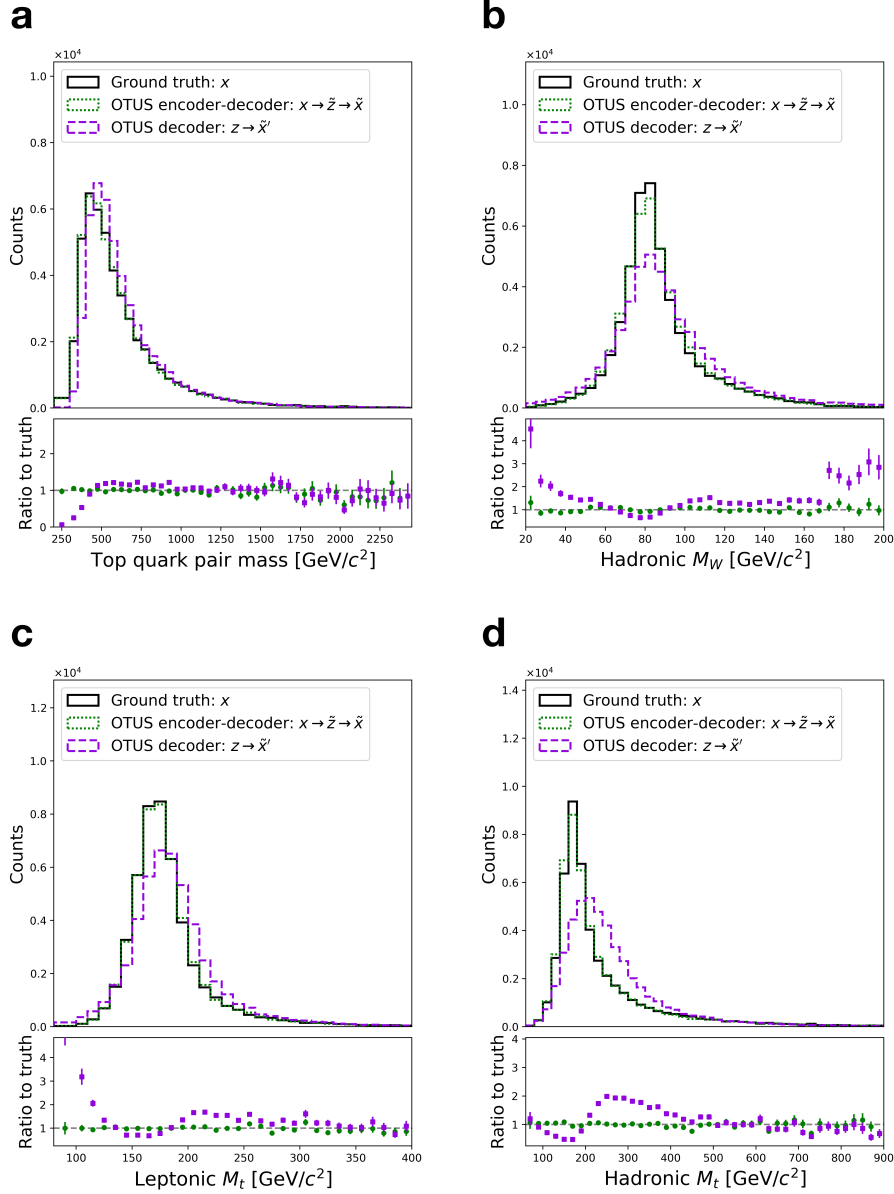


Figure 4.8: **Performance of OTUS for semileptonic $t\bar{t}$ decays in physically important derived quantities in \mathcal{X} .** **a** Matching of the invariant mass of the combined $t\bar{t}$ pair. **b** Matching of the invariant mass of the hadronically decaying W -boson, M_W . **c** Matching of the invariant mass of the top-quark, M_t , reconstructed using information from the leptonically decaying W -boson. **d** Matching of the invariant mass of the top-quark, M_t , reconstructed using information from the hadronically decaying W -boson. These show the testing sample $\{x \sim p(x)\}$ (solid black) in the experimental space, \mathcal{X} , as well as output from the decoder applied to samples drawn from $p(z)$, labeled as $z \rightarrow \tilde{x}'$ (dashed purple). Also shown are samples passed through both the decoder and encoder chain, $x \rightarrow \tilde{z} \rightarrow \tilde{x}$ (dotted green). Dotted green and solid black distributions are matched explicitly during training. Enhanced differences between dashed purple and solid black indicate the encoder's output needs improvement, as $p_E(z)$ does not fully match $p(z)$. Residual plots show bin-by-bin ratios with statistical uncertainties propagated accordingly (see Section 4.6.4).

In total, we generated 491,699 events for $Z \rightarrow e^+e^-$ and 422,761 events for semileptonic $t\bar{t}$. The last 160,000 events in each case were reserved solely for statistical tests after training and validation of OTUS.

4.6.2 Model

4.6.2.1 Model Choice

In this section, we briefly survey the literature of machine learning methods which might be considered for this task. We discuss their features and whether they are compatible choices for this application.

We will primarily focus on OT-based probabilistic autoencoder methods (i.e. WAE [31] and its derivatives) but first we briefly address a derivative of VAEs, β -VAE. This method appears similar to WAE in the form of loss function that is used. Both have a data-space loss and a latent-space loss with a relative hyperparameter weighting β (or λ for the WAE). However, the β -VAE method is not principled in OT and thus is distinct from the WAE method and its derivatives. Most importantly for our application, the β -VAE (like its predecessor VAE) is likelihood-based which precludes it from applications where the latent prior is not analytically known. The interested reader can find more information on these distinctions in the following reference [92].

The WAE method [31] provides a general framework for an autoencoder whose training is based on ideas from OT theory, namely the Wasserstein distance. This work defined a large umbrella under which a rich amount of subsequent literature falls (e.g. SWAE [36], Sinkhorn Autoencoders [56], CWAE [93]). The key difference between these methods and the original WAE method is the fact that each chooses a different $d_z(\cdot, \cdot)$ cost function. Therefore, the choice of method largely comes down to finding a suitable d_z for the given problem.

The original WAE work proposes two specific options for the d_z , defining two versions of WAE: GAN-WAE and MMD-WAE. The first is an adversarial approach in which d_z is the Jensen-Shannon divergence estimated using a discriminator network. The second chooses d_z to be the Maximum Mean Discrepancy (MMD) [31].

The GAN-WAE strategy suffers from the same practical issues as other adversarial methods such as GANs (i.e. mode collapse). This possibility of training instability makes it an undesirable choice. The MMD-WAE does not have this training instability issue but requires an a priori choice of a kernel for the form of latent space prior, $p(z)$. This implies that we analytically know the desired prior form ahead of time, which is not the case for particle physics in general. Therefore, this option will not work for the applications explored in this work.

We now explore WAE derivatives which choose other choices for d_z that might be more amenable to our application. CWAE [93] chooses the Cramer-Wold distance as the d_z cost function. For a Gaussian latent space prior, this provides a computationally efficiency boost due to the existence of a closed-form solution. However, this assumption makes it unsuitable for our current application because our latent prior, $p(z)$, is non-Gaussian and often does not have a form which is known analytically a priori.

Two other derivatives allow for a flexible prior form which would be suitable for the task at hand. SWAE [36] chooses the d_z cost function to be the SW distance and Sinkhorn Autoencoder (SAE) [56] chooses it to be the Sinkhorn divergence which is estimated via the Sinkhorn algorithm. Both have comparable performance with trade-offs in performance and computational efficiency. SAE claims superior performance to SWAEs for Gaussian priors, while it is slightly more computationally intensive ($\mathcal{O}(M^2)$) as opposed to SWAEs best case $\mathcal{O}(M)$ or worst case $\mathcal{O}(M \log M)$). However, both methods are valid choices for this application. Therefore, we suggest that SAE performance on this task be explored in future work.

We also note the existence of other WAE-derivative methods which generalize the underlying OT framework. In our application, the d_z metric always compares distributions in the same ambient space \mathcal{Z} . Additionally, the overall loss function also approximates the Wasserstein distance between two distributions in the same ambient space \mathcal{X} , namely $W_c(p(x), p_D(x))$. However, recent work using the Gromov-Wasserstein distance [94] extends the underlying Optimal Transport (OT) framework to situations where the two probability measures μ and ν are not defined on the same ambient space (e.g. \mathbb{R}^n and \mathbb{R}^m with different dimensions n and m). For this application, this is an over-powered tool since by construction $p(z)$ and $p_E(z)$ ($p(x)$ and $p_D(x)$) always lie in the same ambient space. However, if one were attempting to study the optimal transportation between different spaces, this would be ideal. This would be an interesting direction to follow-up recent related work which connects OT and particle physics [32, 33].

4.6.2.2 Base Model

Both the encoder and decoder models of OTUS are implicit conditional generative models, and operate by concatenating the input with random noise and passing the resulting vector through feedforward neural networks.

For a model, G , mapping from a space, \mathcal{U} , to a space, \mathcal{V} , the steps are as follows. (1) A sample of raw input data, $u \in \mathcal{U}$, is standardized by subtracting the mean and dividing by the standard deviation resulting in the standardized data vector, \bar{u} . (2) A noise neural network computes a conditional noise distribution $p_N(\epsilon | u)$, where the noise vector $\epsilon \sim p_N(\epsilon | u)$ has the same dimensionality as the core network prediction \bar{w} (defined in the next step). (3) The standardized data vector, \bar{u} , and noise vector, ϵ , are then concatenated and fed into a core neural network. This network outputs the 3-momentum, \mathbf{p} , information of each particle in the standardized space, collected into a vector \bar{w} . (4) The vector \bar{w} is then unstandardized by inverting the relationship in step 1, creating a vector w . (5) The Minkowski relation (see

Section 4.4.1) is then enforced explicitly to reinsert the energy information of each particle, transforming w into the final $v \in \mathcal{V}$ which is distributed according to $p_G(v | u)$.

Both the encoding and decoding model’s noise networks produce Gaussian-distributed noise vectors with mean and diagonal covariances $[\mu(x), \sigma^2(x)]$ and $[\mu(z), \sigma^2(z)]$ respectively. For the $Z \rightarrow e^+e^-$ study, the core and noise networks for both the encoder and decoder each used a simple feed-forward neural network architecture with a single hidden layer, with 128 hidden units and ReLU activation.

4.6.2.3 Model for semileptonic top-quark decay study

To better model the complexities in the semileptonic $t\bar{t}$ data, we introduced a restriction to the decoder model and modified the training procedure accordingly (see Section 4.6.3). With these modifications, the base model encountered difficulty during training, so we introduced the following three changes to the architecture for more effective training.

First, the conditionality of the noise network is removed and the noise is instead drawn from a fixed standard normal distribution, $p_N(\epsilon | u) = p_N(\epsilon) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Second, the model now has a residual connection such that the core network now predicts the change from the input u . The 3-momentum sub-vector of u is added to w before proceeding to imposing the Minkowski relation in step 5. This input-to-output residual connection provides an architectural bias towards identity mapping, when the model is initialized with small random weights.

Lastly, the core network itself is augmented with residual connections [95] and batch normalization [96]. An input vector to the core network is processed as follows: (A) A linear transform layer with K units maps the input to a vector $r \in \mathbb{R}^K$. (B) Two series of [BatchNorm, ReLU, Linear] layers are applied sequentially to r , without changing the dimensionality, resulting in $s \in \mathbb{R}^K$. (C) A residual connection from r is introduced, so that $s \rightarrow s + r$. (D) The resulting s is then transformed by a final linear layer with J units to

obtain the output vector $t \in \mathbb{R}^J$. For the $t\bar{t}$ study, the input vector $[\bar{u}, \epsilon]$ is $24 + 18 = 42$ dimensional, the output dimension $J = 18$, and we set $K = 64$ for the core network, in both the encoder and decoder models.

4.6.3 Training

4.6.3.1 Base Training Strategy

As described in Section 4.4.1, the model is trained by minimizing the SWAE loss function augmented with anchor terms

$$\begin{aligned} \mathcal{L}_{\text{SWAE}}(p(x), p_D(x | z), p_E(z | x)) &= \mathbb{E}_{x \sim p(x)} \mathbb{E}_{z \sim p_E(z | x)} \mathbb{E}_{\tilde{x} \sim p_D(x | z)} [c(x, \tilde{x})] + \lambda d_{\text{SW}}(p_E(z), p(z)) \\ &\quad + \beta_E \mathcal{L}_A(p(x), p_E(z | x)) + \beta_D \mathcal{L}_A(p(z), p_D(x | z)), \end{aligned} \tag{4.11}$$

with respect to parameters of the encoder $p_E(z | x)$ and decoder $p_D(x | z)$ distributions.

As each term in the loss function has the form of an expectation, we approximate each with samples and compute the following Monte-Carlo estimate of the loss:

$$\begin{aligned} \hat{\mathcal{L}}_{\text{SWAE}} &= \frac{1}{M} \sum_{m=1}^M c(x_m, \tilde{x}_m) + \lambda \frac{1}{L * M} \sum_{l=1}^L \sum_{m=1}^M c((\theta_l \cdot z_m)_{\text{sorted}}, (\theta_l \cdot \tilde{z}_m)_{\text{sorted}}) \\ &\quad + \beta_E \frac{1}{M} \sum_{m=1}^M c_A(x_m, \tilde{z}_m) + \beta_D \frac{1}{M} \sum_{m=1}^M c_A(z_m, \tilde{x}'_m), \end{aligned} \tag{4.12}$$

where $\{x_m\}_{m=1}^M$ and $\{z_m\}_{m=1}^M$ are M instances of \mathcal{X} and \mathcal{Z} samples, $\{\tilde{z}_m \sim p_E(\cdot | x_m)\}_{m=1}^M$ are drawn from the encoder, $\{\tilde{x}'_m \sim p_D(\cdot | z_m)\}_{m=1}^M$ are drawn from the decoder, and $\{\tilde{x}_m \sim p_D(\cdot | \tilde{z}_m)\}_{m=1}^M$ are drawn from the auto-encoding chain $x \rightarrow \tilde{z} \rightarrow \tilde{x}$.⁴ The estimation of $d_{\text{SW}}(p(z), p_E(z))$ uses L random slicing directions $\{\theta_l\}_{l=1}^L$ drawn uniformly from the unit

⁴This is equivalent to drawing a sample $(x, \tilde{z}, \tilde{x})$ from the joint distribution $p(x)p_E(\tilde{z} | x)p_D(\tilde{x} | \tilde{z})$.

sphere, along which the samples $z_m \sim p(z)$ and $\tilde{z}_m \sim p_E(z)$ are compared; this involves estimating each CDF^{-1} by sorting the two sets of projections in ascending order as $\{(\theta_l \cdot z_m)_{\text{sorted}}\}_{m=1}^M$ and $\{(\theta_l \cdot \tilde{z}_m)_{\text{sorted}}\}_{m=1}^M$, for each direction θ_l ; we refer interested readers to [36] for more technical details of the Sliced Wasserstein distance. We use the squared norm as the cost metric $c(u, v) = \|u - v\|^2$ in the SWAE loss [36]. The anchor cost, c_A , between two observation vectors u, v (which can reside in either \mathcal{X} or \mathcal{Z} space) is defined as $c_A(u, v) := 1 - \hat{\mathbf{p}}_u \cdot \hat{\mathbf{p}}_v$, where $\hat{\mathbf{p}}_u$ is the unit vector of the coordinates of u corresponding to the momentum of a pre-specified particle, and $\hat{\mathbf{p}}_v$ is defined analogously with respect to the same particle; this is chosen as the electron in our experiments. For example, $c_A(x, \tilde{z})$ would be computed as

$$c_A(x, \tilde{z}) = 1 - \hat{\mathbf{p}}_x^{e^-} \cdot \hat{\mathbf{p}}_{\tilde{z}}^{e^-} = 1 - \frac{\mathbf{p}_x^{e^-}}{\|\mathbf{p}_x^{e^-}\|} \cdot \frac{\mathbf{p}_{\tilde{z}}^{e^-}}{\|\mathbf{p}_{\tilde{z}}^{e^-}\|}. \quad (4.13)$$

At a higher level, the computation of $\hat{\mathcal{L}}_{\text{SWAE}}$ based on a mini-batch proceeds as follows. Following the path through the full model, a batch of samples $X \sim p(x)$ from \mathcal{X} space is passed to the encoder model, E , producing $\tilde{Z} \in \mathcal{Z}$ distributed according to $p_E(z | x)$. The encoding anchor loss term $L_{A,E}(X, \tilde{Z}) \equiv \mathcal{L}_A(p(x), p_E(z | x))$ is then computed along with the SW distance latent loss, $\hat{d}_{\text{SW}}(Z, \tilde{Z}) \equiv \hat{d}_{\text{SW}}(p(z), p_E(z))$. The samples \tilde{Z} and $Z \sim p(z)$ are then passed independently in parallel through the decoder model, D , producing \tilde{X} and \tilde{X}' , respectively. The decoding anchor loss term $L_{A,D}(Z, \tilde{X}') \equiv \mathcal{L}_A(p(z), p_D(x | z))$ is then computed. Finally, the data space loss, chosen to be $\text{MSE}(X, \tilde{X})$, is computed. See Fig. 4.9 for a visual representation. We can then minimize the tractable Monte-Carlo estimate of the objective, $\hat{\mathcal{L}}_{\text{SWAE}}$, by stochastic gradient descent with respect to parameters of the encoder and decoder networks.

Since the original (S)WAE aimed to ultimately minimize $d_W(p(x), p_D(x))$ via an approximate variational formulation,⁵ we also consider an auxiliary strategy of directly minimizing

⁵When minimized over all $p_E(z | x)$ that satisfies the constraint $p_E(z) = p(z)$, term A of Equation (4.1) becomes an upper bound on $d_W(p(x), p_D(x))$; the bound is tight for deterministic decoders [31]. The overall

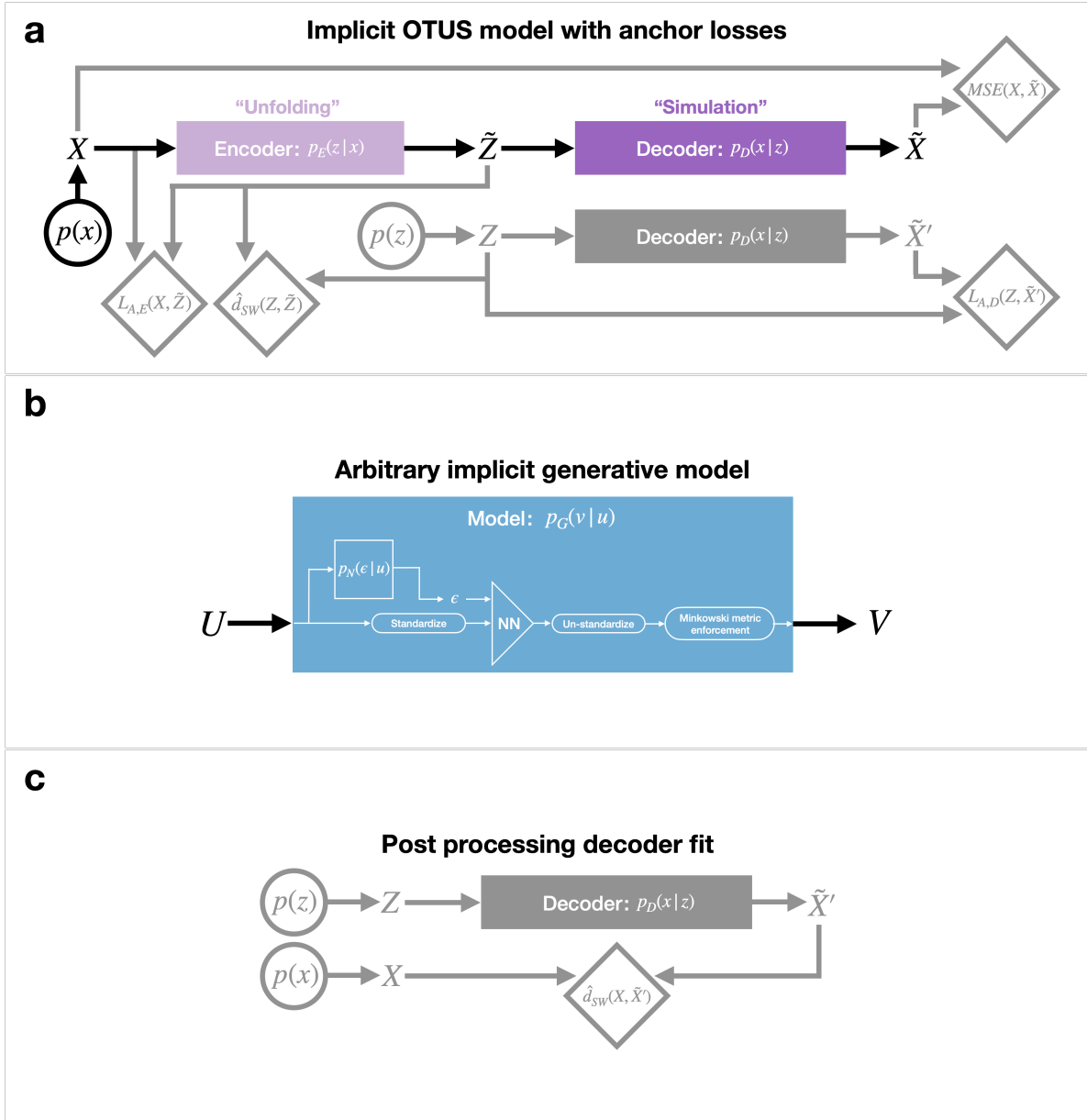


Figure 4.9: **Schematic diagrams of the network and loss structures used in this study for the base training strategy.** **a** Diagram showing the full OTUS model where gray indicates information used in the calculation of losses only. **b** Diagram showing the internal structure present in both the encoder and decoder models. **c** Diagram showing the setup used for the post processing decoder network loss. See the text for more details.

the more computationally convenient SW distance $d_{\text{SW}}(p(x), p_D(x))$ to train a decoder, or minimizing $d_{\text{SW}}(p(z), p_E(z))$ to train an encoder. This can be done by simply optimizing the

WAE loss \mathcal{L}_{WAE} is a relaxation of the exact variational bound, and recovers the latter as $\lambda \rightarrow \infty$.

Monte-Carlo estimates

$$\hat{d}_{\text{SW}}(p(x), p_D(x)) = \frac{1}{L * M} \sum_{l=1}^L \sum_{m=1}^M c((\theta_l \cdot x_m)_{\text{sorted}}, (\theta_l \cdot \tilde{x}'_m)_{\text{sorted}}) \quad (4.14)$$

$$\hat{d}_{\text{SW}}(p(z), p_E(z)) = \frac{1}{L * M} \sum_{l=1}^L \sum_{m=1}^M c((\theta_l \cdot z_m)_{\text{sorted}}, (\theta_l \cdot \tilde{z}_m)_{\text{sorted}}), \quad (4.15)$$

where the samples $\{x_m, z_m, \tilde{x}'_m, \tilde{z}_m\}_{m=1}^M$ are defined the same way as before. Auxiliary training of the encoder was found helpful for escaping local minima when optimizing the joint loss $\hat{\mathcal{L}}_{\text{SWAE}}$, and auxiliary fine-tuning of the decoder in post-processing also improved the decoder's fit to the data.

Note that the idea of training a decoder by itself is similar in spirit to GANs, but again with the major distinction and innovation that we use samples from a physically meaningful prior $p(z)$ instead of an uninformed generic one (e.g. Gaussian), as we are also interested in a physical conditional mapping $p_D(x | z)$ in addition to achieving good fit to the marginal $p(x)$.

4.6.3.2 Training an (S)WAE with a restricted decoder

As was previously explained, experimental limitations in the semileptonic $t\bar{t}$ study require a minimum threshold, so that jets which have $p_T < 20$ [GeV c^{-1}] are discarded and therefore are not represented in the training dataset, as they would not be available in control region data. Denoting the region of \mathcal{X} space which passes this threshold by S , we are faced with the task of fitting a distribution $p_D(x)$ over \mathcal{X} while only having access to data samples in the valid subset $S \subset \mathcal{X}$.

We propose a general method for fitting an (S)WAE such that its marginal data distribution $p_D(x)$, when restricted to the valid set S , matches that of the available data. We first define

the restricted marginal data distribution,

$$\bar{p}_D(x) = \frac{p_D(x)\mathbf{1}_S(x)}{P_D(S)}, \quad (4.16)$$

where $\mathbf{1}_S(x)$ is the indicator function of S so that it equals 1 if $x \in S$, and 0 otherwise, and $P_D(S) := \int dt p_D(t)\mathbf{1}_S(t)$ normalizes this distribution. Note that $P_D(S)$ depends on the decoder parameters, and can be identified as the probability that the data model $p_D(x)$ yields a valid sample $x \in S$.

Our goal is then to minimize $d_W(p(x), \bar{p}_D(x))$. This can be done by minimizing the same variational upper bound as in a typical (S)WAE, but with an adjustment to the data loss function in term A of Equation (4.1), so it becomes

$$\mathbb{E}_{x \sim p(x)} \mathbb{E}_{p_E(z|x)} \mathbb{E}_{\tilde{x} \sim \bar{p}_D(x|z)} [c(x, \tilde{x})] \rightarrow \mathbb{E}_{x \sim p(x)} \mathbb{E}_{p_E(z|x)} \mathbb{E}_{\tilde{x} \sim p_D(x|z)} \left[\frac{\mathbf{1}_S(\tilde{x})}{P_D(S)} c(x, \tilde{x}) \right]. \quad (4.17)$$

Letting θ denote the parameters of the model, it can be shown that the gradient of the modified cost function has the simple form

$$\nabla_{\theta} \frac{\mathbf{1}_S(\tilde{x})}{P_D(S)} c(x, \tilde{x}) = \frac{\mathbf{1}_S(\tilde{x})}{P_D(S)} \nabla_{\theta} c(x, \tilde{x}). \quad (4.18)$$

This means that training an (S)WAE with a restricted decoder by stochastic gradient descent proceeds as in the unrestricted base training strategy, except that only the valid samples in S contribute to the gradient of the data loss term, with the contribution scaled inversely by the factor $P_D(S)$, which can be estimated by drawing samples $\tilde{x}'_m \sim p_D(x)$ ⁶ and forming the Monte-Carlo estimate

$$P_D(S) \approx \frac{1}{M} \sum_{m=1}^M \mathbf{1}_S(\tilde{x}'_m). \quad (4.19)$$

⁶This is equivalent to passing $z_m \sim p(z)$ through the decoder to produce \tilde{x}'_m .

4.6.3.3 Parameter Optimization

For the $Z \rightarrow e^+e^-$ study, we used the base training strategy. We optimized $\hat{\mathcal{L}}_{\text{SWAE}}$ for 80 epochs with anchor penalties $\beta_E = \beta_D = 50$, followed by another 800 epochs with the anchor penalties set to 0. For the semileptonic $t\bar{t}$ study, we modified the base training strategy to accommodate a restricted decoder, substituting all appearances of $p_D(x)$ in the loss $\hat{\mathcal{L}}_{\text{SWAE}}$ by $\bar{p}_D(x)$ (e.g. using the modified data loss term Equation (4.17)). We optimized the resulting loss $\hat{\mathcal{L}}_{\text{SWAE}}$ till convergence, for about 1000 epochs. Then we froze the encoder and fine-tuned the decoder by minimizing $\hat{d}_{\text{SW}}(p(x), \bar{p}_D(x))$ for 10 epochs, with a reduced learning rate. The input-to-output residual connection (see Section 4.6.2) in the $t\bar{t}$ model allowed for sufficiently high $P_D(S) \approx 0.6$ and reliable gradient estimates during training, and the architectural bias towards identity mapping made the anchor losses redundant, so we set $\beta_E = \beta_D = 0$.

In both studies, we found that a sufficiently large batch size significantly improved results. This is likely do to increasing the accuracy of gradient estimates for stochastic gradient descent and also the CDF^{-1} in the SWAE latent loss. In all of our experiments, we used the Adam optimizer [97] with $L = 1,000$ number of slices, a batch size of $M = 20,000$, and learning rate of 0.001. We tuned the λ hyperparameter of the (S)WAE loss $\mathcal{L}_{\text{SWAE}}^{\hat{}}$ on the validation set; we set $\lambda = 1$ for the $Z \rightarrow e^+e^-$ model, and $\lambda = 20$ for the $t\bar{t}$ model.

4.6.4 Evaluation

This section provides details on the various qualitative and quantitative evaluation techniques used in this work.

As common in the literature [26, 68, 70], we visualize our results along informative one-dimensional projections using histograms (e.g. Fig. 4.3 and Fig. 4.5). We choose the bin sizes such that the error on the counts can be approximated as Gaussian distributed. These

histograms are accompanied by residual plots, showing the ratio between the histograms from generated samples and the histogram from true samples, with accompanying statistical errors [98]⁷. We also visualize the generative mappings using transportation plots (e.g. Fig. 4.4) that allow us to confirm the physicality of the learned mappings.

In addition to qualitative comparisons, we also evaluated the results using several quantitative metrics. To this end, we calculate the Monte-Carlo estimate of the SW distance, $\hat{d}_{\text{sw}}(\cdot, \cdot)$, using $L = 1,000$ slices according to the cost metric $c(u, v) = \|u - v\|^2$. The results are reported for each study in the text. In addition, we apply several statistical tests on the considered one-dimensional projections, which we report in Tables 1-4 in Appendix B. First, we calculate the reduced χ^2 , χ_R^2 , for each comparison and report it along with the degrees-of-freedom (dof). Second, we calculate the unbinned two-sample, two-sided Kolmogorov-Smirnov distance. Lastly, we calculate the Monte-Carlo estimate of the Wasserstein distance, $\hat{d}_W(\cdot, \cdot)$, according to the cost metric $c(u, v) = \|u - v\|^2$. All statistical tests were carried-out using two separate test sets not used during training or validation of the networks. The number of samples in each test set were 80,000 in the $Z \rightarrow e + e^-$ study and 47,856 in the semileptonic $t\bar{t}$ study.⁸

4.7 Conclusion

OTUS is a data-driven, machine-learned, predictive simulation strategy which suggests a possible new direction for alleviating the prohibitive computational costs of current Monte-Carlo approaches, while avoiding the inherent disadvantages of other machine-learned approaches. We anticipate that the same ideas can be applied broadly outside of the field of particle

⁷Specifically, for a bin with counts h_1 and h_2 , respectively, the error on the ratio, $r = h_2/h_1$ is $\sigma_r = r\sqrt{\frac{1}{h_2} + \frac{1}{h_1}}$.

⁸Note that the number of samples in the semileptonic $t\bar{t}$ study is lower due to the hard p_T cutoff constraint as described in section 4.5.2. The events present are ones that passed this cutoff constraint.

physics.

In general, OTUS can be applied to any process where unobserved latent phenomena \mathcal{Z} can be described in the form of a prior model, $p(z)$, and are translated to an empirical set of experimental data, \mathcal{X} , via an unknown transformation. For example, in molecular simulations in chemistry observations could be measurements of real-world molecular dynamics, $p(z)$ would represent the model description of the system, and $p(x | z)$ would model the effects of real-world complications [64]. In cosmology, \mathcal{X} could be the distribution of mass in the observed universe, $p(z)$ could describe its distribution in the early universe, and $p(x | z)$ would model the universe’s unknown expansion dynamics (e.g. due to inflation) [66, 99]. In climate simulations, $p(z)$ could correspond to the climate due to a physical model, while $p(x | z)$ takes unknown geography-specific effects into account [65]. Additionally, an immediate and promising application of OTUS is in medical imaging, which uses particle physics simulations to model how the imaging particles (e.g x-rays) interact with human tissue and suffers from the great computational cost of these simulations [67]. We note that our method assumes a high degree of mutual information between \mathcal{Z} and \mathcal{X} in the desired application. Therefore, in situations where such mutual information is low (e.g. chaotic turbulent flows) the transformations learned by this method would likely be less reliable.

Moreover, features of this method can be adapted to suit the particular problem’s needs. For example, in this work we were interested in low-dimensional data, however the method could also be applied to high-dimensional datasets. Moreover, the encoding and decoding mappings can be stochastic, as in this work, or deterministic. Lastly, while this work aimed to be completely unsupervised, and thus data-driven, OTUS can be easily extended to a semi-supervised setting. In this case, the data would consist mostly of unpaired samples but would have a limited number of paired examples (z, x) (e.g. from simulation runs). These pairs sample the joint distribution, $p(z, x)$, which, combined with the decoder $p_D(\tilde{x} | z)$, yields a transportation map γ between $p(x)$ and $p_D(\tilde{x})$, $\gamma(p(x), p_D(\tilde{x})) := \int dz p(z, x) p_D(\tilde{x} | z)$. Since

calculating the Wasserstein distance between $p(x)$ and $p_D(\tilde{x})$ involves finding the optimal transportation map, this particular choice yields an upper bound on the Wasserstein distance. We can similarly construct a transportation map between $p(z)$ and $p_E(\tilde{z})$ using $p(z, x)$ and $p_E(z | x)$. This makes directly optimizing the Wasserstein distances $d_W(p(x), p_D(x))$ and $d_W(p(z), p_E(z))$ tractable in this high-dimensional setting. Therefore, we get the alternative objectives

$$\mathcal{L}_{\text{paired}}(p_D(x | z), p(z, x)) = \mathbb{E}_{(z,x) \sim p(z,x)} \mathbb{E}_{\tilde{x} \sim p_D(x|z)} [c(x, \tilde{x})] \quad (4.20)$$

$$\mathcal{L}_{\text{paired}}(p_E(z | x), p(z, x)) = \mathbb{E}_{(z,x) \sim p(z,x)} \mathbb{E}_{\tilde{z} \sim p_E(z|x)} [c(z, \tilde{z})], \quad (4.21)$$

which are upper bounds on $d_W(p(x), p_D(x))$ and $d_W(p(z), p_E(z))$ respectively. These terms can be incorporated alongside the unsupervised SWAE loss, to leverage paired examples $\{(z, x) \sim p(z, x)\}$ in a semi-supervised setting.

We have demonstrated the ability of OTUS to learn a detector transformation in an unsupervised way. The results, while promising for this initial study, leave room for improvement. Several directions could lead to higher fidelity descriptions of the data and latent spaces.

First, the structure of the latent and data spaces can significantly affect the performance and physicality of the resulting simulations. Particle physics data has rich structures often governed by group symmetries and conservation laws. Our current vector format description of the data omits much of this complicated structure. For example, we omitted categorical characteristics of particles like charge and type. Knowledge of such properties and the associated rules likely would have excluded the necessity of terms like the anchor loss. Therefore, future work should explore network architectures and losses that can better capture the full nature of these data structures [50, 100].

The next technical hurdle is the ability to handle variable input and output states. The same $p(z)$ can lead to different detected states as was described, but not explored, in the

semileptonic $t\bar{t}$ study where the number of jets can vary. Additionally, it should be possible to handle mixtures of underlying priors in the latent space. This can cause the number and types of latent-space particles to vary from one sample to another. For example, the Z boson can decay into $Z \rightarrow \mu^+\mu^-$ in addition to $Z \rightarrow e^+e^-$; a simulator should be able to describe these two cases holistically.

Finally, an essential feature of a predictive simulator is that it learns a general transformation, allowing it to make predictions for points in the latent space which lie outside of the control regions. This would require structuring the latent and data spaces to accommodate data from several control regions, such that the network may learn to interpolate between them. Since networks excel at interpolation we expect that this will be a straightforward step.

4.8 Data Availability

The datasets generated and analysed during the current study are available in the DRYAD repository, doi: 10.7280/D1WQ3R.

4.9 Code Availability

The code used during the current study are available in the Zenodo repository and are linked to the dataset, doi: 10.5281/zenodo.4706055.

4.10 Author Contributions Statement

Using the CASRAI CRediT Contributor Roles Taxonomy: Conceptualization, J.N.H., S.M., D.W., Y.Y.; Data curation, J.N.H.; Formal analysis, J.N.H., Y.Y.; Funding acquisition, J.N.H., S.M., D.W.; Investigation, J.N.H., S.M., D.W., Y.Y.; Methodology, J.N.H., Y.Y.; Project administration, J.N.H., S.M., D.W.; Software, J.N.H., Y.Y.; Supervision, S.M., D.W.; Validation, Y.Y.; Visualization, J.N.H.; Writing – original draft, J.N.H., D.W.; Writing – review & editing, J.N.H., S.M., D.W., Y.Y..

Chapter 5

Dark Matter Freeze-out during $SU(2)_L$ Confinement

This chapter is heavily based on work previously published in collaboration with Seyda Ipek, Tim M.P. Tait, and Jessica Turner [2].

5.1 Introduction

The identity of the dark matter and its role in a theory of fundamental interactions remains one of the most pressing open questions today, and drives a vibrant program of experimental and theoretical research into Physics beyond the Standard Model (SM) [101]. A key property that distinguishes among different possibilities is the nature of the interactions between the dark matter and the ingredients of the Standard Model, typically characterized by the masses and couplings of the mediator particles.

An economical choice is to allow the dark matter to transform under the SM's $SU(2)_L$ weak interaction, repurposing the electroweak bosons of the Standard Model (W , Z , and h) as the

mediators. This results in a prototypical weakly interacting massive particle (WIMP), whose abundance in the Universe can be naturally understood as a result of it freezing out after an initial period of chemical equilibrium with the SM plasma [102]. While attractive, an $SU(2)_L$ -charged WIMP whose abundance is set by freeze-out is highly constrained. The TeV masses favored by the dark matter abundance often predict signals which are expected to have been visible at colliders [103, 104], in searches for ambient dark matter scattering with heavy nuclei [105], and by searches for high energy annihilation products which make their way to the Earth [106]. With dominant couplings typically fixed by $SU(2)_L$ gauge invariance, a specific choice of $SU(2)_L$ -charged WIMP freezes out with the correct abundance for only a very narrow range of masses. While windows of viable parameter space exist (see e.g. Ref. [13]), many types of $SU(2)_L$ -charged WIMPs naively appear to be excluded as relics whose abundance is determined by freeze-out.

An $SU(2)_L$ -charged WIMP typically freezes out at a temperature $\simeq M/20$, which for an electroweak-sized mass corresponds to a period of cosmology that is much earlier than Big Bang Nucleosynthesis, and thus during an epoch that is relatively unconstrained by observational data. At this time, the Universe may deviate dramatically from our extrapolation based on the SM, due to unforeseen Physics beyond the Standard Model. Indeed, explorations of non-standard cosmological histories, including a period of early matter domination [7], late entropy injection [8], and modifications of fundamental parameters such as the strength of the $SU(3)$ coupling [9, 10] have all been shown to lead to dramatically different expectations in the mapping of WIMP parameter space onto its predicted abundance in the early Universe.

This chapter explores a non-standard cosmology that can dramatically change the favored mass range for an $SU(2)_L$ -charged WIMP, which makes up the bulk of the dark matter. We introduce dynamics that modify the value of the $SU(2)_L$ interaction strength very early, causing it to confine [11]. This weak confinement causes the left-chiral quarks and leptons of

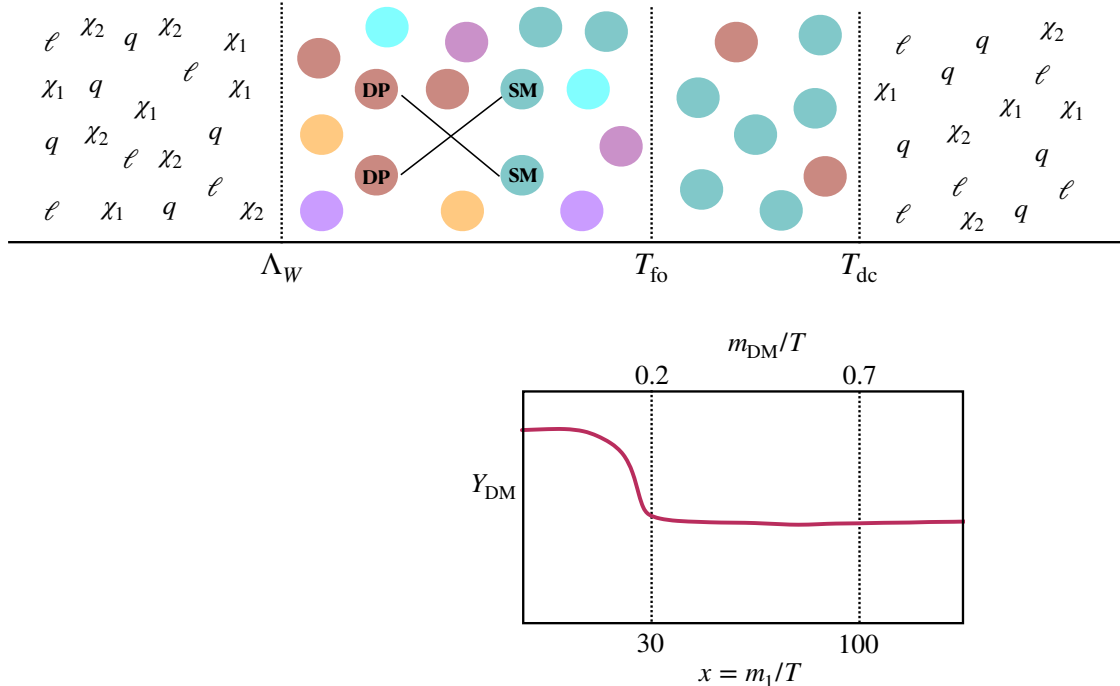


Figure 5.1: **Schematic diagram of the weak confinement and dark pion freeze-out.** Upper panel: A sketch of the cosmological history of the Universe where we assume a period of weak confinement begins at Λ_W , at which point the DM (χ_1 , χ_2) and SM (q , ℓ) doublets are bound into weak pions. During this epoch, the freeze-out of dark pions takes place at T_{fo} , followed by deconfinement at T_{dc} . Lower panel: The evolution of the dark pion abundance for a representative value of the freeze-out temperature $x_{fo} = m_1/T_{fo} \simeq 30$, corresponding to a temperature of $0.2m_{DM}$. In our notation, m_1 and m_{DM} denote the lightest dark pion and the constituent dark matter masses respectively, see Section 5.3 for details.

the SM, and a new vector-like pair of fermionic doublets that plays the role of dark matter, to bind into composite pion-like states that are $SU(2)_L$ neutral. The freeze-out process involves those pions containing the dark matter annihilating into lighter pions composed entirely of SM fermions. At some time after freeze-out, the $SU(2)_L$ interaction returns to its currently observed value, at which point the pions deconfine, leaving behind the frozen out dark matter. A sketch of this cosmological history is shown in Fig. 5.1.

Our work is organized as follows: in Section 5.2, we introduce the description of the Universe during an early period of $SU(2)_L$ confinement, including an additional vector-like pair of fermionic doublets which can play the role of dark matter. In Section 5.3 we discuss the freeze-out process in detail and identify the parameter space leading to the observed

abundance of dark matter today and our results are summarized in Fig. 5.4. The more realistic case including three generations of SM fermions is discussed in Section 5.4. Finally, we conclude in Section 5.6 and provide technical details in the appendices.

5.2 Weak Confinement and Dark Matter

Our dark matter production mechanism involves a temporary cosmological era of $SU(2)_L$ confinement. The possibility that the weak sector was strong in the early universe was initially proposed in [107–110] (see also [111–114]) and the cosmological consequences of such a scenario were studied in Ref. [11]. We refrain from rederiving the complete results of Ref. [11], which gives a detailed discussion of the gauge and global symmetry breaking patterns as well as the particle content of the confined phase, and instead highlight some key results pertinent for this work:

- Weak confinement causes the $SU(2)_L$ doublets to condense into bound states analogous to the mesons and baryons of QCD. The lowest-lying states are mesons, Π and η' , composed of the SM lepton and quark doublets, l and q respectively. These states are contained in the complex antisymmetric scalar field, Σ_{ij} , where $i, j = 1, \dots, 2N_f$ with $2N_f$ of left-chiral Weyl fermion fields. For the Standard Model with three generations, $N_f = 6$.
- Following intuition based on chiral symmetry breaking in QCD [115, 116] and evidence from lattice simulations, there is a chiral condensate spontaneously breaking the global symmetry: $SU(2N_f) \rightarrow Sp(2N_f)$ [117–123]. This pattern of symmetry breaking is encoded by the antisymmetric field Σ_{ij} acquiring a vacuum expectation value $\langle \Sigma_{ij} \rangle = (\Sigma_0)_{ij}$ that satisfies $\Sigma_0^\dagger \Sigma_0 = \Sigma_0 \Sigma_0^\dagger = \mathbf{1}$. Neglecting the other SM gauge interactions and Yukawas, this symmetry breaking results in $2N_f^2 - N_f - 1$ massless Goldstone

bosons (GBs) and a single massive pseudo-Goldstone boson (PGB), analogous to the η' of QCD.

- The dynamics of the confined theory are described by an infrared Lagrangian which is constructed from the scalar field Σ_{ij} that contains the massive PGB and massless GBs:

$$\Sigma = \exp \left[i\eta' / \sqrt{N_f} f \right] \exp \left[\sum_a 2iX^a \Pi^a / f \right] \Sigma_0 , \quad (5.1)$$

where X_{ij}^a are the $2N_f^2 - N_f - 1$ broken generators of $\text{Sp}(2N_f)$ and f is the decay constant. Considering the three SM generations of $\text{SU}(2)_L$ doublets, there are 65 massless pions. However, loop-induced corrections from the SM gauge and Yukawa interactions provide masses to 58 of the 65 pions.

- Weak confinement breaks the gauge symmetry of the Standard Model from $\text{SU}(3)_C \times \text{U}(1)_Y$ to $\text{SU}(2)_C \times \text{U}(1)_Q$, resulting in four massless gauge bosons ($G^{1,2,3}, A'$) and five massive gauge bosons, which can be arranged into a pair of complex gauge bosons (W'^{\pm}) and single real vector boson (Z').

We augment the SM particle content by two $\text{SU}(2)_L$ doublets, χ_1 and χ_2 (of hypercharges $\pm 1/2$, respectively), which play the role of dark matter. They are assembled into a pseudo-Dirac state,

$$\mathcal{L}_\chi = i\chi_1^\dagger \bar{\sigma}^\mu D_\mu \chi_1 + i\chi_2^\dagger \bar{\sigma}^\mu D_\mu \chi_2 + m_{\text{DM}} \chi_1 \chi_2 + \text{h.c.} , \quad (5.2)$$

where D_μ is a covariant derivative of the unconfined phase and m_{DM} is the mass of the constituent dark matter. This Lagrangian is invariant under a $\text{U}(1)_\chi$ symmetry under which χ_1 (χ_2) are charged ± 1 that ensures their stability.

The infrared Lagrangian, describing the dynamics of the confined theory, has the form

$$\mathcal{L}_{\text{IR}} \supset \frac{f^2}{4} \text{Tr} [D_\mu \Sigma^\dagger D^\mu \Sigma] + \Lambda_W^3 \text{Tr} [M \Sigma + \Sigma^\dagger M^T] + \kappa \Lambda_W^2 f^2 \text{Re}[\det \Sigma] + \Delta \mathcal{L}, \quad (5.3)$$

where D_μ is a covariant derivative of the confined phase, $\Lambda_W \sim 4\pi f$ is the weak confinement scale, κ is an $\mathcal{O}(1)$ dimensionless number, and M is the mass matrix, treated as an $\text{SU}(2N_f)$ -breaking spurion in the limit $m_{\text{DM}} \ll \Lambda_W$. In the simplified case where we consider a single generation of $\text{SU}(2)_L$ doublets together with the dark matter, $2N_f = 6$ and the mass matrix, defined in the basis $\{\ell, q^R, q^G, q^B, \chi_1, \chi_2\}$ where R , G , and B denote the colors of $\text{SU}(3)_C$, is:

$$M = \frac{m_{\text{DM}}}{2} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}. \quad (5.4)$$

The infrared Lagrangian also contains operators reflecting the explicit breaking of $\text{SU}(2N_f)$ by the gauging of $\text{SU}(3)_C$ and $\text{U}(1)_Y$:

$$\begin{aligned} \Delta \mathcal{L} = & C_G \Lambda_W^2 f^2 \frac{g_s^2}{16\pi^2} \sum_{a=1,2,3} \text{Tr}[L^a \Sigma^\dagger L^{aT} \Sigma] + C_A \Lambda_W^2 f^2 \frac{e_Q^2}{16\pi^2} \text{Tr}[Q \Sigma^\dagger Q \Sigma] \\ & + C_W \Lambda_W^2 f^2 \frac{g_s^2/2}{16\pi^2} \sum_{\pm} \sum_{i=1,2} \text{Tr}[L^{i\pm} \Sigma^\dagger L^{i\pm} \Sigma] + C_Z \Lambda_W^2 f^2 \frac{e_Q^2/s_Q^2 c_Q^2}{16\pi^2} \text{Tr}[J \Sigma^\dagger J \Sigma], \end{aligned} \quad (5.5)$$

where the dimensionless coefficients, C_G , C_A , C_W , and C_Z , encode the non-perturbative $\text{SU}(2)_L$ dynamics, and are expected to be $\mathcal{O}(1)$ [124, 125]. The $\text{SU}(2)_C$ and hypercharge couplings are denoted as g_s and g' , respectively, and $\sin \theta_Q = g'/\sqrt{3g_s^2 + g'^2}$ with $e_Q \approx g'$ in the limit, $g' \ll g_s$. The generators of the $\text{SU}(2)_C$ and $\text{U}(1)_Q$ are denoted as L^a and Q ,

respectively, and L^\pm is a combination of $SU(3)_C$ generators which couple to the massive vector fields W'^\pm . Finally, J is a combination of an $SU(3)_C$ and an $U(1)_Q$ generator which couple to the massive Z' gauge boson (see Appendix D and Ref. [11] for further details).

For the remainder of this Section, we consider a simplified toy model consisting of one SM generation of fermionic doublets together with χ_1 and χ_2 (corresponding to $N_f = 3$, for which there are 14 broken generators of the $SU(6)$ flavor symmetry). This allows us to extract the most important points in a framework that is simpler to analyze. We return to the more realistic case of three generations plus $\chi_{1,2}$ (corresponding to $N_f = 14$) in Section 5.4.

5.2.0.1 Pion Masses and Mass Eigenstates

The mass spectrum of the pions during weak confinement is determined from the terms of Eq. (5.3) that, after plugging in the expression for $\Sigma(\Pi)$ in Eq. (5.1), are quadratic in the meson fields, $\mathcal{L}_{\text{IR}} \rightarrow -(1/2)(M_\Pi^2)_{ab}\Pi^a\Pi^b$. Following Ref. [11], we define M_Π^2 in the basis $\Pi = \{\eta', \Pi^a\}$ where $a = 1\dots 14$. In contrast to the case studied in [11], the resulting mass matrix contains non-diagonal entries mixing the η' with the meson dominantly composed of $\chi_1\chi_2$:

$$M_\Pi^2 = \begin{pmatrix} M_{0,0}^2 & \cdot & \cdot & \cdot & M_{0,14}^2 \\ \cdot & M_{1,1}^2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & M_{13,13}^2 & \cdot \\ M_{0,14}^2 & \cdot & \cdot & \cdot & M_{14,14}^2 \end{pmatrix}, \quad (5.6)$$

and thus the interaction and mass eigenstates are not aligned. We rotate to the mass basis via the unitary transformation $\Pi \rightarrow W\Pi$, for which

$$M_{\text{diag}}^2 = WM_{\Pi}^2W^{-1}, \quad (5.7)$$

where W is a unitary matrix

$$W = \begin{pmatrix} \cos\theta & . & . & . & \sin\theta \\ . & 1 & . & . & . \\ . & . & . & . & . \\ . & . & . & 1 & . \\ -\sin\theta & . & . & . & \cos\theta \end{pmatrix}, \quad (5.8)$$

with

$$\tan 2\theta = 2\frac{M_{0,14}^2}{(M_{0,0}^2 - M_{14,14}^2)}, \quad (5.9)$$

and

$$M_{0,0}^2 = 24\kappa\Lambda_W^2 + \frac{2\Lambda_W^3 m_{\text{DM}}}{3f^2}, \quad M_{0,14}^2 = -\frac{2\sqrt{2}\Lambda_W^3 m_{\text{DM}}}{3f^2}, \quad M_{14,14}^2 = \frac{4\Lambda_W^3 m_{\text{DM}}}{3f^2}. \quad (5.10)$$

Substituting Eq. (5.10) into Eq. (5.9), we find that

$$\tan 2\theta = \frac{2\sqrt{2}\pi m_{\text{DM}}}{\pi m_{\text{DM}} - 9\kappa f} \approx -\frac{2\sqrt{2}\pi m_{\text{DM}}}{9\kappa f} + \mathcal{O}\left(\frac{m_{\text{DM}}^2}{f^2}\right), \quad (5.11)$$

where we have taken $\Lambda_W = 4\pi f$ (see Appendix F for more details). Throughout we assume that $m_{\text{DM}} \ll f$ and this implies that the mixing between η' (which we label as Π_0) and the

Pion	Mass ²	U(1) _Q	SU(2) _C	content
Π_0^{mass}	$384\pi^2 f^2 \kappa$	0	1	χ_1, χ_2
$\Pi_{1,2,3,4}^{\text{mass}}$	$-\frac{1}{2}C_A e_Q^2 f^2 - \frac{3}{2}C_G f^2 g_s^2 + C_W f^2 g_s^2 + \frac{C_Z e_Q^2 f^2}{6s_Q^2} + \frac{1}{2}C_Z e_Q^2 f^2$	± 1	2	ℓ, q_D, q_S
$\Pi_{5,8}^{\text{mass}}$	$64\pi^3 f m_{\text{DM}}$	0	1	χ_1, χ_2, q_S
$\Pi_{6,7}^{\text{mass}}$	$-2C_A e_Q^2 f^2 - 2C_Z e_Q^2 f^2 s_Q^2 + \frac{2}{3}C_Z e_Q^2 f^2 + 64\pi^3 f m_{\text{DM}}$	± 1	1	$\ell, \chi_1, \chi_2, q_S$
$\Pi_{9,10,11,12}^{\text{mass}}$	$-\frac{1}{2}C_A e_Q^2 f^2 - \frac{3}{2}C_G f^2 g_s^2 + \frac{C_Z e_Q^2 f^2}{18s_Q^2} + 64\pi^3 f m_{\text{DM}}$	± 1	2	χ_1, χ_2, q_D
Π_{13}^{mass}	0	0	1	ℓ, q_S
Π_{14}^{mass}	$\frac{256}{3}\pi^3 f m_{\text{DM}}$	0	1	χ_1, χ_2

Table 5.1: Masses of the pions (for the one SM generation case) in the small mixing limit, along with their U(1)_Q×SU(2)_C charges and constituent SU(2)_L doublet content.

$\chi_1 \chi_2$ (which we label as Π_{14}) state is small, $\cos \theta \approx 1$ and $\sin \theta \approx \theta$, and this leads to:

$$\Pi_0^{\text{mass}} \approx \Pi_0^{\text{int}} + \theta \Pi_{14}^{\text{int}}, \quad (5.12a)$$

$$\Pi_{14}^{\text{mass}} \approx \Pi_{14}^{\text{int}} - \theta \Pi_0^{\text{int}}, \quad (5.12b)$$

where $\Pi_i^{\text{mass}} = \Pi_i^{\text{int}}$ for $i = 1, \dots, 13$. The masses of Π_0^{mass} and Π_{14}^{mass} are:

$$\begin{aligned} M_0^2 &\approx 384\pi^2 f^2 \kappa \left(1 + \frac{\pi m_{\text{DM}}}{9\kappa f} + \mathcal{O}\left(\frac{m_{\text{DM}}^2}{f^2}\right) \right), \\ M_{14}^2 &\approx \frac{256\pi^3}{3} f m_{\text{DM}} \left(1 - \frac{\pi m_{\text{DM}}}{9\kappa f} + \mathcal{O}\left(\frac{m_{\text{DM}}^2}{f^2}\right) \right). \end{aligned} \quad (5.13)$$

Table 5.1 shows the approximate masses of the 15 mesons for the one generation SM case, as well as their representations under the residual U(1)_Q × SU(2)_C gauge symmetries, in the small mixing limit.

The specific pion masses depend on the non-perturbative coefficients C_G, C_A, C_Z, C_W , and κ . These could in principle be determined from lattice simulations, and are expected to be $\mathcal{O}(1)$ based on arguments from naive dimensional analysis [126]. We proceed under the assumption that $C_G = C_A = C_Z = -1$ and $C_W = \kappa = 1$. As is evident from Table 5.1, the masses of $\Pi_{1,2,3,4}^{\text{mass}}$ are independent of m_{DM} , reflecting the fact that they are purely composed of SM quark and lepton doublets, with masses generated via SM gauge interactions, Eq. (5.5),

and are typically the lightest of the massive pions. The Π_0^{mass} is significantly heavier than the other mesons, rendering it unimportant for the freeze-out dynamics due to Boltzmann suppression. We observe that Π_{14}^{mass} is 4/3 times heavier than $\Pi_{5,8}^{\text{mass}}$ and hence, we can ignore the effect of Π_{14}^{mass} in calculating the dark matter dynamics. In Fig. 5.2, we show the pion masses as a function of m_{DM} for $f = 65$ TeV, corresponding to $\Lambda_W \approx 800$ TeV (this choice is motivated by discussions of DM abundance in Section 5.3). We examine two benchmark cases: BP1 where g_s, g' and $s_Q = g' / \sqrt{3g_s^2 + g'^2}$ are found by evaluating the running SM coupling constants to approximately Λ_W and BP2, which is similar to a regime of interest from Ref. [11]. More specifically:

$$\begin{aligned} \text{BP1} \quad & g_s = 0.8, \quad e_Q = 0.5, \quad s_Q^2 = 0.12, \\ \text{BP2} \quad & g_s = 0.1, \quad e_Q = 0.01, \quad s_Q^2 = 3.3 \times 10^{-3}. \end{aligned}$$

Fig. 5.2 indicates that $M_{5,8}, M_{6,7}$ and $M_{9,10,11,12}$ differ slightly due to the loop contributions, and that $M_{5,8}$ are the lightest massive states. BP2 has values of g_s, e_Q which are smaller than those in BP1, leading to much smaller differences between $M_{5,8}, M_{6,7}$ and $M_{9,10,11,12}$, resulting in a more compressed spectrum.

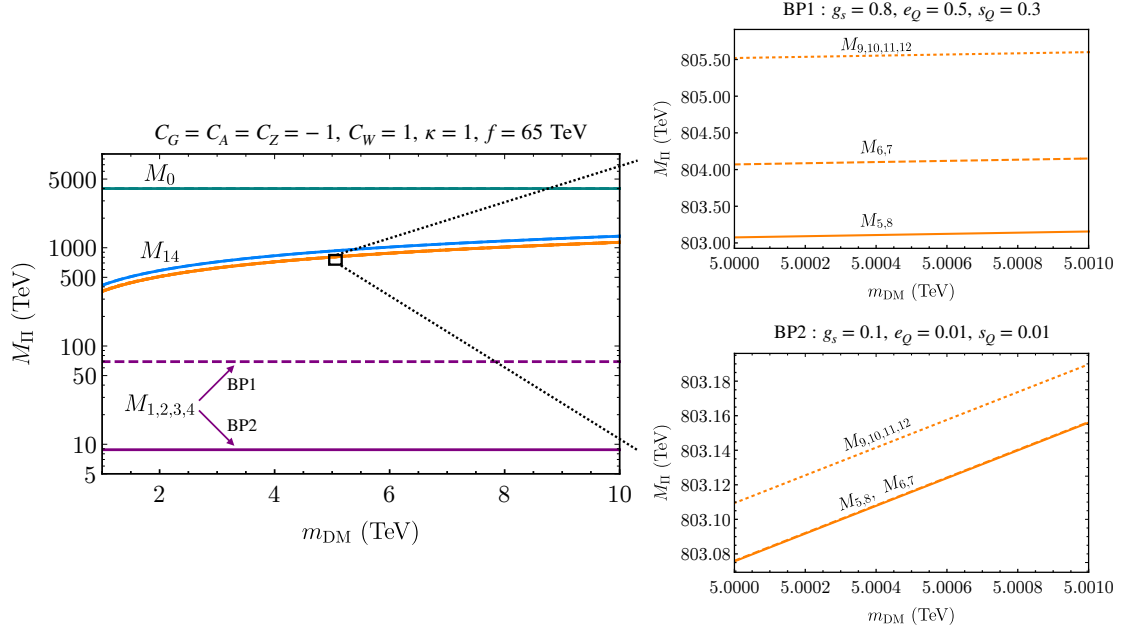


Figure 5.2: **Pion masses as a function of m_{DM} , assuming $C_G = C_A = C_Z = -1$, $C_W = 1$ and $\kappa = 1$, for two benchmark points: BP1 where $g_s, e_Q \simeq g'$ and $s_Q \simeq g'/\sqrt{3g_s^2 + g'^2}$ are found by running g_s and g' to $\Lambda_W = 4\pi f \simeq 800$ TeV; and BP2 where we take $g_s = 0.1$ and $e_Q = 0.01$. $M_{13} = 0$ is not shown.**

5.2.0.2 $U(1)_\chi$ Eigenstates

$U(1)_\chi$ remains unbroken during the confined phase, and it is convenient to organize the pions based on their $U(1)_\chi$ charges. This is evident from the fact that the $U(1)_\chi$ generator,

$$Q_\chi = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}, \quad (5.14)$$

leaves the vacuum invariant: $Q_\chi \Sigma_0 + \Sigma_0 Q_\chi = 0$. To infer the $U(1)_\chi$ charges of the pions, we transform Σ by an infinitesimal $U(1)_\chi$ rotation:

$$\Sigma \xrightarrow{U(1)_\chi} e^{iQ_\chi \theta_\chi} \Sigma (e^{iQ_\chi \theta_\chi})^T \approx \Sigma + i\theta_\chi (Q_\chi \Sigma + \Sigma Q_\chi) + \dots, \quad (5.15)$$

and expand Σ to first order, $\Sigma \simeq \Sigma_0 + \frac{i}{f} \Pi_a X_a \Sigma_0 + \dots$, from which we can extract the transformation of each pion:

$$\Pi_b \xrightarrow{U(1)_\chi} \Pi_b + i\theta_\chi \underbrace{2\Pi_a \text{Tr}[[Q_\chi, X_a], X_b]}_{\delta\Pi_b}. \quad (5.16)$$

Using the specific form of the generators X_a and Q_χ we can explicitly evaluate $\delta\Pi_a$ for each $a = 0, \dots, 14$, and construct complex linear combinations of pion fields that have definite $U(1)_\chi$ charge:

$$\begin{aligned} \tilde{\Pi}_1^\pm &\equiv \frac{1}{\sqrt{2}} (\Pi_5^{\text{mass}} \mp i\Pi_8^{\text{mass}}), \\ \tilde{\Pi}_2^\pm &\equiv \frac{1}{\sqrt{2}} (\Pi_6^{\text{mass}} \mp i\Pi_7^{\text{mass}}), \\ \tilde{\Pi}_3^\pm &\equiv \frac{1}{\sqrt{2}} (\Pi_9^{\text{mass}} \mp i\Pi_{12}^{\text{mass}}), \\ \tilde{\Pi}_4^\pm &\equiv \frac{1}{\sqrt{2}} (\Pi_{10}^{\text{mass}} \mp i\Pi_{11}^{\text{mass}}), \end{aligned} \quad (5.17)$$

and $\tilde{\Pi}_i^0 \equiv \Pi_i^{\text{mass}}$ for $i \in \{0, 1, 2, 3, 4, 13, 14\}$ are left as zero-charge real scalar fields. Note that these redefinitions commute with the mass basis, as expected.

5.2.0.3 Pion Interactions

The most important interactions of the pions, for our purposes, are four-point vertices arising as residual strong interactions from the confined $SU(2)_L$ force. These are encoded in the infrared Lagrangian as higher order terms (in powers of Π/f). Expanding Σ to second

order:

$$\begin{aligned}\Sigma(x) &= \exp\left[\frac{i\eta'}{\sqrt{N_f f}}\right] \exp\left[i\frac{2\Pi_a(x)X_a}{f}\right] \Sigma_0 \\ &\approx \left[1 + i\left(\frac{2\Pi_a(x)X_a}{f}\right) - \frac{1}{2}\left(\frac{2\Pi_a(x)X_a}{f}\right)^2 + \mathcal{O}\left(\frac{\Pi^3}{3!f^3}\right)\right] \Sigma_0,\end{aligned}\tag{5.18}$$

where the relevant terms from Eq. (5.3) take the form:

$$\mathcal{L}_4 = \frac{4}{f^2} \text{Tr}_1(a, b, c, d) \Pi_a \Pi_b \partial^\mu \Pi_c \partial_\mu \Pi_d + \frac{2m_{\text{DM}}\Lambda_W^3}{3f^4} \text{Tr}_2(a, b, c, d) \Pi_a \Pi_b \Pi_c \Pi_d,\tag{5.19}$$

with flavor tensors Tr_1 and Tr_2 defined by

$$\begin{aligned}\text{Tr}_1(a, b, c, d) &\equiv \frac{1}{4} \left(\text{Tr} [X_c X_a X_d X_b] + \text{Tr} [X_a X_c X_d X_b] \right) \\ &\quad - \frac{1}{12} \left(\text{Tr} [X_c X_a X_b X_d] + \text{Tr} [X_a X_c X_b X_d] \right) - \frac{1}{3} \text{Tr} [X_a X_b X_c X_d], \\ \text{Tr}_2(a, b, c, d) &\equiv -\text{Tr} [A X_a X_b X_c X_d],\end{aligned}\tag{5.20}$$

where $A \equiv \text{diag}(\mathbb{0}_{2 \times 2}, \dots, \mathbb{0}_{2 \times 2}, \mathbb{1}_{2 \times 2})$. These expressions are written in the interaction (mass) basis, and can be transformed into states of definite $U(1)_X$ charge via Eq. (5.17).

The pions charged under $SU(2)_C$ and $U(1)_Q$ will also have gauge interactions with those gauge bosons, contained in the kinetic terms of Eq. (5.3). However, we have verified that these couplings are small enough at the scales of interest (leading to cross sections of $\mathcal{O}(10^{-3})$ smaller than those characterizing annihilation into SM pions) that they can be neglected in our freeze-out analysis.

5.3 Dark Matter freeze-out

At the time of freeze-out, the dark matter particles are bound into *dark pion* (DP) states, and the final abundances of $\chi_{1,2}$ are determined by the frozen-out densities of $\tilde{\Pi}_{1,2,3,4}^\pm$ (each of which contains one χ) and $\tilde{\Pi}_0^0$ and $\tilde{\Pi}_{14}^0$ (each of which contains two χ s). In practice, because of the large mass hierarchy between $\tilde{\Pi}_{0,14}^0$ and $\tilde{\Pi}_{1,2,3,4}^\pm$, it is sufficient to neglect the contributions from the two neutral states and to focus on the $U(1)_\chi$ -charged ones.

The relic abundance of the $\tilde{\Pi}_i^\pm$ is controlled by the temperature, T_{fo} , at which their number-changing interactions freeze-out from thermal equilibrium, which in turn depends sensitively on their annihilation cross sections into the lightest neutral pions comprised of SM doublets: $\tilde{\Pi}_i^+ \tilde{\Pi}_j^- \rightarrow \tilde{\Pi}_{13}^0 \tilde{\Pi}_{13}^0$. The charged states are typically sufficiently close in mass ($\Delta m/T_{\text{fo}} \sim 10^{-2}$) that coannihilation processes can be important [15, 16], and are included in our calculations. Nonetheless, the relic abundance is dominated by the annihilation of the lightest DP state into the zero-mass SM pion: $\tilde{\Pi}_1^+ \tilde{\Pi}_1^- \rightarrow \tilde{\Pi}_{13}^0 \tilde{\Pi}_{13}^0$.

5.3.1 Annihilation Cross Section

The rate for $\Pi_i \Pi_j \rightarrow \Pi_c \Pi_d$ is determined by the Feynman diagrams shown in Fig. 5.3, where the dashed (solid) lines indicate legs on which derivatives do (do not) act in the corresponding operator. We define the incoming legs to correspond to the pion flavors i, j , and outgoing to c, d . The resulting matrix element, \mathcal{M} , takes the form

$$\begin{aligned}
 i\mathcal{M} = & -i \frac{4(p_c \cdot p_d)}{f^2} G_1 + i \frac{4(p_i \cdot p_c)}{f^2} G_2 - i \frac{4(p_i \cdot p_j)}{f^2} G_3 + i \frac{4(p_j \cdot p_d)}{f^2} G_4 \\
 & + i \frac{4(p_j \cdot p_c)}{f^2} G_5 + i \frac{4(p_i \cdot p_d)}{f^2} G_6 + i \frac{128\pi^3 m_{\text{DM}}}{3f} G_7,
 \end{aligned} \tag{5.21}$$

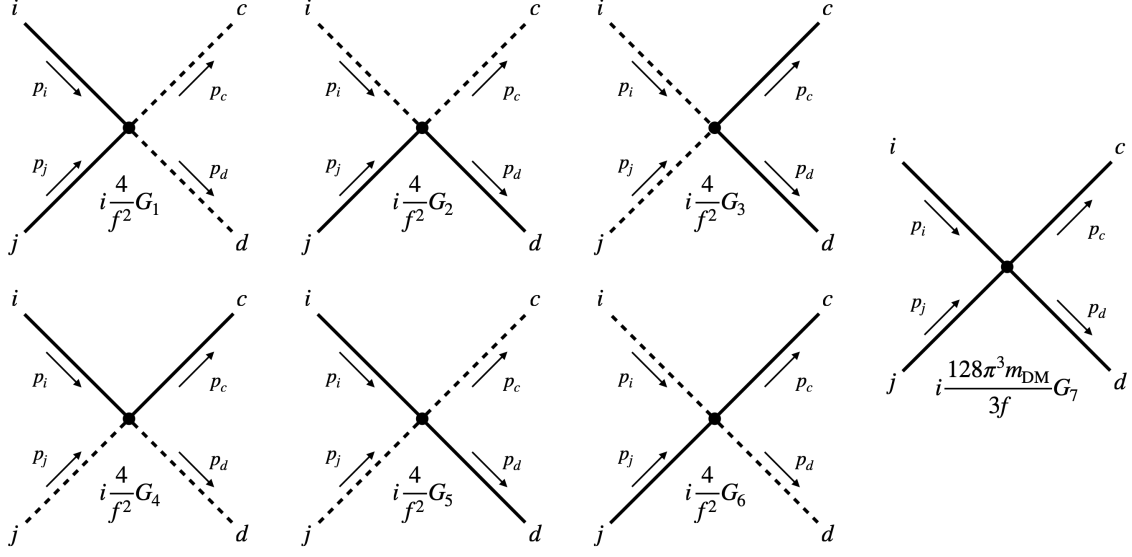


Figure 5.3: **Four-point pion interaction diagrams contributing to the process $\Pi_a \Pi_b \rightarrow \Pi_c \Pi_d$.** The dashed lines denote fields on which derivatives act, contributing a factor of the corresponding momentum. An incoming (outgoing) field contributes a negative (positive) momentum factor to the matrix element.

where we used $\Lambda_W = 4\pi f$ and define:

$$\begin{aligned}
 G_1 &= \text{Tr}_1(i, j, c, d), & G_2 &= \text{Tr}_1(d, j, c, i), & G_3 &= \text{Tr}_1(c, d, i, j), & G_4 &= \text{Tr}_1(i, c, j, d), \\
 G_5 &= \text{Tr}_1(i, d, c, j), & G_6 &= \text{Tr}_1(c, j, i, d), & G_7 &= \text{Tr}_2(i, j, c, d),
 \end{aligned}$$

where Tr_1 and Tr_2 are given in Eq. (5.20). Subsequently, the annihilation cross section can be expressed as,

$$\sigma_{ij}(s) = \frac{1}{16\pi} \frac{1}{\lambda(s, m_i^2, m_j^2)} \left[C_{\text{const}}(t_+ - t_-) + \frac{1}{2} C_{\text{lin}}(t_+^2 - t_-^2) + \frac{1}{3} C_{\text{quad}}(t_+^3 - t_-^3) \right], \quad (5.22)$$

where λ is the well-known Källén function, $\lambda(x, y, z) \equiv (x - y - z)^2 - 4yz$, and

$$\begin{aligned}
 t_+ &= m_c^2 + m_i^2 - 2E_c E_i + 2|\vec{p}_c||\vec{p}_i|, \\
 t_- &= m_c^2 + m_i^2 - 2E_c E_i - 2|\vec{p}_c||\vec{p}_i|.
 \end{aligned} \quad (5.23)$$

The above traces define the coefficients inside the square brackets as:

$$\begin{aligned}
C_{\text{const}} &\equiv \left(\frac{128\pi^3 m_{\text{DM}}}{3f} \right)^2 |C|^2 + \frac{4s^2}{f^4} |G_{13,56}|^2 - \frac{2s}{f^2} \left(\frac{128\pi^3 m_{\text{DM}}}{3f} \right) [C^* G_{13,56} + C G_{13,56}^*] , \\
C_{\text{lin}} &\equiv \frac{4s}{f^4} (G_{13,56} G_{24,56}^* + G_{13,56}^* G_{24,56}) - \frac{2}{f^2} \left(\frac{128\pi^3 m_{\text{DM}}}{3f} \right) [C^* G_{24,56} + C G_{24,56}^*] , \\
C_{\text{quad}} &\equiv \frac{4}{f^4} |G_{24,56}|^2 ,
\end{aligned} \tag{5.24}$$

with

$$\begin{aligned}
C &\equiv G_7 + \frac{3m_i^2}{64\pi^3 f m_{\text{DM}}} (G_2 + G_3 - G_5) + \frac{3m_j^2}{64\pi^3 f m_{\text{DM}}} (G_3 + G_4 - G_6) \\
&\quad + \frac{3m_c^2}{64\pi^3 f m_{\text{DM}}} (G_1 + G_2 - G_6) + \frac{3m_d^2}{64\pi^3 f m_{\text{DM}}} (G_1 + G_4 - G_5) ,
\end{aligned} \tag{5.25}$$

$$G_{13,56} \equiv G_1 + G_3 - G_5 - G_6 ,$$

$$G_{24,56} \equiv G_2 + G_4 - G_5 - G_6 .$$

Note that $64\pi^3 f m_{\text{DM}}$ is the mass squared of the lightest pion with DM constituent. In the non-relativistic limit, the $2 \rightarrow 2$ scattering cross section can be expanded in terms of the relative velocity, $v = |\vec{v}_i - \vec{v}_j|$, of the incoming particles,

$$\langle \sigma v \rangle = \sigma_0 + \sigma_2 \langle v^2 \rangle + .. \tag{5.26}$$

At freeze-out, the leading (s -wave) term of this expansion dominates over the order higher terms and hence the velocity averaged cross section is:

$$\langle \sigma_{ij} v \rangle_{s\text{-wave}} = \frac{\lambda^{1/2}(s, m_c^2, m_d^2)}{32\pi E_a E_b s} [C_{\text{const}} + C_{\text{lin}} W_1 + C_{\text{quad}} W_1^2] , \tag{5.27}$$

where

$$W_1 = m_i^2 + m_c^2 - \frac{1}{2s}(s + m_i^2 - m_j^2)(s + m_c^2 - m_d^2). \quad (5.28)$$

Further, assuming that the incoming particles are non-relativistic implies that $s = (m_i + m_j)^2$. For the most significant annihilation processes, $\tilde{\Pi}_1^+ \tilde{\Pi}_1^- \rightarrow \tilde{\Pi}_{13}^0 \tilde{\Pi}_{13}^0$, $m_c = m_d = 0$ and $m_i^2 \simeq m_j^2 \simeq 64\pi^3 f m_{\text{DM}}$. In this limit, the parametric dependence of the s -wave annihilation cross section is:

$$\langle \sigma_{ijv} \rangle_{s\text{-wave}} \propto \text{constant} \times \frac{m_{\text{DM}}}{f^3}, \quad (5.29)$$

where the overall constant is a combination of various traces and found to be $O(1)$ from numerical analysis.

5.3.2 Freeze-out

The number density of the dark pions, $n_{\text{DP}} = n_{\Pi_1^\pm} + n_{\Pi_2^\pm} + n_{\Pi_3^\pm} + n_{\Pi_4^\pm}$, evolves according the Boltzmann equation [16]:

$$\dot{n}_{\text{DP}} + 3Hn_{\text{DP}} = -\langle \sigma_{\text{eff}} v \rangle (n_{\text{DP}}^2 - n_{\text{DP,eq}}^2), \quad (5.30)$$

where $H = \sqrt{8\pi^3 g_*/90} T^2 / M_{\text{Pl}}$ is the Hubble rate during radiation domination, $n_{\text{DP,eq}} = g_* m_1^2 T / (2\pi^2) K_2(m_1/T)$ is the equilibrium number density of the lightest dark pion and m_1 is the mass of the lightest DP freezing out, more specifically the mass of $\tilde{\Pi}_1^\pm$. The effective

co-annihilation cross section is defined as

$$\sigma_{\text{eff}} = \sum_{i,j=1}^4 \sigma_{ij} \frac{g_i g_j}{g_{\text{eff}}^2} (1 + \Delta_i)^{3/2} (1 + \Delta_j)^{3/2} e^{-x(\Delta_i + \Delta_j)},$$

$$\text{with } g_{\text{eff}} = \sum_{i=1}^4 g_i (1 + \Delta_i)^{3/2} e^{-x\Delta_i},$$
(5.31)

where $x = m_1/T$, σ_{ij} is the cross section for the reaction $\tilde{\Pi}_i^\pm \tilde{\Pi}_j^\mp \rightarrow \tilde{\Pi}^0 \tilde{\Pi}^0$ given in Eq. (5.27) (summed over all kinematically accessible SM pions in the final state), $g_i = 2$ is the number of degrees of freedom of $\tilde{\Pi}_i^\pm$, and $\Delta_i \equiv (m_i - m_1)/m_1$ is the mass difference between the heavier dark pions and $\tilde{\Pi}_1^\pm$.

In Fig. 5.4 we present $\langle \sigma_{\text{eff}} v \rangle$ for a range of f and m_{DM} . By fitting our numerical results to the approximation given in Eq. (5.29), we find the velocity-averaged effective cross section to be

$$\langle \sigma_{\text{eff}} v \rangle \simeq (1.5 - 2) \times 10^{-11} \text{ GeV}^{-2} \left(\frac{m_{\text{DM}}}{5 \text{ TeV}} \right) \left(\frac{65 \text{ TeV}}{f} \right)^3,$$
(5.32)

where the lower and higher values correspond to one or three generations of SM fermions respectively. For smaller f and larger constituent DM mass, $\langle \sigma_{\text{eff}} v \rangle$ is larger, resulting in too much annihilation and hence not enough dark pions left over to produce the observed abundance of the dark matter. Conversely, a lighter constituent dark matter mass and higher confinement scale result in a lower dark pion annihilation cross section and an overabundance of dark matter.

5.3.3 Deconfinement

The freeze-out of the dark pions determines the final comoving number density of dark pions, which has an associated energy density, $\rho_{\text{DP}} = m_1 n_{\text{DP}}$. At the time of deconfinement, each

dark pion flies apart into one χ as well as SM radiation. At that point, the dark matter consists of freely streaming χ particles, with energy density,

$$\rho_{\text{DM}} = \frac{m_{\text{DM}}}{m_1} \times \rho_{\text{DP}} = m_{\text{DM}} \times n_{\text{DP}} , \quad (5.33)$$

which is to be compared with the observed abundance of dark matter from cosmological measurements, $\Omega h^2 = 0.1200 \pm 0.0012$ [127].

We assume that the weak sector deconfines at temperature T_{dc} , where $T_{\text{dc}} \sim m_1/100$. In estimating the relic density of χ , we assume that the entropy dump into the thermal plasma from the deconfinement process is negligible¹. After deconfinement, the free χ particles could begin to annihilate into SM through the now unbound weak interactions, for which the cross section is parametrically $\sigma_{\text{W}} \approx \alpha_{\text{W}}^2 \pi / m_{\text{DM}}^2$, where $\alpha_{\text{W}} \sim 0.1$ has presumably returned to the value measured by experiments today. In our numerical scans, we verify that $\sigma_{\text{W}} n_{\text{DP}} \ll H$ at $x = 100$ for the regions of (m_{DM}, f) of interest, ensuring that no period of thermalization after deconfinement occurs and therefore alters the dark matter relic density from Eq. (5.33).

5.3.4 Numerical Results

We numerically solve Eq. (5.30), adapting the infrastructure of ULYSSES [128], a publicly available PYTHON package developed to solve Boltzmann equations associated with leptogenesis. For each benchmark point, we determine the regions of the parameter space, (m_{DM}, f) that are consistent with the measured relic abundance. To perform this task, we use ULYSSES in conjunction with MULTINEST [129–131] (more precisely, PYMULTINEST [132], a wrapper around MULTINEST written in PYTHON). We place flat priors on the parameters

¹The vacuum energy in the confined phase is $\sim c_0 \Lambda_W^4$, where c_0 is a constant. We require that this energy is always smaller than the contribution from relativistic degrees of freedom in the Universe, $g_* T^4$. Assuming deconfinement happens at a temperature $T_{\text{dc}} = 10^{-4} \Lambda_W$, requiring $c_0 \Lambda_W^4 < g_* T_{\text{dc}}^4$ would imply that $c_0 \lesssim 10^{-14}$.

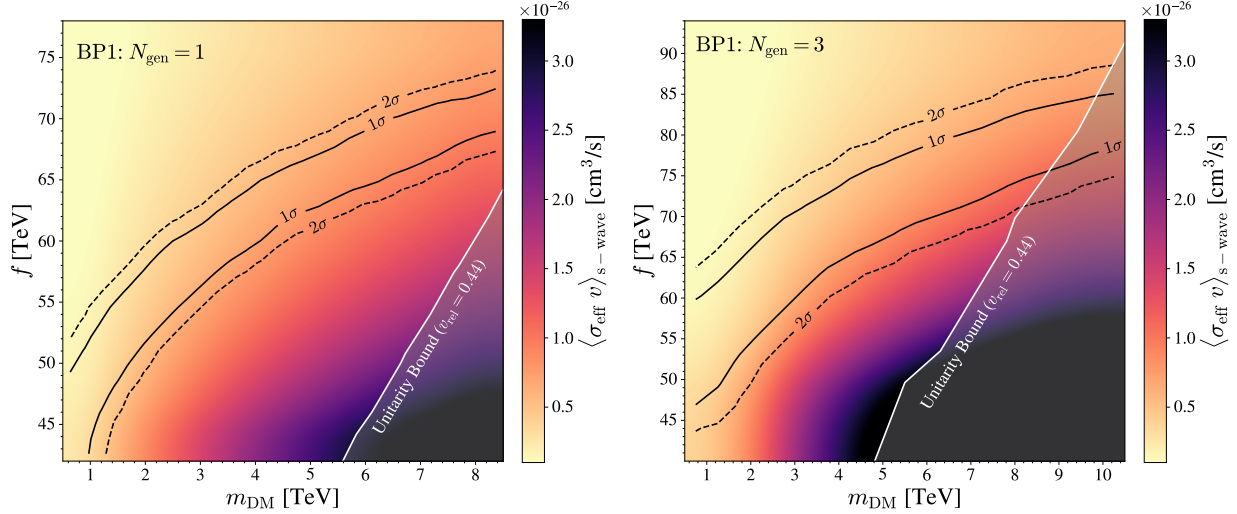


Figure 5.4: **The region of interest for the constituent dark matter mass, m_{DM} , and the weak confinement scale, f , for one generation (left) and three generation (right) cases.** The solid and dashed lines show where the DM relic density is consistent with observations at 1 and 2 σ respectively. We show the velocity-averaged effective cross section during freeze-out given in Eq. (5.31). The grey shaded area is inconsistent with unitarity constraints. Note that for both cases we start our scan at $m_{\text{DM}} = 500$ GeV and that the highest points for our scans are $m_{\text{DM}} = 8.5$ TeV and 10.5 TeV for one generation and three generation case respectively. For the benchmark shown above, BP1, $g_s = 0.8$, $e_Q = 0.5$ and $s_Q^2 = 0.12$.

(m_{DM}, f) and employ the log-likelihood as the MULTINEST objective function:

$$\log L = -\frac{1}{2} \left(\frac{\Omega h^2(m_{\text{DM}}, f) - \Omega h_{\text{PDG}}^2}{\Delta \Omega h^2} \right)^2, \quad (5.34)$$

where $\Omega h^2(m_{\text{DM}}, f)$ is the calculated relic density for a point in the model parameter space, Ωh_{PDG}^2 is the best-fit value of the relic density and $\Delta \Omega h^2$ is the 1- σ experimental uncertainty range of the relic abundance [127]. In the left panel of Fig. 5.4, we show the regions for which the predicted relic abundance of dark matter is consistent with the observed abundance at the one and two sigma. We find that multi-TeV χ masses (and $f \sim 60$ TeV) are favored and consistent with the perturbative unitarity bound [133], which, using the approximate

analytic form of $\langle\sigma_{\text{eff}}v\rangle$ Eq. (5.29), takes the form:

$$\langle\sigma_{\text{eff}}v\rangle_{\text{s-wave}} \approx \frac{0.8m_{\text{DM}}}{f^3} \lesssim \frac{4\pi}{64\pi^3 f m_{\text{DM}} v} \Rightarrow m_{\text{DM}}^2 \lesssim \frac{5f^2}{64\pi^2 v}, \quad (5.35)$$

where we substitute $m_{\text{DP}}^2 = 64\pi^3 f m_{\text{DM}}$. For a freeze-out temperature of $T_{\text{fo}} \simeq m_1/30$, the unitarity limit constrains $m_{\text{DM}} \lesssim 1.3f$, which cuts into the parameter regime favored by the relic density at around $m_{\text{DM}} \sim 10$ TeV. Fig. 5.4 shows the unitarity limit on the region of interest using the numerical results for $\langle\sigma_{\text{eff}}v\rangle_{\text{s-wave}}$. The numerical results for BP1 and BP2 are qualitatively very similar. Our code, which calculates the effective cross section and solves the Boltzmann equations, for both the one- and three-generation case, is publicly available at [🔗](#).

5.4 Three Generations of Standard Model doublets and Dark Matter

For simplicity, we have outlined the freeze-out dynamics in the case of a single generation of SM doublets together with the pair of vector-like fermionic $\text{SU}(2)_L$ doublets ($\{\ell, q^r, q^g, q^b, \chi_1, \chi_2\}$). In this Section, we generalize to three generations ($\{\ell_i, q_i^r, q_i^g, q_i^b, \chi_1, \chi_2\}$ with $i = 1, 2, 3$) where there are 90 pseudo-Goldstone bosons and an η' . The mass matrix is 91×91 and, due to the added complexity of three generations of SM doublets, the mass² matrix contains off-diagonal entries which depend non-trivially on the scan parameters (m_{DM}, f). Therefore, unlike in the one generation case, where we could perform the diagonalization of the mass squared matrix analytically, in the three-generation case, we instead rely on a numerical diagonalization of the mass-squared matrix to transform from the interaction to the mass basis for each parameter scan point. We perform the same procedure outlined in Section 5.2.0.1 to transform from the mass to the $\text{U}(1)_\chi$ basis, and compute annihilation

Pion (mass basis)	#	Mass squared value	U(1) _Q charge	SU(2) _C charge
Π_1^{mass}	1	$64\pi^2 f \left(7f\kappa + \pi m_{\text{DM}} + \sqrt{49f^2\kappa^2 - 10\pi f\kappa m_{\text{DM}} + \pi^2 m_{\text{DM}}^2} \right)$	0	1
Π_2^{mass}	24	$-\frac{1}{2}C_A e_Q^2 f^2 - \frac{3}{2}C_G f^2 g_s^2 + C_W f^2 g_s^2 - \frac{1}{2}C_Z e_Q^2 f^2 s_Q^2 + \frac{C_Z e_Q^2 f^2}{6s_Q^2} + \frac{1}{3}C_Z e_Q^2 f^2$	± 1	2
Π_3^{mass}	14	0	0	1
Π_4^{mass}	6	$-2C_A e_Q^2 f^2 - 2C_Z e_Q^2 f^2 s_Q^2 - \frac{2C_Z e_Q^2 f^2}{9s_Q^2} + \frac{4}{3}C_Z e_Q^2 f^2$	± 1	1
Π_5^{mass}	12	$-\frac{1}{2}C_A e_Q^2 f^2 - \frac{3}{2}C_G f^2 g_s^2 - C_W f^2 g_s^2 - \frac{1}{2}C_Z e_Q^2 f^2 s_Q^2 + \frac{C_Z e_Q^2 f^2}{6s_Q^2} + \frac{1}{3}C_Z e_Q^2 f^2$	± 1	2
Π_6^{mass}	6	$64\pi^3 f m_{\text{DM}}$	0	1
Π_7^{mass}	6	$-2C_A e_Q^2 f^2 - 2C_Z e_Q^2 f^2 s_Q^2 + \frac{2}{3}C_Z e_Q^2 f^2 + 64\pi^3 f m_{\text{DM}}$	± 1	1
Π_8^{mass}	9	$-4C_G f^2 g_s^2$	0	3
Π_9^{mass}	12	$-\frac{1}{2}C_A e_Q^2 f^2 - \frac{3}{2}C_G f^2 g_s^2 - \frac{1}{2}C_Z e_Q^2 f^2 s_Q^2 + \frac{C_Z e_Q^2 f^2}{18s_Q^2} + 64\pi^3 f m_{\text{DM}}$	± 1	2
Π_{10}^{mass}	1	$64\pi^2 f \left(7f\kappa + \pi m_{\text{DM}} - \sqrt{49f^2\kappa^2 - 10\pi f\kappa m_{\text{DM}} + \pi^2 m_{\text{DM}}^2} \right)$	0	1

Table 5.2: **Table of mass squared values corresponding to mass basis states along with the relevant $SU(2)_C \times U(1)_Q$ charges. Three SM generations with χ_1 and χ_2 are included.**

cross section as described in Section 5.3.1, but in the three-generation case, there are 12 charged dark pion states. We find that there are ten distinct pion masses as shown in Table. 5.2. Rather than provide the complete indexing of states, we provide the number of pions (second column) with each mass eigenvalue. Interestingly, several new states, such as the color triplet, appear in the multi-generational case.

In the right panel of Fig. 5.4, we show the regions for which the predicted relic abundance of dark matter in the three generation case is consistent with the observed abundance. The favored region that explains the DM abundance in the three-generation case is approximately the same as the simplified one generation case but favors slightly higher f values for a given m_{DM} . Another slight difference is that the unitarity constraint is more stringent due to the higher values of $\langle \sigma_{\text{eff}} v \rangle_{\text{s-wave}}$ in the three-generation case.

5.5 Methods

This section expands upon the calculation steps leading to the results in this chapter. In particular, we give an outline of the code's [134] organization and structure, as well as more statistical details for the parameter scan.

5.5.1 Outline of Code

The majority of the code is in the format of PYTHON [135] script files (.py) and JUPYTER [136] notebook files (.ipynb), with some initial information supplemented by analytic MATHEMATICA [137] calculations. The code considers two cases, a toy model with only the first generation of Standard Model (SM) particle content (denoted $N_{\text{gen}} = 1$) and the full model with three generations of SM particle content (denoted $N_{\text{gen}} = 3$).

There are 5 main stages that the code progresses through:

1. Converting the vector of Π states from the interaction basis to the definite Dark Matter (DM) charge basis.
2. Calculating the velocity averaged effective cross section of interactions which deplete the constituent DM abundance, $\langle\sigma_{\text{eff}}v\rangle$. Namely, interactions which convert pions containing the constituent DM particles (Π_{DM}) into pions containing only SM particles (Π_{SM}).
3. Solving the Boltzmann equations using this effective cross section to determine the evolution of the number density of DM during the freeze-out of the Π_{DM} 's.
4. Comparing the final constituent DM abundance (after Π_{DM} freeze-out and subsequent deconfinement) to the experimentally measured value, $\Omega h^2 = 0.1200 \pm 0.0012$ [127].

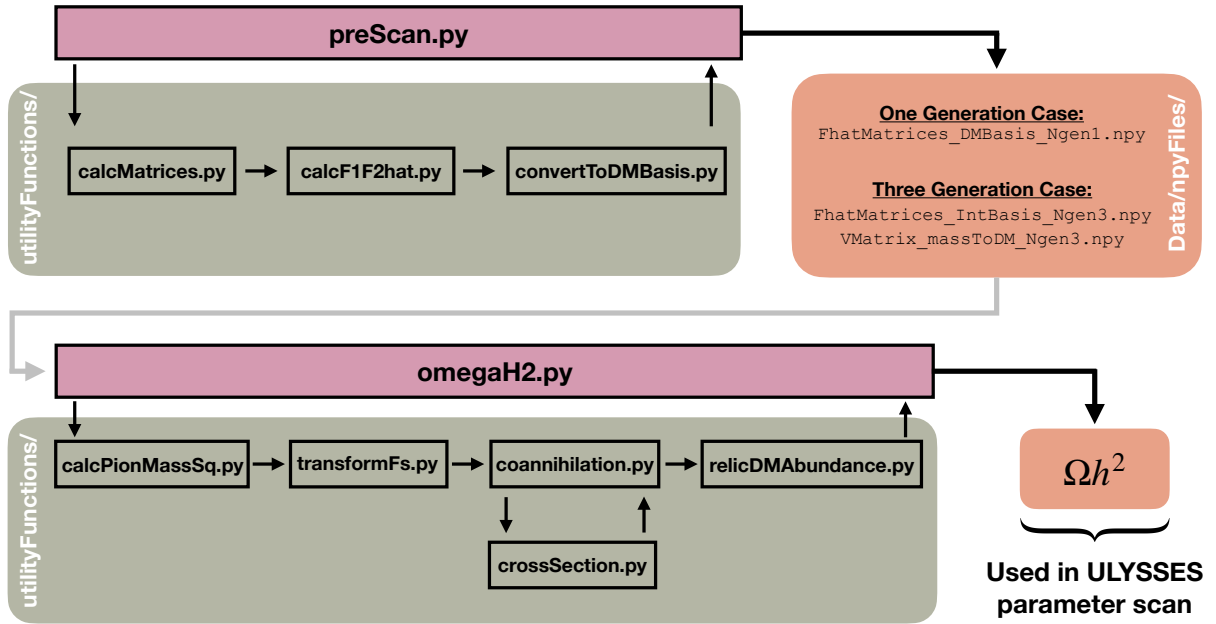



Figure 5.5: **Schematic outline of the code showing the dependencies of the various script files.** There are two main script files `preScan.py` and `omegaH2.py` which call functions in several helper script files located in the directory `utilityFunctions/`. The arrows indicate the order in which functions are called in the helper script files. Relevant file outputs of `preScan.py` and `omegaH2.py` are shown in the orange sections. The code is publicly available  [134].

5. Performing a parameter space scan to determine the region of parameter space which agrees with the experimentally measured value at the 1σ and 2σ levels. The scan utilizes the software package ULYSSES [128] which gives an efficient method for scanning the parameter space with a PYMULTINEST [132] backend.

The $N_{\text{gen}} = 1$ and $N_{\text{gen}} = 3$ cases progress through these stages similarly, differing only in the exact details of stage 1 (analytic mass-squared matrix diagonalization versus numeric, respectively) and the difference in the number of Π states to keep track of throughout (15 versus 91, respectively). Fig. 5.5 outlines the dependencies of the core calculation script files as well as the key outputs.

5.5.2 Statistical Handling of Parameter Scan Results

We do not scan directly in $\{f, m_{\text{DM}}\}$ parameter space, but rather in $\{f_{\text{pow}}, b_{\text{small,pow}}\}$ parameter space. These spaces are related in the following way

$$f_{\text{pow}} := \log_{10}\left(\frac{f}{\text{GeV}}\right) \quad (5.36)$$

$$b_{\text{small,pow}} := \log_{10}(b_{\text{small}}) \quad (5.37)$$

$$b_{\text{small}} := \frac{m_{\text{DM}}}{\Lambda_W} \quad (5.38)$$

$$\Lambda_W := 4\pi f. \quad (5.39)$$

The definition of b_{small} allows us to easily ensure that the constituent DM mass, m_{DM} , is well below the weak scale, Λ_W , which is acting as our EFT cutoff (i.e. $b_{\text{small}} \ll 1$). Additionally, scanning in \log_{10} -parameter space, $\{f_{\text{pow}}, b_{\text{small,pow}}\}$, rather than directly in the parameter space, $\{f, b_{\text{small}}\}$, is computationally favorable.

The ULYSSES parameter scan results in a text file (`ULSNEST.txt`). Each row of this file corresponds to a different parameter space point, $\theta := \{f_{\text{pow}}, b_{\text{small,pow}}\}$. The first and second columns correspond to the posterior and negative 2 log-likelihood ($-2 \ln L$) values at each parameter space point, respectively. The final columns correspond to the scan parameter values at which the posterior and $-2 \ln L$ were evaluated. In this case there are two columns corresponding to f_{pow} and $b_{\text{small,pow}}$.

The goal is to use the sampled posterior and $-2 \ln L$ information from the scan to discover the viable regions of parameter space (so-called *interval* or, in the case of multiple parameters, *region estimation*). Whether we use the posterior or $-2 \ln L$ information depends on whether we would like to adopt a Bayesian or frequentist approach, respectively. In the former we would have *credible* regions and in the latter we would have *confidence* regions. In what follows we adopt the latter, which is more typical [138]. For more details on this topic and

how these intervals are drawn in general see Appendix E which draws heavily from Chapter 9 of [138].

We define our likelihood function (in general form) as

$$-2 \ln L_g = \left(\frac{g(\theta) - \mu}{\sigma} \right)^2 \quad (5.40)$$

where $\mu := \Omega h_{\text{PDG}}^2 = 0.1200$ and $\sigma := \Delta \Omega h^2 = 0.0012$ are the experimentally measured Ωh^2 value and its 1σ error. In what follows, statistically, these are treated as fixed, known parameters. The function $g(\theta) := \Omega h^2(\theta)$ converts the scan parameters into a prediction for the Ωh^2 value. The parameters are $\theta = \{b_{\text{small,pow}}, f_{\text{pow}}\}$ (or, equivalently, $\theta = \{m_{\text{DM}}, f\}$ since g is unaffected by this reparameterization). A scan is performed to find the θ such that $-2 \ln L_g$ is minimized. By the definition above and since μ and σ are considered known, the theoretical minimum of $-2 \ln L_g$ will occur at $g(\theta) = \mu$.

By the likelihood's property of invariance, the regions where $-2 \ln L_g$ is minimized will correspond directly to the regions where $-2 \ln L_\theta$ is minimized. Because of this we need not know the analytic form of L_θ , which is likely quite complicated. This invariance property is why using likelihood information to draw confidence regions is so common.² Because of this, and to simplify the notation, we now simply denote the likelihood as L .

The parameter space has been sampled irregularly, with samples preferentially falling in regions of minimal $-2 \ln L$. But we need to use this information to obtain a general function of L over the parameter space scan region. We therefore estimate the two-dimensional profile likelihood (following the methodology here [139]). We begin by binning the desired region of parameter space to create a grid. Next we use the (approximate) definition that $\chi^2 = -2 \ln(L/L_{\text{max}}) = -2 \ln L$, where we have taken a frequentist perspective in assuming $L_{\text{max}} = 1$ (the theoretically maximum value). For each bin, we find the minimum χ^2 value

²Note that this invariance is practically very useful but is only approximately true.

that falls in that bin. If a bin is empty we set its value to infinity (`inf`). This defines the grid of the profile- χ^2 . From this we can invert the relationship to get the gridded profile-likelihood (i.e. $\text{profile-likelihood} = \exp(-0.5 \text{ profile-}\chi^2)$). The final step is to smoothly interpolate between these grid points in order to draw confidence intervals over parameter space. Given the likelihood's (approximate) relationship to χ^2 , a 1σ (2σ) confidence interval is drawn by finding the contour where the likelihood is equal to 0.32 (0.05). This corresponds to requiring that $\chi^2 \leq F_{\chi^2}^{-1}(1 - \alpha, \text{dof} = 2) \Rightarrow L/L_{\text{max}} = L \leq \alpha$ for $\alpha = 0.32$ ($\alpha = 0.05$), where $F_{\chi^2}^{-1}(\cdot)$ is the Cumulative Distribution Function (CDF) of the χ^2 distribution. These regions are shown in Fig. 5.4.

5.6 Discussion

Our results indicate that a modification to the strength of the $SU(2)_L$ weak coupling dramatically transforms the nature of the freeze-out process for an $SU(2)_L$ -charged WIMP. For a vector-like pair of doublets, we find that the weak confinement scenario favors a range of masses (depending on the early $SU(2)_L$ confinement scale) around $O(1 - 10)$ TeV and can be much larger than the $\simeq 1.1$ TeV favored by a standard cosmological history [102]. This highlights the possibility that the physics of the dark matter itself could be drastically different at the time of freeze-out from today. In particular, the constraints on a several TeV WIMP are quite different from those restricting a ~ 1 TeV mass particle.

Direct searches for WIMPs scattering with heavy nuclei remain an important challenge. At ~ 10 TeV, XENON1T data restricts the cross section to scatter with a nucleon to be smaller than about $\sim 10^{-44}$ cm² [105], which is still incompatible with the cross section mediated by full strength Z boson exchange ($\sim 10^{-38}$ cm²). However, this bound can be avoided by

introducing Majorana masses via a dimension-5 operator of the form,

$$\mathcal{L}_{\Delta M} = \frac{1}{M_1}(H^\dagger\chi_1)(H^\dagger\chi_1) + \frac{1}{M_2}(H\chi_2)(H\chi_2) + \text{h.c.} \quad (5.41)$$

where $M_{1,2}$ parameterize the interaction strength and parentheses indicate how $SU(2)_L$ indices are contracted. After electroweak breaking, these operators result in Majorana masses of order $v^2/M_{1,2}$, which split the Dirac χ into two Majorana fermions, in close analogy with the see-saw mechanism for generating neutrino masses. The Majorana particles have vanishing vector currents, and thus Z boson exchange mediates inelastic scattering, which is kinematically suppressed once the mass splitting is larger than the typical kinetic energy of the WIMPs in the Galactic halo [140, 141]. Provided the scales M_1 and M_2 are sufficiently large, these operators play essentially no role in freeze-out, and do not themselves mediate an observable scattering with nuclei via Higgs boson exchange.

Despite its full-strength electroweak interactions, a multi-TeV dark matter particle is too heavy to be accessible at the LHC. Even when kinematically accessible, unless there is mixing with another nearby state via electroweak symmetry-breaking, the signatures at colliders are challenging because the charged state is expected to be degenerate with its neutral counterpart to within a few hundred MeV [142], and thus requires mono-jet or disappearing track analyses. As a result, even a future 100 TeV hadron collider is expected to struggle to reach sensitivity to TeV mass electroweak doublets [143].

Indirect searches for the annihilation products of WIMP annihilation, for example, from observation of high energy γ -rays, can reach sensitivity to around 10 TeV for electroweak-sized annihilation cross sections [106], particularly for masses for which the annihilation experiences a Sommerfeld-like enhancement due to the exchange of weak bosons. These bounds exhibit a considerable sensitivity to the distribution profile of the dark matter around the Galactic center, which is not well constrained by observation (see, e.g. Ref. [13] for

discussion). Despite these challenges, a future gamma-ray observatory such as the Cherenkov Telescope Array [144] could offer the best chance of a direct observation of dark matter in such a scenario.

Looking forward, it would be interesting to explore further the consequences of a period of early $SU(2)_L$ confinement. It may be that such an epoch could enable new possibilities to understand other mysteries of the early Universe, such as the primordial asymmetry between baryons and anti-baryons. And more widely, our results illustrate the general truth that the early Universe may well turn out to have been more weird and wonderful than simply extrapolating the SM to high temperatures would lead us to expect. Exploring the space of possibilities and how to constrain them with experimental measurements will remain an essential task for particle physics.

Chapter 6

Conclusion and Outlook

In this thesis we have demonstrated two ways in which mathematical and computational methods can work together with physics insight to make valuable contributions to the field. Moreover, these works demonstrate how questioning prior assumptions and incorporating tools from other fields can lead to novel results.

In Chapter 4 we saw how mathematical techniques from optimal transport (OT) theory have augmented powerful unsupervised machine learning methods and enabled the design of a physics-informed, data-driven particle simulator: OTUS. OTUS has been shown to work on proof-of-principle cases, but should be developed further before being applied generally. In particular, it is worth taking a step back to examine how to formulate a general description of LHC events which also incorporates physically-motivated constraints. It may be that OT is the answer for this as well. Interesting recent work has been able to naturally express many jet observables in terms of OT [32, 33], and is even being proposed as a means to formulate a general metric on the space of particle collisions [34]. Having a robust description of the data will undoubtedly improve the network’s ability to learn simulation and unfolding mappings by allowing it to focus in on physically meaningful features. Other

immediate future directions include investigating semi-supervised applications of OTUS, in which some simulated samples are used to ground the learned mappings, and investigating other Wasserstein distance estimation techniques for the latent loss (such as Sinkhorn distance [56]). Finally, future work should investigate the ability for OTUS to interpolate over \mathcal{Z} space (see Section 4.7 for more details). This last step is crucial to achieve the goal of applying OTUS to LHC searches for BSM physics.

In Chapter 5 we saw how questioning assumptions about the early universe can widen the range of possible WIMP DM masses which could produce the observed DM relic abundance. We also saw how investigating the effects of changing these assumptions could be aided by computational tools. Considering the effects of Electroweak force ($SU(2)_L$) confinement presented a computational challenge even in the simple toy model. Tackling this problem analytically with software like MATHEMATICA [137] was computationally intractable, and thus would have required introducing simplifying assumptions to move forward. Therefore, efficient numerical PYTHON [135] computational tools made investigating this scenario with relatively few assumptions possible. Future work should investigate what other effects a phase of $SU(2)_L$ confinement in the early universe could have had. For example, it would be useful to investigate the extent to which this model (with possible minor modifications) may serve as a mechanism for baryogenesis — the process by which the matter/anti-matter asymmetry in the Universe arose. Similar work considering a phase of early QCD force confinement showed that such a change could lead to a novel baryogenesis mechanism [12]. Additionally, this work remained agnostic to the dynamics of the scalar field ϕ whose vev alters the strength of the Electroweak force; investigating these dynamics in future work may also lead to interesting implications.

6.1 Thoughts on Applying Machine Learning to Problems in Particle Theory

In recent years, it seems like every scientific field has seen some kind of machine learning (ML) revolution. In experimental and phenomenological particle physics, such a revolution has been especially fruitful. State-of-the-art ML methods are being applied to a wide variety of problems [145]. For example, on the experimental side ML has been applied to problems in data analysis, trigger algorithms, pileup subtraction, feature identification, unfolding, simulations [146–148]; and on the phenomenological side, ML has been applied to problems like solving differential equations [149, 150] and parameter searches [151, 152]. Many of these particle physics problems have natural analogs to common ML tasks, making matching problems with the correct ML tool straightforward. Additionally, the traditional methods used to solve these problems are already numerical in nature, so the ideological jump to using ML (a very powerful numerical method) is not so large.

Applications of ML to problems in theoretical physics more broadly (and also mathematics) have seen a comparative lag, but this lag is already disappearing rapidly [145]. ML tools which can handle more abstract data-types such as natural language processing and, more generally, graph networks have already been applied to finding relations between abstract objects [50, 153, 154], in problems in string theory [155], and even helped to discover new mathematical relations [154]. Another interesting area of recent work is pushing to connect ML and particle physics, more specifically QFT, on a mathematical level [156–161]. It has been shown that many network structures are Gaussian processes (freely interacting fields) in the limit of infinite nodes, indicating that a finite network can be modeled as a field with interactions. Moreover, the evolution of parameters in a common gradient descent training algorithm has been connected with the evolution of a scalar field in early-universe cosmology. Such connections immediately provide more insight into how ML algorithms function, which

is often considered a black-box, and may allow particle theorists to investigate these systems in new ways. Another fascinating, and promising, theoretical insight came indirectly from applying ML particle physics problems. In particular, applications of optimal-transport based ML methods to problems in LHC phenomenology have led people to consider whether the mathematics of OT are a natural description for particle interactions. The history of OT also foreshadows its potential utility, with historical studies applying OT to find solutions to Boltzmann equations and particle system dynamics [14]. This potentially powerful method has only resurfaced as a result of applications of OT-based ML methods. These blossoming directions of ways ML can help particle physics (and vice versa) are sure to only become more fruitful in the coming years. But just as people twenty years ago could not have foreseen the extent to which ML is being used in LHC experiments, the variety of possible applications of ML to problems in particle theory will no doubt expand from what we see today.

When attempting to push the boundaries of how ML might be applied to problems in theoretical particle physics, it is easy to feel like ML is an over-powered tool. Often the answer to whether you could use ML to solve a problem is, "Sure, I guess you could. But couldn't you just do... ⟨insert analytical method, approximation, or simpler numerical technique⟩?" So while you *can* open a peanut with a sledgehammer, it is fair to ask whether you *should*.¹ At this point, it seems that you have hit a wall and there are two options: stick to using ML to improve upon numerical methods or change your perspective. It is not surprising that many of the problems in theoretical particle physics seem better suited to analytical approximations or simple numerical methods as this is all that has historically been available. ML is great at solving difficult problems, but often fails miserably at simple ones. So when finding a home for ML in theoretical particle physics, it is best to start from the beginning, consider the problem in its most general, difficult form, with as few prior assumptions as possible. Then ask yourself, 1) What mapping am I trying to learn? and 2) How will I know

¹Credit for this analogy goes to Dr. Gopolang "Gopi" Mohlabeng in a very fun conversation about this topic.

when I have succeeded i.e. what is the objective? And keep in mind that ML need not solve the problem in its entirety to be useful. If it can point us in the right direction, or help us in a specific case, this already brings us closer to our goals. And perhaps a bit of AI assistance is all we need to bring about the next paradigm shift in theoretical physics.

Bibliography

- [1] “Learning to simulate high energy particle collisions from unlabeled data”. In: *Scientific Reports* 12.1 (2022), p. 7567. DOI: 10.1038/s41598-022-10966-7. URL: <https://doi.org/10.1038/s41598-022-10966-7>.
- [2] Jessica N. Howard et al. “Dark matter freeze-out during SU(2)L confinement”. In: *Journal of High Energy Physics* 2022.2 (Feb. 2022). DOI: 10.1007/jhep02(2022)047. URL: <https://doi.org/10.1007%5C%2Fjhep02%5C%282022%5C%29047>.
- [3] Edward W. Kolb and Michael S. Turner. *The Early Universe*. Vol. 69. 1990. ISBN: 978-0-201-62674-2. DOI: 10.1201/9780429492860.
- [4] Planck Collaboration. “iPlanck/i2018 results”. In: *Astronomy & Astrophysics* 641 (Sept. 2020), A1. DOI: 10.1051/0004-6361/201833880. URL: <https://doi.org/10.1051%5C%2F0004-6361%5C%2F201833880>.
- [5] Particle Data Group. “Review of Particle Physics”. In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020). 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptaa104. eprint: <https://academic.oup.com/ptep/article-pdf/2020/8/083C01/34673722/ptaa104.pdf>. URL: <https://doi.org/10.1093/ptep/ptaa104>.
- [6] Seyda Ipek and Tim M. P. Tait. “Early Cosmological Period of QCD Confinement”. In: *Physical Review Letters* 122.11 (Mar. 2019). DOI: 10.1103/physrevlett.122.112001. URL: <https://doi.org/10.1103%5C%2Fphysrevlett.122.112001>.

- [7] Saleh Hamdan and James Unwin. “Dark Matter Freeze-out During Matter Domination”. In: *Modern Physics Letters A* 33.29 (2018), p. 1850181. DOI: 10.1142/S021773231850181X. arXiv: 1710.03758 [hep-ph].
- [8] Graciela B. Gelmini and Paolo Gondolo. “Neutralino with the right cold dark matter abundance in (almost) any supersymmetric model”. In: *Phys. Rev. D* 74 (2006), p. 023510. DOI: 10.1103/PhysRevD.74.023510. arXiv: hep-ph/0602230.
- [9] Dillon Berger et al. “Dark Matter Freeze Out during an Early Cosmological Period of QCD Confinement”. In: *JHEP* 07 (2020), p. 192. DOI: 10.1007/JHEP07(2020)192. arXiv: 2004.06727 [hep-ph].
- [10] Lucien Heurtier, Fei Huang, and Tim M. P. Tait. “Resurrecting Low-Mass Axion Dark Matter Via a Dynamical QCD Scale”. In: (Apr. 2021). arXiv: 2104.13390 [hep-ph].
- [11] Joshua Berger, Andrew J. Long, and Jessica Turner. “Phase of confined electroweak force in the early Universe”. In: *Phys. Rev. D* 100.5 (2019), p. 055005. DOI: 10.1103/PhysRevD.100.055005. arXiv: 1906.05157 [hep-ph].
- [12] Djuna Croon et al. “QCD Baryogenesis”. In: (2019). arXiv: 1911.01432 [hep-ph].
- [13] Jason Arakawa and Tim M. P. Tait. “Is a Miracle-less WIMP Ruled Out?” In: (Jan. 2021). arXiv: 2101.11031 [hep-ph].
- [14] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN: 9783540710509. URL: https://books.google.com/books?id=hV8o5R7%5C_5tkC.
- [15] Geraldine Servant and Timothy M. P. Tait. “Is the lightest Kaluza-Klein particle a viable dark matter candidate?” In: *Nucl. Phys. B* 650 (2003), pp. 391–419. DOI: 10.1016/S0550-3213(02)01012-X. arXiv: hep-ph/0206071.
- [16] Kim Griest and David Seckel. “Three exceptions in the calculation of relic abundances”. In: *Phys. Rev. D* 43 (1991), pp. 3191–3203. DOI: 10.1103/PhysRevD.43.3191.

- [17] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. *The frontier of simulation-based inference*. 2020. arXiv: 1911.01429 [stat.ML].
- [18] Igor Volobouev. *Matrix Element Method in HEP: Transfer Functions, Efficiencies, and Likelihood Normalization*. 2011. DOI: 10.48550/ARXIV.1101.2259. URL: <https://arxiv.org/abs/1101.2259>.
- [19] Sébastien Wertz. “The Matrix Element Method in the LHC era”. In: *EPJ Web Conf.* 137 (2017). Ed. by Y. Foka, N. Brambilla, and V. Kovalenko, p. 11010. DOI: 10.1051/epjconf/201713711010.
- [20] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. “PYTHIA 6.4 Physics and Manual”. In: *JHEP* 0605 (2006), p. 026. DOI: 10.1088/1126-6708/2006/05/026. arXiv: hep-ph/0603175 [hep-ph].
- [21] et al. S. Agostinelli. “Geant4 - a simulation toolkit”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: [https://doi.org/10.1016/S0168-9002\(03\)01368-8](https://doi.org/10.1016/S0168-9002(03)01368-8). URL: <http://www.sciencedirect.com/science/article/pii/S0168900203013688>.
- [22] Johan Alwall et al. *MadGraph 5 : Going Beyond*. arxiv:1106.0522. 2011. URL: <http://arxiv.org/abs/1106.0522>.
- [23] J. de Favereau et al. “DELPHES 3, A modular framework for fast simulation of a generic collider experiment”. In: *JHEP* 02 (2014), p. 057. DOI: 10.1007/JHEP02(2014)057. arXiv: 1307.6346 [hep-ex].
- [24] G. Aad et al. “The ATLAS Simulation Infrastructure”. In: *The European Physical Journal C* 70.3 (Sept. 2010), pp. 823–874. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-010-1429-9. URL: <http://dx.doi.org/10.1140/epjc/s10052-010-1429-9>.

- [25] CMS Collaboration. *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical Design Report CMS. There is an error on cover due to a technical problem for some items. Geneva: CERN, 2006. URL: <https://cds.cern.ch/record/922757>.
- [26] Anja Butter, Tilman Plehn, and Ramon Winterhalder. “How to GAN LHC Events”. In: *SciPost Phys.* 7 (6 2019), p. 75. DOI: 10.21468/SciPostPhys.7.6.075. arXiv: 1907.03764 [hep-ph]. URL: <https://scipost.org/10.21468/SciPostPhys.7.6.075>.
- [27] Bobak Hashemi et al. *LHC analysis-specific datasets with Generative Adversarial Networks*. 2019. arXiv: 1901.05282 [hep-ex].
- [28] H. Wussing et al. *The Genesis of the Abstract Group Concept: A Contribution to the History of the Origin of Abstract Group Theory*. MIT Press, 1984. ISBN: 9780262231091. URL: https://books.google.com/books?id=9%5C_%5C_uAAAAMAAJ.
- [29] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. 1781.
- [30] Martin Arjovsky, Soumith Chintala, and Léon Bottou. *Wasserstein GAN*. 2017. arXiv: 1701.07875 [stat.ML].
- [31] Ilya Tolstikhin et al. *Wasserstein Auto-Encoders*. 2017. arXiv: 1711.01558 [stat.ML].
- [32] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. “The hidden geometry of particle collisions”. In: *Journal of High Energy Physics* 2020.7 (July 2020). ISSN: 1029-8479. DOI: 10.1007/jhep07(2020)006. URL: [http://dx.doi.org/10.1007/JHEP07\(2020\)006](http://dx.doi.org/10.1007/JHEP07(2020)006).
- [33] Tianji Cai et al. “Linearized optimal transport for collider events”. In: *Physical Review D* 102.11 (Dec. 2020). ISSN: 2470-0029. DOI: 10.1103/physrevd.102.116019. URL: <http://dx.doi.org/10.1103/PhysRevD.102.116019>.

- [34] Tianji Cai et al. “Which metric on the space of collider events?” In: *Physical Review D* 105.7 (Apr. 2022). DOI: 10.1103/physrevd.105.076003. URL: <https://doi.org/10.1103%5C%2Fphysrevd.105.076003>.
- [35] Leonid N. Vaserstein. *Markov processes over denumerable products of spaces describing large systems of automata*. 1969.
- [36] Soheil Kolouri et al. *Sliced-Wasserstein Autoencoder: An Embarrassingly Simple Generative Model*. 2018. arXiv: 1804.01947 [cs.LG].
- [37] Soheil Kolouri et al. *Generalized Sliced Wasserstein Distances*. 2019. DOI: 10.48550/ARXIV.1902.00434. URL: <https://arxiv.org/abs/1902.00434>.
- [38] M. Loeve. *Probability Theory: Third Edition*. Dover Books on Mathematics. Dover Publications, 2017. ISBN: 9780486814889. URL: <https://books.google.com/books?id=sKvPDgAAQBAJ>.
- [39] Tatsuya Morita, Toshihiko Sasaki, and Izumi Tsutsui. *Complex Probability Measure and Aharonov’s Weak Value*. 2012. DOI: 10.48550/ARXIV.1210.1298. URL: <https://arxiv.org/abs/1210.1298>.
- [40] Cebraïl Hasimi. Oktar. *Complex Probability Theory*. 2011.
- [41] G.G. Roussas. *An Introduction to Measure-theoretic Probability*. EBSCO ebook academic collection. Elsevier Science, 2005. ISBN: 9780125990226. URL: https://books.google.com/books?id=smP02%5C_aPVPMC.
- [42] D.H. Fremlin. *Measure Theory*. v. 5, pt. 2. Torres Fremlin, 2000. ISBN: 9780953812967. URL: https://books.google.com/books?id=1nI%5C_n18CZVgC.
- [43] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015. ISBN: 9783319208282. URL: <https://books.google.com/books?id=UOHHCgAAQBAJ>.

- [44] Gabriel Peyré and Marco Cuturi. “Computational Optimal Transport”. In: (2018). DOI: 10.48550/ARXIV.1803.00567. URL: <https://arxiv.org/abs/1803.00567>.
- [45] Nicolas Bonnotte. *PhD thesis: Unidimensional and Evolution Methods for Optimal Transportation*. 2013. URL: <https://www.normalesup.org/~bonnotte/doc/phd-bonnotte.pdf>.
- [46] Kimia Nadjahi et al. *Fast Approximation of the Sliced-Wasserstein Distance Using Concentration of Random Projections*. 2021. DOI: 10.48550/ARXIV.2106.15427. URL: <https://arxiv.org/abs/2106.15427>.
- [47] David Kirkby. *Machine Learning and Statistics for Physicists, Lectures*. 2019. URL: <https://github.com/dkirkby/MachineLearningStatistics>.
- [48] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.
- [49] Stephen Welch. *Neural Networks Demystified, Lectures*. 2013. URL: <https://github.com/stephencwelch/Neural-Networks-Demystified>.
- [50] Peter W. Battaglia et al. *Relational inductive biases, deep learning, and graph networks*. 2018. arXiv: 1806.01261 [cs.LG].
- [51] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. DOI: 10.48550/ARXIV.1406.2661. URL: <https://arxiv.org/abs/1406.2661>.
- [52] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2013. arXiv: 1312.6114 [stat.ML].
- [53] Mark A. Kramer. “Nonlinear principal component analysis using autoassociative neural networks”. In: *AIChE Journal* 37.2 (1991), pp. 233–243. DOI: <https://doi.org/10.1002/aic.690370209>. eprint: <https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1002/aic.690370209>. URL: <https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690370209>.

- [54] Mehdi Mirza and Simon Osindero. *Conditional Generative Adversarial Nets*. 2014. DOI: 10.48550/ARXIV.1411.1784. URL: <https://arxiv.org/abs/1411.1784>.
- [55] Tim Salimans et al. *Improved Techniques for Training GANs*. 2016. arXiv: 1606.03498 [cs.LG].
- [56] Giorgio Patrini et al. *Sinkhorn AutoEncoders*. 2019. arXiv: 1810.01118 [cs.LG].
- [57] Mark Srednicki. *Quantum Field Theory*. Cambridge University Press, 2007. ISBN: 9781139462761. URL: <https://books.google.com/books?id=50epxIG42B4C>.
- [58] A. Zee. *Quantum Field Theory in a Nutshell: Second Edition*. In a Nutshell. Princeton University Press, 2010. ISBN: 9781400835324. URL: <https://books.google.com/books?id=n8Mmbjtco78C>.
- [59] Nathaniel Craig, Isabel Garcia Garcia, and Seth Koren. “The weak scale from weak gravity”. In: *Journal of High Energy Physics* 2019.9 (Sept. 2019). DOI: 10.1007/jhep09(2019)081. URL: <https://doi.org/10.1007%5C%2Fjhep09%5C%282019%5C%29081>.
- [60] Steven Abel and Keith R. Dienes. “Calculating the Higgs mass in string theory”. In: *Physical Review D* 104.12 (Dec. 2021). DOI: 10.1103/physrevd.104.126032. URL: <https://doi.org/10.1103%5C%2Fphysrevd.104.126032>.
- [61] David B. Kaplan. *Lectures on effective field theory, delivered at the ICTP-SAFIR, Sao Paulo*. 2016. URL: https://archive.int.washington.edu/users/dbkaplan/572_16/EFT.pdf.
- [62] Tim M.P. Tait. *Lectures on Chiral Perturbation Theory*.
- [63] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to quantum field theory*. Reading, USA: Addison-Wesley, 1995. ISBN: 978-0-201-50397-5.

- [64] Frank Noé et al. “Machine Learning for Molecular Simulation”. In: *Annual Review of Physical Chemistry* 71.1 (2020). PMID: 32092281, pp. 361–390. DOI: 10.1146/annurev-physchem-042018-052331. eprint: <https://doi.org/10.1146/annurev-physchem-042018-052331>. URL: <https://doi.org/10.1146/annurev-physchem-042018-052331>.
- [65] David Rolnick et al. *Tackling Climate Change with Machine Learning*. 2019. arXiv: 1906.05433 [cs.CY].
- [66] Arnaud Delaunoy et al. *Lightning-Fast Gravitational Wave Parameter Inference through Neural Amortization*. 2020. arXiv: 2010.12931 [astro-ph.IM].
- [67] Jiajia Zhou et al. “Emerging role of machine learning in light-matter interaction”. In: *Light: Science & Applications* 8.1 (2019), p. 84. DOI: 10.1038/s41377-019-0192-4. URL: <https://doi.org/10.1038/s41377-019-0192-4>.
- [68] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. “CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks”. In: *Phys. Rev. D* 97 (1 Jan. 2018), p. 014021. DOI: 10.1103/PhysRevD.97.014021. arXiv: 1712.10321 [hep-ex]. URL: <https://link.aps.org/doi/10.1103/PhysRevD.97.014021>.
- [69] Sydney Otten et al. *Event Generation and Statistical Sampling for Physics with Deep Generative Models and a Density Information Buffer*. 2019. arXiv: 1901.00875 [hep-ph].
- [70] Luke de Oliveira, Michela Paganini, and Benjamin Nachman. “Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis”. In: *Computing and Software for Big Science* 1.1 (2017), p. 4. DOI: 10.1007/s41781-017-0004-6. arXiv: 1701.05927 [stat.ML]. URL: <https://doi.org/10.1007/s41781-017-0004-6>.

- [71] Erik Buhmann et al. *Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed*. 2020. arXiv: 2005.05334 [physics.ins-det].
- [72] Kamil Deja et al. *End-to-end Sinkhorn Autoencoder with Noise Generator*. 2020. arXiv: 2006.06704 [cs.LG].
- [73] Yadong Lu et al. “Sparse autoregressive models for scalable generation of sparse images in particle physics”. In: *Physical Review D* 103.3 (Feb. 2021). ISSN: 2470-0029. DOI: 10.1103/physrevd.103.036012. URL: <http://dx.doi.org/10.1103/PhysRevD.103.036012>.
- [74] Anders Andreassen et al. “JUNIPR: a framework for unsupervised machine learning in particle physics”. In: *The European Physical Journal C* 79.2 (Feb. 2019). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-019-6607-9. URL: <http://dx.doi.org/10.1140/epjc/s10052-019-6607-9>.
- [75] Artidoro Pagnoni, Kevin Liu, and Shangyan Li. *Conditional Variational Autoencoder for Neural Machine Translation*. 2018. arXiv: 1812.04405 [cs.CL].
- [76] DØCollaboration. “A precision measurement of the mass of the top quark”. In: *Nature* 429.6992 (2004), pp. 638–642. DOI: 10.1038/nature02589. URL: <https://doi.org/10.1038/nature02589>.
- [77] Marco Bellagente et al. *Invertible Networks or Partons to Detector and Back Again*. 2020. arXiv: 2006.06685 [hep-ph].
- [78] Anders Andreassen et al. “OmniFold: A Method to Simultaneously Unfold All Observables”. In: *Physical Review Letters* 124.18 (May 2020). ISSN: 1079-7114. DOI: 10.1103/physrevlett.124.182001. URL: <http://dx.doi.org/10.1103/PhysRevLett.124.182001>.
- [79] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (Sept.

- 2012), pp. 1–29. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.020. URL: <http://dx.doi.org/10.1016/j.physletb.2012.08.020>.
- [80] S. Chatrchyan et al. “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”. In: *Physics Letters B* 716.1 (Sept. 2012), pp. 30–61. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.021. URL: <http://dx.doi.org/10.1016/j.physletb.2012.08.021>.
- [81] Andrea Castro. *Top Quark Mass Measurements in ATLAS and CMS*. 2019. arXiv: 1911.09437 [hep-ex].
- [82] Cheng Zhang et al. “Advances in variational inference”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 2008–2026.
- [83] Christopher P. Burgess et al. *Understanding disentangling in β -VAE*. 2018. arXiv: 1804.03599 [stat.ML].
- [84] Pierre Baldi et al. “Parameterized neural networks for high-energy physics”. In: *The European Physical Journal C* 76.5 (Apr. 2016). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-016-4099-4. URL: <http://dx.doi.org/10.1140/epjc/s10052-016-4099-4>.
- [85] Joshua Batson et al. “Topological obstructions to autoencoding”. In: *Journal of High Energy Physics* 2021.4 (Apr. 2021). ISSN: 1029-8479. DOI: 10.1007/jhep04(2021)280. URL: [http://dx.doi.org/10.1007/JHEP04\(2021\)280](http://dx.doi.org/10.1007/JHEP04(2021)280).
- [86] R. Aaij et al. “Searches for low-mass dimuon resonances”. In: *Journal of High Energy Physics* 2020.10 (Oct. 2020). ISSN: 1029-8479. DOI: 10.1007/jhep10(2020)156. URL: [http://dx.doi.org/10.1007/JHEP10\(2020\)156](http://dx.doi.org/10.1007/JHEP10(2020)156).
- [87] Michael Fenton et al. “Permutationless Many-Jet Event Reconstruction with Symmetry Preserving Attention Networks”. In: (Oct. 2020). arXiv: 2010.09206 [hep-ex].

- [88] J. Erdmann et al. “From the bottom to the top—reconstruction of $t\bar{t}$ events with deep learning”. In: *Journal of Instrumentation* 14.11 (Nov. 2019), P11015–P11015. ISSN: 1748-0221. DOI: 10.1088/1748-0221/14/11/p11015. URL: <http://dx.doi.org/10.1088/1748-0221/14/11/P11015>.
- [89] Johannes Erdmann et al. “A likelihood-based reconstruction algorithm for top-quark pairs and the KLFitter framework”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 748 (June 2014), pp. 18–25. ISSN: 0168-9002. DOI: 10.1016/j.nima.2014.02.029. URL: <http://dx.doi.org/10.1016/j.nima.2014.02.029>.
- [90] R. Brun and F. Rademakers. “ROOT: An object oriented data analysis framework”. In: *Nucl. Instrum. Meth. A* 389 (1997). Ed. by M. Werlen and D. Perret-Gallix, pp. 81–86. DOI: 10.1016/S0168-9002(97)00048-X.
- [91] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. “The anti-ktjet clustering algorithm”. In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 063–063. ISSN: 1029-8479. DOI: 10.1088/1126-6708/2008/04/063. URL: <http://dx.doi.org/10.1088/1126-6708/2008/04/063>.
- [92] Olivier Bousquet et al. *From optimal transport to generative modeling: the VEGAN cookbook*. 2017. arXiv: 1705.07642 [stat.ML].
- [93] Szymon Knop et al. *Cramer-Wold AutoEncoder*. 2019. arXiv: 1805.09235 [cs.LG].
- [94] Titouan Vayer et al. *Sliced Gromov-Wasserstein*. 2020. arXiv: 1905.10124 [stat.ML].
- [95] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].
- [96] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *arXiv preprint arXiv:1502.03167* (2015).

- [97] Diederik P Kingma and Jimmy Lei Ba. “Adam: A method for stochastic gradient descent”. In: *International Conference on Learning Representations*. 2015.
- [98] Philip R Bevington and D Keith Robinson. *Data reduction and error analysis for the physical sciences; 3rd ed.* New York, NY: McGraw-Hill, 2003. URL: <https://cds.cern.ch/record/1305448>.
- [99] Uroš Seljak et al. “Towards optimal extraction of cosmological information from non-linear data”. In: *Journal of Cosmology and Astroparticle Physics* 2017.12 (Dec. 2017), pp. 009–009. DOI: 10.1088/1475-7516/2017/12/009. URL: <https://doi.org/10.1088/1475-7516/2017/12/009>.
- [100] Alexander Bogatskiy et al. *Lorentz Group Equivariant Neural Network for Particle Physics*. 2020. arXiv: 2006.04780 [hep-ph].
- [101] Gianfranco Bertone and Tim Tait M. P. “A new era in the search for dark matter”. In: *Nature* 562.7725 (2018), pp. 51–56. DOI: 10.1038/s41586-018-0542-z. arXiv: 1810.01668 [astro-ph.CO].
- [102] Marco Cirelli, Nicolao Fornengo, and Alessandro Strumia. “Minimal dark matter”. In: *Nucl. Phys. B* 753 (2006), pp. 178–194. DOI: 10.1016/j.nuclphysb.2006.07.012. arXiv: hep-ph/0512090.
- [103] Georges Aad et al. “Search for new phenomena in events with an energetic jet and missing transverse momentum in pp collisions at $\sqrt{s}=13$ TeV with the ATLAS detector”. In: *Phys. Rev. D* 103.11 (2021), p. 112006. DOI: 10.1103/PhysRevD.103.112006. arXiv: 2102.10874 [hep-ex].
- [104] CMS Collaboration. “Search for new particles in events with energetic jets and large missing transverse momentum in proton-proton collisions at $\sqrt{s}=13$ TeV”. In: (2021).
- [105] E. Aprile et al. “Dark Matter Search Results from a One Ton-Year Exposure of XENON1T”. In: *Phys. Rev. Lett.* 121.11 (2018), p. 111302. DOI: 10.1103/PhysRevLett.121.111302. arXiv: 1805.12562 [astro-ph.CO].

- [106] M. L. Ahnen et al. “Limits to Dark Matter Annihilation Cross-Section from a Combined Analysis of MAGIC and Fermi-LAT Observations of Dwarf Satellite Galaxies”. In: *JCAP* 02 (2016), p. 039. DOI: 10.1088/1475-7516/2016/02/039. arXiv: 1601.06590 [astro-ph.HE].
- [107] L. F. Abbott and E. Farhi. “Are the Weak Interactions Strong?” In: *Phys. Lett.* 101B (1981), pp. 69–72. DOI: 10.1016/0370-2693(81)90492-5.
- [108] L. F. Abbott and E. Farhi. “A Confining Model of the Weak Interactions”. In: *Nucl. Phys. B* 189 (1981), pp. 547–556. DOI: 10.1016/0550-3213(81)90580-0.
- [109] M. Claudson, E. Farhi, and R. L. Jaffe. “The Strongly Coupled Standard Model”. In: *Phys. Rev. D* 34 (1986), p. 873. DOI: 10.1103/PhysRevD.34.873.
- [110] Gerard 't Hooft. “Topological aspects of quantum chromodynamics”. In: *From the Planck length to the Hubble radius. Proceedings, International School of Subnuclear Physics, Erice, Italy, August 29-September 7, 1998*. 1998, pp. 216–236. arXiv: hep-th/9812204 [hep-th].
- [111] Xavier Calmet and Harald Fritzsch. “The Electroweak interaction as a confinement phenomenon”. In: *Phys. Lett.* B496 (2000), pp. 161–168. DOI: 10.1016/S0370-2693(00)01299-5. arXiv: hep-ph/0008243 [hep-ph].
- [112] Xavier Calmet and Harald Fritzsch. “Calculation of the Higgs boson mass using the complementarity principle”. In: *Phys. Lett. B* 525 (2002), pp. 297–300. DOI: 10.1016/S0370-2693(01)01449-6. arXiv: hep-ph/0107085.
- [113] Xavier Calmet and Harald Fritzsch. “Electroweak D waves”. In: *Phys. Lett. B* 526 (2002), pp. 90–96. DOI: 10.1016/S0370-2693(01)01471-X. arXiv: hep-ph/0103333.
- [114] Nakin Lohitsiri and David Tong. “If the Weak Were Strong and the Strong Were Weak”. In: *SciPost Phys.* 7.5 (2019), p. 059. DOI: 10.21468/SciPostPhys.7.5.059. arXiv: 1907.08221 [hep-th].

- [115] John Preskill. “Subgroup Alignment in Hypercolor Theories”. In: *Nucl. Phys. B* 177 (1981), pp. 21–59. DOI: 10.1016/0550-3213(81)90265-0.
- [116] D. A. Kosower. “SYMMETRY BREAKING PATTERNS IN PSEUDOREAL AND REAL GAUGE THEORIES”. In: *Phys. Lett. B* 144 (1984), pp. 215–216. DOI: 10.1016/0370-2693(84)91806-9.
- [117] Randy Lewis, Claudio Pica, and Francesco Sannino. “Light Asymmetric Dark Matter on the Lattice: SU(2) Technicolor with Two Fundamental Flavors”. In: *Phys. Rev. D* 85 (2012), p. 014504. DOI: 10.1103/PhysRevD.85.014504. arXiv: 1109.3513 [hep-ph].
- [118] Rudy Arthur et al. “SU(2) gauge theory with two fundamental flavors: A minimal template for model building”. In: *Phys. Rev. D* 94.9 (2016), p. 094507. DOI: 10.1103/PhysRevD.94.094507. arXiv: 1602.06559 [hep-lat].
- [119] Tuomas Karavirta et al. “Determining the conformal window: SU(2) gauge theory with $N_f = 4, 6$ and 10 fermion flavours”. In: *JHEP* 05 (2012), p. 003. DOI: 10.1007/JHEP05(2012)003. arXiv: 1111.4104 [hep-lat].
- [120] M. Hayakawa et al. “Lattice Study on quantum-mechanical dynamics of two-color QCD with six light flavors”. In: *Phys. Rev. D* 88.9 (2013), p. 094506. DOI: 10.1103/PhysRevD.88.094506. arXiv: 1307.6696 [hep-lat].
- [121] Alessandro Amato et al. “Approaching the conformal window: systematic study of the particle spectrum in SU(2) field theory with $N_f = 2, 4$ and 6.” In: *PoS LATTICE2015* (2016), p. 225. DOI: 10.22323/1.251.0225. arXiv: 1511.04947 [hep-lat].
- [122] Viljami Leino et al. “Infrared fixed point of SU(2) gauge theory with six flavors”. In: *Phys. Rev. D* 97.11 (2018), p. 114501. DOI: 10.1103/PhysRevD.97.114501. arXiv: 1707.04722 [hep-lat].

- [123] Viljami Leino, Kari Rummukainen, and Kimmo Tuominen. “Slope of the beta function at the fixed point of SU(2) gauge theory with six or eight flavors”. In: *Phys. Rev. D* 98.5 (2018), p. 054503. DOI: 10.1103/PhysRevD.98.054503. arXiv: 1804.02319 [hep-lat].
- [124] T. Das et al. “Electromagnetic mass difference of pions”. In: *Phys. Rev. Lett.* 18 (1967), pp. 759–761. DOI: 10.1103/PhysRevLett.18.759.
- [125] Venkitesh Ayyar et al. “Radiative Contribution to the Composite-Higgs Potential in a Two-Representation Lattice Model”. In: *Phys. Rev. D* 99.9 (2019), p. 094504. DOI: 10.1103/PhysRevD.99.094504. arXiv: 1903.02535 [hep-lat].
- [126] Aneesh Manohar and Howard Georgi. “Chiral Quarks and the Nonrelativistic Quark Model”. In: *Nucl. Phys. B* 234 (1984), pp. 189–212. DOI: 10.1016/0550-3213(84)90231-1.
- [127] N. Aghanim et al. “Planck 2018 results. VI. Cosmological parameters”. In: *Astron. Astrophys.* 641 (2020), A6. DOI: 10.1051/0004-6361/201833910. arXiv: 1807.06209 [astro-ph.CO].
- [128] Alessandro Granelli et al. “ULYSSES: Universal LeptogeneSiS Equation Solver”. In: *Comput. Phys. Commun.* 262 (2021), p. 107813. DOI: 10.1016/j.cpc.2020.107813. arXiv: 2007.09150 [hep-ph].
- [129] F. Feroz, M. P. Hobson, and M. Bridges. “MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics”. In: *Mon. Not. Roy. Astron. Soc.* 398 (2009), pp. 1601–1614. DOI: 10.1111/j.1365-2966.2009.14548.x. arXiv: 0809.3437 [astro-ph].
- [130] Farhan Feroz and M. P. Hobson. “Multimodal nested sampling: an efficient and robust alternative to MCMC methods for astronomical data analysis”. In: *Mon. Not. Roy. Astron. Soc.* 384 (2008), p. 449. DOI: 10.1111/j.1365-2966.2007.12353.x. arXiv: 0704.3704 [astro-ph].

- [131] F. Feroz et al. “Importance Nested Sampling and the MultiNest Algorithm”. In: *ArXiv e-prints* (June 2013). arXiv: 1306.2144 [astro-ph.IM].
- [132] Buchner, J. et al. “X-ray spectral modelling of the AGN obscuring region in the CDFS: Bayesian model selection and catalogue”. In: *A&A* 564 (2014), A125. DOI: 10.1051/0004-6361/201322971. URL: <https://doi.org/10.1051/0004-6361/201322971>.
- [133] Kim Griest and Marc Kamionkowski. “Unitarity Limits on the Mass and Radius of Dark Matter Particles”. In: *Phys. Rev. Lett.* 64 (1990), p. 615. DOI: 10.1103/PhysRevLett.64.615.
- [134] jnhoward. *jnhoward/SU2LDM_public: Initial public version of code*. Version v0.1.0. Feb. 2022. DOI: 10.5281/zenodo.5965537. URL: <https://doi.org/10.5281/zenodo.5965537>.
- [135] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.
- [136] Thomas Kluyver et al. “Jupyter Notebooks – a publishing format for reproducible computational workflows”. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87–90.
- [137] Wolfram Research Inc. *Mathematica, Version 13.1*. Champaign, IL, 2022. URL: <https://www.wolfram.com/mathematica>.
- [138] F. James. *Statistical Methods in Experimental Physics*. World Scientific, 2006. ISBN: 9789812567956. URL: <https://books.google.com/books?id=QbBm2VhV5TQC>.
- [139] Andrew Fowlie and Michael Hugh Bardsley. “Superplot: a graphical interface for plotting and analysing MultiNest output”. In: *Eur. Phys. J. Plus* 131.11 (2016), p. 391. DOI: 10.1140/epjp/i2016-16391-0. arXiv: 1603.00555 [physics.data-an].
- [140] David Tucker-Smith and Neal Weiner. “Inelastic dark matter”. In: *Phys. Rev. D* 64 (2001), p. 043502. DOI: 10.1103/PhysRevD.64.043502. arXiv: hep-ph/0101138.

- [141] Joseph Bramante et al. “Inelastic frontier: Discovering dark matter at high recoil energy”. In: *Phys. Rev. D* 94.11 (2016), p. 115026. DOI: 10.1103/PhysRevD.94.115026. arXiv: 1608.02662 [hep-ph].
- [142] Richard J. Hill and Mikhail P. Solon. “Universal behavior in the scattering of heavy, weakly interacting dark matter on nuclear targets”. In: *Phys. Lett. B* 707 (2012), pp. 539–545. DOI: 10.1016/j.physletb.2012.01.013. arXiv: 1111.0016 [hep-ph].
- [143] Matthew Low and Lian-Tao Wang. “Neutralino dark matter at 14 TeV and 100 TeV”. In: *JHEP* 08 (2014), p. 161. DOI: 10.1007/JHEP08(2014)161. arXiv: 1404.0682 [hep-ph].
- [144] A. Acharyya et al. “Sensitivity of the Cherenkov Telescope Array to a dark matter signal from the Galactic centre”. In: *JCAP* 01 (2021), p. 057. DOI: 10.1088/1475-7516/2021/01/057. arXiv: 2007.16129 [astro-ph.HE].
- [145] Matthew Feickert and Benjamin Nachman. *A Living Review of Machine Learning for Particle Physics*. 2021. DOI: 10.48550/ARXIV.2102.02770. URL: <https://arxiv.org/abs/2102.02770>.
- [146] Julian Collado et al. “Learning to identify electrons”. In: *Phys. Rev. D* 103.11 (2021), p. 116028. DOI: 10.1103/PhysRevD.103.116028.
- [147] Simon Badger et al. “Machine Learning and LHC Event Generation”. In: (Mar. 2022). Ed. by Anja Butter, Tilman Plehn, and Steffen Schumann. arXiv: 2203.07460 [hep-ph].
- [148] Pierre Baldi et al. “How to GAN Higher Jet Resolution”. In: (Dec. 2020). arXiv: 2012.11944 [hep-ph].
- [149] Maria Laura Piscopo, Michael Spannowsky, and Philip Waite. “Solving differential equations with neural networks: Applications to the calculation of cosmological phase transitions”. In: *Physical Review D* 100.1 (July 2019). DOI: 10.1103/physrevd.100.016002. URL: <https://doi.org/10.1103/physrevd.100.016002>.

- [150] Jack Y. Araz, Juan Carlos Criado, and Michael Spannowsky. *Elvet – a neural network-based differential equation and variational problem solver*. 2021. DOI: 10.48550/ARXIV.2103.14575. URL: <https://arxiv.org/abs/2103.14575>.
- [151] Sascha Caron et al. “Constraining the parameters of high-dimensional models with active learning”. In: *The European Physical Journal C* 79.11 (Nov. 2019). DOI: 10.1140/epjc/s10052-019-7437-5. URL: <https://doi.org/10.1140%5C%2Fepjc%5C%2Fs10052-019-7437-5>.
- [152] Jacob Hollingsworth et al. “Efficient sampling of constrained high-dimensional theoretical spaces with machine learning”. In: *The European Physical Journal C* 81.12 (Dec. 2021). DOI: 10.1140/epjc/s10052-021-09941-9. URL: <https://doi.org/10.1140%5C%2Fepjc%5C%2Fs10052-021-09941-9>.
- [153] Sergei Gukov et al. *Learning to Unknot*. 2020. DOI: 10.48550/ARXIV.2010.16263. URL: <https://arxiv.org/abs/2010.16263>.
- [154] Alex Davies et al. “Advancing mathematics by guiding human intuition with AI”. In: *Nature* 600.7887 (2021), pp. 70–74. DOI: 10.1038/s41586-021-04086-x. URL: <https://doi.org/10.1038/s41586-021-04086-x>.
- [155] James Halverson, Brent Nelson, and Fabian Ruehle. “Branes with brains: exploring string vacua with deep reinforcement learning”. In: *Journal of High Energy Physics* 2019.6 (June 2019). DOI: 10.1007/jhep06(2019)003. URL: <https://doi.org/10.1007%5C%2Fjhep06%5C%282019%5C%29003>.
- [156] Daniel A. Roberts, Sho Yaida, and Boris Hanin. *The Principles of Deep Learning Theory*. 2021. DOI: 10.48550/ARXIV.2106.10165. URL: <https://arxiv.org/abs/2106.10165>.
- [157] James Halverson, Anindita Maiti, and Keegan Stoner. “Neural networks and quantum field theory”. In: *Machine Learning: Science and Technology* 2.3 (Apr. 2021),

- p. 035002. DOI: 10.1088/2632-2153/abeca3. URL: <https://doi.org/10.1088%5C%2F2632-2153%5C%2Fabeca3>.
- [158] James Halverson. *Building Quantum Field Theories Out of Neurons*. 2021. DOI: 10.48550/ARXIV.2112.04527. URL: <https://arxiv.org/abs/2112.04527>.
- [159] Sven Krippendorf and Michael Spannowsky. *A duality connecting neural network and cosmological dynamics*. 2022. DOI: 10.48550/ARXIV.2202.11104. URL: <https://arxiv.org/abs/2202.11104>.
- [160] Johanna Erdmenger, Kevin T. Grosvenor, and Ro Jefferson. “Towards quantifying information flows: relative entropy in deep neural networks and the renormalization group”. In: *SciPost Phys.* 12.1 (2022), p. 041. DOI: 10.21468/SciPostPhys.12.1.041. arXiv: 2107.06898 [hep-th].
- [161] Kevin T. Grosvenor and Ro Jefferson. “The edge of chaos: quantum field theory and deep neural networks”. In: *SciPost Phys.* 12.3 (2022), p. 081. DOI: 10.21468/SciPostPhys.12.3.081. arXiv: 2109.13247 [hep-th].
- [162] Stephen J Wright and Jorge Nocedal. *Springer Series in Operations Research and Financial Engineering*. New York, NY: Springer, 1999. DOI: 10.1007/978-0-387-40065-5.
- [163] Aaron Pierce, Nausheen R. Shah, and Stefan Vogl. “Stop co-annihilation in the minimal supersymmetric standard model revisited”. In: *Physical Review D* 97.2 (Jan. 2018). DOI: 10.1103/physrevd.97.023008. URL: <https://doi.org/10.1103%2Fphysrevd.97.023008>.

Appendix A

Chiral transformation invariance of QCD lagrangian

As a fun exercise, we will see the $SU(2)_L \times SU(2)_R$ invariance of

$$\mathcal{L}_{QCD} = -\frac{1}{4}F_{\mu\nu}^a F^{a\mu\nu} + i\bar{u}\not{D}u + i\bar{d}\not{D}d = -\frac{1}{4}F_{\mu\nu}^a F^{a\mu\nu} + i\bar{Q}\not{D}Q$$

play out explicitly. As a reminder, we are only considering the first generation of quarks (u and d) and are neglecting their masses.

For simplicity, we will focus on the term $\bar{Q}\not{D}Q$. We can split $Q := (u, d)^T$ into two parts $Q_{L,R}$. Each of these parts will have a definite transformation under $SU(2)_L \times SU(2)_R$. Specifically,

$$Q = P_L Q + (\mathbb{1} - P_L)Q = P_L Q + P_R Q =: Q_L + Q_R \tag{A.1}$$

where $P_L := (1 - \gamma^5)/2$ is the left-handed projection operator. These left and right handed

parts, $Q_{L,R}$, transform as follows

$$Q_L \xrightarrow{\text{SU}(2)_L} LQ_L \quad \text{and} \quad Q_L \xrightarrow{\text{SU}(2)_R} Q_L \quad (\text{A.2})$$

$$Q_R \xrightarrow{\text{SU}(2)_R} RQ_R \quad \text{and} \quad Q_R \xrightarrow{\text{SU}(2)_L} Q_R. \quad (\text{A.3})$$

And, recalling that $\bar{Q}_{L,R} = Q_{L,R}^\dagger \gamma^0$, $\bar{Q}_{L,R}$ transforms as

$$\bar{Q}_L = Q_L^\dagger \gamma^0 \xrightarrow{\text{SU}(2)_L} Q_L^\dagger L^\dagger \gamma^0 = \bar{Q}_L L^\dagger \quad (\text{A.4})$$

$$\bar{Q}_R = Q_R^\dagger \gamma^0 \xrightarrow{\text{SU}(2)_R} Q_R^\dagger R^\dagger \gamma^0 = \bar{Q}_R R^\dagger, \quad (\text{A.5})$$

where we have used the fact that γ^0 commutes (trivially) with L^\dagger, R^\dagger .

We can expand out the desired term

$$\bar{Q} \not{D} Q = (\bar{Q}_L + \bar{Q}_R) \not{D} (Q_L + Q_R) \quad (\text{A.6})$$

$$= \bar{Q}_L \not{D} Q_L + \bar{Q}_R \not{D} Q_R + \bar{Q}_L \not{D} Q_R + \bar{Q}_R \not{D} Q_L \quad (\text{A.7})$$

$$= \bar{Q}_L \not{D} Q_L + \bar{Q}_R \not{D} Q_R. \quad (\text{A.8})$$

Where in the last step we have used the fact that the left and right projection operators will annihilate. To see this in more detail, first use the fact that $\{\gamma^\mu, \gamma^5\} = 0$, we then know that $\bar{Q}_L = Q_L^\dagger \gamma^0 = Q^\dagger P_L \gamma^0 = Q^\dagger \gamma^0 P_R = \bar{Q} P_R$. This then implies

$$\bar{Q}_L \not{D} Q_R = \bar{Q} P_R \gamma^\mu D_\mu P_R Q = \bar{Q} \gamma^\mu D_\mu P_L P_R Q = 0. \quad (\text{A.9})$$

Eq. (A.8) transforms as

$$\bar{Q}_L \not{D} Q_L + \bar{Q}_R \not{D} Q_R \xrightarrow{\text{SU}(2)_L \times \text{SU}(2)_R} \bar{Q}_L L^\dagger \not{D} L Q_L + \bar{Q}_R R^\dagger \not{D} R Q_R. \quad (\text{A.10})$$

Therefore, the invariance of this term depends on how \not{D} transforms. We need $L^\dagger \not{D} L = R^\dagger \not{D} R = \not{D}$. This is true because we are assuming $D^\mu = \partial^\mu - ig_s T^a A_a^\mu$ which is invariant under $SU(2)_{L,R}$ transformations.

Appendix B

OTUS Statistical Matching

	z vs \tilde{z}		
	W [GeV²]	(χ_R^2, dof)	KS
Fig. 3a (p_y)	$1.34 \times 10^{+00}$	(50.583, 23)	1.61×10^{-02}
Fig. 3a (p_z)	$1.59 \times 10^{+00}$	(1.325, 26)	4.90×10^{-03}
Fig. 3a (E)	$1.29 \times 10^{+00}$	(8.814, 26)	1.47×10^{-02}
Fig. 5a	$2.73 \times 10^{+01}$	(822.762, 39)	2.46×10^{-01}

Table B.1: **Table showing \mathcal{Z} space statistical test results for the $Z \rightarrow e^+e^-$ dataset.** These tests were performed on the distributions in the referenced figures in the main text. W is the Wasserstein distance, χ_R^2 is the reduced χ^2 and dof is the degrees-of-freedom, and KS is the value of the Kolmogorov-Smirnov statistical test. See the Evaluation section in the main text for detailed information about the calculations of these statistics.

	x vs \tilde{x}			x vs \tilde{x}'		
	W [GeV²]	(χ_R^2, dof)	KS	W [GeV²]	(χ_R^2, dof)	KS
Fig. 3b (p_y)	4.22×10^{-01}	(1.391, 23)	3.48×10^{-03}	$1.05 \times 10^{+00}$	(37.560, 23)	1.22×10^{-02}
Fig. 3b (p_z)	$3.71 \times 10^{+00}$	(1.523, 26)	1.03×10^{-02}	$9.53 \times 10^{+00}$	(4.775, 26)	7.49×10^{-03}
Fig. 3b (E)	6.64×10^{-01}	(0.489, 26)	3.19×10^{-03}	$3.64 \times 10^{+00}$	(9.370, 26)	2.00×10^{-02}
Fig. 5b	7.28×10^{-01}	(5.055, 39)	2.61×10^{-02}	7.15×10^{-01}	(12.821, 39)	3.14×10^{-02}

Table B.2: **Table showing \mathcal{X} space statistical test results for the $Z \rightarrow e^+e^-$ dataset.** These tests were performed on the distributions in the referenced figures in the main text. W is the Wasserstein distance, χ_R^2 is the reduced χ^2 and dof is the degrees-of-freedom, and KS is the value of the Kolmogorov-Smirnov statistical test. See the Evaluation section in the main text for detailed information about the calculations of these statistics.

	z vs \tilde{z}		
	W [GeV ²]	(χ_R^2 , dof)	KS
Fig. 6a (p_y)	$1.58 \times 10^{+01}$	(7.418, 49)	1.25×10^{-02}
Fig. 6a (p_z)	$5.52 \times 10^{+01}$	(4.613, 55)	1.65×10^{-02}
Fig. 6a (E)	$6.20 \times 10^{+01}$	(31.228, 31)	4.04×10^{-02}

Table B.3: **Table showing \mathcal{Z} space statistical test results for the semileptonic $t\bar{t}$ dataset.** These tests were performed on the distributions in the referenced figures in the main text. W is the Wasserstein distance, χ_R^2 is the reduced χ^2 and dof is the degrees-of-freedom, and KS is the value of the Kolmogorov-Smirnov statistical test. See the Evaluation section in the main text for detailed information about the calculations of these statistics.

	x vs \tilde{x}			x vs \tilde{x}'		
	W [GeV ²]	(χ_R^2 , dof)	KS	W [GeV ²]	(χ_R^2 , dof)	KS
Fig. 6b (p_y)	$2.40 \times 10^{+01}$	(2.395, 49)	1.66×10^{-02}	$1.23 \times 10^{+02}$	(34.021, 49)	4.59×10^{-02}
Fig. 6b (p_z)	$1.08 \times 10^{+02}$	(0.828, 55)	9.90×10^{-03}	$3.42 \times 10^{+02}$	(1.980, 55)	6.94×10^{-03}
Fig. 6b (E)	$4.11 \times 10^{+01}$	(1.281, 30)	1.02×10^{-02}	$3.24 \times 10^{+02}$	(50.072, 30)	4.80×10^{-02}
Fig. 8a	$1.60 \times 10^{+02}$	(1.192, 43)	7.63×10^{-03}	$1.03 \times 10^{+03}$	(54.598, 43)	1.03×10^{-01}
Fig. 8b	9.30×10^{-01}	(3.974, 35)	1.66×10^{-02}	$1.04 \times 10^{+02}$	(68.392, 35)	1.09×10^{-01}
Fig. 8c	$8.83 \times 10^{+00}$	(1.579, 30)	5.91×10^{-03}	$7.41 \times 10^{+01}$	(92.533, 30)	1.31×10^{-01}
Fig. 8d	$2.21 \times 10^{+01}$	(2.455, 41)	1.72×10^{-02}	$1.11 \times 10^{+03}$	(160.712, 41)	2.35×10^{-01}

Table B.4: **Table showing \mathcal{X} space statistical test results for semileptonic $t\bar{t}$ dataset.** These tests were performed on the distributions in the referenced figures in the main text. W is the Wasserstein distance, χ_R^2 is the reduced χ^2 and dof is the degrees-of-freedom, and KS is the value of the Kolmogorov-Smirnov statistical test. See the Evaluation section in the main text for detailed information about the calculations of these statistics.

Appendix C

OTUS Ablation Study

In this section we show the results of an ablation study to demonstrate the effect of the various hyperparameters. As seen in our final loss function Equation (11), the main hyperparameters of our approach are the λ coefficient in front of the latent space loss, as well as the β_E and β_D coefficients weighing the anchor losses for the encoder and decoder, respectively. For the semileptonic $t\bar{t}$ study the only hyperparameter is λ , as the anchor loss is redundant with the choice of a ResNet [95] architecture (see Section 6.2.3). We performed ablations by retraining the models as in Section 6.3.3 but with different values of the hyperparameters on a grid, and comparing the results on validation data.

For studying the effect of λ , we reran both the $Z \rightarrow e^+e^-$ and the semileptonic $t\bar{t}$ studies with λ in $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, while keeping all other hyperparameters unchanged (specifically, in the $Z \rightarrow e^+e^-$ study we kept $\beta_E = \beta_D = 50$). For the effect of the anchor loss coefficients, we always assume that $\beta_E = \beta_D$ and define a shared hyperparameter $\beta := \beta_E = \beta_D$. We reran the $Z \rightarrow e^+e^-$ study with β in $\{0, 10, 20, 50, 100, 200\}$, while keeping $\lambda = 1$ as in the original experiment. We did not repeat this for the semileptonic $t\bar{t}$ study as it did not use an anchor loss.

We first consider how the hyperparameters on the anchor loss terms, $\beta_E = \beta_D$, affect performance. The anchor losses are direct constraints on the learned encoding and decoding mappings which are based on physical concerns. Namely, the anchor loss penalizes networks which would map electron/positron (e^\mp) information in \mathcal{Z} to positron/electron (e^\pm) information in \mathcal{X} , and vice versa. We impose this constraint because we know that misidentification of charge in the process of data reconstruction is extremely rare in particle experiments. Therefore, for our simulation to be physical, it should not make these unphysical inversions. Unsurprisingly, without this constraint we can see that these inversions can occur during training (see Supplementary Figure C.1). On the other hand, if the values of $\beta_E = \beta_D$ are too high we observe unphysical behavior. This is likely due to the fact that the anchor loss is only a proxy for enforcing charge conservation.

We next consider the hyperparameter λ which is present in both case studies. The behavior of λ has theoretical motivations. The WAE method aims to minimize $W_c(p(x), p_D(x))$ by converting its calculation into a constrained optimization problem. It was shown [31] that $W_c(p(x), p_D(x)) = \inf_{p_E(z|x): p_E(z)=p(z)} \mathbb{E}[c(X, D(Z))]$ for a deterministic decoder $p_D(x|z) = \delta_{D(z)}(x)$,¹. Namely, we need to minimize a reconstruction error over all probabilistic encoders, $p_E(z|x)$, satisfying the latent-space matching condition, $p(z) \stackrel{!}{=} p_E(z)$ where $p_E(z) := \int_x p_E(z|x)p(x)dx$. To make the constrained optimization computationally tractable, the WAE method only softly enforces this constraint via a penalty term $\lambda d_z(p(z), p_E(z))$, and considers minimizing the surrogate penalty loss $\mathbb{E}_{p(x)p_E(z|x)p_D(\tilde{x}|z)}[c(x, \tilde{x})] + \lambda d_z(p(z), p_E(z))$ instead.

Standard results on penalty methods [162] conclude that for a fixed decoder, $p_D(x|z)$, globally minimizing the penalty loss with respect to the encoder $p_E(z|x)$ results in a lower bound on $W_c(p(x), p_D(x))$, and solving a sequence of such penalized problems while annealing λ towards infinity results in the exact $W_c(p(x), p_D(x))$. However, when training a WAE, it

¹We can show that more generally, for a stochastic decoder, we have an upper bound $W_c(p(x), p_D(x)) \leq \inf_{p_E(z|x): p_E(z)=p(z)} \mathbb{E}_{p(x)p_E(z|x)p_D(\tilde{x}|z)}[c(X, \tilde{X})]$

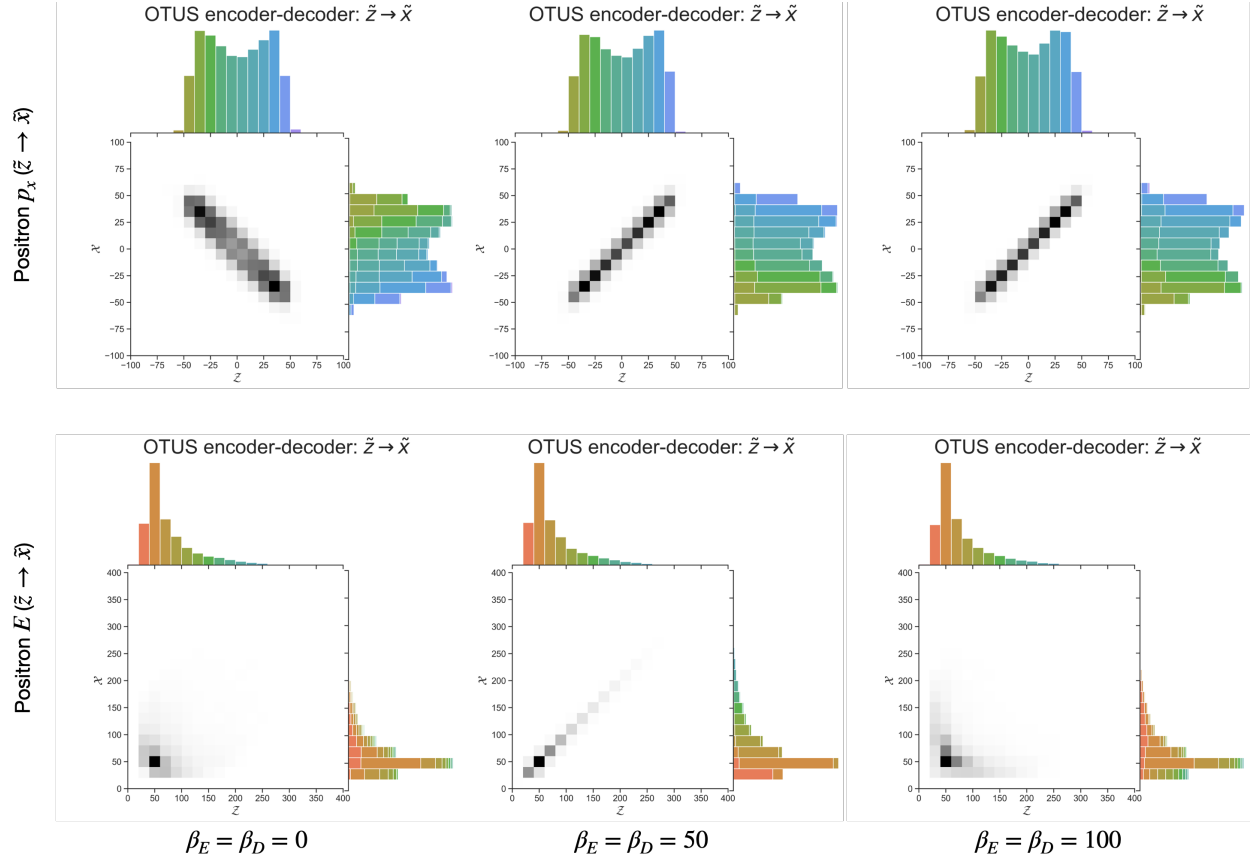


Figure C.1: **Results of anchor loss ablation study in the $Z \rightarrow e^+e^-$ study.** For $\beta_E = \beta_D = 0$ we can see that unphysical transformations can arise. In p_x , negative values in \mathcal{Z} are being mapped to positive values in \mathcal{X} . This is a result of e^\pm information being swapped in the learned transformation. For $\beta_E = \beta_D = 50$, this effect goes away; we also see more physical behavior in E as well. For $\beta_E = \beta_D = 100$, we observe that high values of β_E and β_D inadvertently encourage unphysical behavior in E . This is likely due to the fact that the anchor loss is only a proxy for enforcing charge conservation.

is expensive to repeat this inner optimization procedure after every decoder update, so in practice both the encoder and decoder are optimized jointly on a penalty loss, keeping λ fixed throughout the entire training [31].

While the theoretical guarantees of the penalty method no longer applies to the joint Stochastic Gradient Descent training procedure used in practice, it does suggest that λ should be set to be as large as possible (and perhaps annealed during training) to better enforce the latent space matching, and consequently offer a better approximation of the ideal objective $W_c(p(x), p_D(x))$. Indeed, recently it was proven [56] that perfect latent space matching $p_E(z) == p(z)$ is a necessary condition for $W(p(x), p_D(x)) = 0$.

Overall, our ablation experiments confirmed this notion and showed that when λ is too small and thus the penalty on latent space matching too weak, neither the encoder or the decoder’s marginal distribution $(p_E(z), p_D(x))$ could capture the ground truth $p(z)$ or $p(x)$ well, despite minimal reconstruction error. We see this behavior in both test cases, however we note that in the semileptonic $t\bar{t}$ the behavior is somewhat less dramatic because of the heavy initial bias towards an identity mapping due to the ResNet [95] architecture (see Supplementary Figure C.2).

We see performance in matching principal axes improve as λ grows larger, possibly plateauing in the case of the semileptonic $t\bar{t}$ study. This plateau is potentially due to issues with optimization and poor numerical conditioning with overly large λ . However, we find that too large of a value of λ results in unphysical mappings.

Specifically, we find unphysical behavior when we view the transport plots and derived quantities (see Fig C.3). Again, we note that this is less noticeable for the semileptonic $t\bar{t}$ study due to the ResNet [95] architecture. We find that the ideal choice is $\lambda \approx 1$ for the $Z \rightarrow e^+e^-$ study and $\lambda \approx 20$ for the semileptonic $t\bar{t}$ study; this retains acceptable principal axis matching while not introducing unphysical transformation characteristics. We suspect

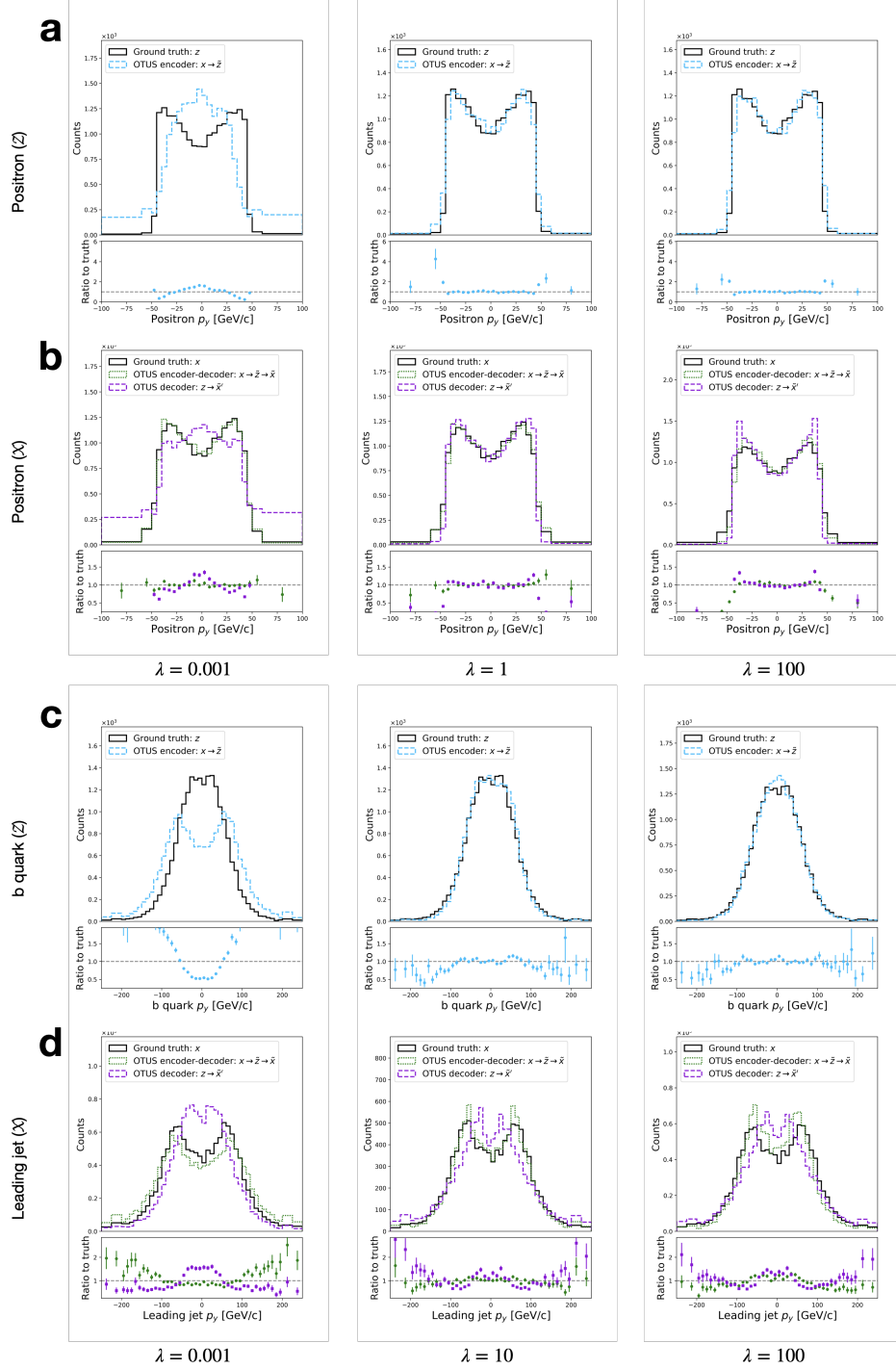


Figure C.2: **Results of λ ablation study on principal axis matching.** **a** Matching of the positron's p_y distribution for $\lambda = 0.001$, $\lambda = 1$, and $\lambda = 100$ in \mathcal{Z} for the $Z \rightarrow e^+e^-$ study. **b** Matching of the positron's p_y distribution for $\lambda = 0.001$, $\lambda = 1$, and $\lambda = 100$ in \mathcal{X} for the $Z \rightarrow e^+e^-$ study. **c** Matching of the b quark's p_y distribution for $\lambda = 0.001$, $\lambda = 10$, and $\lambda = 100$ in \mathcal{Z} for the semileptonic $t\bar{t}$ study. **d** Matching of the leading jet's p_y distribution for $\lambda = 0.001$, $\lambda = 10$, and $\lambda = 100$ in \mathcal{X} for the semileptonic $t\bar{t}$ study. For small values of λ ($\lambda = 0.001$) we find that performance suffers as latent space matching is not enforced. This improves as we increase λ but eventually plateaus.

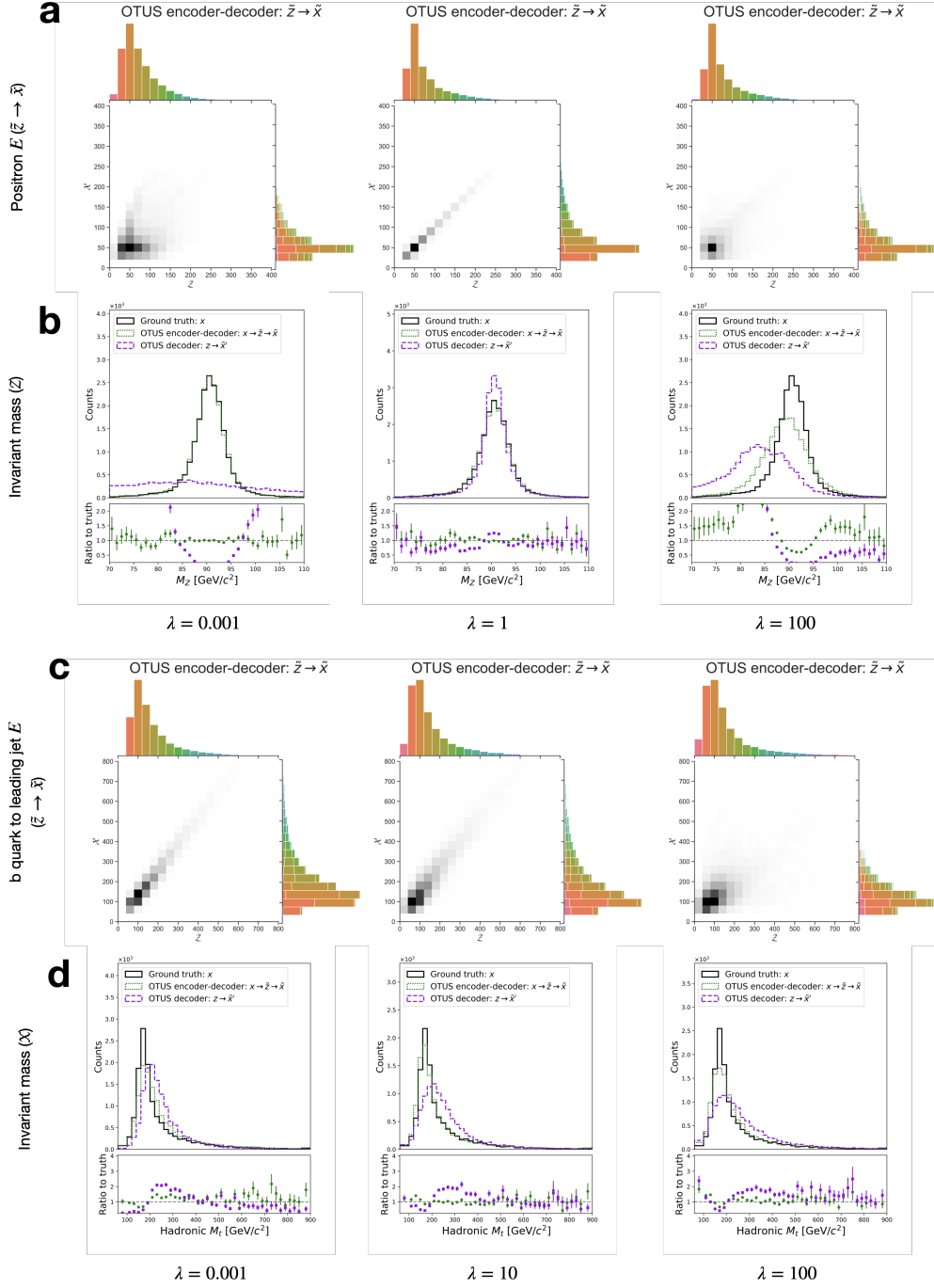


Figure C.3: **Results of λ ablation study on transport plots and derived quantity matching.** **a** Transport plans from $\tilde{z} \rightarrow \tilde{x}$ of the positron's E distribution for $\lambda = 0.001$, $\lambda = 1$, and $\lambda = 100$ for the $Z \rightarrow e^+e^-$ study. **b** Matching of the invariant mass of the Z -boson for $\lambda = 0.001$, $\lambda = 1$, and $\lambda = 100$ for the $Z \rightarrow e^+e^-$ study. **c** Transport plans from $\tilde{z} \rightarrow \tilde{x}$ of the b quark's E distribution in \mathcal{Z} to the leading jet's E distribution in \mathcal{X} for $\lambda = 0.001$, $\lambda = 10$, and $\lambda = 100$ for the semileptonic $t\bar{t}$ study. **d** Matching of the invariant mass of the top-quark, M_t , reconstructed using information from the hadronically decaying W -boson for $\lambda = 0.001$, $\lambda = 10$, and $\lambda = 100$ for the semileptonic $t\bar{t}$ study.

that if the choice of λ is too large, it over-constrains the optimization problem and should instead be annealed.

As discussed, instead of an expensive double-loop procedure where we train the encoder to optimality with the penalty method before updating the decoder, we forego theoretical considerations by jointly optimizing the encoder and decoder of a WAE on a surrogate loss as in [31]. The resulting loss is neither an upper nor a lower bound on the ideal objective $W_c(p(x), p_D(x))$, and we choose λ by experimentation. An alternative would be to use the Sinkhorn Autoencoder [56] approach, which only needs a large enough λ for its loss to be a proper upper bound on $W_c(p(x), p_D(x))$. This is further motivation that this method should be explored in future work.

Appendix D

Matrices for One Generation

First, recall the definitions for the Pauli matrices

$$\sigma_1 := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 := \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (\text{D.1})$$

Additionally, define the following matrices

$$l_1 := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad l_2 := \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \quad (\text{D.2})$$

The explicit form of the matrices in Eq. (5.3) for a single generation of Standard Model doublets in addition to χ_1 and χ_2 are:

$$L_j = \frac{1}{2} \begin{pmatrix} \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \\ \mathbb{0}_{2 \times 2} & \sigma_j & \mathbb{0}_{2 \times 2} \\ \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \end{pmatrix} \quad \text{where } j = 1, 2, 3, \quad (\text{D.3})$$

$$Q = \frac{1}{2} \begin{pmatrix} \sigma_3 & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \\ \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \\ \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & \sigma_3 \end{pmatrix}, \quad (\text{D.4})$$

$$L^{j,+} = \begin{pmatrix} \mathbb{0}_{2 \times 2} & l_j & \mathbb{0}_{2 \times 2} \\ \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \\ \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \end{pmatrix} \quad \text{where } j = 1, 2, \quad (\text{D.5})$$

$$L^{j,-} = \begin{pmatrix} \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \\ l_j^T & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \\ \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \end{pmatrix} \quad \text{where } j = 1, 2, \quad (\text{D.6})$$

$$J = \text{diag} \left(-\frac{s_Q^2}{2}, \frac{s_Q^2}{2} - \frac{1}{3}, \frac{1}{6}, \frac{1}{6}, -\frac{s_Q^2}{2}, \frac{s_Q^2}{2} \right) \quad (\text{D.7})$$

$$M = \frac{m_{\text{DM}}}{2} \begin{pmatrix} \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \\ \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} \\ \mathbb{0}_{2 \times 2} & \mathbb{0}_{2 \times 2} & i\sigma_2 \end{pmatrix}. \quad (\text{D.8})$$

Appendix E

Statistics Supplement

In this appendix we review the statistics of how one draws valid regions in parameter space. There are generally two kinds of regions which can be drawn *credible* regions and *confidence* regions, which correspond to whether you are operating in a Bayesian or frequentist setting.

The frequentist setting assumes that there is a single, constant true (but unknown) parameter, θ_0 . The goal then is to try to draw a confidence region that contains this parameter with some probability. Whereas the Bayesian setting assumes only that there exists a *distribution* of possible true values. In the Bayesian setting, the posterior $p(\theta \mid D)$ is the probability density distribution that encodes all knowledge about the parameters, θ . Namely, it incorporates prior knowledge given by the aptly named prior distribution, $p(\theta)$, and information provided by the observed data, D . The posterior can be used to draw credible regions, but these do not come with guarantees of *coverage*. Namely, we cannot say that the credible regions contain the true parameter with some probability. This makes sense because in a Bayesian framework there is no notion of a constant true parameter. Both approaches have their advantages and disadvantages, but adopting the frequentist setting (i.e. drawing confidence regions) is more common.

For this analysis, we therefore focus on using $-2 \ln L$ to draw confidence regions in parameter space. This is quite effortless due to $-2 \ln L$ having some very nice statistical properties which we outline below. Since our parameter space is multidimensional (2 dimensional) this process is called finding the *profile likelihood*.

As is generally true in statistics, everything is easier for normally distributed data. Unfortunately, often times the nice properties of normally distributed data do not carry over to general distributions. However, the likelihood is a special case where we can retain much of the inference potential of the normally distributed case.

As an illustrative example, consider the case of a single data sample, D , which is normally distributed with the known parameter variance, σ^2 , but unknown parameter mean, μ . The likelihood for this single observation, as a function of the unknown parameter μ , is given by

$$\ln L(D | \mu) = \ln c - \frac{(\mu - D)^2}{2\sigma^2}. \quad (\text{E.1})$$

The typical convention is to set $c = 1$ such that the maximum $\ln L$ (minimum $-2 \ln L$) occurs at $D = \mu$. Rearranging we have

$$-2 \ln L(D | \mu) = \frac{(\mu - D)^2}{\sigma^2}. \quad (\text{E.2})$$

We are ultimately interested in ensuring that the interval $[D - \Delta_L, D + \Delta_U]$ for some $\Delta_{L,U}$ contains the unknown μ with probability β . By convention we are usually most interested in $\beta = 68.3\%$ and $\beta = 95.5\%$ which define 1σ and 2σ intervals respectively. Explicitly, a 1σ interval implies that there is a 68.3% chance that the true unknown parameter μ falls in the interval.

From the properties of the normal distribution, we know that $P[(D - \mu)^2 \leq \sigma] = 68.3\%$. This can be rearranged such that we find $P[(D - \mu)^2 \leq \sigma] = P[D - \sigma \leq \mu \leq D + \sigma] = 68.3\%$.

Similarly, we find that $P[D - 2\sigma \leq \mu \leq D + 2\sigma] = 95.5\%$. Therefore the 1σ and 2σ intervals are $[D - \sigma, D + \sigma]$ and $[D - 2\sigma, D + 2\sigma]$ respectively.

Notice that we can also arrive at this result another way. If we draw a line at $-2 \ln L = 1$ we find that it will intersect $\ln L$ at $\mu = D \mp \sigma$, thereby defining the edges of the 1σ interval. A line at $-2 \ln L = 4$ will similarly define the 2σ interval. Therefore, we can obtain the interval directly from $-2 \ln L$.

This story generalizes to multiple unknown parameters. Consider a k -dimensional parameter space denoted by θ . The confidence region with probability content β are given by the hypersurface

$$-2 \ln L_\theta(D | \theta) = -2 \ln L_{\max} + \chi_\beta^2(k). \quad (\text{E.3})$$

Where $\ln L_{\max} := L(D | \hat{\theta})$, namely the value of the maximum of the likelihood function, occurring at $\theta = \hat{\theta}$, where $\hat{\theta}$ is the true parameter space point. One can also define the likelihood ratio $\ell := L(D | \theta)/L(D | \hat{\theta})$

$$-2 \ln \ell(\theta) = \chi_\beta^2(k). \quad (\text{E.4})$$

This generalization can be justified by the fact that -2ℓ is asymptotically distributed as $\chi^2(k)$. Therefore, by the properties of $\chi_\beta^2(k)$

$$\beta = P[\chi^2(k) \leq \chi_\beta^2(k)] \approx P[-2 \ln \ell(\theta) \leq \chi_\beta^2(k)] \quad (\text{E.5})$$

Note that this statement is only approximately true.

The above is technically still assuming normally distributed data, however the likelihood function has the property of *invariance* which allows you to make statistical inferences about functions $g(\theta)$. Therefore, even if your likelihood as a function of θ , $L(D | \theta)$, is very different

than it would be in the normally distributed case (i.e. non-parabolic), you can perform a change of parameter space variables to get $L(D | g(\theta))$ which is consistent with the normally distributed case (i.e. parabolic). In this $g(\theta)$ parameter space, you can use all of the above to draw confidence regions. Then, crucially, the property of invariance allows you to use the regions in $g(\theta)$ parameter space to immediately find the corresponding regions in θ space. This is important because it allows you to define the confidence regions for an arbitrarily distributed likelihood function. Note that the confidence regions in θ space are exact only to order $1/N$ where N is the number of observed data samples.

For example, consider a one-dimensional case ($k = 1$). Assume that $\ln L_\theta$ has some non-parabolic shape in θ parameter space. It has a peak at the true value, $\hat{\theta}$, and its 1σ confidence interval is defined as $[\theta_L, \theta_U]$, such that $\ln L_\theta(\theta_{L,U}) = -1/2$. Assume that there is a function, $g(\theta)$, such that in this new space $\ln L_g$ is parabolic. It has a peak at the true value, \hat{g} , and its 1σ confidence interval is defined as $[g_L, g_U]$, such that $\ln L_g(g_{L,U}) = -1/2$. The property of invariance suggests that the confidence interval in g parameter space will, to order $1/N$, correspond to the interval in θ parameter space (i.e. $g(\theta_{L,U}) = g_{L,U}$).

Appendix F

Small Angle Approximation

In this appendix we explicitly work through the calculation leading to the result in Eq. (5.11).

For convenience, we repeat Eq. (5.9) and Eq. (5.10) below

$$\tan 2\theta = 2 \frac{M_{0,14}^2}{(M_{0,0}^2 - M_{14,14}^2)},$$

$$M_{0,0}^2 = 24\kappa\Lambda_W^2 + \frac{2\Lambda_W^3 m_{\text{DM}}}{3f^2}, \quad M_{0,14}^2 = -\frac{2\sqrt{2}\Lambda_W^3 m_{\text{DM}}}{3f^2}, \quad M_{14,14}^2 = \frac{4\Lambda_W^3 m_{\text{DM}}}{3f^2}.$$

We begin by simplifying the denominator of Eq. (5.9)

$$M_{0,0}^2 - M_{14,14}^2 = 24\kappa\Lambda_W^2 + \frac{2\Lambda_W^3 m_{\text{DM}}}{3f^2} - \frac{4\Lambda_W^3 m_{\text{DM}}}{3f^2} = 24\kappa\Lambda_W^2 - \frac{2\Lambda_W^3 m_{\text{DM}}}{3f^2} \quad (\text{F.1})$$

$$= \frac{72\kappa\Lambda_W^2 f^2 - 2\Lambda_W^3 m_{\text{DM}}}{3f^2}. \quad (\text{F.2})$$

We now substitute this and the expression for $M_{0,14}^2$ into Eq. (5.9), canceling the $3f^2$ terms, to get

$$\tan 2\theta = 2 \left[\frac{-2\sqrt{2}\Lambda_W^3 m_{\text{DM}}}{72\kappa\Lambda_W^2 f^2 - 2\Lambda_W^3 m_{\text{DM}}} \right] = \frac{-2\sqrt{2}\Lambda_W m_{\text{DM}}}{36\kappa f^2 - \Lambda_W m_{\text{DM}}}. \quad (\text{F.3})$$

We now use $\Lambda_W = 4\pi f$ to get

$$\tan 2\theta = -2\sqrt{2} \left[\frac{(4\pi f) m_{\text{DM}}}{36\kappa f^2 - (4\pi f) m_{\text{DM}}} \right] = -2\sqrt{2} \pi m_{\text{DM}} \left[\frac{1}{9\kappa f - \pi m_{\text{DM}}} \right]. \quad (\text{F.4})$$

We would now like to expand this in some small parameter. Writing things suggestively as

$$\tan 2\theta = \frac{-2\sqrt{2} \pi m_{\text{DM}}}{9\kappa f} \left[\frac{1}{1 - \frac{\pi m_{\text{DM}}}{9\kappa f}} \right]. \quad (\text{F.5})$$

we can expand this using $\frac{1}{1-x} = 1 + x + x^2 + \dots$ if $x := \left| \frac{\pi m_{\text{DM}}}{9\kappa f} \right| < 1$. Rearranging this (and using the fact that all of these variables are positive) we find that in order for this to be true

$$\pi m_{\text{DM}} < 9\kappa f \quad \Rightarrow \quad m_{\text{DM}} < \frac{9}{\pi} \kappa f. \quad (\text{F.6})$$

Given the fact that $\kappa \sim 1$ and we are only considering parameter space regions such that $m_{\text{DM}} \ll f$, we can see that this relation is indeed satisfied. This allows us to write

$$\tan 2\theta \approx \frac{-2\sqrt{2} \pi m_{\text{DM}}}{9\kappa f} \left[1 + \mathcal{O}\left(\frac{m_{\text{DM}}}{f}\right) \right] \quad (\text{F.7})$$

$$\approx -\frac{2\sqrt{2}\pi m_{\text{DM}}}{9\kappa f} + \mathcal{O}\left(\frac{m_{\text{DM}}^2}{f^2}\right). \quad (\text{F.8})$$

For the parameter space point $(m_{\text{DM}}, f) = (5 \text{ TeV}, 65 \text{ TeV})$ used in Eq. (5.32), $\tan 2\theta \approx 0.0759$.

Appendix G

Handling Derivative Field Interactions

In this work, there are two main lagrangian terms contributing to 2-to-2 pion interactions ($\Pi_i \Pi_j \rightarrow \Pi_c \Pi_d$). These are given in Eq. (5.19): $\Pi_i \Pi_j \partial^\mu \Pi_c \partial_\mu \Pi_d$ and $\Pi_i \Pi_j \Pi_c \Pi_d$. The goal is to write down the contributions to $i\mathcal{M}$ for all unique Feynman diagrams that come from these two lagrangian terms. In other words, we need to figure out all possible ways to connect the end-points labeled i, j, c, d using the terms in Eq. (5.19). Both terms connect these points with one vertex, so there are no propagators to consider.

For the $\Pi_i \Pi_j \Pi_c \Pi_d$ term this is straightforward. There is only one (non-redundant) diagram that can contribute to $\Pi_i \Pi_j \rightarrow \Pi_c \Pi_d$. And the corresponding contribution to $i\mathcal{M}$ is just related to the vertex factor ($i\mathcal{M} \supset i \frac{128\pi^3 m_{\text{DM}}}{3f} G_7$).

For the $\Pi_i \Pi_j \partial^\mu \Pi_c \partial_\mu \Pi_d$ term, things are a bit more complicated. Naively, there are $4! = 24$ diagrams one could draw. This comes from the fact that there are 4 possible slots in how you can place the labels on the Lagrangian terms (e.g. $\Pi_{(\text{slot } 1)} \Pi_{(\text{slot } 2)} \partial^\mu \Pi_{(\text{slot } 3)} \partial_\mu \Pi_{(\text{slot } 4)}$). You have 4 options (i, j, c, d) for slot 1, then 3 options for slot 2, and so on. This means the total number of combinations is $4 * 3 * 2 * 1 = 4!$. However some of these diagrams are equivalent. For example, in the term $\Pi_{(\text{slot } 1)} \Pi_{(\text{slot } 2)} \partial^\mu \Pi_{(\text{slot } 3)} \partial_\mu \Pi_{(\text{slot } 4)}$ we can freely interchange

(slot 1) \leftrightarrow (slot 2) and (slot 3) \leftrightarrow (slot 4) and still get the same result. Each of these exchanges represents a factor of 2 over-counting. Therefore, the number of unique diagrams we have is instead $4!/(2 * 2) = 6$. Explicitly, these correspond to the terms (1) $\Pi_i \Pi_j \partial^\mu \Pi_c \partial_\mu \Pi_d$, (2) $\Pi_d \Pi_j \partial^\mu \Pi_c \partial_\mu \Pi_i$, (3) $\Pi_c \Pi_d \partial^\mu \Pi_i \partial_\mu \Pi_j$, (4) $\Pi_i \Pi_c \partial^\mu \Pi_j \partial_\mu \Pi_d$, (5) $\Pi_i \Pi_d \partial^\mu \Pi_c \partial_\mu \Pi_j$, and (6) $\Pi_c \Pi_j \partial^\mu \Pi_i \partial_\mu \Pi_d$ which are also depicted in Fig. 5.3.¹ So all together, we should have 7 terms contributing to $i\mathcal{M}$, 6 of which arise from the first term in Eq. (5.19).

Since the first term of Eq. (5.19) contains derivatives of fields, the contributions to $i\mathcal{M}$ will be more complicated. This appendix discusses how to handle interaction terms in a lagrangian which contain derivatives (i.e $\Pi_i \Pi_j \partial^\mu \Pi_c \partial_\mu \Pi_d$) in more detail. Note that we are restricting ourselves to only consider the relevant case of scalar fields in the Lagrangian, (i.e. Π_a).

The derivatives introduce factors of momentum with signs corresponding to whether it is acting on an incoming or outgoing field (leg of the diagram). Since we are considering the physical process $\Pi_i \Pi_j \rightarrow \Pi_c \Pi_d$, when a derivative acts on Π_i or Π_j it is acting on an incoming leg and when a derivative acts on Π_c or Π_d it is acting on an outgoing leg. When a derivative acts on an leg, Π_a , which is incoming to a vertex (being “destroyed”), it will contribute a factor of $-ip_a$ to $i\mathcal{M}$. When a derivative acts on an leg, Π_a , which is outgoing from a vertex (being “created”), it will contribute a factor of $+ip_a$ to $i\mathcal{M}$. Each diagram will always have two derivative legs, so these factors of momentum are dotted together in each term. For example, for the term $\Pi_i \Pi_j \partial^\mu \Pi_c \partial_\mu \Pi_d$ the derivatives are both acting on outgoing legs so there will be a factor $+ip_c \cdot +ip_d = -p_c \cdot p_d$ multiplying the vertex factor $i \frac{4}{f^2} G_1$. So $i\mathcal{M} \supset -i \frac{4(p_c \cdot p_d)}{f^2} G_1$.

¹Note that, despite the shuffling of which fields the derivatives act on, all of these terms still contribute to the physical 2-to-2 process $\Pi_i \Pi_j \rightarrow \Pi_c \Pi_d$.

Appendix H

Neglecting Gauge Interactions

In this work we are interested in interactions which could deplete the abundance of pions containing Dark Matter (Π_{DP}). The dominant depletion interaction is the 2-to-2 interaction $\Pi_{\text{DP}}\Pi_{\text{DP}} \rightarrow \Pi_{\text{SM}}\Pi_{\text{SM}}$; this is what is considered explicitly in the work of Chapter 5.

However, one might wonder whether we should also include interactions which deplete into gauge bosons ($\Pi_{\text{DP}}\Pi_{\text{DP}} \rightarrow GG$). This turns out to be negligible compared to $\Pi_{\text{DP}}\Pi_{\text{DP}} \rightarrow \Pi_{\text{SM}}\Pi_{\text{SM}}$ interactions. This appendix gives a sketch of how we can see that this process should be negligible but also outlines how one would go about finding the Lagrangian terms leading to such interactions.

As shown in Chapter 5, we can estimate the cross-section of the process $\text{III} \rightarrow \text{III}$ via the following

$$\sigma_{\text{III} \rightarrow \text{III}} = \mathcal{C}_{\Pi} \frac{m_{\text{DM}}}{f^3} \tag{H.1}$$

where \mathcal{C}_{Π} can be estimated from the numerical results but is ~ 0.8 for $(m_{\text{DM}}, f) = (5 \text{ TeV}, 65 \text{ TeV})$ in BP1 for $N_{\text{gen}} = 1$.

To get a similar estimate for $\text{III} \rightarrow GG$, we can turn to the literature on analogous interactions in QCD. In particular, a calculation of stops to gluons in Ref. [163] indicates that the analogous version in this case should be

$$\sigma_{\text{III} \rightarrow GG} = \tilde{\mathcal{C}}_G \frac{g_s^4}{\pi m_{\text{II}}^2} = \mathcal{C}_G \frac{g_s^4}{f m_{\text{DM}}} \quad (\text{H.2})$$

where in the last step we used the fact that for the $\text{II}_{\text{DP}}\text{II}_{\text{DP}} \rightarrow \text{II}_{\text{SM}}\text{II}_{\text{SM}}$ reaction of interest $m_{\text{II}}^2 = 64\pi^3 f m_{\text{DM}}$, thus $\mathcal{C}_G = \tilde{\mathcal{C}}_G / (64\pi^4)$. The analogy of $\tilde{\mathcal{C}}_G$ in Ref. [163] is 7/216. Translating it to the case where we have SU(2) instead of SU(3) should reduce this by approximately half due to the differences in gauge factors. Therefore, we estimate $\tilde{\mathcal{C}}_G = 7/432$.

In order to safely neglect $\text{III} \rightarrow GG$ interactions we want

$$\frac{\sigma_{\text{III} \rightarrow GG}}{\sigma_{\text{III} \rightarrow \text{III}}} \ll 1 \quad (\text{H.3})$$

$$\Rightarrow \left(\frac{\mathcal{C}_G g_s^4}{f m_{\text{DM}}} \right) \left(\frac{f^3}{\mathcal{C}_{\text{II}} m_{\text{DM}}} \right) \ll 1 \quad (\text{H.4})$$

$$\Rightarrow \frac{g_s^4 \mathcal{C}_G}{\mathcal{C}_{\text{II}}} \frac{f^2}{m_{\text{DM}}^2} \ll 1 \quad (\text{H.5})$$

Let us now estimate whether this condition is satisfied for our standard point of interest, $(m_{\text{DM}}, f) = (5 \text{ TeV}, 65 \text{ TeV})$, under BP1 ($g_s = 0.8$),

$$g_s^4 \frac{\mathcal{C}_G}{\mathcal{C}_{\text{II}}} \frac{f^2}{m_{\text{DM}}^2} = (0.8)^4 \left(\frac{7}{432 * 64\pi^4} \right) \left(\frac{1}{0.8} \right) \left(\frac{65}{5} \right)^2 \approx 2.25 \times 10^{-4}, \quad (\text{H.6})$$

which is much less than 1, and thus can be safely neglected. Fig. H.1 shows this ratio directly for other points in (m_{DM}, f) parameter space using the corresponding numerical values of \mathcal{C}_{II} for each point. All values of this ratio are less than or equal to $\mathcal{O}(10^{-3})$ in the parameter space regions of interest.

If one wanted to calculate $\sigma_{\text{III} \rightarrow GG}$ in this theory directly, the process would proceed as

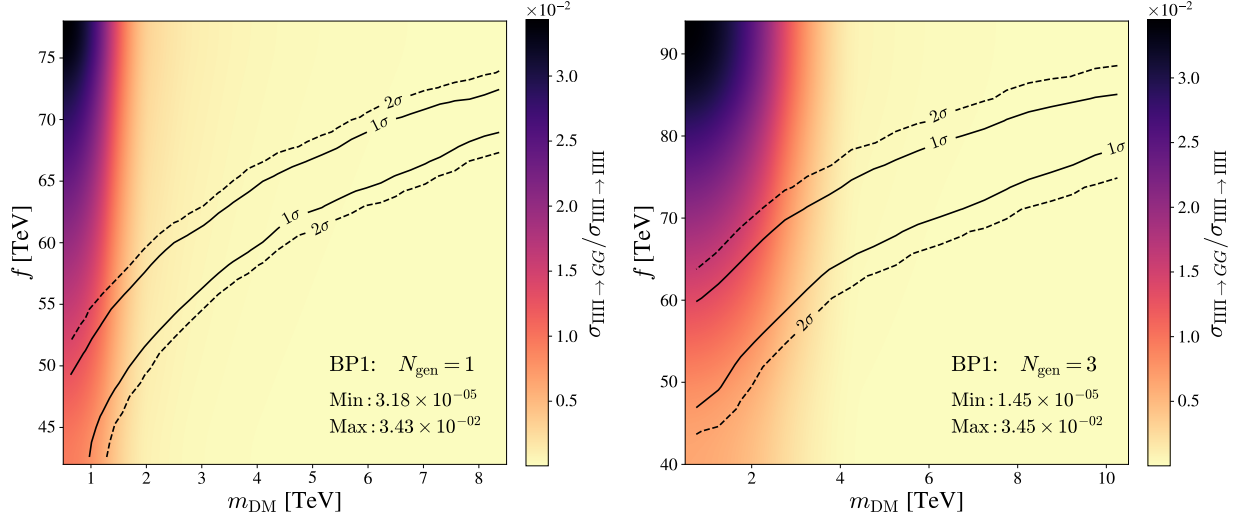


Figure H.1: **Depiction of the ratio of the cross section for the $\text{IIII} \rightarrow GG$ reaction as compared to our main $\text{IIII} \rightarrow \text{IIII}$ interaction.** We can see that for the majority of our valid parameter space, $\sigma_{\text{IIII} \rightarrow GG}$ is a factor $\mathcal{O}(10^{-5})$ to $\mathcal{O}(10^{-3})$ smaller than $\sigma_{\text{IIII} \rightarrow \text{IIII}}$. For low masses, this factor increases to being around $\mathcal{O}(10^{-2})$, which is still much less than 1. We can therefore safely neglect these contributions in our results.

follows. Note that we will not actually perform this calculation, as the back-of-the-envelope result is sufficiently convincing, but, for completeness, we sketch how it would be done. Neglecting gauge interactions (at the Lagrangian level) amounts to the approximation that $D_\mu \approx \partial_\mu$. If we want to include gauge interactions explicitly, this relationship will change. Specifically for this model, it will become

$$D_\mu \Sigma = \partial_\mu \Sigma - ig_s \sum_{a=1}^3 G_\mu^a [L^a \Sigma + \Sigma (L^a)^T] - ie_Q A'_\mu [Q \Sigma + \Sigma Q] \quad (\text{H.7})$$

Therefore, the kinetic term of the IR Lagrangian will now be

$$\mathcal{L}_{\text{IR}} \supset \frac{f^2}{4} \text{Tr}[D^\mu \Sigma^\dagger D_\mu \Sigma] \quad (\text{H.8})$$

Plugging in Eq. (H.7) into this and expanding will wind up looking like the following

$$\mathcal{L}_{\text{IR}} \supset \frac{f^2}{4} \text{Tr} [(\partial_\mu \Sigma^\dagger)(\partial_\mu \Sigma)] + \mathcal{L}_{(\text{gauge interactions})}. \quad (\text{H.9})$$

The first term is what we have already considered explicitly in Chapter 5, while $\mathcal{L}_{(\text{gauge interactions})}$ will contain terms that contribute to $\text{III} \rightarrow GG$ interactions. From here, the goal would be to find what terms in $\mathcal{L}_{(\text{gauge interactions})}$ contribute to $\text{III} \rightarrow GG$ interactions and how this cross-section, $\sigma_{\text{III} \rightarrow GG}$, relates to $\sigma_{\text{III} \rightarrow \text{III}}$ analytically.