THE UNIVERSITY OF CHICAGO

ANALYTIC AND MACHINE LEARNING METHODS FOR CONTROLLING NONLINEARITIES IN PARTICLE ACCELERATORS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHYSICS

BY
LIPI GUPTA

CHICAGO, ILLINOIS
AUGUST 2021

For Mom, Dad, Didi, and Ranna.

Had I the heavens' embroidered cloths,
Enwrought with golden and silver light,
The blue and the dim and the dark cloths
Of night and light and the half light,
I would spread the cloths under your feet:
But I, being poor, have only my dreams;
I have spread my dreams under your feet;
Tread softly because you tread on my dreams.

- W. B. Yeats

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to my committee for their guidance, excellent questions, and interest in my work. I am honored to have had the opportunity to work with you.

I am very grateful to my wonderful mentors: Young-Kee Kim, Sergei Nagaitsev, Nicole Neveu, Chris Mayes, Auralee Edelen, Adi Hanuka, Stas Baturin, and Ryan Roussel. Without their guidance, help, and wisdom this thesis would not have been possible. Throughout the many obstacles I faced, both academic and personal, knowing that I had the support of these amazing mentors helped me persevere. In particular, thank you to Chris for handing me a copy of Wille's book and the BMAD manual in 2013, which launched me into the challenging and beautiful world of charged particle beam optics. Thank you to Auralee for helping me secure DOE funding and for teaching me about the world of ML for science. Thank you to Stas for being honest and tough with me, but also hilarious and kind. And a big thank you to Nicole and Adi for their accelerator expertise, moral support during my many stress-induced crises, and for lending their ears and hearts when I vented my frustrations or celebrated my victories.

Thank you to my cohort members Karthik Ramanathan, Dani Scheff, Evan Angelico, and many others for their friendship, homework help, and moral support through the years.

Thank you to Matt Gordon and Joshua Paul for always being my greatest allies.

Thank you to the cohort that adopted me as their own and invited me to paint nights, movie nights, karaoke, and all manner of fun activities. You all are such bright, intelligent, kind people and incredible scientists.

Thank you to the Women and Gender Minorities in Physics community for all of the lovely seminars, meals, and conversations. It is a joy to find one's community in graduate school, and WaGmIP was mine. To my closest friends Bruce Winkelman, Lesya Horyn, Ginna Roach, Emily Smith, Anne Driscoll, and Jacob Rueben: I love you all so much. Each of you has supported me, tolerated me, lifted me up, and brought me back down to reality so many times. It has been a privilege to call you all my friends, and I hope I can be there

for each of you the same way you always were for me. Thank you to Ryan Thomas for being so caring and patient with me. Without your love and support, I am sure I would have stumbled many times.

Finally, this Ph. D. is dedicated to my mom, my dad, my sister, and Ranna. Without them, none of my achievements thus far in life would have been possible. My mom's strength is the bedrock upon which my whole life is built. My dad's love of science is the reason I pursued my love of accelerator physics. My sister's passion and drive are the reason I continued when I thought I could not. And Ranna is the angel who took care of us during the hardest time in our lives. You four are my whole world, and I love you all more than words can express. Thank you for everything you have done for me, and for your continuing love and support.

ABSTRACT

Particle accelerators are versatile machines which make both particle physics research and probing nature at microscopic scales possible. At advanced light source facilities, high energy charged particles are used to produce energetic photons. The light generated can resolve atomic structures at Angstrom length scales and observing processes occurring at femtosecond time scales. To achieve this, advancements in methods for designing and controlling nonlinear phenomena in the accelerator is crucial. In machines with periodic structures such as storage rings, nonlinear effects can compound and result in beam loss limiting the radiation power emitted for experimental use. In this thesis, a method for understanding how resonance elimination can be done without relying on extensive numerical optimization, as well as how to design such lattices is developed. In cases when numerical methods are necessary, such as for machines which are already built, machine-learning methods can be used to accurately model the machine behavior. These models can then be used for model-based operation and control of the accelerator. A novel approach to training machine-learning measurement-based models was developed and demonstrated, furthering the goal of improving methods for modelling and controlling accelerators for scientific purposes.

CHAPTER 1

PARTICLE ACCELERATORS IN THE WORLD

While many of the best known particle accelerators around the world are large machines spanning miles through underground tunnels, particle accelerators are all around us. Any system that can accelerate a charged particle is a particle accelerator. Therefore, particle accelerators are ubiquitous and universal. A particle accelerator can be any series of components within which a charged particle can be accelerated. The earliest accelerator was the simple cathode ray tube, in which a heater would stimulate electron emission from a cathode, and the potential difference between the cathode and a downstream anode would accelerate the electrons. The use of magnetic structures allowed the electron beam to be deflected as needed. These are the same principles were developed and applied in new and creative ways, thus advancing particle acceleration technology from its humble beginnings.

Accelerators used for the purpose of creating bright radiation are often called light sources. These accelerators specifically accelerate charged particles to create radiation, which can be used for a variety of scientific experimental purposes. By grouping particles into packets, called bunches, these particles can be made to radiate together to increase the radiated power. However, some of the fundamental challenges faced by accelerator physicists are: how can we produce high charge bunches? How can we better harness radiated power to create brighter photon packets, or higher energy photons? Particles with the same charge polarity repel one another, but at high energies near the speed of light, this repulsion becomes minimal. Thus packing particles into bunches and accelerating them such that they maintain certain qualities requires significant research. For example, radiation power is maximized by tightly focusing the charged particle bunches. However, nature does not allow us to simultaneously focus particles in both transverse directions unless energy is added or removed from the particle bunch. But what if sacrificing energy, by adding or removing it, is not an option, as this can change or limit the radiated photon energies. These are just some of the fundamental physical limitations that are addressed by accelerator physicists. We do this by

finding novel, creative, and strategic applications of the mathematics and physics of classical particle mechanics and electrodynamics. Thus, beyond the fact that particle accelerators are incredible feats of engineering, they also present interesting physics and mathematical challenges.

Only after it is clear that demanded beam dynamics are allowed by nature, can the engineering challenge of building and controlling these components begin. In this thesis, the challenge of designing a series of machine components which obey mathematical constraints, yet combine together to eliminate a highly mitigating sextupole magnet coupling is explored. The mathematics of nonlinear charged particle beam dynamics is an area of active research, with dedicated facilities such as the IOTA project at Fermi National Accelerator Laboratory, designed to investigate how to apply mathematical theory in unique ways to improve particle accelerators(1).

1.1 A Brief History of Particle Accelerators

The history of particles accelerators is extensive, but the impetus for each advancement was the need for creating charged particle beams of higher energies, for collision and target experiments. High energy particles are necessary for collision and target experiments, in order to achieve particle production for studying fundamental physics. To create a particle of mass m, one must produce enough energy to satisfy the relation $E = mc^2$. Collision experiments allow that energy to produce other particles, as long as momentum is conserved. Thus, colliding higher energy beams can result in the production of increasingly massive particles. This fact has long been a motivating factor in improving particle sources and acceleration methods, in order to advance nuclear physics and study fundamental particle physics.

Similarly, after the discovery of x-rays by Röntgen in 1895, there was also increased interest in producing this high energy radiation in controlled methods. In 1906, J. J. Thompson published his findings on the emission of radiation from accelerating particles upon colliding

with other atoms, followed by G.A. Schott's formulation of a theory of synchrotron radiation in 1907. During this time, Rutherford's landmark gold-foil experiment also propelled many others to pursue methods for creating higher energy charged particle beams, to continue to probe molecules and discover atomic structures. Several years of experiments between 1907 and 1932, including Van der Graaf's first high voltage generator, eventually lead to Lawrence and Livingston accelerating protons to 1.2MeV in a cyclotron (2; 3). This achievement was one of many which lead directly to the modern accelerators used today.

As these experiments continued, by the end of the 1940's, there was sufficient interest and knowledge to commission and build particle accelerators. It was noticed that emitting particles with the precise momenta needed to retain particles in orbit (in circular machine geometries) was challenging. Any particle with slightly higher or lower momentum would end up on a spiraling trajectory, and eventually be lost due to collision with the walls of the vacuum chamber within the machine. In 1950, N. Christofilos formulated methods for strong focusing, called alternate gradient focusing, which could aid in preventing particle loss by focusing the beam. With this advancement and the work of E. Courant, who also worked on the theory of strong focusing, in 1954 R. R. Wilson successfully demonstrated an alternating-gradient electron synchrotron at Cornell University, operating at 1.1GeV. This method now allowed the transverse beam cross section to be significantly smaller, meaning the aperture within the magnetic structure could be reduced in size. This led directly to the ability to produce high-field magnets, pushing particle accelerators into the GeV energy frontier. However, as is the nature of scientific advancement, while strong focusing made it possible to create higher energy synchrotrons without significant beam loss, this method introduced further challenges. Strong focusing can correct geometric aberrations, such as the spiraling trajectories, but introduce chromatic aberration. These mitigating challenges are still in need of creative solutions, to continue to push circular particle accelerators into higher energy and bunch charge regimes (2).

In the pursuit of creating higher energy particles by designing new particle acceleration

schemes, another phenomena was discovered. This phenomena was synchrotron radiation, first discovered by D. Ivanenko and I. Ya. Pomeranchuk, and independently by J. Schwinger in 1944. They noted that there would be an energy limit in circular electron accelerators due to radiative losses, which was confirmed experimentally by 1945 by J. P. Blewett. Based on Lienard's previous work, it was known that a relativistic charged particle would radiate power by the following relation during transverse acceleration:

$$P = \frac{e^2 c}{6\pi\epsilon_0} \frac{1}{(m_0 c^2)^4} \frac{E^4}{R^2} \tag{1.1}$$

where e is the particle charge, ϵ_0 is the permittivity of free space, c is the speed of light, m_0 is the mass of the particle, E is the energy of the particle, and R is bending radius of the particle trajectory (3). It was clear that the power is damped by the mass of the particle. For example, a proton is 1800 times as massive as an electron, therefore the power radiated from accelerating a proton to the same energy as an electron would be significantly smaller. Thus, synchrotron radiation was primarily noticed at electron facilities. Furthermore, the radiated power increases as the fourth power of the particle energy. It was becoming clear to experimentalists that accelerating electrons to higher beam energies, would produce powerful radiation. Techniques for refining the creation and use of this radiation are an integral part of current accelerator physics research.

1.2 X-rays for Science

In the same way that high energy charged particles are used to probe the fundamental particles of nature, high energy photons are needed in order to study molecular structures. High energy particles accelerators, where electrons are made to radiate, can be a source for the necessary photons.

In order to resolve a microscopic structure, the wavelength of the light used to probe the structure must be smaller than the structure. To resolve molecular structures, sub-Angstrom

photons are needed. Photons with energies ranging from 100 to 1000 eV, classified as X-rays, are therefore ideal for probing molecular and other microscopic samples (3). Researchers from fields ranging from biology, material science, to medicine and history can use X-rays to learn about samples including geometries and other properties.

There are many examples of how powerful X-ray facilities have helped provide the radiation to improve our understanding of natural phenomena. One such facility is the Stanford Linear Accelerator Center (SLAC) National Accelerator Laboratory Linac Coherent Light Source (LCLS). At present, the LCLS facilitates over a thousand scientific experiments per year, and many of these have fundamentally improved our understanding of important processes, such as photosynthesis (4) and electron-phonon interactions that could aid the design of new materials (5; 6).

1.3 Accelerator Geometries

There are several particle accelerators around the world, which are used for a variety of applications. The application often enforces the kind of accelerator that is used. For high-power X-ray sources (like those discussed in this thesis), the charged particle of choice is the electron. This is due to the many factors, the most important of which is the relationship between radiated power and the mass of the particle emitting, discussed and shown in Eqn. 1.1. From this relationship it is clear that the radiated power is limited by the mass of the charged particle. Therefore, electrons are the charged particle of choice to maximize the radiated power at X-ray sources. Further, methods for effectively emitting electrons from cathode materials to produce high charge bunches was also known. Thermal cathodes, like that in the CRT, cause electron emission via heating of the cathode material, whereas photocathodes use the photoelectric effect. Thus, there are many options for producing electrons for acceleration, which can then produce powerful radiation.

1.3.1 Circular Machines

Many early particle accelerators relied on circular geometries which allow for acceleration over repeated passes. This allowed for achieving high energy particle bunches, while minimizing physical space requirements. Early machines, called cyclotrons, consisted of permanent magnets for bending and focusing the electrons. The next generation of circular accelerators, called synchrotrons, used electromagnetic structures which could successfully control and focus higher energy particle bunches. The current generation of circular machines are related to early synchrotrons, but through technological advances, can achieve higher energies.

Circular accelerators produce a fan of synchrotron radiation as the beam circulates through the machine. This is because the radiation power is largely concentrated tangential to the direction of motion. As the beam circulates, the radiation is emitted, and continues to emit as the beam moves through a small arc. This creates a fan of radiation in all directions. Furthermore, this radiation is not emitted with a single frequency, due to the fact that the beam of particles is not mono-energetic. Therefore the synchrotron radiation spectrum is broad. This spectrum is distributed around a characteristic frequency which is determined primarily by the frequency of the circulating particles. In terms of the bending radius and beam energy in terms of the rest mass,$\gamma = E/m_0c^2$ the frequency is:

$$\omega_{char} = \frac{3\pi c\gamma^3}{2R}. \tag{1.2}$$

While the total power P radiated from an electron bunch of N particles scales linearly with the current circulating in the machine:

$$P = \frac{e\gamma^4}{3\epsilon_0 R} I_{beam}, \tag{1.3}$$

the power at a given photon energy is limited due to the broad spectrum radiation that is produced (3).

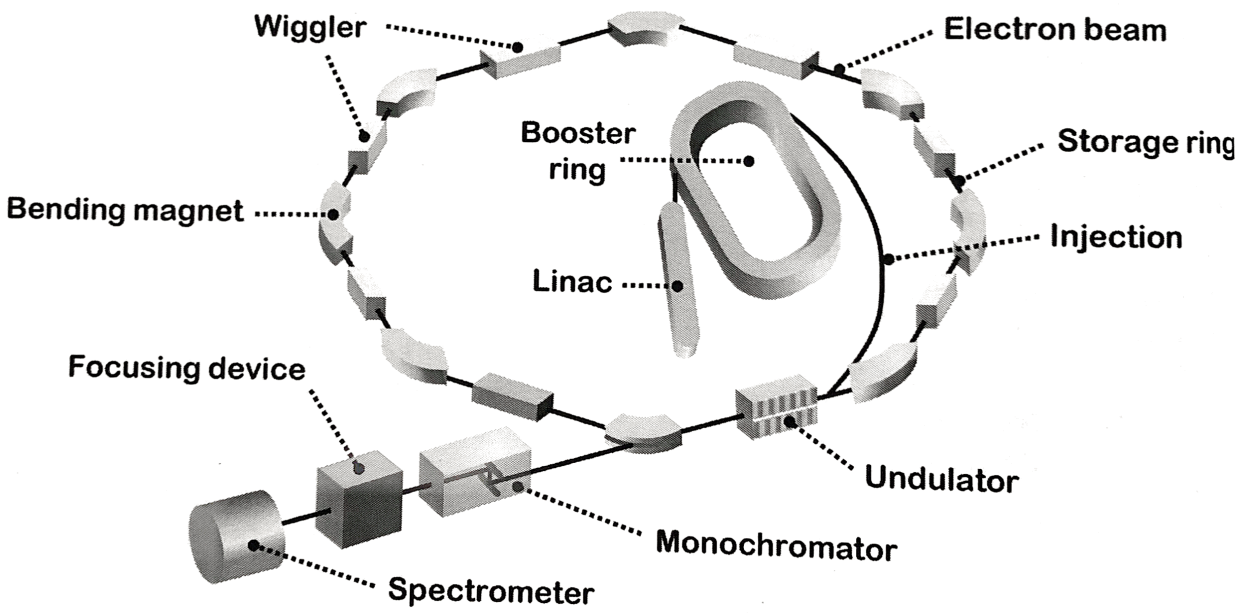


Figure 1.1: This diagram shows how particles are injected into a booster, which accelerates the particles to the desired energy and then passes the bunches into a storage ring. This storage ring then maintains the beam quality by restoring radiated power, and confines the particles as they circulate through accelerating structures such as undulators. The radiation produced in the wigglers or undulators can then be harnessed and used by experimenters. Figure from (2).

To build coherence and thus collapse the photon spectrum around a chosen frequency, special magnets called wigglers or undulators can be inserted into a circular machine. These magnets consist of a straight portion of several short magnets, with alternating polarity. As the name suggests, this causes the electron bunch to wiggle within the magnet, while continuing on a trajectory through it. These wiggles have an incredibly small radius of curvature, which allows the small cone of radiation emitted to build coherence and produce radiation which is highly peaked around a given frequency.

With this technology, machines which were initially built as colliders can be upgraded to include wigglers and undulators. Tangential to these portions are often experimental stations where the powerful radiation can be harnessed by an experimenter. This is shown in Fig. 1.1, where a small linear accelerator injects particles into a booster ring, or synchrotron. The particles are then sent to the storage ring where the created radiation can be harnessed.

Currently, several premier synchroton facilities provide powerful X-rays to experimenters. The Advanced Photon Source at Argonne National Lab in Lemont, IL is one such facility. The APS uses a 7 GeV electron beam to provide 19.5keV photons to more than 30 beamlines (7). This has resulted in over 30,000 published results (8).

1.3.2 Linear Machines

While circular machines can use repeated acceleration structures to achieve high energy particle bunches, linear acceleration geometries are also advantageous in other ways. Linear machines are valuable as both colliders, as well as light sources. The primary technology used when making linear light sources, is repeated undulator magnets. This configuration, called a free electron laser (FEL), was first developed by John Madey in 1971. This technology allows for the creation of high energy photons with very narrow frequency spectra, thus earning the label "laser." In particular, the phenomena of microbunching, which provides a density modulation within the bunch structure, allows for coherence to build such that the power at the peak frequency scales as N^2 instead of N as seen in Eq. 1.3.

The Linac Coherent Light Source (LCLS) is one of the premier facilities where this technology was demonstrated and improved upon. The LCLS is one of the premier linear light sources in the world, and the upgraded LCLS-II will include superconducting technology and several other advancements to provide extremely powerful X-rays. With a pulse repetition rate of 1MHz, increased from 120 Hz, and a larger range of photon energies, the LCLS-II will allow for measurements of structures on the atomic scale (9).

1.4 Challenges in Designing a Particle Accelerator

While the first synchrotrons were built for particle physics experiments, synchrotron radiation was used parasitically. These machines are considered the first generation of synchrotron light sources. The second generation consists of machines which were dedicated radiation

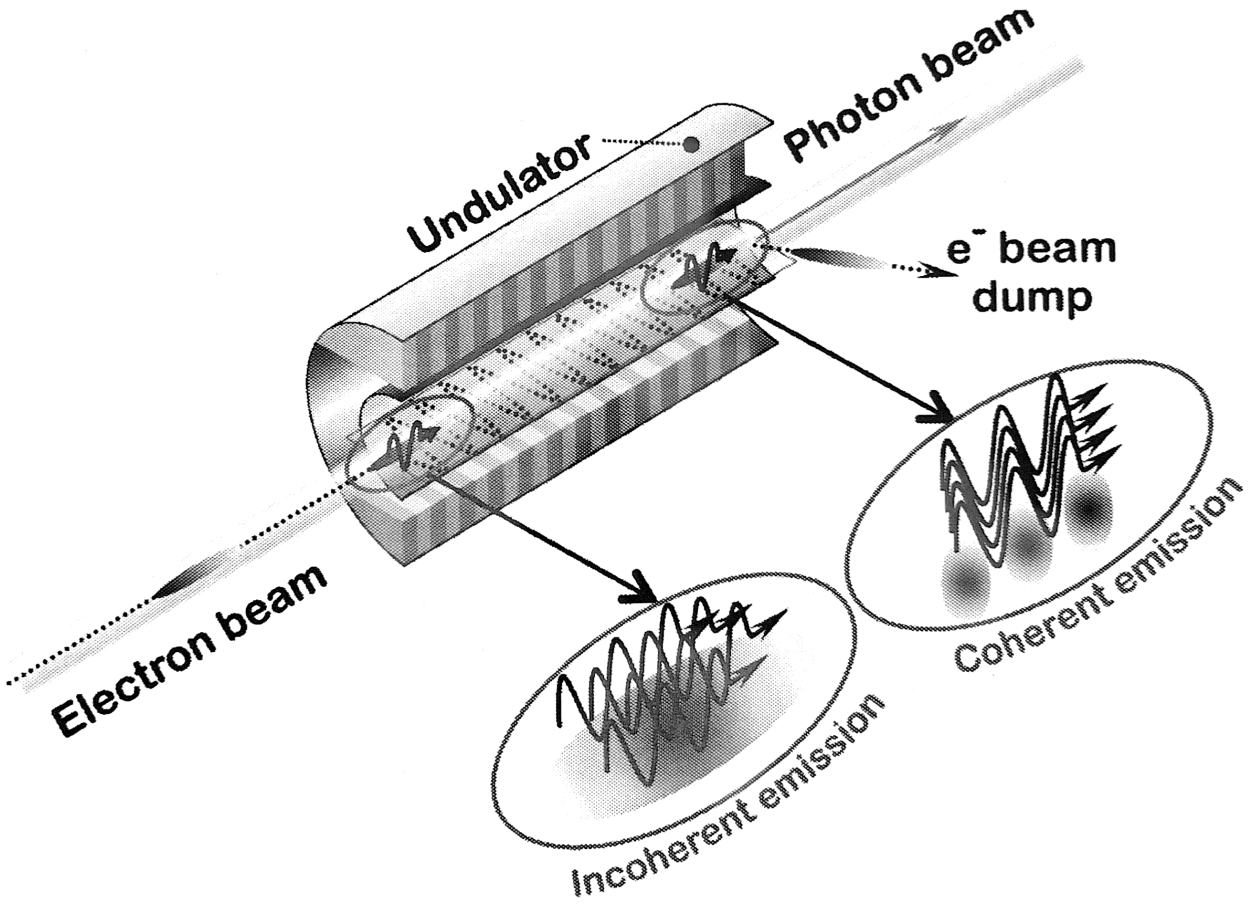


Figure 1.2: A simple demonstration of the microbunching phenomena which can take place in an undulator, which is the building block of a free electron laser (2).

facilities. Currently, we are seeing 3rd and 4th generation machines, which are improvements on the dedicated facilities, or facilities which have reached new, higher photon energies and photon fluxes. However, some of these machines are still in their nascence, thus solutions and novel approaches for improving machine designs and other technical challenges are topics of current research.

Machine parameters of interest include the electron energy and repetition rate (which is the number of bunches available to radiate per unit time). Another important measure is the beam emittance, which quantifies how compact the beam is in phase space (position and momentum space). For light sources, it is ideal to minimize the emittance as much as possible, as this improves the brightness of the radiation, which is a measure of how many photons are emitted by the accelerated particles. Because experimenters rely on bright light sources, the emittance and associated beam brightness are often the most important quantitative measures.

Related to brightness is another quantity called brilliance, B. This quantity describes the photon flux within a 0.1% bandwidth of a useful photon energy and normalized by beam current, per unit phase space area:

$$B = \frac{F}{4\pi^2 \epsilon_{\text{trans}} \epsilon_{\text{ang}}}, \tag{1.4}$$

where,

$$F = \frac{\text{photons}}{0.1\%\text{BW A s}}, \tag{1.5}$$

and ϵ_{trans} and ϵ_{ang} are the transverse and angular emittances respectively, with the factor of $4\pi^2$ from integrating the solid angle. It is clear from these definitions that increasing the brilliance can be done in two main ways: increasing the photon flux within the useable bandwidth, or decreasing the beam emittances. Advancements in both approaches lead to current accelerators and the incredible brilliance that can be achieved, shown in Fig. 1.3. Clearly, 4th generation light sources sacrifice pulse repetition rate, to provide extremely

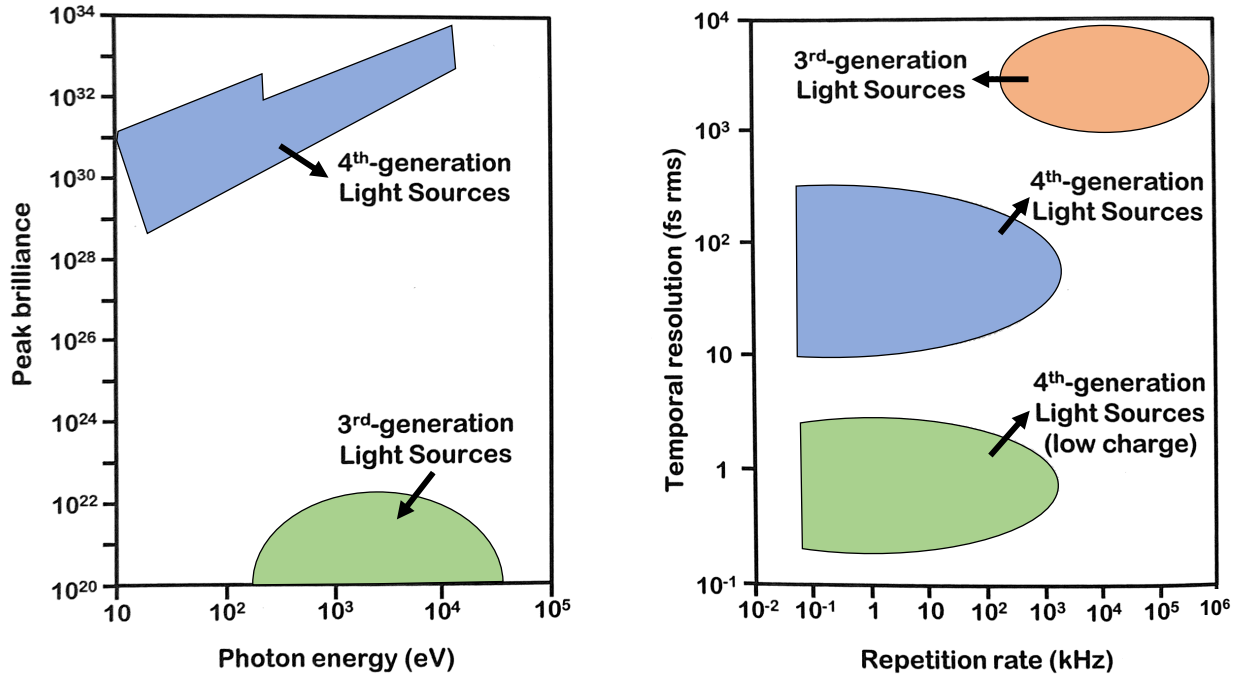


Figure 1.3: A comparison of different generation light sources and the photon brilliance that can be achieved, for a given photon energy or repetition rate. Figure from (2).

bright pulses.

Currently, light sources are designed to reach specific beam parameters, such as high electron bunch charges, low beam emittances, or high repetition rates. In some cases, an existing machine is upgraded to achieve new, previously unavailable beam parameters. In order to design new machines or upgrade existing ones, the design process requires both analytic, semi-analytic, and simulated calculations. These calculations can help explore novel machine concepts and their capabilities, or verify that a known improvement will achieve the desired beam.

1.5 Addressing Design and Control Challenges at Light Sources

Particles accelerators are already vastly improved since the mid 1900s when they were first developed. However, many challenges remain unaddressed. As the push for higher photon energies and brighter beams rages on, the challenge of designing and controlling these pow-

erful machines arises. In storage rings, methods for containing and stabilizing high energy, high current beams must be addressed. Similarly, as the electron energies are increased, radiative losses lead to emittance degradation, which diminishes the radiation quality (10). Further, because storage rings must preserve a beam for long periods of time, stability issues arise if nonlinearities in the magnetic structure are not controlled or mitigated (1). These are just a few of the challenges that face the synchrotron and storage ring accelerator community today.

While nonlinear effects due to periodic beam trajectories are not a concern at linear machines, control tasks such as adjusting upstream machine parameters to achieve a desired downstream output is very challenging. At linear and circular machine facilities, operators rely on empirical control methods or slow parameter scans to diagnose machine performance or adjust tunable parameters. This is entirely because nonlinear effects compound and misalignment, noise, or complex machine interactions can result in unpredictable beam behavior.

Thus, it is prudent to address methods for designing machines such that nonlinear effects are managed or mitigated as much as possible. If this is not possible, because the machine is already commissioned and built to specification, numerical and machine learning methods can be employed to provide models, which could aid in machine control.

1.5.1 Analytic and Semi-Analytic Approaches

One way to address the challenges mentioned previously is by designing a machine to minimize the effects of the mitigating factors by applying known physical principles from classical mechanics and electrodynamics. These methods are considered analytic or semi-analytic, as they may rely on a combination of solving equations and numerical integration. There is a spectrum of methods from differential equation solvers, to beam dynamics simulations which use a combination of close-form physics equations and numerical approximations to calculate beam parameters given component parameters and geometries. Such methods are powerful because they can also illuminate or demonstrate physical phenomena which have

not previously been seen.

In this thesis, analytic methods were used to understand how to improve beam stability in a storage rings. Chapter 3 will build the necessary mathematical background needed to understand how linear and nonlinear beam optics. In Chapters 4 and 5, new mathematical approaches to controlling nonlinearities which could be physically realized in an accelerator are introduced.

1.5.2 Numerical and Machine Learning-based Approaches

A primary challenge for studying beam dynamics, and designing components to achieve new beam parameters at particle accelerator facilities, is simulating the accelerator and associated phenomena. Simulation software are often developed due to the need for high-fidelity numerical calculations. However, these simulations can be computationally expensive, which can be prohibitive during the design stage as well as for online use in accelerators. In many cases, a single simulation can require minutes to hours to complete; this is a time-scale which is too long for interactive online use in the accelerator control room. Further, machine-time, for the sole purpose of characterizing the machine is rare, and often certain measurements cannot be done simultaneously or taken without disrupting operation. Therefore simulations are vitally important in design and experiment planning.

However, simulations introduce some challenges as well. In many cases, there are large discrepancies between the simulated phenomena and measurements at the machine due to simplifications or limitations within the simulations. Even with calibration, the simulation results may be quite different than reality. In addition, the physics simulation often predicts outputs based on the designed machine geometry and settings, and not the as-built machine. Further, some simulations are very time consuming due to the number of operations and the complexity of operations that must be completed, regardless of the computational power available. When simulations are used to aid in design work or to study physical phenomena, waiting for results may not be an issue, but this is the primary reason that simulations are

not used while operating the machine. If simulations were faster, operators would be able to leverage the physics knowledge of simulations by running an optimizer over the simulation to find desired machine settings. This is not currently possible because each simulation can take minutes to complete. Without real-time feedback of beam parameters given rapidly changing machine parameters, simulations are not feasible for us in model-based control in particle accelerators (11). Thus, there is interest in expanding the numerical methods that are used in particle accelerator physics to aid in both online and offline use (12).

Machine learning methods are a broad class of numerical and statistical methods, which can be applied to a variety of applications. Because these methods are easily adapted to the types of regression tasks that could be used to model accelerator physics, there has been a recent push to apply these methods where appropriate. Many applications of machine-learning methods have been demonstrated with initial success (11), (13), (14). However, methods for creating fast-executing surrogate models based on machine data are limited. As mentioned, because machine misalignment can limit how well models are able to predict machine behavior, using machine-learning methods trained using machine data is an important research avenue for accelerator physics.

CHAPTER 2

CHARGED PARTICLE BEAM DYNAMICS

A particle accelerator consists of many magnets and non-magnetic elements. Different elements in the accelerator allow for the transport of many bunches of electrons through the accelerator structure, and the sequence of these elements is called the lattice. Drift spaces, which are non-magnetic segments of the lattice, are used to transport the beam between magnetic elements. The primary element used in circular machines is the dipole magnet which consists of a single pair for north and south poles, and are used to provide a constant magnetic field that bends the beam radially. Quadrupole magnets consist of two pairs of north and south poles, and are used for focusing and defocusing the beam particles, in order to maintain a compact bunch of electrons. The magnetic fields in these elements have linear spatial dependence, and are therefore called linear elements. The next type of accelerator magnet is the sextupole magnet, which consists of three pairs of north and south poles, and is used for off-energy particle corrections. Sextupole magnets introduce nonlinearity due to the quadratic spatial dependence in the magnetic field, and are therefore referred to as non-linear elements. A schematic of these elements is shown in Fig. 2.1.

Table 2.1: Accelerator magnets have magnetic fields which can depend on the spatial coordinate within the magnet, relative to the center of the magnet. Based on this dependence, some magnets contribute to the linear beam optics, such as dipole and quadrupole magnets, and some contribute to the non-linear optics, such as sextupole magnets. See (3) for details.

Magnetic Field	Magnet Type
$B_z = \frac{-\mu_0 I_0}{2a}$	dipole
$B_z = \frac{-\mu_0 I_0}{2a} x$	quadrupole
$B_z = \frac{-\mu_0 I_0}{2a} x^2$	sextupole

While the basic principles of electrostatic and electromagnetic interactions allow for a qualitative understanding of particle acceleration, the mathematical expression of these relationships illuminates the mechanisms and their limitations. The derivation of the equation of motion is available in many accelerator physics texts including (3). These equations can

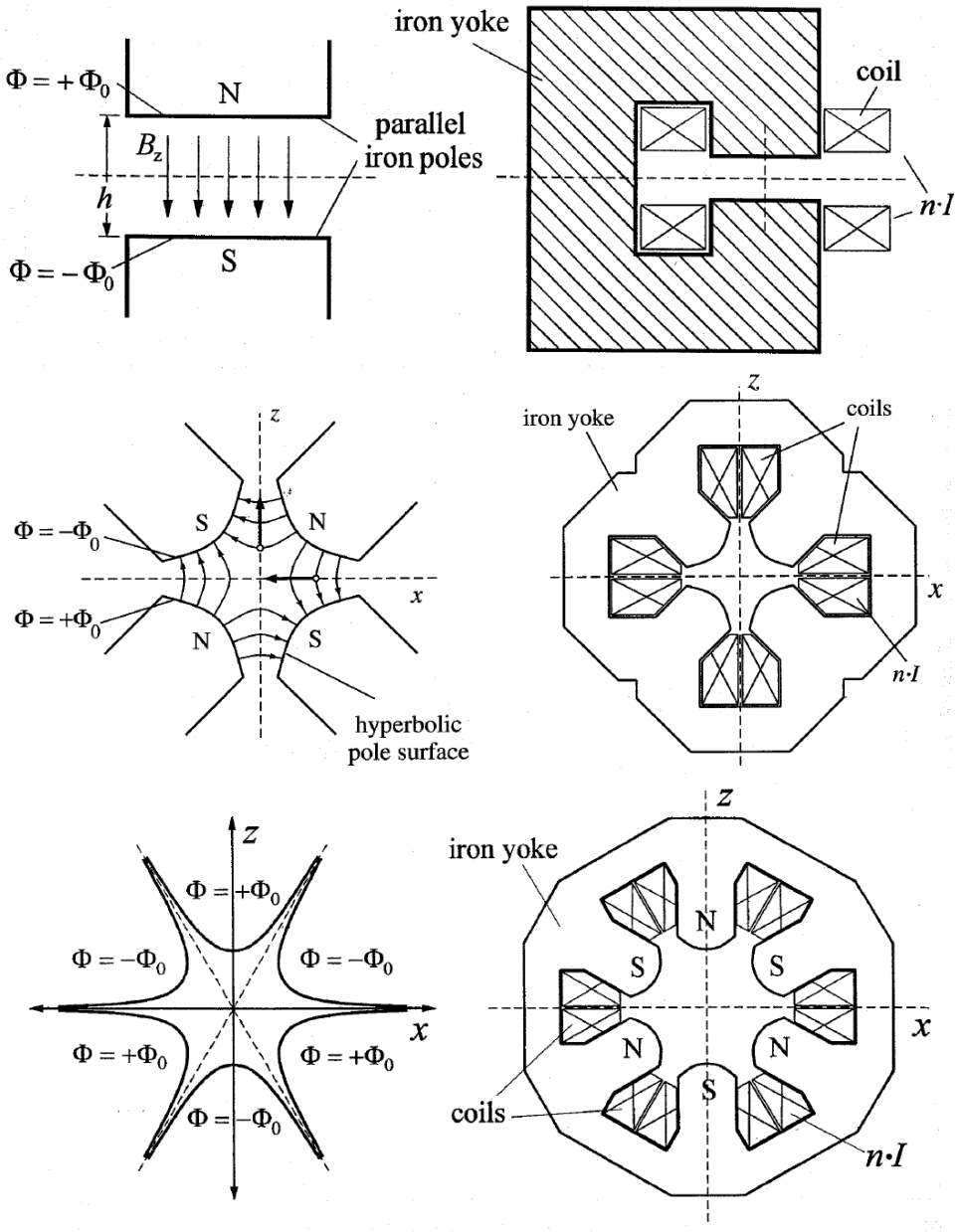


Figure 2.1: Schematic drawings showing the structure and pole distributions for a dipole, quadrupole, and sextupole magnet (3).

be written in matrix form, which make describing the linear evolution of a particle through the lattice very easy.

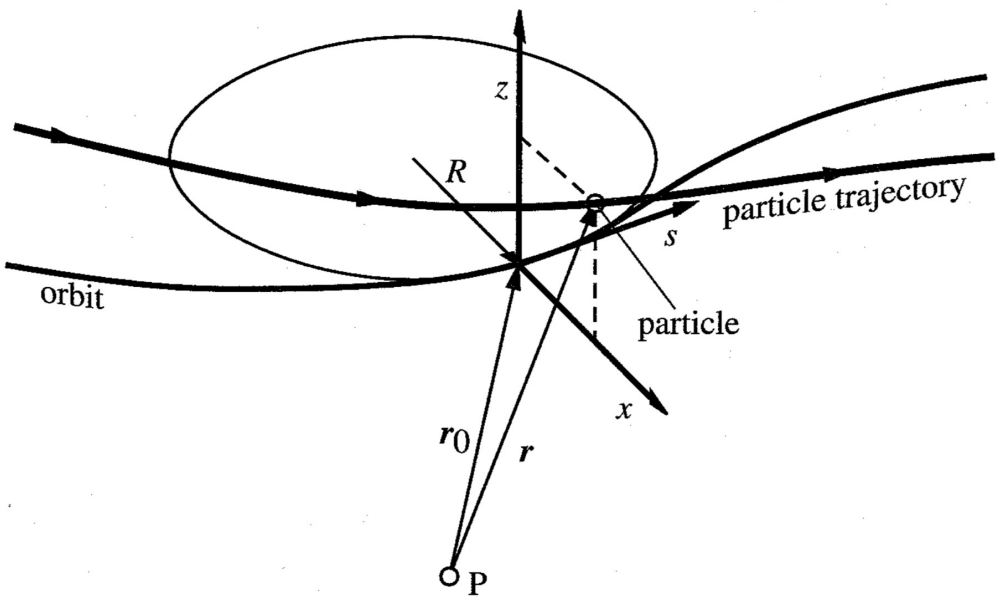


Figure 2.2: Coordinate system for particle accelerator physics. Adapted from (3).

2.1 Linear Optics

2.1.1 Matrix Formalism for Linear Elements

The basic linear accelerator elements include drift spaces (vacuum without applied magnetic field), bending magnets, and quadrupole magnets. The formalism of ray optics can be used to express the action of these elements on the phase space coordinates of a charged particle as it traverses the lattice. The phase space coordinates of a charged particle will be labeled (x, p_x, y, p_y), with the coordinate system shown in 2.2.

These elements act as lenses, thus the following transfer matrices which act on the phase space of a particle can be formulated:

$$\mathrm{M_D} = \begin{pmatrix} 1 & L_1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & L_1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.1}$$

where L_1 is the length of the drift space. Next, the bending magnet which applies a constant magnetic field to the charged particle and via the Lorentz force bends the particles in a circular path of radius R, as longitudinal position in the accelerator s, with length L_2. The transfer matrix which describes these dipole magnets is:

$$\mathrm{M_B} = \begin{pmatrix} \cos(\frac{\mathrm{s}}{\mathrm{R}}) & R\sin(\frac{\mathrm{s}}{\mathrm{R}}) & 0 & 0 \\ -\frac{1}{R}\sin(\frac{\mathrm{s}}{\mathrm{R}}) & \cos(\frac{\mathrm{s}}{\mathrm{R}}) & 0 & 0 \\ 0 & 0 & 1 & L_2 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.2}$$

Lastly, a thin quadrupole focusing magnets with horizontal (x direction) focusing gradient k at longitudinal position s along the accelerator:

$$\mathrm{M_Q} = \begin{pmatrix} \cos(\sqrt{|\mathrm{k}|}\mathrm{s}) & \frac{1}{\sqrt{|k|}}\sin(\sqrt{|\mathrm{k}|}\mathrm{s}) & 0 & 0 \\ -\frac{1}{\sqrt{|k|}}\sin(\sqrt{|\mathrm{k}|}\mathrm{s}) & \cos(\sqrt{|\mathrm{k}|}\mathrm{s}) & 0 & 0 \\ 0 & 0 & \cosh(\sqrt{|\mathrm{k}|}\mathrm{s}) & \frac{1}{\sqrt{|k|}}\sinh(\sqrt{|\mathrm{k}|}\mathrm{s}) \\ 0 & 0 & -\frac{1}{\sqrt{|k|}}\sinh(\sqrt{|\mathrm{k}|}\mathrm{s}) & \cosh(\sqrt{|\mathrm{k}|}\mathrm{s}) \end{pmatrix}. \tag{2.3}$$

By expressing the linear elements of an accelerator in this manner, the full transformation of the phase space coordinates of a particle in the accelerator can be described easily by acting the transfer matrix of the each element successively on the phase space vector. For example, for a particle with initial coordinates $(x^i, {p_x}^i, y^i, {p_y}^i)$, the final coordinates after passing through a drift space, then a quadrupole, then a dipole could be calculated by:

$$\begin{pmatrix} x_f \\ p_{x,f} \\ y_f \\ p_{y,f} \end{pmatrix} = M_B M_Q M_D \begin{pmatrix} x_i \\ p_{x,i} \\ y_i \\ p_{y,i} \end{pmatrix}. \tag{2.4}$$

Similarly, the entire linear transformation applied by an entire accelerator consisting of n linear elements, can be easily determined by pre-multiplying each transfer matrix based on the design, which is called the one-turn map for circular machines. This is shown in 2.5

$$M_{otm} = M_n M_{n-1} M_{n-2} ... M_3 M_2 M_1. \tag{2.5}$$

2.1.2 Betatron Oscillations and Tune in Periodic Machines

From the equations of motion, a few important quantities which can describe the periodic motion of particles within a circular machine with only drifts, dipoles, and quadrupoles are as follows.

The primary equation of motion is Hill's equation, shown here in one direction:

$$\frac{d^2x}{ds^2} = -k_1(s)x. \tag{2.6}$$

where $k_1(s)$ is the linear focusing function. The solution is:

$$x(s) = \sqrt{2J_x\beta_x(s)}\cos\phi_x. \tag{2.7}$$

Introduced here are is the beta function, which is a periodic function and is an important quantity used to describe particle motion in an accelerator. The beta function has the same period as $k_1(s)$, with s-dependent phase ϕ along the reference trajectory:

$$\frac{d\phi_x}{ds} = \frac{1}{\beta_x(s)} \tag{2.8}$$

These results are similar in the vertical direction, but with the solution to Hill's equation having the opposite sign, due to the focusing and defocusing in orthogonal directions respectively. Clearly, this motion is oscillatory, thus these oscillations are called betatron oscillations, where J_x is the betatron amplitude and ϕ_x is the betatron phase. In a bunch, the

particles will have different amplitudes and phases, therefore the solution to Hill's equation is better expressed by averages:

$$< x^2 >= 2\beta_x(s) < J_x \cos^2(\phi_x) > . \tag{2.9}$$

Assuming that the amplitudes and phases of individual particles are not correlated, and that the phases are randomly distributed between 0 and 2π, the previous expression can be averaged giving:

$$< x^2 >= \beta_x(s) < J_x > . \tag{2.10}$$

Next, applying a derivative with respect to position s to Eq. 2.7 gives the momentum:

$$p_x(s) = -\sqrt{\frac{2J_x}{\beta_x(s)}}(\sin(\phi_X) + \alpha_x \cos\phi_x), \tag{2.11}$$

where

$$\alpha_x = \frac{-1}{2}\frac{d\beta_x}{ds}. \tag{2.12}$$

Applying the same assumptions about the correlation of the phases and amplitudes gives:

$$< xp_x >= -\alpha_x < J_x > \tag{2.13}$$

$$< {p_x}^2 >= \alpha_x < J_x > \tag{2.14}$$

The β function and α function are related to each other by the following γ function:

$$\beta_x(s)\gamma_x(s) - \alpha_x(s)^2 = 1 \tag{2.15}$$

where β, α, and γ are the Courant-Snyder parameters(15), or Twiss parameters(16), and are used to describe collective beam behavior within a particle accelerator. They will be

critical for understanding the stability of a periodic system,

Matched Orbit

In a periodic system, like a storage ring, we can define a single-turn transfer matrix constructed from the individual transfer matrices for each element in the lattice, as shown in Eq. 2.5. From this expression, it is clear that as the particles move through the accelerator, their positions will change. Despite this, the second-order moments of the bunch tend towards an equilibrium distribution, and remain constant once the motion has stabilized.

A covariance matrix, $\Sigma(s)$ can be constructed to track how the second order moments $< x^2 >$, $< xp_x >$, $< {p_x}^2 >$ stabilize.

$$\Sigma(s) = \begin{pmatrix} < x^2 > & < xp_x > \\ < xp_x > & < {p_x}^2 > \end{pmatrix} = B(s) < J_x > \tag{2.16}$$

where,

$$B(s) = \begin{pmatrix} \beta(s) & -\alpha_x(s) \\ -\alpha_x(s) & \gamma_x(s) \end{pmatrix}. \tag{2.17}$$

Then, for one pass through the accelerator from initial coordinates at s_0 to $s_0 + C_0$,

$$\begin{pmatrix} x \\ p_x \end{pmatrix}_{s_0+C_0} = M_{otm} \begin{pmatrix} x \\ p_x \end{pmatrix}_{s_0}. \tag{2.18}$$

and similarly the covariance matrix is transformed by:

$$\Sigma(s_0 + C_0) = M_{otm} \Sigma(s_0) {M_{otm}}^T \tag{2.19}$$

where M^T is the transpose of the one-turn map. The optics in a circular machine are designed to be matched, in order to enforce the following conditions. The matched optics

results in a covariance matrix which we can call $\Sigma_{matched}$. By definition, this matrix must remain unchanged through one-turn as well, such that:

$$M_{\mathrm{otm}}\Sigma(s_0)_{\mathrm{matched}}{M_{\mathrm{otm}}}^T = \Sigma(s_0)_{\mathrm{matched}}. \tag{2.20}$$

In order for this condition to be satisfied, the one-turn map must satisfy:

$$M = I\cos(\mu_x) + B(s_0)S\sin(\mu_x), \tag{2.21}$$

where I is the identity matrix, S is:

$$S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{2.22}$$

and μ_x is called the betatron phase advance.

2.1.3 Beam Emittance

Mentioned in Chapter 1, was a quantity called emittance. The emittance is a measure of the "size" of the beam in phase space. A small emittance suggests that the particles are tightly configured, at least in one-dimension. A mathematical description of the emittance illuminates this further.

Starting with the solutions to Hill's equations, shown in Eq. 2.7, we can see that the amplitude of the motion J_x is constant. We can invert Eq. 2.7 to represent the amplitude and phase of a single particle in terms of the Twiss parameters to achieve the following relationships:

$$2J_x = \gamma_x x^2 + 2\alpha_x x p_x + \beta_x {p_x}^2. \tag{2.23}$$

If the amplitude of each particle is constant, then the average of all amplitudes should

also be a constant. As shown previously, the Twiss parameters are related to the the second order moments of the bunch. If we define the average amplitude $< J_x >\equiv \epsilon_x$, then the emittance is:

$$\epsilon_x = \frac{1}{2}\gamma_x x^2 + \alpha_x x p_x + \frac{1}{2}\beta_x {p_x}^2. \tag{2.24}$$

This expression also turns out to give the area of an ellipse, suggesting that the particles will cover an elliptical region in each projected phase space direction. The dimensions of this ellipse can be expressed by the Twiss parameters as well, shown in Fig. 2.3. A single particle will trace the ellipse in phase space as it traverses the accelerator. Collectively, all of the particles will trace out their own ellipses with amplitude according to their betatron amplitude.

A further useful definition of the emittance is a geometric definition. By using Eq. 2.15, and the definitions of the second order moments, Eq.2.24 becomes:

$$\epsilon_x = \sqrt{< x^2 >< {p_x}^2 > - < x p_x >^2}. \tag{2.25}$$

The geometric emittance can be used to define the normalized emittance $\epsilon_{x,n}$, where $\epsilon_{x,n} = \gamma\epsilon_x$, and γ is the relativistic Lorentz factor.

Due to Liouville's theorem in the absence of changes in energy, the bunch emittance is a constant. For a storage ring, because there is energy loss from synchrotron radiation, and energy restoration through RF cavities, there is an equilibrium emittance that can be reached, based on the design specifications of the machine.

Betatron Phase Advance

Due to the periodic nature of the motion, imposed by the matched conditions, the Twiss parameters will be unchanged as the particles circulate. But the betatron phase will increase by: $\phi(s_0 + C_0) = \phi(s_0) + \mu_x$. The oscillatory nature of Twiss parameters can further be

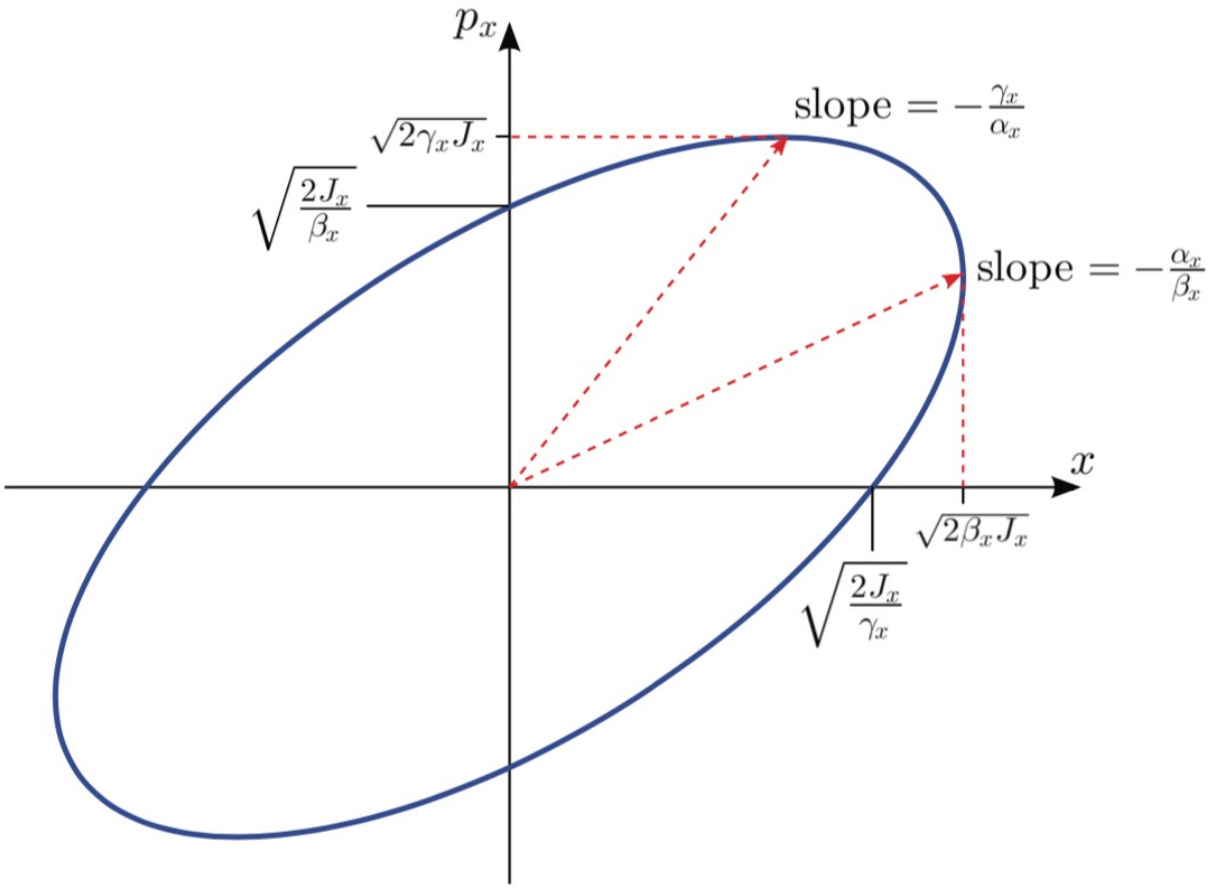


Figure 2.3: Shown is the phase space ellipse defined by the emittance relationship, in which the Twiss parameters describe the dimensions and important features of the ellipse. The area of the ellipse is the emittance for the given projection, such as ϵ_x which is shown in the figure. Figure adapted from (17).

parameterized by the number of betatron oscillations that are completed in one turn, called the betatron tune, ν_x, and is related to the phase advance by $\mu_x = 2\pi\nu_x$.

The stability condition therefore depends on the phase advance, and is calculated by:

$$2\cos(\mu_x) = \text{Trace}(M_{otm}). \tag{2.26}$$

This condition corresponds to the following as well:

$$B(s_0) = \frac{S}{\sin(\mu_x)}(I\cos(\mu_x) - M_{otm}). \tag{2.27}$$

From these two conditions, a variety of limitations become clear. Eq. 2.26, the condition that the trace of the one turn map must be less than two, otherwise the phase advance becomes imaginary. If that were the case, the solutions to the equations of motions become hyperbolic cosine and sine functions, which do not oscillate. Thus, the trajectories would grow exponentially rather than oscillating about a reference orbit, and particles could be lost by hitting the physical aperture of the vacuum chamber. From Eq. 2.27, if the phase advance is an integer or half integer multiple of π, the denominator goes to zero and the expression becomes undefined. This behavior is what is called a resonance. Thus, resonances are associated with values of the tune which can result in this kind of undefined behavior, and can result in particle loss.

2.2 Nonlinearity and Resonances in Periodic Lattices

As mentioned, periodic lattices result in the particles circulating through the machine over and over again, which can potentially lead to deleterious results if field errors are allowed to compound. For example, take a particle which begins on the reference orbit, and has an integer tune. If there is a dipole field error at a certain location along the trajectory, the particle will always pass through it at the same position and momentum (due to the integer tune). This will allow the misalignment to compound over repeated turns, and lead to the

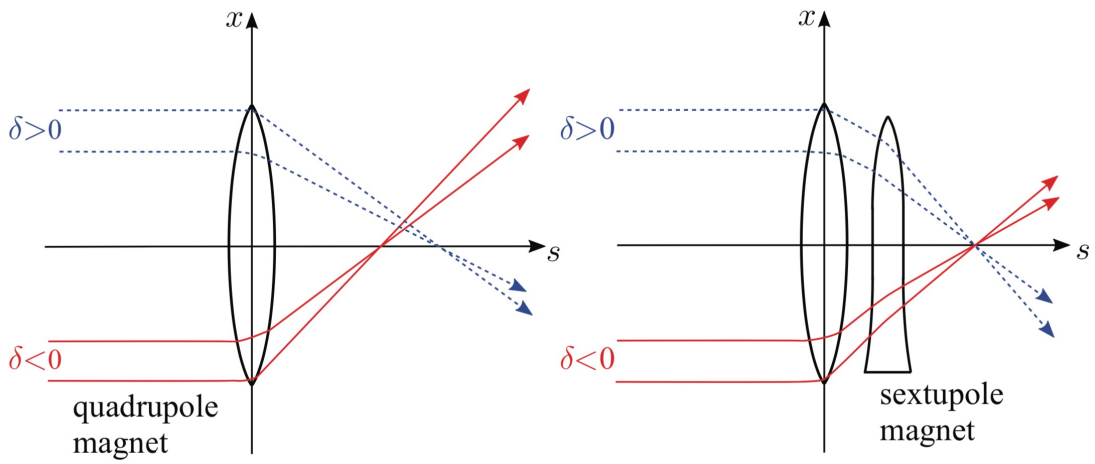


Figure 2.4: Left: Particles with energy deviations shown entering a quadrupole, with the path due to their incoming energies shown. Right: With proper sextupole strength, the focal length depends on the energy deviation of the incoming particle, in order to correctly focus particles. Figure from (17).

particles eventually becoming unstable. Similarly, a quadrupole field error can compound a half integer tune resonance. However, even a perfectly tuned machine with no misalignments can still be victim to resonances. High-order resonances from sextupole and octopole magnets can excite 3rd and 4th order resonances due to 1/3 and 1/4 tunes.

2.2.1 Chromaticity Correction with Sextupole Magnets

Not explicitly mentioned in the previous section is the effect of energy deviation on the trajectory of a particle. For both linear elements, the amount a particle is bent as it goes through the element will be different relative to another particle of different energy. For example, the focal length from a quadrupole for a particle with energy deviation δ greater than the reference energy, is slightly less. Therefore, this particle will not be focused as much as an on-energy particle. Similarly, a particle with a $-\delta$ energy deviation will be bent more than an on-energy particle.

The focal length is therefore longer for particles at energies higher than the reference energy, and shorter for particles with lower energies. As the higher energy particles are focused less and less, the number of oscillations that they complete during one-turn will change. This resulting tune change is called the chromaticity. The horizontal and vertical chromaticities are defined as:

$$\xi_{x,y} = \frac{d\nu_{x,y}}{d\delta}. \tag{2.28}$$

The chromaticity is a dimensionless number, and tends to be a negative quantity in storage rings. Sextupole magnets are nonlinear elements that provide chromatic correction, and produce a magnetic field which varies quadratically with the spatial coordinates of the magnetic aperture.

The transformation to the phase space coordinates through a sextupole magnet is:

$$\begin{pmatrix} x_f \\ p_{x,f} \\ y_f \\ p_{y,f} \end{pmatrix} = \begin{pmatrix} x_i \\ p_{x,i} + K_2({x_i}^2 - {y_i}^2) \\ y_f \\ p_{y,i} - 2K_2(x_i y_i) \end{pmatrix}, \tag{2.29}$$

where K_2 is the magnetic gradient, or strength. Further, this transformation shows that a sextupole couples the motion of both transverse directions. This coupling is a major concern which makes designing machines with small vertical emittance challenging. In general, coupling introduces the challenge of not being able to design and optimize the sextupole distributions to correct each transverse direction independently.

Despite this drawback, sextupoles magnets are necessary for chromatic correction due to the negative effects chromaticity introduces. First, a large chromaticity means that particles with different energies have significantly different tunes. This can mean that it is more likely that some of these particles will operate close to a resonance, despite attempting to find

a sufficient nominal tune for the machine. Second, chromaticity can results in a head-tail instability. This instability occurs when particles at the head of the bunch create EM fields that can drive betatron oscillations in the tail. This instability can also lead to significant beam loss, if not corrected with sextupoles or damped properly during operation (18).

2.2.2 Coupled Resonances

The coupling introduced by sextupoles has the added effect of producing coupled resonances. As mentioned previously, certain values of the tune can result in particle loss due to repeated momentum kicks which can add up if the particle always returns to the same location in the periodic structure. Further complications arise when sextupoles are added, as they also limit the operating tune.

The previous discussion of resonances described a particle which experiences a dipole or quadrupole field error, which could compound if the tune is a half or quarter integer multple of π. Conversely, even in a perfectly tuned and aligned machine, a particle which loses energy to synchrotron radiation may deflect slightly in a quadrupole, and those deflections can add up over successive passes. Thus, through no fault of the magnetic structure, the particle trajectory can become unstable. Lastly, the addition of sextupoles introduces betatron coupling, as well as 1/3 integer resonances. A concise description of all resonances can be given by:

$$m\nu_x + n\nu_y = l, \tag{2.30}$$

where m, n, and l are positive or negative integers, and ν_x and ν_y are the transverse tunes. A resonance diagram is often used to show where these resonances will occur, and to find a suitable parameter range to operate at.

Shown in Fig. 2.5 are all of the tune values which would result in a certain order of resonance. This visualization highlights the challenge of designing a machine within which energy

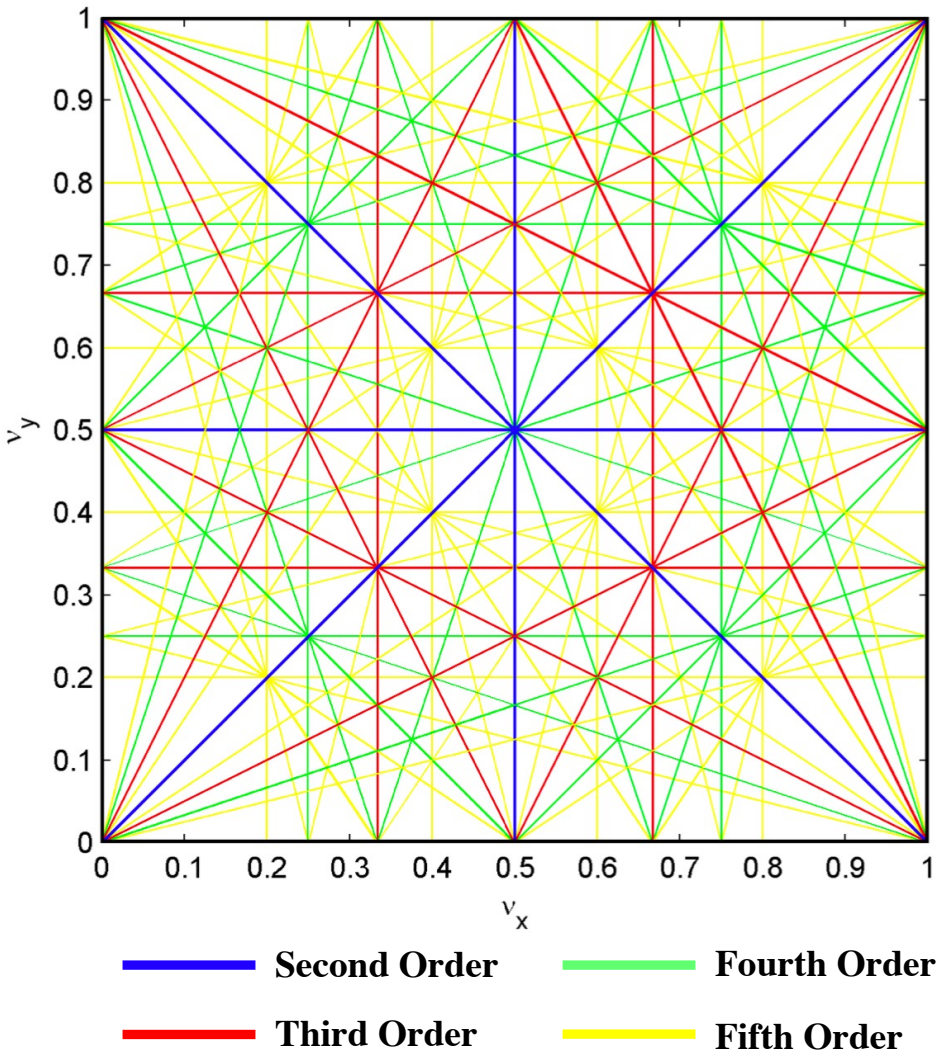


Figure 2.5: Shown is a visual representation of Eq.2.30, where the lines of different colors represent all tune values that would result in the associated resonance. The operating tune for a storage ring is usually selected such that it falls between lines, in any remaining white spaces, such that resonances are avoided. Figure adapted from (17).

deviations are compensated, but the tune is not allowed to operate near any resonances.

2.2.3 Dynamic Aperture

Clearly, there are several phenomena which, while necessary to ensure stable, focused particles, also limit the operation of the machine. Further, despite chromatic corrections, there are higher order chromatic effects that can still contribute to tune shift, despite the use of sextupole or higher multipole magnets. Thus, the operating parameters for the beam are limited to a certain range, called the dynamic aperture. This aperture is not a physical aperture, but a 6-dimensional volume in which the stable phase space trajectories exist.

In one projection, the area of phase space with stable orbits can be shown by tracking particles through a lattice, and plotting their phase space coordinate after each full rotation. Particles in a completely linear lattice would then trace out the phase space ellipse described by Eq. 2.24. Each particle can be initialized with a different amplitude, to show how particles in a bunch would fill the phase space. A lattice with any field errors, or nonlinearities, may include particles which exhibit these resonances by getting trapped in "islands." This behavior is not an issue necessarily, but may indicate that certain nonlinearities are larger in magnitude than is ideal. Further, particles that begin their trajectory too close to a resonance condition may immediately diverse and not exhibit a close trajectory, but a chaotic one.

These behaviors are shown in Fig. 2.6. An ideal operating tune would allow for many closed trajectories, at large amplitudes. The two examples of resonance behavior are seen when the horizontal tune is 0.252π and 0.33π, which are near a quarter and third resonance respectively. It is also noticeable that even a small change in the tune from 0.248π to 0.252π can significantly distort the phase space trajectories. Higher order resonances are visible as well, including 5th order resonance islands.

The region of stable trajectories is separated from the region of chaotic trajectories by the separatrix. Finding operating conditions that can enlarge the separatrix can extend the dynamic aperture, and ensure more of the beam survives in the storage ring.

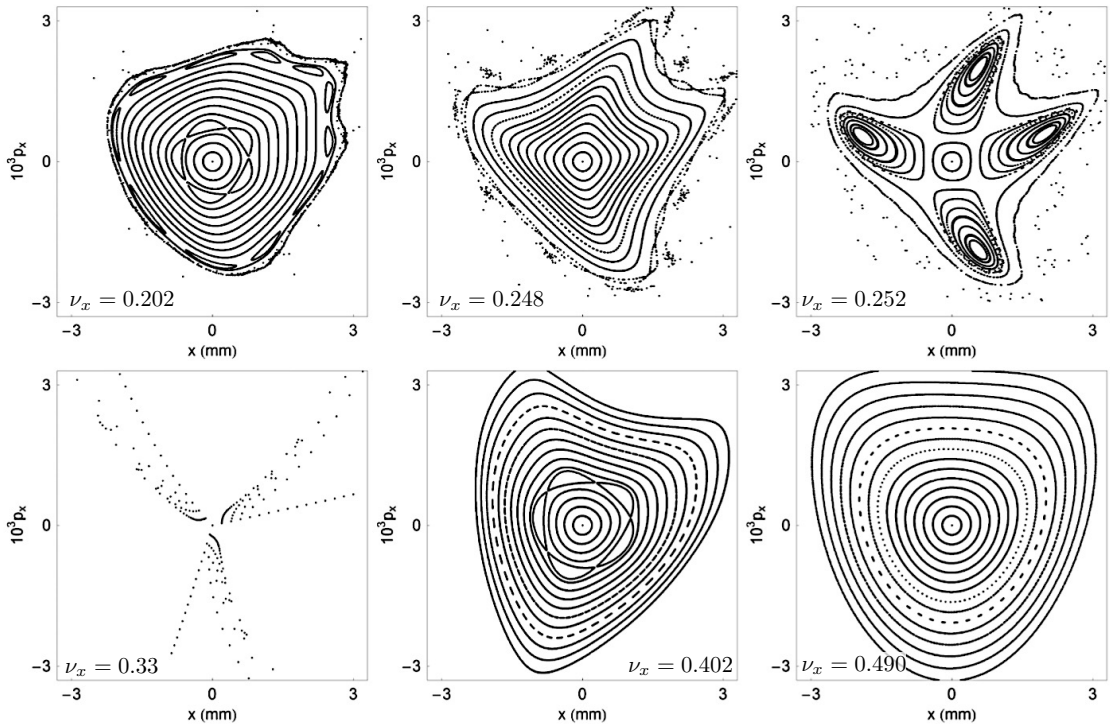


Figure 2.6: Shown are six examples of phase space trajectories, for a lattice with dipole, quadrupole, and sextupole magnets. The trajectories represent 1000 turns through the lattice. For each plot, the tune is shown. Figure adapted from (17).

2.3 Approaches to Extending Dynamic Aperture in Periodic Lattices

2.3.1 Challenges in Designing Periodic Particle Accelerators

The main goal at a facility with a periodic accelerator lattice, such as a storage ring of a collider, is to maintain the bunch quality as long as possible. At light sources, the quality of the bunch must be maintained in order to ensure radiation for use in diffraction imaging or other types of radiation-based measurements. Thus, the design of these machines focuses primarily on producing and maintaining high beam currents, small emittances, and stable orbits. Significant work is put into designing and optimizing periodic lattices in which the chromaticity and nonlinear resonances that cause beam loss are controlled and minimized.

Chromatic correction in storage rings is challenging due to the coupling introduced by sextupole magnets. The chromaticity in each transverse direction, therefore, cannot be corrected independently. Further, without proper chromatic correction, the dynamic aperture of the beam is limited. As described in the previous chapter, the dynamic aperture describes the volume in phase space in which stable particle orbits can exist. Maximizing the dynamic aperture, while minimizing chromaticity, is therefore the primary challenge when designing the nonlinear optics in periodic lattices.

Currently, the primary method for designing nonlinear optics, such as placement along the lattice and magnet strengths, is by using evolutionary computation algorithms. This includes the genetic algorithm and particle swarm optimization, which are both global optimization methods. Both methods are heuristic search methods, which start by randomly initializing a population of solutions and evaluating a fitness function to iterate the solutions. The reason heuristic methods are used is because there is no closed-form, or approximation methods for determining how resonances effect the dynamic aperture. Thus, heuristic searches of magnet settings, with length particle tracking simulations which evaluate the dynamic aperture, and iterate accordingly are the current best solution.

In the case of dynamic aperture optimization, the fitness function generally calculates one of two things. One method is to calculate the phase space volume of stable orbits. This is done by initializing particles at a variety of starting amplitudes and phases, then numerically integrating the trajectory through the lattice. In a purely linear machine, this could be done very quickly by multiplying each initial particle position and momentum by the one turn matrix, to determine the final position and momentum after a single turn. Successive application of the matrix results in the final position and momentum after N turns. However, because nonlinear magnets cannot be expressed as easily by matrices, this integration becomes a numerical challenge. A host of sympletic methods exist for integrating trajectories with the presence of nonlinear potentials in multiple dimensions. However, these methods are computationally expensive, and must be iteratively applied to each trajectory. Thus, a single calculation of the dynamic aperture for hundreds of particles can take significant computational resources and time to complete.

In 2006, an algorithm for chromatic compensation was described by Levichev and Pimiov, for correcting the transverse chromaticities simultaneously by placing and optimizing the strengths of a N pairs of sextupoles. While this method is a fairly standard form of constrained optimization, the fitness evaluation still requires tracking particles to calculate the dynamic aperture after each optimization step(19). Even with improvements on this method, the fundamental rate-determining procedure is the dynamic aperture calculation (20). Thus, this method is similar and prone to the same challenges as the previously mentioned numerical optimization techniques.

Another method for dynamic aperture optimization is to calculate and strategically minimize resonance driving terms (RDTs). These terms quantify the strength of the various nonlinear resonances that can appear in a periodic lattice and are derived in detail in (21). The RDTs can be likened to Fourier coefficients, describing the nonlinear contributions to the accelerator Hamiltoninan. These components are then minimized as musch as possible, though often this means strategically increasing the strength of higher order resonances that

occur at large particle amplitudes. In general, the optimization process requires some empirically applied strategy, as well as the heuristic methods such as multi-objective genetic algorithm optimization(22).

A Physics-Informed Approach

The methods described above, while powerful and established, accept physical limitations such as the existence of sextupole coupling as inevitable. Therefore, a slightly different approach to solve this problem would be reexamine if these coupling is, in fact, impossible to eliminate. If it is possible to eliminate or even diminish significantly, then the use of such a magnet or system of magnets could decrease the need for computationally expensive and time-consuming optimizations.

The One-Dimensional Sextupole One such approach would be to determine a method for cancelling out the sextupole coupling by using a channel of linear magnets, namely quadrupoles. The theoretical construction of such a channel was developed by S. Nagaitsev, and has been further explored by J. Ögren and V. Ziemann(23). In this thesis, this method is further explored, with particular attention to how these channels could be designed for an arbitrary lattice.

Further, the effects of beam energy spread through a 1-D sextupole channel is studied to determine how the dynamics would change. This helps determine the feasibility of 1-D sextupoles.

Resonance Elimination Another approach to aid in nonlinear beam design would be to determine novel techniques for eliminating resonances entirely. This can be done by determining how a series of sextupole magnets and a focusing element, can be configured to provide a sufficient phase advance such that the resonance driving term associated with the primary 3rd order resonance vanishes. This could make it easier to design accelerators, by simply removing resonances which can be tricky to control otherwise.

Sextupole Calibration Progress toward lower emittance in storage ring synchrotron light

sources to achieve diffraction limited beam quality at angstrom wavelengths will necessarily require stronger chromaticity correction (24). Thus, online diagnosis and correction of storage ring sextupole distributions may become an important tool in the preservation and maximization of dynamic aperture. Methods for locating and measuring the magnitude of sextupole field errors have been demonstrated with good success by using turn-by-turn beam position monitor (BPM) measurements (25; 26). These methods employ a spectral analysis of BPM data for the calculation of RDTs that encode sextupole field error information(27). Coherent displacement of the beam is either performed with a single pulsed kicker (pinger) or an ac dipole. In the case of a pinged beam, the number of useful turns in the analysis is limited by radiation damping. The ac dipole technique can accumulate many more turns of turn-by-turn data (28), and at the Cornell Electron Storage Ring (CESR), tune trackers are used to synchronously drive the beam at the betatron tunes (in both planes) even in the presence of phase jitter arising from guide field fluctuations on time scales of ˜100 turns (29). This configuration serves as an ideal test-bed for a robust sextupole field error mapping method to be developed.

Thus, the method described in this work is meant to function at a facility in which this kind of driving takes place. The method was tested experimentally, and the results show that the method can be further developed to provide sextupole calibration.

CHAPTER 3

DESIGN FOR A 1-DIMENSIONAL SEXTUPOLE

As mentioned in Chapter 3, the use of sextupoles for chromatic correction is necessary to maintain beam quality in periodic accelerators. However, the use of sextupole magnets also introduces restrictions to the dynamic aperture, by introducing 3rd order tune resonances. Further, sextupole energy corrections are coupled in the transverse directions, which further complicates how they can be used in accelerators to provide momentum corrections.

Studies of integrable systems for particle accelerators by Sergei Nagaitsev (1) suggest that there is an analytic way to create transfer matrices which limit the momentum kick to one dimension. Thus, one could create an accelerator lattice using linear elements (dipoles and quadrupoles), in series with sextupole magnets to create a "one-dimensional sextupole." While still nascent in application, the theory of such magnets is not. The methods developed for 1-D sextupole are analagous to the work of creating a nonlinear magnet for the IOTA lattice, which can provide the octopole potential needed to damp high order collective instabilities, while maintaining lattice linearity (1).

While the theoretical demonstration of this semi-analytic method has been done, and is shown in this chapter, the following contributions are made by this thesis:

- A demonstration of how a 1-D sextupole channel within a machine could be designed using optics simulation tools, such as Tao and BMAD (30).
- An analytic method for calculating 1-D sextupole channel parameters including quadrupole lengths and focusing gradients, based on requirements for a physical machine.
- A analytic and numerical analysis of how energy spread within a bunch would change the 1-D sextupole dynamics.

These contributions are necessary when considering how to explore a 1-D sextupole channel in experiment, but also the applications and feasibility of this method for use in light sources facilities.

3.1 Theory

The one-dimensional sextupole was developed by considering a simple transfer matrix, which can be created using linear elements such as drift spaces and quadrupole magnets, and has the following properties. The matrix must be diagonal, and invertible, with diagonal elements as shown in Eq. 3.1.

$$D = \begin{pmatrix} a_1 & & & \\ & \frac{1}{a_1} & & \\ & & b_1 & \\ & & & \frac{1}{b_1} \end{pmatrix} \tag{3.1}$$

where both a_1 and b_1 are any real scalar numbers. Work by Prof. Nagaitsev shows that this form of transfer matrix and the inverse of this matrix, when applied to two consecutive sextupole magnets results in an entirely one-dimensional momentum kick:

$$DS_1D^{-1}S_2 = \begin{bmatrix} x \\ p_x + \Delta p_x \\ y \\ p_y \end{bmatrix}. \tag{3.2}$$

where S_1 and S_2 are non-linear transfer matrices whose action is given by:

$$S_1 = \begin{bmatrix} x \\ p_x + q_1(x^2 - y^2) \\ y \\ p_y - 2q_1xy \end{bmatrix} \tag{3.3}$$

$$S_2 = \begin{bmatrix} a_1 x \\ p_x + \frac{q_1(x^2-y^2)}{a_1} + q_2[(a_1 x)^2 - (b_1 y)^2] \\ b_1 y \\ \frac{p_y - 2q_1 xy}{b_1} - 2q_2 a_1 b_1 xy \end{bmatrix} \tag{3.4}$$

and q_1 and q_2 are the currents (or focusing gradients) of the sextupoles.

The transformation in Eq. 3.2 is a composition of transformations, and is not completed by simple matrix multiplication due to the non-linear nature of sextupole transformation. In order to do this calculation, the sextupole kicks are added to the initial state vector, then the matrix multiplication of D on the state is carried out. The second sextupole kick is then added to the resulting state vector, and then the final matrix multiplication $D^{-}1$ is carried out; thus the transformation shown in Eq. 3.2 results in:

$$DS_1 D^{-1} S_2 = \begin{bmatrix} x \\ p_x + (q_1 + a^3 q_2)x^2 + (q_1 + ab^2 q_2)y^2 \\ y \\ p_y - 2(q_1 + ab^2 q_2)xy \end{bmatrix}. \tag{3.5}$$

The condition that must be satisfied in order for the coupling term in the y-direction to vanish is:

$$2(q_1 + ab^2 q_2)xy = 0 \tag{3.6}$$

$$q_2 = \frac{-q_1}{a_1 {b_1}^2} \tag{3.7}$$

By setting q_2 as shown in Eq.3.7, the transformation yields:

$$DS_1D^{-1}S_2 = \begin{bmatrix} x \\ p_x + q_1x^2 - \frac{a^2q_1x^2}{b^2} \\ y \\ p_y \end{bmatrix}. \tag{3.8}$$

Therefore, one can see from the final state vector, Eq. 3.8, the change in the momentum in the x-direction, due to the 1-d sextupole transfer matrices is:

$$\Delta p_x = q_1[1 - (\frac{a_1}{b_1})^2]x^2. \tag{3.9}$$

From these calculations, we have shown that if the transfer matrices in Eq. 3.2 could created by an accelerator lattice, then the momentum kick in the transverse direction would be given by Eq. 3.9 where the currents q_1 and q_2 are related by Eq. 3.7. This would eliminate the x-y coupling in the y-direction, though in practice we hope this would simply render the coupling negligible.

Generally, any lattice that satisfies these conditions could produce a 1-D sextupole, but a short lattice with minimal quadrupoles is preferable, as it could be easily inserted into an existing lattice. Original work by Prof. Nagaitsev used precisely this prescription to determine a general solution for a system of 4 quadrupole magnets with focusing gradients, k_i and 3 drift lengths of length l_i. The general lattice transfer matrices for this system is:

$$\mathrm{A} = \begin{pmatrix} 1 & 0 \\ k_4 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ k_3 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ k_2 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ k_1 & 1 \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix} \tag{3.10}$$

and for the orthogonal direction:

$$\mathrm{B} = \begin{pmatrix} 1 & 0 \\ -k_4 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -k_3 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -k_2 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -k_1 & 1 \end{pmatrix} = \begin{pmatrix} b & 0 \\ 0 & 1/b \end{pmatrix} \tag{3.11}$$

By solving this system, we can determine formulas for the focusing gradients and the drift lengths:

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} = \begin{pmatrix} a & 0 & 0 & 0 \\ 0 & 1/a & 0 & 0 \\ 0 & 0 & b & 0 \\ 0 & 0 & 0 & 1/b \end{pmatrix} \tag{3.12}$$

Solving this system for the focusing gradients and drift lengths yields:

$$l_1 = l_1[\mathrm{m}] \tag{3.13}$$

$$l_2 = -l_1 \frac{(a-b)^2}{2(-2ab+a+b)}[\mathrm{m}] \tag{3.14}$$

$$l_3 = l_1 \frac{ab(a+b-2)}{-2ab+a+b}[\mathrm{m}] \tag{3.15}$$

$$k_1 = \frac{1}{l_1} \frac{a^3 + a^2((7-8b)b-4) + 7ab^2 + (b-4)b^2}{2(a-b)(a+b-2)(a+b)} \tag{3.16}$$

$$k_2 = \frac{1}{l_1} \frac{a+b-2}{b-a} \tag{3.17}$$

$$k_3 = \frac{1}{l_1} \frac{(-2ab+a+b)^2}{ab(a-b)(a+b-2)} \tag{3.18}$$

$$k_4 = \frac{1}{l_1} \frac{a^3(4b-1) - 7a^2b + ab(b(4b-7)+8) - b^3}{2ab(a-b)(a+b-2)(a+b)} \tag{3.19}$$

These results confirm Nagaitsev's calculations. This system will be used as an "ideal" solution for a 1-d sextupole, in order to consider the effect of energy spread in the system.

3.2 Approaches for Designing a 1-D Sextupole

In order to begin designing a 1-D sextupole, it was prudent to determine which parameters were free, and which ones need to be controlled. Because, in order to determine if such a system could be built, and that a transfer matrix like Eq. 3.1 can be achieved using linear elements, I took the following approach.

All design and simulation work was done using the Tool for Accelerator Optics (Tao), which is a program that uses the Bmad library, developed at Cornell University by David Sagan. This tool can be used for several accelerator physics applications, including single and multi-particle tracking, as well as lattice design, simulation, and analysis (30). Tao can calculate a variety of beam and lattice parameters, and allows for dynamic optimization of simulated lattices. Tao provides several optimization options, which are carried out by specifying the desired value of certain beam or lattice parameters, and the variables which can be varied in order to achieve those values. The optimization procedure minimizes the merit function, M:

$$M \equiv \sum_i w_i[\text{data\_model(i)} - \text{data\_meas(i)}]^2 + \sum_{\text{j}} \text{w}_\text{j}[\text{var\_model(j)} - \text{var\_meas(j)}]^2 \quad (3.20)$$

which is a measure of how close the current simulated lattice is to the desired lattice. Here, the data_model(i) value is the i^{th} data value from the currently modeled lattice, and data_meas(i) is the desired value of the same datum. For example, one could require the β function at the beginning of the lattice be 15m, but if the current model has a beginning β value of 13m, the data_model(β_{beg}) would be 13m and data_meas(i) would be 15m. The second sum is similar, but the the variable quantities which are used in the optimized, which can be set to have "measured" valued that one wants the variable value to be, and the model value which is set during the optimization process. Finally, the w_i and w_j terms represent weights, which can be used to emphasize a particular data value or variable setting over other ones, in order to accurately represent the user's design goals.

For this research, the optimization scheme used was the differential evolution optimizer (31), which uses a genetic algorithm. The genetic algorithm proceeds by selecting a portion of the variables, then randomly steps those variables and calculates the merit function for the new parameters. The merit function is compared to the previous merit function, and the solution is updated if the merit function value decreased. This method is very powerful, as it can be used for constrained and unconstrained optimization, and can be used to find the global minimum of the merit function.

For designing the 1-D sextupole, I constructed the lattice in Tao and specified the element parameters as variables to be changed and used as variables for optimization. I was able to simulate the lattice with as many quadrupole magnets and drift spaces as needed, assign beginning optics parameters, and propagate the beam through the lattice. In Tao, the transfer matrix elements were used as data values, and focusing gradients and drift lengths as variables, such that the optimization process could search for a diagonal solution. In order to do this, the initial lattice and beam parameters were carefully set up, which facilitated the optimization process.

First, a simple focusing, open (drift space), defocusing, open (FODO) lattice cell (3) was designed, with an arbitrary transfer matrix, R. An example is shown in fig. 3.1 The cell was designed initially to have reasonable beginning β functions. From this, the transfer matrix for a general cell consisting of nine quadrupoles and 10 drift spaces, which would have an arbitrary transfer matrix, U. The transfer matrix, U was determined by the relation:

$$UR = D \tag{3.21}$$

$$U = DR^{-1} \tag{3.22}$$

The matrix U was determined, and the second cell of quadrupoles and drift spaces was optimized by varying the initial quadrupole focusing gradients and drift lengths, until the U transfer matrix was attained. The same procedure was used for constructing the inverse

matrix, $D^{-}1$ in order to construct the entire 1-d sextupole matrix show in Eq. 3.2. The variables used during optimization were the drift lengths, the focusing gradients,and the beginning optics.

An initial solution was a transformation in which $a_1 = 1$ and $b_1 = -1$, shown in Fig. 3.2.

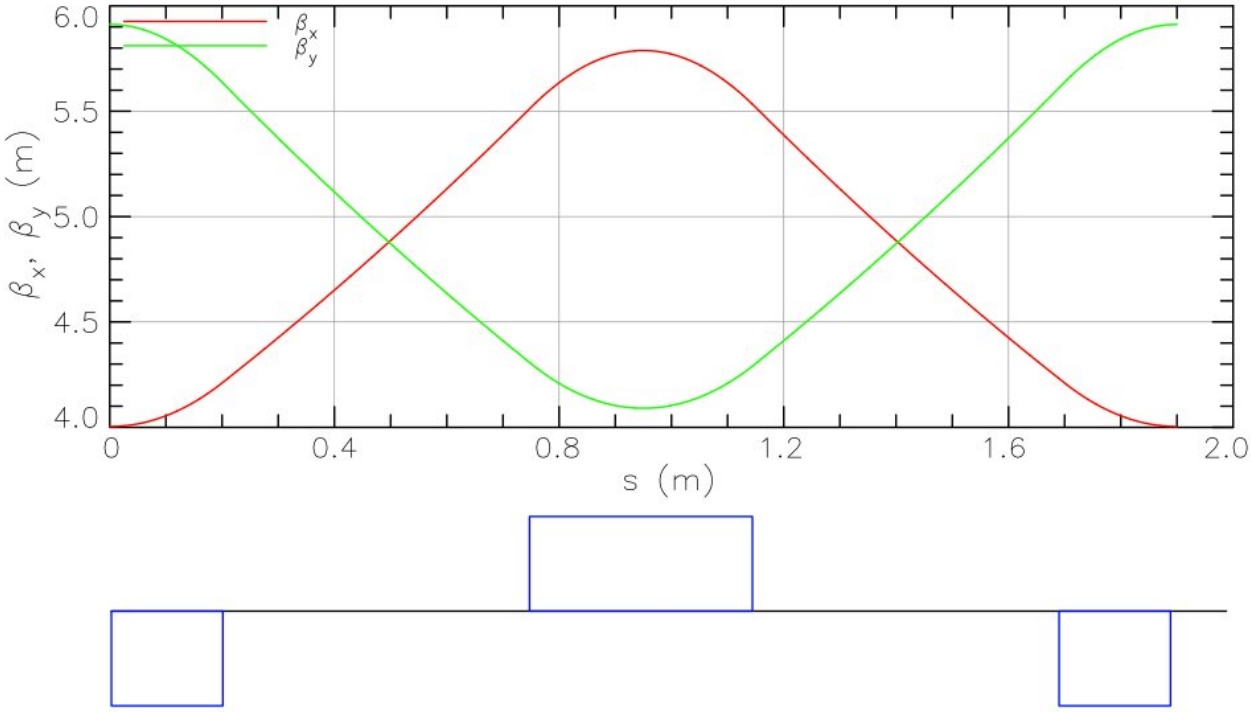


Figure 3.1: The beam optics for an example FODO cell used to create the matrix R, which was then used to design a lattice which would result in a diagonal final transfer matrix.

The unity transformation is a fairly special one, because the nonlinearity is canceled out entirely because $a_1 = -b_1$. As can be seen from Eq. 3.9, $\Delta p = 0$ for $a_1 = -b_1$. This is an interesting result, and can be used in systems in which one wants all nonlinearity to vanish (32)(33)(18). The same methodology and optimization was used to simulate the full lattice for $a_1 = 2$, $b_1 = 3$, shown in Fig. 3.3

This solution shows that a lattice in which $\Delta p \neq 0$ can be created, because it consists of D and D^{-1} matrices which are not unity transformations The solution shown in fig. 3.3, is the best solution that was obtained by this method, but still has px-py coupling terms that are on the order of $10^{-}3$. Despite the small coupling terms which still exist in this

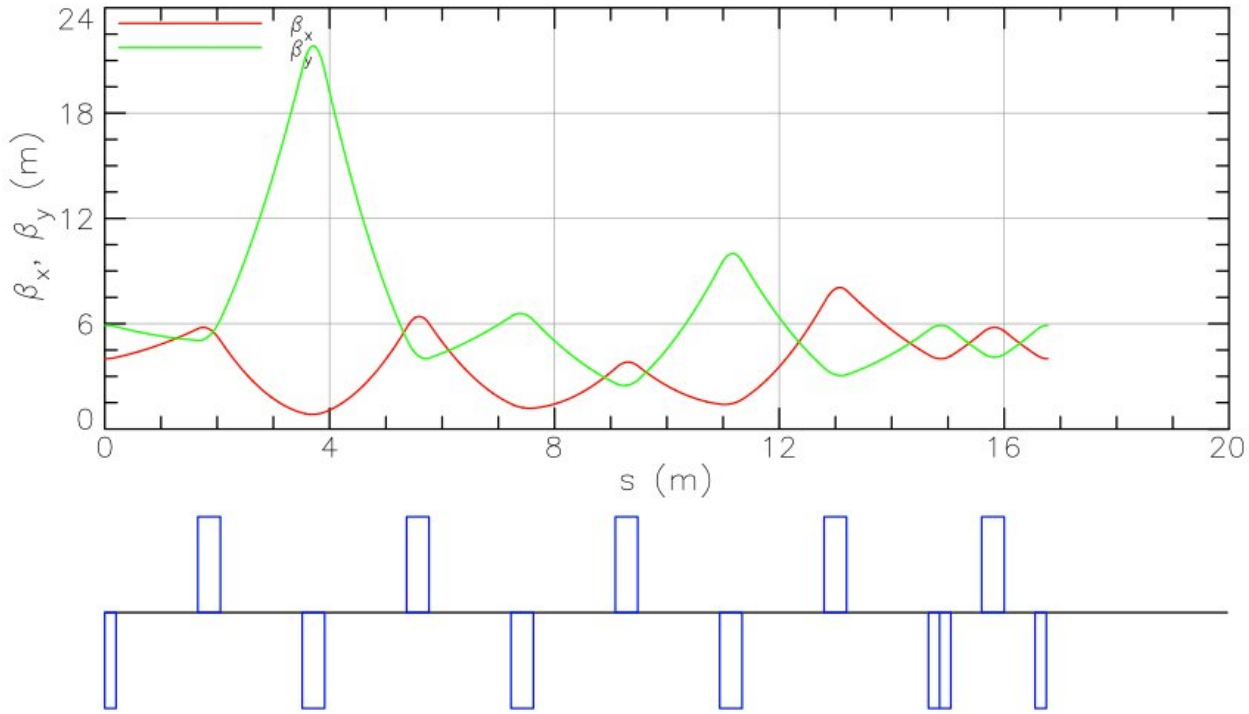


Figure 3.2: The lattice with the resulting final transfer matrix D, where $a_1 = 1$ and $b_1 = -1$.

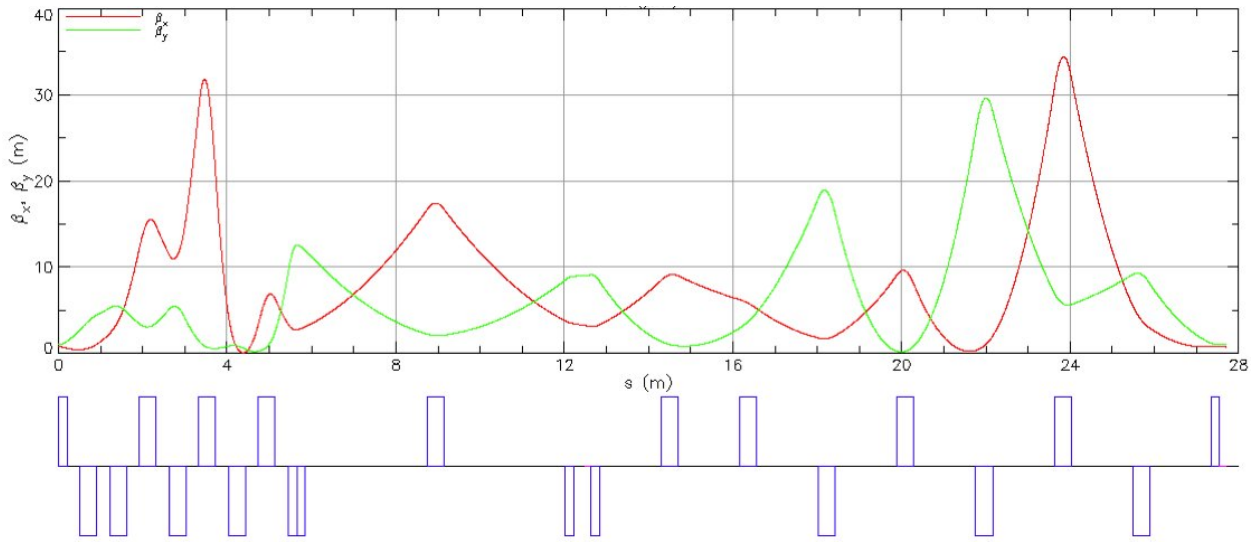


Figure 3.3: The full lattice solution resulting in an identity transform matrix, where $a_1 = 2$ and $b_1 = 3$.

solution, which is fairly specific, it shows that such a lattice can be created and that other such solutions may exist.

This result is significant because it could be used to inform an experiment in order to

test whether the theory of a 1-D sextupole could come to fruition in currently operating accelerators and storage rings. In particular, because this lattice solution requires only basic accelerator elements, an experiment could be built without the need to buy or build special elements, but simply by repurposing existing elements. However, before an experiment could be proposed, it was important to determine how realistic this solution is. Some considerations included:

- the coupling terms have not vanished entirely, but are relatively small. This was expected, but could be improved.
- It is not optimized for the number of quadrupoles, it may be possible that such a lattice can be created with as little as three quadrupoles.
- During the optimization process, the quadrupole focusing gradients were not constrained, in order to determine if a solution exists. While a solution does exist, the focusing gradients are very large ($k_i > 0.5\text{m}^{-1}$).
- The β functions are too small at the beginning of the lattice. The initial conditions used more reasonable β functions, but during the optimization process these values changed.

After considering these issues, it was prudent to reach out to the members of the CBB collaboration, in particular those interested in working on non-linear beam dynamics, to discuss another approach. Because we still hope to be able to set up an experiment to test this sextupole scheme, another method for designing a 1-D sextupole with greater control over certain parameters was needed.

3.2.1 4-F Imaging

In order to be able to design the 1-D sextupole with realistic parameters, we needed a way to generalize the procedure without relying on optimization, due to the nature of the

constraints. In order to do this, an idea was presented by Jared Maxson to apply the 4f imaging system used in Fourier optics to this problem. This method uses two lenses and four drift spaces, which are the same length as the focal length of the lenses. The 4f imaging system, also called a 4f correlator creates the FT of an image at the midway point of the system, which can be used to determine the correlation between the image and a mask placed at the midway point (34). The complete transfer matrix of this system is a negative identity transformation. The one-dimensional sextupole also consists of an identity transformation in two parts, consisting of a diagonal matrix and its inverse, with sextupole magnets in between. In this way, the 4f correlator could be considered in analogy to the one-dimensional sextupole. However, in order to create the lens used for optical wavelengths, two accelerator magnets must be used; one each for focusing in the x and y directions. Therefore, the accelerator lattice equivalent system will consist of four quadrupole magnets, instead of just two.

Below are the steps taken to determine an analytic solution to creating a diagonal matrix D with the necessary properties outline in previous theory section.

Analytic Solution

A general formula for a diagonal matrix with diagonal elements related as show in Eq. 3.1, can be calculated by the following method. The general lattice will be constructed using the transfer matrices for drift and quadrupole elements, with variable lengths/element widths, and focusing gradients. The general transfer matrices are:

$$\text{drift} = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$$

$$\text{quad}_{\text{focusing}} = \begin{pmatrix} \cos(\sqrt{|k|}s) & \frac{1}{|k|}\sin(\sqrt{|k|}s) \\ -|k|\sin(\sqrt{|k|}s) & \cos(\sqrt{|k|}s) \end{pmatrix}$$

$$\text{quad}_{\text{defocusing}} = \begin{pmatrix} \cosh(\sqrt{|k|}s) & \frac{1}{|k|}\sinh(\sqrt{|k|}s) \\ -|k|\sinh(\sqrt{|k|}s) & \cosh(\sqrt{|k|}s) \end{pmatrix}$$

where s is the length or width of the element, and k is the focusing gradient, which is negative for a focusing quadrupole and positive in a defocusing quadrupole. The full lattices are:

The first half of the lattice consists of a series of drift matrices and two quadrupoles. While the focusing gradient is always a positive quantity, in order to create a general solution in which the defocusing and focusing magnets could be in any order, the property

$$\cos(ix) = \cosh(x) \tag{3.23}$$

was exploited. Mathematically, by using an imaginary value for the focusing gradient k, we could create an extremely general solution without having to resolve the equations for a different quadrupole configuration.

$$\text{lat}_1 = \begin{pmatrix} 1 & L_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\sqrt{k_{1,1}}L_{1,1}) & \frac{\sin(\sqrt{k_{1,1}}L_{1,1})}{k_{1,1}} \\ -k_{1,1}\sin(\sqrt{k_{1,1}}L_{1,1}) & \cos(\sqrt{k_{1,1}}L_{1,1}) \end{pmatrix} \begin{pmatrix} 1 & L_{mid,1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\sqrt{k_{1,2}}L_{1,2}) & \frac{\sin(\sqrt{k_{1,2}}L_{1,2})}{k_{1,2}} \\ -k_{1,2}\sin(\sqrt{k_{1,2}}L_{1,2}) & \cos(\sqrt{k_{1,2}}L_{1,2}) \end{pmatrix} \begin{pmatrix} 1 & L_2 \\ 0 & 1 \end{pmatrix} \tag{3.24}$$

The second half:

$$\text{lat}_2 = \begin{pmatrix} 1 & L_3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\sqrt{|k_{2,1}|}L_{2,1}) & \frac{\sin(\sqrt{k_{2,1}}L_{2,1})}{|k_{2,1}|} \\ -|k_{2,1}|\sin(\sqrt{|k_{2,1}|}L_{2,1}) & \cos(\sqrt{|k_{2,1}|}L_{2,1}) \end{pmatrix} \begin{pmatrix} 1 & L_{mid,2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\sqrt{|k_{2,2}|}L_{2,2}) & \frac{\sin(\sqrt{|k_{2,2}|}L_{2,2})}{|k_{2,2}|} \\ -|k_{2,2}|\sin(\sqrt{|k_{2,2}|}L_{2,2}) & \cos(\sqrt{|k_{2,2}|}L_{2,2}) \end{pmatrix} \begin{pmatrix} 1 & L_4 \\ 0 & 1 \end{pmatrix} \tag{3.25}$$

where the complete lattice consisting both lattices consecutively, such that:

$$lat_1 lat_2 = \begin{pmatrix} a_1 & 0 \\ 0 & 1/a_1 \end{pmatrix}. \tag{3.26}$$

Each lattice individually results in a general matrix with both diagonal and off-diagonal terms. In order to force that the resulting complete transfer matrix, from both parts of the lattice is diagonal, we can force that the transfer matrices of lat_1 and lat_2 are entirely off-

diagonal, thus resulting in a diagonal final matrix. By picking a few parameters (a minimum of two) that have less strict design constraints, we can solve this system for those parameters in terms of the highly constrained parameters. In our case, there is a very narrow range of physically attainable quadrupole focusing gradients, as well as the range of quadrupole widths. Therefore, we can solve this system to obtain formulas for the drift lengths, in terms of the focusing gradients and the quadrupole widths.

The first step to obtain the desired formulas, is to solve for the drift lengths that will result in the diagonal terms in both lat_1 and lat_2 vanishing, so the transfer matrices are entirely off-diagonal, shown in Eq. 3.27. For lat_1 we solved the system for L_1 and $L_{mid,1}$, and for lat_2, we solved for L_3 and $L_{mid,2}$.

$$lat_i = \begin{pmatrix} 0 & f(L_{i,j}, k_{i,j}, L_n) \\ g(L_{i,j}, k_{i,j}, L_n) & 0 \end{pmatrix} \tag{3.27}$$

where f and g are elements which are functions of the parameters introduced in Eq. 3.24 and 3.25. Solving these equations using the Wolfram Mathematica software (**?**), yields the following results (where we have made the change of variables $\sqrt{|k_{i,j}|} \to k_{i,j}$ for simplicity in notation):

$$L_1 = -\frac{\operatorname{csch}^2(k_{1,2}L_{1,2})\big(k_{1,1}\big(\sin(2k_{1,1}L_{1,1}) - 2k_{1,1}L_2\sin^2(k_{1,1}L_{1,1})\big) + k_{1,2}\sinh(2k_{1,2}L_{1,2})\big)}{2k_{1,2}^2} \tag{3.28}$$

$$L_{mid,1} = \frac{\sin(k_{1,1}L_{1,1})\Big(k_{1,2} - k_{1,1}^2 L_2\coth(k_{1,2}L_{1,2})\Big) + k_{1,1}\cos(k_{1,1}l_{1,1})(\coth(k_{1,2}l_{1,2}) + k_{1,2}L_2)}{k_{1,1}k_{1,2}(k_{1,1}L_2\sin(k_{1,1}L_{1,1}) - \cos(k_{1,1}L_{1,1}))} \tag{3.29}$$

$$L_3 = -\frac{\operatorname{csch}^2(k_{2,2}L_{2,2})\big(k_{2,1}\big(\sin(2k_{2,1}L_{2,1}) - 2k_{2,1}L_4\sin^2(k_{2,1}L_{2,1})\big) + k_{2,2}\sinh(2k_{2,2}L_{2,2})\big)}{2k_{2,2}^2} \tag{3.30}$$

$$L_{mid,2} = \frac{\sin(k_{2,1}L_{2,1})\Big(k_{2,2} - k_{2,1}^2 L_4\coth(k_{2,2}L_{2,2})\Big) + k_{2,1}\cos(k_{2,1}l_{2,1})(\coth(k_{2,2}l_{2,2}) + k_{2,2}L_4)}{k_{2,1}k_{2,2}(k_{2,1}L_4\sin(k_{2,1}L_{2,1}) - \cos(k_{2,1}L_{2,1}))} \tag{3.31}$$

Once we solved for the drift lengths that will force the diagonal terms in lat_1 and lat_2

to vanish, we plugged these values back into the transfer matrices, and determined the new lattice matrices:

$$lat_1 = \begin{pmatrix} 0 & \frac{\operatorname{csch}(k_{1,2}l_{1,2})(k_{1,1}L_2\sin(k_{1,1}l_{1,1})-\cos(k_{1,1}l_{1,1}))}{k_{1,2}} \\ \frac{k_{1,2}\sinh(k_{1,2}l_{1,2})}{\cos(k_{1,1}l_{1,1})-k_{1,1}L_2\sin(k_{1,1}l_{1,1})} & 0 \end{pmatrix} \tag{3.32}$$

and similarly for the second part:

$$lat_2 = \begin{pmatrix} 0 & \frac{\operatorname{csch}(k_{2,2}l_{2,2})(k_{2,1}L_4\sin(k_{2,1}l_{2,1})-\cos(k_{2,1}l_{2,1}))}{k_{2,2}} \\ \frac{k_{2,2}\sinh(k_{2,2}l_{2,2})}{\cos(k_{2,1}l_{2,1})-k_{2,1}L_4\sin(k_{2,1}l_{2,1})} & 0 \end{pmatrix} \tag{3.33}$$

The resulting diagonal matrix is therefore:

$$lat_1 lat_2 = \begin{pmatrix} \frac{k_{1,2}\operatorname{csch}(k_{2,2}L_{2,2})(k_{2,1}L_4\sin(k_{2,1}L_{2,1})-\cos(k_{2,1}L_{2,1}))\sinh(k_{1,2}L_{1,2})}{k_{2,2}(\cos(k_{1,1}l_{1,1})-k_{1,1}L_2\sin(k_{1,1}L_{1,1}))} & 0 \\ 0 & \frac{k_{2,2}\operatorname{csch}(k_{1,2}l_{1,2})(k_{1,1}L_2\sin(k_{1,1}l_{1,1})-\cos(k_{1,1}L_{1,1}))\sinh(k_{2,2}L_{2,2})}{k_{1,2}(\cos(k_{2,1}L_{2,1})-k_{2,1}L_4\sin(k_{2,1}L_{2,1}))} \end{pmatrix} \tag{3.34}$$

Numerical Solution

We can confirm that in the "symmetric" case, where the quadrupole focusing strengths are all the same, D is the negative identity transform. This can be easily confirmed by setting the parameters as shown in Table 3.1

Table 3.1: From the listed parameter values, the full diagonal transfer matrix resulting from Eq. 3.34 results in the negative identity transformation.

Parameter	Value
$k_{i,j}$	$\sqrt{0.5}m^{-1}$
$l_{i,j}$	$0.4m$
L_2, L_4	$1m$

The diagonal elements can be tuned by either solving the elements for a specific value of a, or by setting the quadrupole strengths, quadrupole widths, and drift lengths to the

necessary lengths.

We can look at a specific example, with physically reasonable values for the focusing gradients, widths, and drift lengths. We can assign these values, and calculate the resulting drift values L_1, L_3, $L_{mid,1}$, and L_{mid_2}, as well as the resulting a_1 value, shown in Table 3.2.

Table 3.2: From the listed parameter values, we can calculate an example of the necessary drift lengths, and the resulting a_1 value.

Parameter	Value
$k_{1,1}$	$0.3m^{-1}$
$k_{1,2}$	$-0.3m^{-1}$
$k_{2,1}$	$0.25m^{-1}$
$k_{2,2}$	$-0.25m^{-1}$
$l_{i,j}$	$0.4m$
L_2, L_4	$1m$

Using the parameters listed above, the necessary drift lengths are: $L_1 = 1$m, $L_3 = 1$m, and $L_{mid,1} = 6.6$m, $L_{mid,2} = 8.3$m. The resulting a_1 value is $a_1 = -1.2315$; alternatively we could tune these values and plug them into Eq. 3.34 to reach a desired value for a_1. The diagonal elements of Eq. 3.34 could be solved further such that the value of a_1 could be set, and a free parameter such as L_2 or L_4 could be set accordingly.

This formulation, while it allows us to control the focusing gradients and quadrupole widths, the resulting drift lengths are slightly longer than ideal. It is unclear if this kind of a 1-d sextupole could be viable, thus further study into the feasibility of a 1-d sextupole, by studying other realistic beam parameters, was necessary.

3.3 Calculations Including Energy Spread

In order to gauge if a 1-D sextupole is a practical option, it is prudent to determine how the system will change when we consider realistic beam parameters. Figure 4.2 demonstrates how the phase space coordinates for the centroid of a bunch would be transformed in a 1-D sextupole, and the effect of beam energy spread which results in a non-zero coupling between the horizontal and vertical motion. It is important to understand the effect of beam energy

spread and how it can be mitigated. The calculations shown earlier only consider on-axis ($\Delta x/x = 0$) and on-energy ($\Delta p/p = 0$) particles. In order to determine the magnitude of the effects energy spread has in a 1-D sextupole, we recalculated the theoretical values by introducing a detuning parameter, δ. This analysis and evaluation of a numerical examples using realistic beam and lattice parameters is presented here.

3.3.1 Change in the Transfer Matrix

Starting from Nagaitsev's analysis for a lattice solution consisting of 4 quadrupole magnets and 3 drift lengths, the general solution where k_i is the i^{th} magnet focusing strength ($\propto 1/f$), and l_i is the i^{th} drift length:

$$\mathrm{A} = \begin{pmatrix} 1 & 0 \\ k_4 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ k_3 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ k_2 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ k_1 & 1 \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix} \tag{3.35}$$

and for the orthogonal direction:

$$\mathrm{B} = \begin{pmatrix} 1 & 0 \\ -k_4 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -k_3 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -k_2 & 1 \end{pmatrix} \begin{pmatrix} 1 & L_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -k_1 & 1 \end{pmatrix} = \begin{pmatrix} b & 0 \\ 0 & 1/b \end{pmatrix} \tag{3.36}$$

We will use the solutions that Nagaitsev determined, that were confirmed earlier, for the focusing gradients, k_i, and drift lengths l_i:

$$l_1 = l_1[\mathrm{m}] \tag{3.37}$$

$$l_2 = -l_1 \frac{(a-b)^2}{2(-2ab+a+b)}[\mathrm{m}] \tag{3.38}$$

$$l_3 = l_1 \frac{ab(a+b-2)}{-2ab+a+b}[\mathrm{m}] \tag{3.39}$$

$$k_1 = \frac{1}{l_1}\frac{a^3+a}{a+b} - \frac{(a-1)^3}{a+b-2} + \frac{a(a-1)}{b-a} - \frac{1}{2} \tag{3.40}$$

$$k_2 = \frac{1}{l_1}\frac{a+b-2}{b-a} \tag{3.41}$$

$$k_3 = \frac{1}{l_1}\frac{(-2ab+a+b)^2}{ab(a-b)(a+b-2)} \tag{3.42}$$

$$k_4 = \frac{1}{l_1}\frac{a^3(4b-1) - 7a^2b + ab(b(4b-7)+8) - b^3}{2ab(a+b-2)\left(a^2-b^2\right)} \tag{3.43}$$

In order to determine the effect of detuning on the transfer matrix needed to create a 1-D sextupole, we can introduce a detuning parameter,δ, representing the change in the focusing gradient seen by an off-energy particle ($\frac{\Delta p}{p} \neq 0$):

$$\mathrm{A}_\delta = \begin{pmatrix} 1 & 0 \\ k_4(1+\delta) & 1 \end{pmatrix} \begin{pmatrix} 1 & L_3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ k_3(1+\delta) & 1 \end{pmatrix} \begin{pmatrix} 1 & L_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ k_2(1+\delta) & 1 \end{pmatrix} \begin{pmatrix} 1 & L_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ k_1(1+\delta) & 1 \end{pmatrix} \tag{3.44}$$

To determine the effect of the detuning, the residual, $\Delta T = A - A_\delta$, was calculated in one direction (the calculations were done in 2 dimensions, but we can study one direction to determine the general effect and validity of this method):

$$\Delta T = A - A_\delta = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \tag{3.45}$$

then, by multiplying out each element, fulling expanding the expressions and keeping

only the terms linear in δ, the results are:

$$r = \frac{2a\delta\left(a^2 - 2a((b-1)b+1) + b^2\right)}{(a-b)(a+b)} \tag{3.46}$$

$$s = -\delta(-2+a+b)l_1 \tag{3.47}$$

$$t = \frac{\delta}{4l_1 ab(a+b-2)\left(a^2-b^2\right)^2}(a^6(1-8b) + 30a^5 b - 3a^4 b(b(16b-21)+24) \tag{3.48}$$

$$+ 4a^3 b(b(b(16(b-2)b+49)-32)+16) - 3a^2 b^3(3b(8b-7)+16) + 30ab^5 + (b-8)b^5) \tag{3.49}$$

$$u = \frac{2\delta\left(a^2 - 2a((b-1)b+1) + b^2\right)}{a(a-b)(a+b)} \tag{3.50}$$

The matrix *res* is the change in the transfer matrix, and could be used to determine whether the 1-D sextupole could be a viable option even when realistic beam parameters are considered. Calculations using the results Eqd. 3.47-3.50 with potential beam parameters will be done below.

3.3.2 Change in the State Vector

It is also prudent to have an estimate of how the state vector of a particle traversing through a one-dimensional sextupole. Using the same lattice, A, defined earlier, and the lattice with the detuning parameter,A_δ , we can calculate:

$$L_0 = AS_1A^{-1}S_2 \tag{3.51}$$

$$L_\delta = A_\delta S_1 A_\delta^{-1} S_2 \tag{3.52}$$

where:

$$S_1 = \begin{bmatrix} x \\ p_x + q_1(x^2 - y^2) \\ y \\ p_y - 2q_1xy \end{bmatrix} \tag{3.53}$$

$$S_2 = \begin{bmatrix} a_1x \\ p_x + \frac{q_1(x^2-y^2)}{a_1} + q_2[(a_1x)^2 - (b_1y)^2] \\ b_1y \\ \frac{p_y-2q_1xy}{b_1} - 2q_2a_1b_1xy \end{bmatrix}. \tag{3.54}$$

After calculating both L_0 and L_δ, and keeping on the terms first order in the detuning parameter we obtain an expression for the change in the state vector. The change due to the energy spread, $\Delta S = L_0 - L_{delta}$ is:

$$\Delta S = \begin{pmatrix} \delta q_2(a+b-2)\left(x^2+y^2\right) \\ -\frac{2a\delta q_2\left(a^2-2a((b-1)b+1)+b^2\right)\left(x^2+y^2\right)}{(a-b)(a+b)} \\ -2\delta q_2xy(a+b-2) \\ -\frac{4b\delta q_2xy\left(a^2-2((a-1)a+1)b+b^2\right)}{(a-b)(a+b)} \end{pmatrix} \tag{3.55}$$

In order to confirm that keeping on terms that are first order in delta is valid, we use numerical values for a, b, δ, l_1, and compare the results. Further, looking at the numerical values of the changes in the transfer matrix and the state vector will give us an idea of the significance of adding energy spread to this theory.

3.3.3 Numerical Optimization to Minimize Energy Spread Effects

Because the specific values of a and b in the diagonal transformation are not as important as maintaining the conditions for the diagonal transformation, they can be varied as needed. In order to minimize the contribution from beam energy spread, the parameters a and b were varied such that the elements in ΔT were minimized.

By optimizing the system for various initial a and b settings, and for $L_1 = 1\text{m}, 0.75\text{m}, 0.5\text{m}, 0.25\text{m}$ and energy detuning $\delta = 10^{-3}$, which are reasonable drift lengths for a short portion of a long linear or circular machine, the following settings were determined to minimize the energy spread contribution to the diagonal transformation matrix. The merit of these solutions is demonstrated by the percent increase of Δp_y from 0, when the parameters are applied to the theory which considers beam energy spread. The percent change was calculated by considering the solutions shown in Table 1, for the following initial phase space parameters: $x_0 = 0.5\text{mm}, \text{y}_0 = 0.05\text{mm}, \text{p}_\text{x} = 0, \text{p}_\text{y} = 0$, and the sextupole currents scaled such that $q_1 = 1, q_2 = \frac{-1}{ab^2}$.

The parameters listed in Table 1 show that the drift lengths are reasonable, though the focusing gradients may not be. The ideal solution would be the one in which $\Delta p_y/\Delta p_x$ is minimal, such as the solution for $a = -7.5, b = -0.4$. However, the focusing gradients in the quadrupoles for this solution are very high and may not be realizable in a normal-conducting machine. Still, this solution is further ideal due the short drift lengths needed to construct it. A short 1-D sextupole would be desirable because it could be easily inserted into an existing machine.

Table 3.3: Shown are the thin lens solution parameters which minimize the change in the diagonal 1D sextupole transformation matrix, for $\delta = 10^{-3}$.

a	**b**	L_1	L_2	L_3	k_1	k_2	k_3	k_4	$\Delta p_y/\Delta p_x(\%)$
-2.9	-0.2	1.00m	0.78m	0.78m	$0.90m^{-1}$	$-1.94m^{-1}$	$2.18m^{-1}$	$-1.17m^{-1}$	0.68%
-3.5	-0.3	0.75m	0.66m	0.76m	$1.21m^{-1}$	$-2.40m^{-1}$	$2.39m^{-1}$	$-1.19m^{-1}$	0.86%
-4.8	-0.4	0.50m	0.57m	0.72m	$1.77m^{-1}$	$-3.24m^{-1}$	$2.72m^{-1}$	$-1.23m^{-1}$	0.87%
-7.5	-0.4	0.25m	0.44m	0.55m	$3.24m^{-1}$	$-5.61m^{-1}$	$3.67m^{-1}$	$-1.53m^{-1}$	0.49%

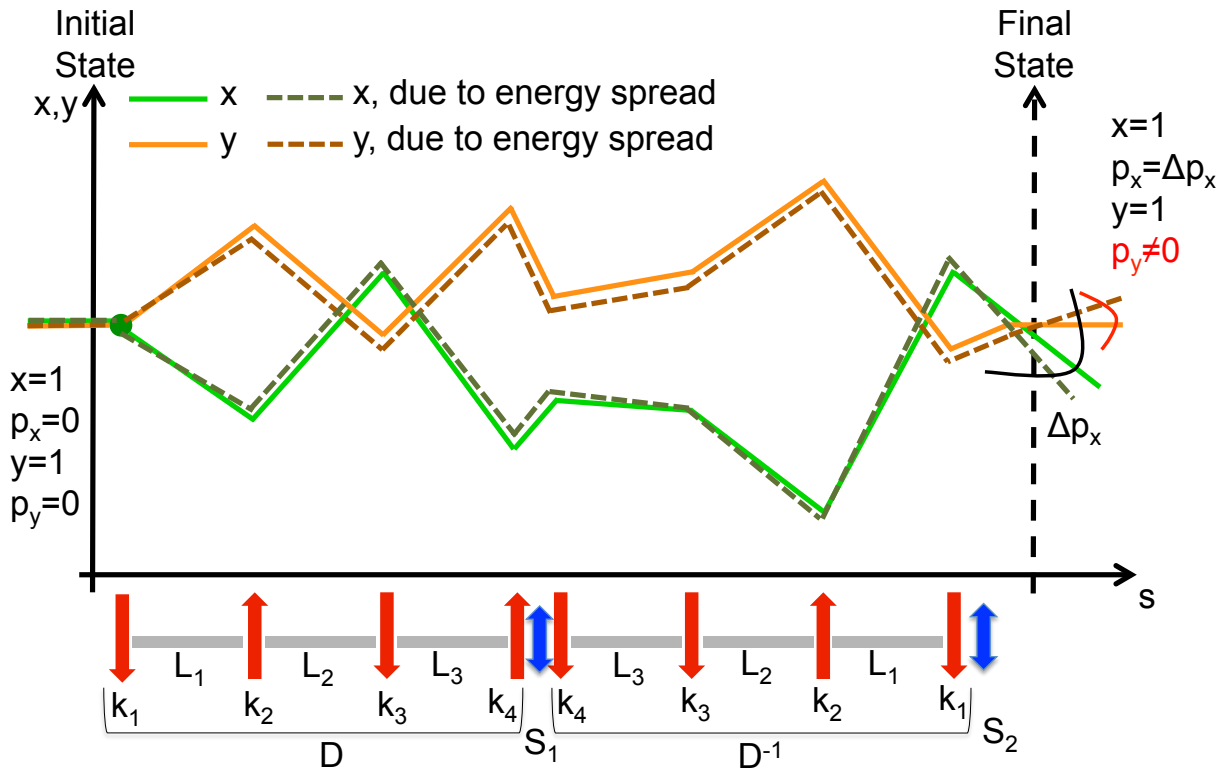


Figure 3.4: A simple example of how a 1-D sextupole transforms the phase space coordinates for the centroid of a bunch, with and without beam energy spread. A particle with $x_0 = y_0,\ p_x = p_y = 0$ entering the 1-D sextupole would exit with $x_0 = y_0,\ p_x = \Delta p_x,\ p_y = 0$, but particles that are off-energy would end up with $x_0 = y_0,\ p_x = \Delta p_x,\ p_y \neq 0$.

However, further optimization could yield other similar solutions with physically realizable quadrupole strengths, that still minimize the momentum coupling due to beam energy spread. The solutions presented therefore display the broad solution space available for 1-D sextupoles parameters.

3.4 Applications of the 1-D Sextupole

The 1-D sextupole is a unique and simple method for mitigating the sextupole coupling introduced by traditional sextupoles. Within this chapter, this method has been expanded, making it easier to design a 1-D sextupole channel for a real machine. The goal was to start with the theory of the 1-D sextupole, and determine a general method or algorithm for creating a 1-D sextupole. The general approach was used to create some initial 1-D

sextupole solutions. In order to create lattice solutions which allowed for more control over lattice parameters, in order to approach a realistic and experimentally feasible solution, we developed another method of designing the sextupole. This was done by using the 4f imaging scheme to inform a procedure in which the focusing gradients of the magnets could specified, without limiting the generality of the 1-D sextupole. This method, while very useful, still suggested that 1-D sextupole lattices may end up being much longer than is feasible in a real accelerator. Continuing in understanding the viability of a 1-D sextupole, the theory was recalculated with an added energy spread term. These calculations suggest that a realistic beam in which there is a spread in the particle energies, might disrupt the trajectory of the beam significantly, thus this method in its current form may not be viable.

Despite these limitations, the expansion of this method makes it possible to attempt an experimental verification of the uncoupled sextupole behavior. At facilities such as IOTA, which are built for the purpose of studying nonlinear beam optics, it may be possible to build a sextupole channel and measure to what extent the coupling is eliminated.

CHAPTER 4

METHODS FOR RESONANCE MITIGATION

4.1 Resonance Elimination Via Semi-Analytic Methods

Methods for eliminating or minimizing nonlinear resonances in accelerators are of interest to reduce the challenges associated with nonlinear beam design. Sextupole magnets are necessary but excite the 3rd order harmonic due to their $\propto x^3$ potential, developing methods to reduce this resonance is of particular interest. The phase space of an accelerator which uses many sextupole magnets is therefore not only prone to the 3rd order harmonic resonance, but also higher order resonances which may appear within the volume of allowed trajectories. This volume is determined by the position of the hyperbolic point, which creates a separatrix beyond which all trajectories diverge to infinity.

While the purpose of using sextupoles in accelerator lattices is primarily to correct chromaticity, here we explore the way in which the third order resonance can be eliminated in a 1-dimensional toy model, without regard to chromaticity compensation. This exercise is based on the findings of Danilov and Nagaitsev (35), in which they showed how to achieve a time independent Hamiltonian in normalized coordinates in the limit of a smooth distribution of nonlinear magnets.

4.1.1 Theoretical Background

As shown in Danilov and Nagaitsev (DN) (35), a transformation exists which can results in a time-dependent integrable Hamiltonian for a simple accelerator lattice consisting of drifts, thin sextupoles, and a focusing element. The full derivation, shown in (35), reveals that appropriately scaling the magnet strength distribution by a power of the β function through the channel will result in resonance elimination. In the case of sextupoles, once the transformation to normalized coordinates:

$$x_N = \frac{x_N}{\sqrt{\beta_x}}, \quad p_{x,N} = p\sqrt{\beta_x} + \frac{\beta_x{}' x}{2\sqrt{\beta_x}}, \tag{4.1}$$

is performed, the effective Hamiltonian is then:

$$H_N = \frac{{p_{x,N}}^2}{2} + \frac{{x_N}^2}{2} + \beta_x(s) V(x_N\sqrt{\beta_x}, s). \tag{4.2}$$

Therefore, for a sextupole potential $V(x) \sim \frac{\alpha x^3}{3}$,

$$V(x_N) \propto \beta_x(s) \frac{(\sqrt{\beta(s)}x)^3}{3} K_3(s). \tag{4.3}$$

From here we can see that by choosing a sextupole distribution $K_3(s)$ to cancel out the s dependence in the potential:

$$K_3(s) \sim \frac{{K_3}^{(0)}}{\beta_x(s)^{\frac{5}{2}}}, \tag{4.4}$$

the desired time-independent $V(x_N)$ can be achieved.

From this theory, we can see that by distributing sextupoles in a drift space with their strength proportional to $\beta_x(s)^{-5/2}$, we can eliminate the time-dependence of the potential in the Hamiltonian, and create an invariant.

Beside the DN prescription, another suggestion for eliminating the 3rd order resonance comes from simply observing the nature of the 3rd order resonance driving term (RDT). The magnitude of the RDT associated with the 3rd order sextupole resonance is defined as:

$$h_3 = \frac{-1}{24} \sum_{i=1}^{N} K_3(s_i) \beta(s_i)^{3/2} e^{i3\mu(s_i)} \tag{4.5}$$

From this expression, we became curious if modifying the DN solution of $K_3(s) \propto \beta_x(s)^{-5/2}$ to $K_3(s) \propto \beta_x(s)^{-3/2}$ to cancel out the β function dependence in the RDT would also help with resonance elimination. In order to achieve similar time-independence as in the DN solution, it was also determined that the sextupoles should be spaced such that the

phase advance between sextupoles through the channel, $\mu(s_i)$, was constant. In contrast, the DN solution does not have spatial symmetry requirements (defined as an equidistant distribution of sextupoles) as described in the original work. Therefore, to demonstrate the DN solution we constructed the channel with the sextupoles placed $\frac{L}{N_{sex}}$, where N_{sex} is the number of sextupoles in the channel, as long as the strength of the magnets is determined by the $\beta_x(s)^{-5/2}$ proportionality.

To illustrate this better, the following figure shows general schematics for these two toy model solutions:

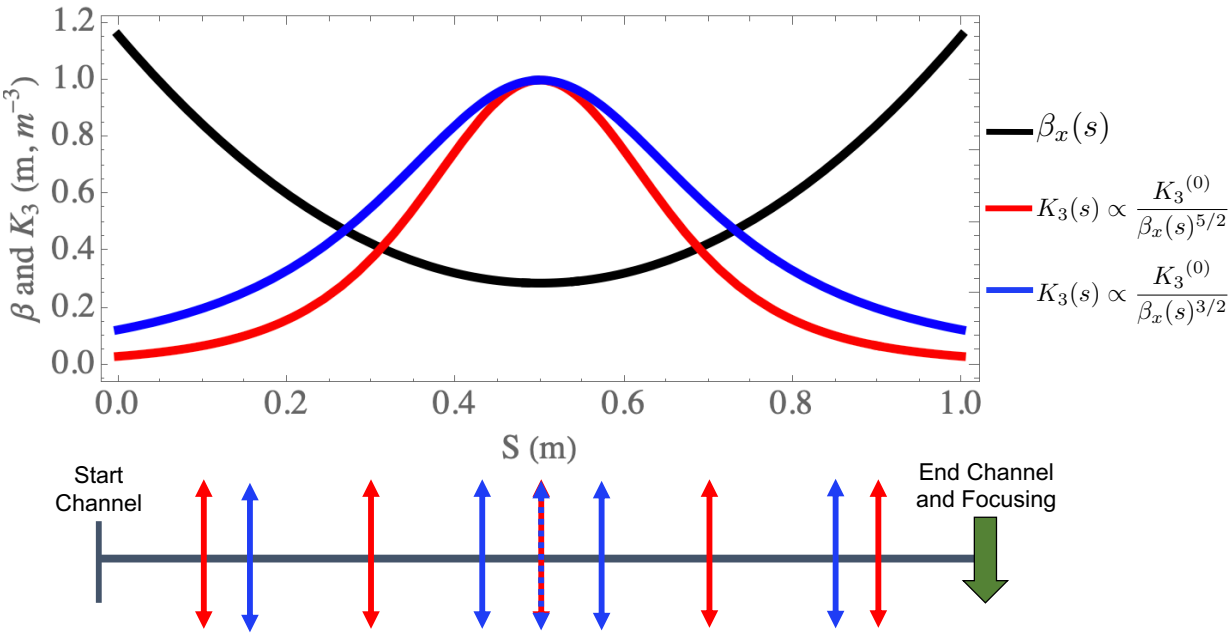


Figure 4.1: A schematic representation of the sextupole channel, where the β function and the two suggested sextupole strength distributions $K_3(s) \propto \beta_x(s)^{-5/2}$ and $K_3(s) \propto \beta_x(s)^{-3/2}$ are shown in red and blue respectively. Beneath the plot is a visualization of the sextupole positioning in real space along the length of the channel. The red double-ended arrows represent the spatially symmetric sextupole distribution, and the blue doubled-ended arrows represent the constant phase advance spacing; the colors also correspond to the magnetic strength distribution used.

4.1.2 1-D Matched Optics Channel

Here, we demonstrate how the DN "matched-optics" prescription is constructed at tunes, ν, very close to $2\pi\frac{1}{3}$. To demonstrate this, a drift-kick lattice was constructed in Mathematica (**?**) in the following way, to track particles in (x, p_x) phase space.

First, the channel consists of a drift space which can be modified to included the desired number of thin, zero-length, sextupoles. At the end of the channel is a thin element which provides linear focusing. While we consider only a 1-D example here, in the 2-D extension the beta functions are matched exactly and there is linear focusing provided in both transverse planes.

The beta function for a drift length L, and focusing k is given by:

$$\beta_x(s) = \frac{L - sk(L-s)}{\sqrt{1-(1-\frac{Lk}{2})^2}}. \tag{4.6}$$

The DN map is constructed in a drift-kick model, in which the drift map is the simple 1-D drift transfer map, and the focusing at the end of the channel is a standard focusing transformation. Details for these maps can be found in (3). The linear focusing constant k is determined by the desired betatron phase advance and tune in the following way. The phase advance is calculated $\phi = (\frac{1}{3} + \delta)2\pi$ for detuning δ away from the 1/3 resonance. The linear focusing is then calculated from the phase advance as:

$$k = -2\frac{\cos(\nu) - 1}{L}. \tag{4.7}$$

The non-linear kick provided by a sextupole at position s_i along the channel is shown in Eq. 4.8, with the sextupole strength given by the function $K_3(s)$.

$$\begin{pmatrix} x_{i+1} \\ p_{x,i+1} \end{pmatrix} = \begin{pmatrix} x_i \\ p_{x,i} + {x_i}^2 K_3(s_i) \end{pmatrix}. \tag{4.8}$$

A tracking code was written in Mathematica to transform a particle initialized in (x, p_x) phase space through the channel.

4.1.3 Resonance Eliminated

Applying the DN solution directly, various channels were created and phase space portraits calculated, in order to determine the least number of sextupoles possible for resonance elimination. It was observed that the resonance was not eliminated as initial suggested.

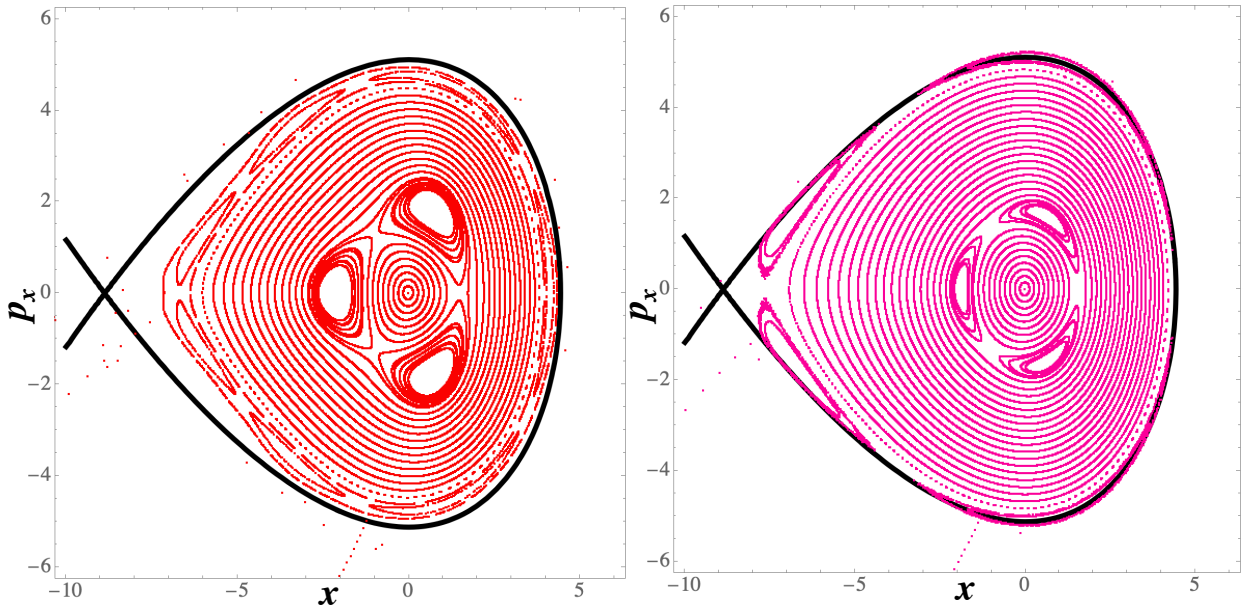


Figure 4.2: Phase space portraits for 1-dimensional particle tracking through a sextupole channel of unit length, for 5 thin sextupoles; (left) the DN solution, (right) the DN solution with RDT optimization for $K_3(s) \propto \beta_x(s)^{-\alpha}$, and $\alpha = 2.12$. The tune ν_x is set to be $1/3+\delta$ for $\delta = 0.005$, which results in a total phase advance for the cell $\phi = 2.12$.

This can be seen clearly in fig. 4.2, where large resonance islands are visible near the origin, and the region of closed orbits near the origin is small and slightly deformed due to the presence of the resonance. We applied RDT minimization to further reduce the magnitude of the resonance. This minimization was done in Mathematica by searching for the appropriate exponent α for $K_3(s) \propto \beta_x(s)^{-\alpha}$ which would further reduce the RDT of the system. The

results of this RDT optimization are shown in fig. 4.2. The resonance is still present, but it's strength is reduced, as shown by the decrease in the size of the islands. Therefore, it is clear that there is yet an optimal configuration.

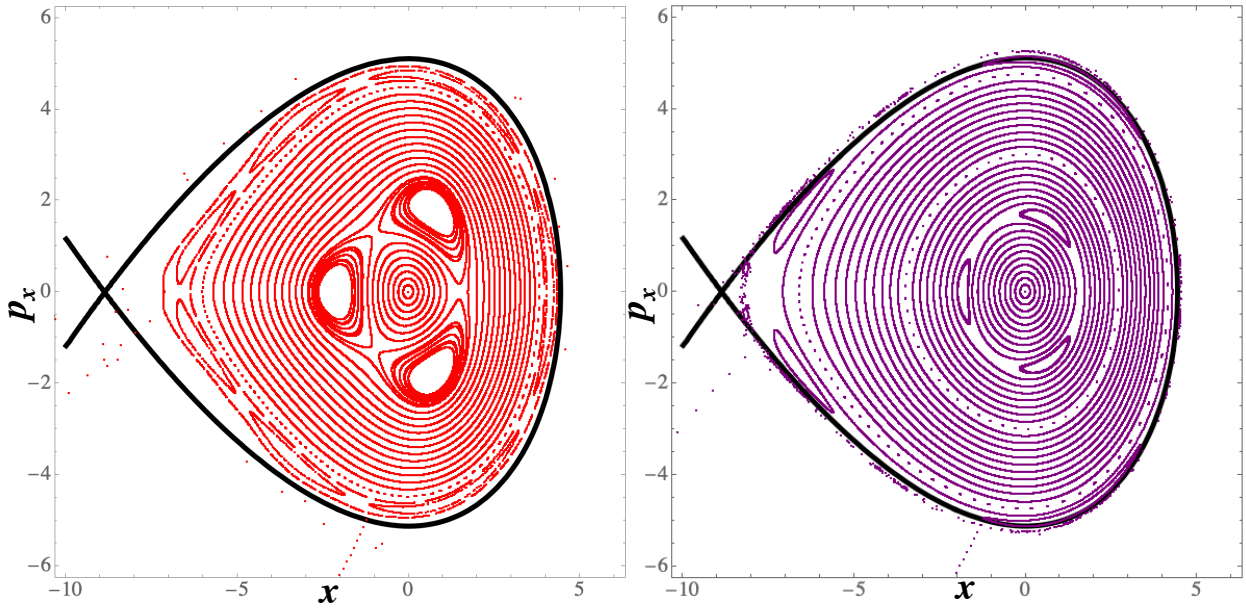


Figure 4.3: Phase space portraits for 1-dimensional particle tracking through a sextupole channel of unit length, for 5 thin sextupoles; (left) the DN solution with asymmetric sextupole spacing and, (right) the DN solution with restored spatial symmetry as shown in fig. 6.20. The tune ν_x is set to be $1/3 + \delta$ for $\delta = 0.005$, which results in a total phase advance for the cell $\phi = 2.12$.

Due to the observation that $K_3(s)$ must be changed significantly in order to improve the dynamics, we tried adjusting the physical spacing of the sextupoles in the channel such that the periodicity is maintained with each cycle. While the sextupoles were placed with equidistant spacing, the symmetry was broken at the periodic point, where the distance was $2\frac{L}{N_{sex}}$ instead of $\frac{L}{N_{sex}}$. After restoring the spatial symmetry, but retaining $K_3(s) \propto \beta_x(s)^{-5/2}$, it was determined that the symmetry of the channel is vital for resonance mitigation. The results of this adjustment are shown in fig. 4.3.

With further observation, and by exploring the behavior Eq. 4.5, and based on the determination that the spatial positions of the sextupoles did contribute to the resonance

mitigation, it was clear that the phase advance between the sextupoles was an important parameter. Similarly, from Eq. 4.5, it is clear that the time-dependence (or s dependence) can be cancelled out by matching the power of the β function dependence, and setting $K_3(s) \propto \beta_x(s)^{-3/2}$. By calculating a sextupole mesh along the channel such that the phase advance between the sextupoles was constant, the following results were observed:

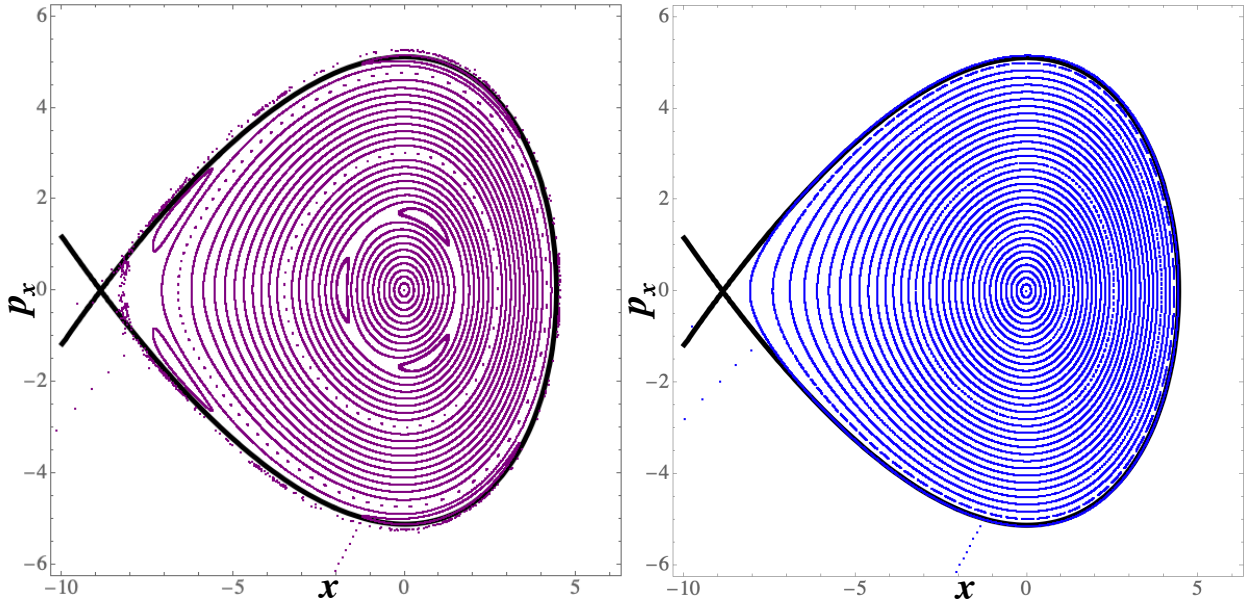


Figure 4.4: Phase space portraits for 1-dimensional particle tracking through a sextupole channel of unit length, for 5 thin sextupoles; (left) the DN solution with asymmetric sextupole spacing and, (right) the equal phase advance sextupole placement as shown in fig. 6.20 and $K_3(s) \propto \beta_x(s)^{-3/2}$. The tune ν_x is set to be $1/3 + \delta$ for $\delta = 0.005$, which results in a total phase advance for the cell $\phi = 2.12$.

From the results shown in fig. 4.4, it is clear that by maintaining periodicity by ensuring that the phase advance between sextupoles is constant and by applying a matching condition for $K_3(s)$, the resonance is entirely eliminated. For this cell with total phase advance $0.3383x2\pi$, with only 3 sextupoles (thus phase advance between sextupoles is no more than $0.12x2\pi$), the resonance is entirely eliminated.

Figure 4.5 shows that with the new prescriptions of $K_3(s) \propto \beta_x(s)^{-3/2}$, with only 3 sextupoles, the third order harmonic resonance is entirely eliminated.

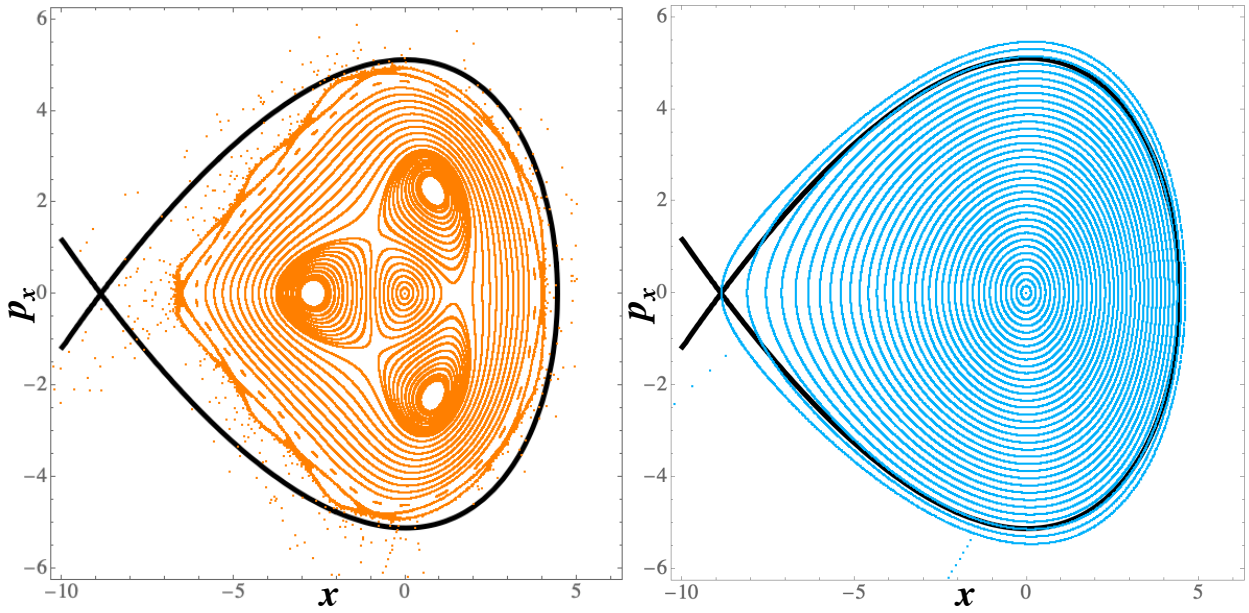


Figure 4.5: Phase space portraits for 1-dimensional particle tracking through a sextupole channel of unit length, (left) for 5 thin sextupoles with DN solution for asymmetric sextupole spacing and, (right) 3 thin sextupoles and the equal phase advance sextupole placement as shown in fig. 6.20 and $K_3(s) \propto \beta_x(s)^{-3/2}$. The tune ν_x is set to be $1/3 + \delta$ for $\delta = 0.005$, which results in a total phase advance for the cell $\phi = 2.12$.

The analysis shown is purely 1-dimensional, with no coupling between the horizontal and vertical coordinates. In a realistic system this cannot be achieved, though it can be emulated by (unmatched beta functions). Therefore, while this method has been demonstrated using particle tracking in phase space, it is prudent to extend this analysis into 2-dimensions. The feasibility of such a system in practice can therefore be determined by constructing realistic lattice in simulation software in order to understand how robust this method is.

4.2 Sextupole Calibration with Beam-Based Measurements

4.2.1 A Simple Model of Beam Response

To develop the mathematical model for determining how turn-by-turn BPM measurements can be analyzed to measure sextupole field errors in the presence of turn by turn tune jitter,

a simplified model (as compared to full RDT analysis) of the spectral response to beam shaking was developed.

We start by considering a simplified nearly-linear CESR lattice with a single sextupole with gradient k_2 and length L. Thus the angular kick in the x-direction is $\delta\theta_x = \frac{K}{2}(x_S^2 + y_S^2)$, where x_S and y_S are the spatial coordinates of the centroid of the beam in the sextupole, and $K = k_2 L$. We define the unperturbed oscillations at the sexutpole to have the form

$$x_S(t) = \sqrt{2A_x\beta_{x,S}}\cos(\omega_x t), \tag{4.9}$$

for action A_x and horizontal beta function at the sextupole, $\beta_{x,S}$. Then, we compute δx_B, the linear order change in the x position at a beam position monitor by summing the angular kicks exerted on the beam at each turn:

$$\begin{aligned}\delta x_B(t) &= \delta\theta(t)\sqrt{\beta_{x,B}\beta_{x,S}}\sin(\Delta\phi_{S\to B}) + \\ &\quad \delta\theta(t-1)\sqrt{\beta_{x,B}\beta_{x,S}}\sin(\omega_x + \Delta\phi_{S\to B}) + \\ &\quad \delta\theta(t-2)\sqrt{\beta_{x,B}\beta_{x,S}}\sin(2\omega_x + \Delta\phi_{S\to B}) + ...,\end{aligned}$$

given the beta functions $\beta_{x,B}$ and $\beta_{x,S}$ at the BPM and sextupole respectively,the betatron phase advance between the sextupole and the BPM $\Delta\phi_{S\to B}$, and ω_x, the one turn phase advance. The sum, written compactly, is

$$\delta x_B(t) = \sum_{k=0}^{\infty}\delta\theta(t-k)\sqrt{\beta_{x,B}\beta_{x,S}}\sin(k\omega_x + \Delta\phi_{S\to B}), \tag{4.10}$$

and has a closed form solution, but is omitted here for compactness. After computing this sum, the Fourier transform of the beam's response at the BPM is performed to determine the magnitude of the resonant tune line in frequency-space. The horizontal response will contain resonances at $2Q_x$ and $2Q_y$, while the vertical direction has resonant lines at $Q_x \pm Q_y$. The computed magnitudes $|\widetilde{C}_{2\omega_x}|$ and $|\widetilde{C}_{\omega_x+\omega_y}|$ are linearly proportional to K. Therefore, in

order to determine the magnitude of K, resonance amplitude measurements can be fit by this model. Note that this analysis is trivially extended to the case in which the tune varies on each turn, replacing ω_x with the instantaneous tune on a given turn k, $\omega_{x,k}$. The tune tracker system at CESR reports this phase each turn, and thus may be directly incorporated in the model.

4.2.2 Simulation results

The above model was applied to simulated BPM measurements for a CESR lattice with a single, strong sextupole. The simulations were run using the Bmad charged particle subroutine library(**?**), in which the beam is displaced horizontally and vertically by 0.5mm in the absence of radiation damping, then tracked for 10,000 turns. In order generate sufficient signal in the betatron harmonics, the single sextupole was set at four times its nominal strength, thus overwhelming the extraneous signals from other nonlinearities in the ring. Alternatively, given the linearity of the above model, the zero-sextupole harmonic response could otherwise be removed as a baseline from the complex response of the beam at a given harmonic (27).

The simulated data was processed by taking the Fourier transform at each BPM and recording the $2Q_x$ harmonic line amplitudes, $|\widetilde{Q}_{2\omega_x}|$ from the horizontal spatial data, and $|\widetilde{Q}_{\omega_x+\omega_y}|$ from the vertical data. In order to fit the data to the model the following merit function was constructed:

$$\begin{aligned}\chi^2 &= \sum^{N} \frac{(|\widetilde{C}_{2\omega_x}| - |\widetilde{Q}_{2\omega_x}|)^2}{\max((|\widetilde{C}_{2\omega_x}| - |\widetilde{Q}_{2\omega_x}|)^2)} + \\ &\quad \frac{(|\widetilde{C}_{\omega_x+\omega_y}| - |\widetilde{Q}_{\omega_x+\omega_y}|)^2}{\max((|\widetilde{C}_{\omega_x+\omega_y}| - |\widetilde{Q}_{\omega_x+\omega_y}|)^2)}\end{aligned} \tag{4.11}$$

A simple fitting routine was used to minimize Eq. 4.11, by varying the amplitude and phase of the spectral line. The fitted Fourier amplitudes for the simulation are shown in Fig. 4.6 and Fig. 4.7.

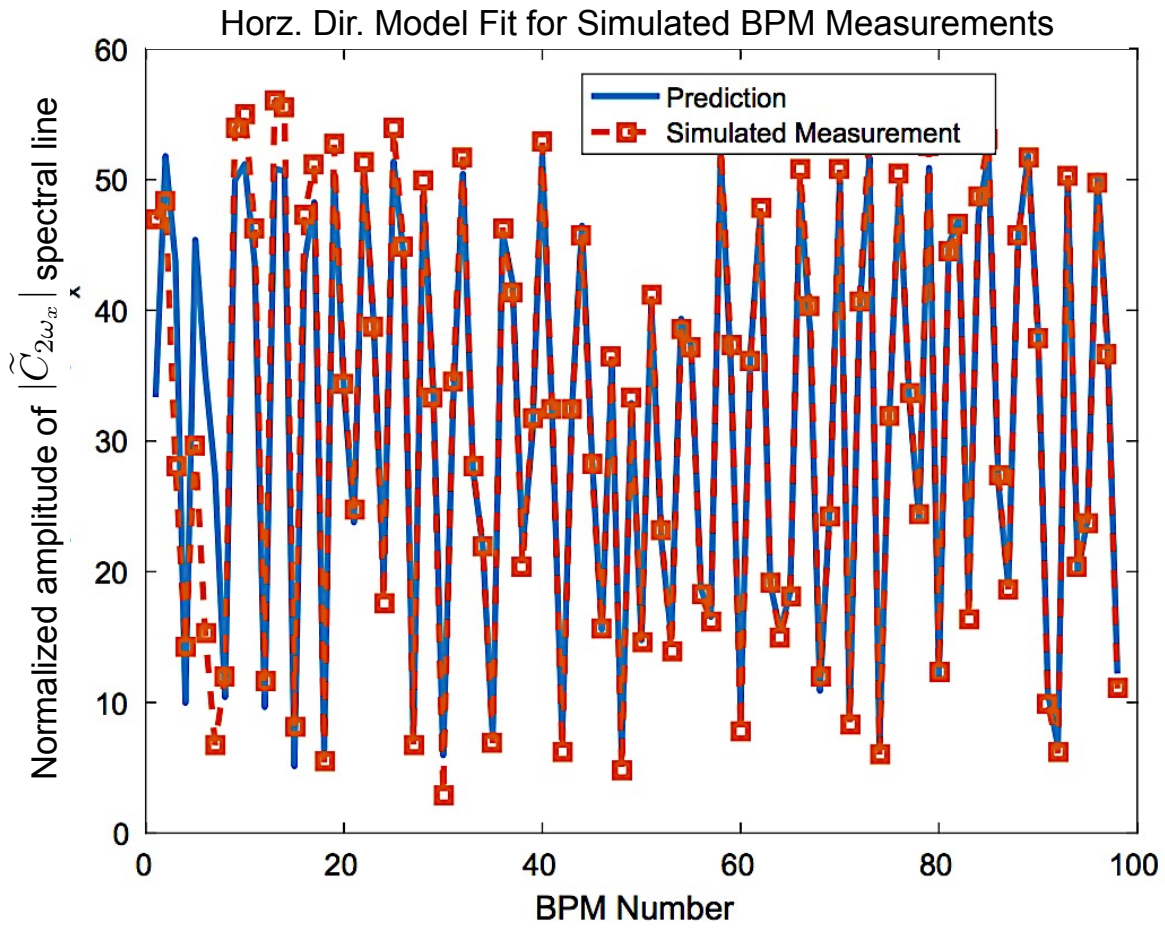


Figure 4.6: Comparison of $|\widetilde{C}_{2\omega_x}|$ spectral line amplitudes at each BPM between simulated BPM measurements and amplitudes calculated from the fitted model, in order to determine the field gradient k_2 and location of the single sextupole magnet.

The fitted model predicted that the sextupole was located at 40.6431m from the beginning of the lattice, and had $K = 1.0381\text{m}^{-2}$. After fitting, the model located the end of the sextupole accurately. The correct simulation setting for K was $K = 1.0336\text{m}^{-2}$, thus the sextupole field gradient was determined to within 0.5%. It will be important to understand both in simulation and experiment the noise to signal ratio, and what level of accuracy can be achieved consistently.

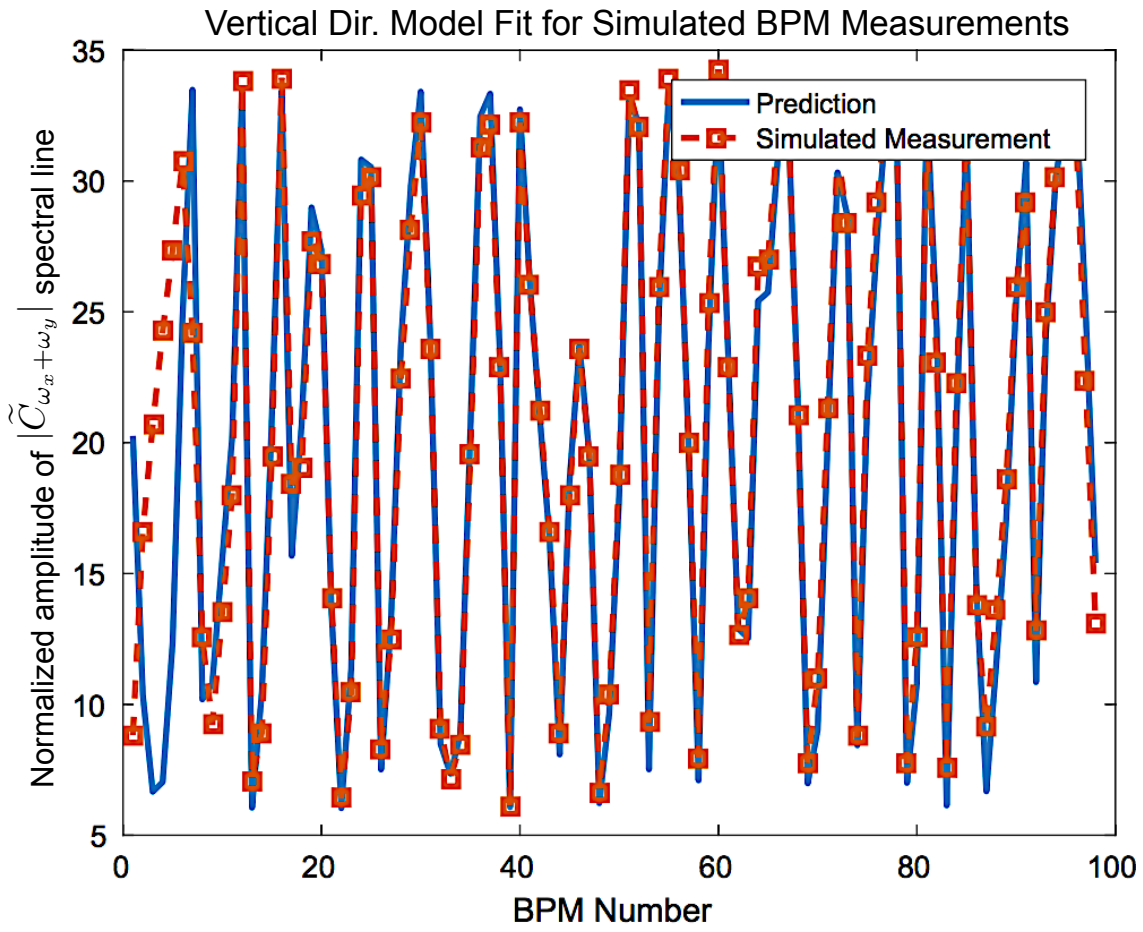


Figure 4.7: Comparison of $|\widetilde{C}_{\omega_x+\omega_y}|$ spectral line amplitudes at each BPM between simulated BPM measurements and amplitudes calculated from the fitted model, in order to determine the field gradient k_2 and location of the single sextupole magnet.

4.2.3 Experimental Progress

TbT measurement comparable to the simulated data were collected durin CESR machine studies in March 2018. During these experiments, the beam is resonantly driven by vertical and horizontal tune trackers and BPM data was acquired on a TbT basis using the CBPM system (36). The tune trackers were driven maximally without incurring beam scraping, in order to maximize the resonant signals in the BPM measurement spectra. Further, unlike the simulation, CESR was configured to have a 2-family sextupole distribtuion, starting with chromaticity compensated to near zero. The field gradient at a single sextupole was doubled until the maximum strength was reached. These measurements were repeated for

two different sextupoles in the lattice.

Equation 4.11 can be applied to this data as well as simple spectral analysis tools. An example spectrum is shown in Fig. 4.8. It is useful to remove the baseline signal in order to study changes in the signal which reflect nonlinearities such as sextupole field errors. Following Ref. (27), the nominal machine setting signal was subtracted from the signals measured with a given sextupole was increased in field strength. From these signals the horizontal and fundamental second harmonic phase advance were calculated, and a discontinuity was observed in both, near the known location of the sextupole. The discontinuity in the betatron phase advance is caused by a quadrupole error due arising from nonzero orbit in the sextupole, and is not unexpected. However, additional discontinuities are observed in the phase of the second harmonic tune line beyond that generated at the position of the sextupole, and thus in the current analysis this phase was not predictive in the determining the location of the sextupole. Work is ongoing to determine and mitigate the source of extraneous phase discontinuities in the second harmonic.

The change in the $2Q_x$ harmonic spectral amplitudes were studied from a sextupole which was increased from $k_{2,\mathrm{nom}} = 0.56\mathrm{m}^{-2}$ (where the "nom" the subscript on k_2 signifies the nominal settings) to $k_2 = 1.9k_{2,\mathrm{nom}} = 1.04\mathrm{m}^{-2}$ and then to $k_2 = 2.8k_{2,\mathrm{nom}} = 1.56\mathrm{m}^{-2}$ in two discrete steps. Figure 4.9 shows the measurement at each sextupole setting; it is clear that the amplitude scales linearly at each point. The inset histogram of the ratio between the amplitudes at each BPM shows that the distribution is highly peaked, and has a mean at 1.98. This ratio is a measurement of δk_2, because the nominal is subtracted, and the exact value of the ratio should therefore be 2.08, based on the readback value of the power supplies. A beam-based method for determining sextupole field gradients is precisely the tool needed in order to ensure these values are accurate, which our study hopes to provide as we continue our work.

A simple model for CESR, in which there is a single sextupole and single BPM, was considered in order to develop a spectral analysis method for BPM measurements, which

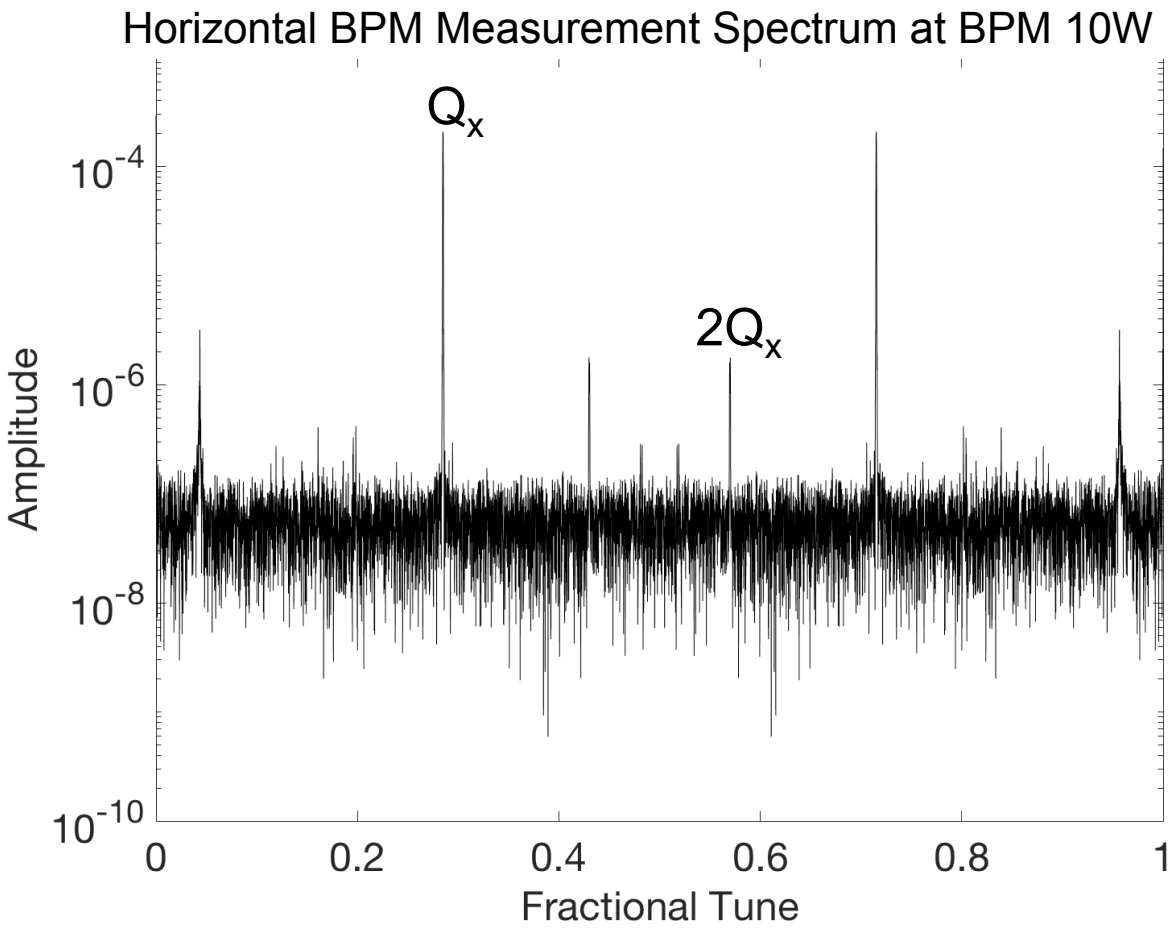


Figure 4.8: Example spectrum from experimental data collected at CESR, with nominal sextupole settings and one sextupole increase in field gradient. The fundamental harmonic and second harmonic spectral lines are labeled as Q_x and $2Q_x$ respectively.

could be used to locate and measure sextupole field errors in the presence of tune jitter or initial offset. This model was shown to work in simulation, in which BPM measurements from a single sextupole left on the lattice were simulated, analyzed, and fit by the model. The fit located the sextupole element, and fit the field gradient K to within 0.5%.

Experimental data were collected to test the spectral analysis methods we developed. BPM measurements were collected at CESR in which the beam was resonantly driven by tune trackers, thus maintaining the phase of the beam excitation relative to the drive. The sextupoles were set to nominal settings which correct chromaticity, and a single sextupole gradient was increased in discrete steps. From the initial analysis, phase discontinuities have

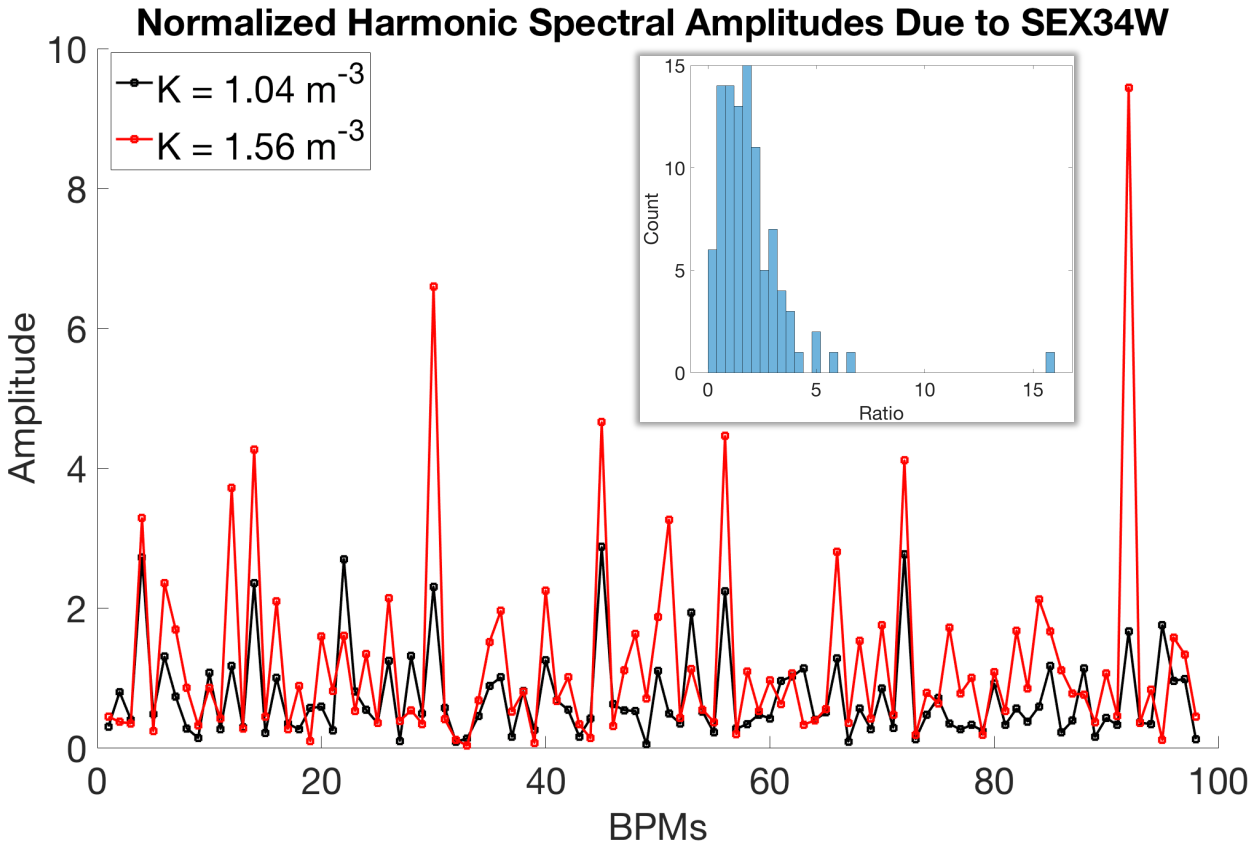


Figure 4.9: Second harmonic spectral line amplitudes for two increased sextupole field gradient settings, with inset displaying the distribution of the ratio of spectral line amplitude at each BPM. The ratio is highly peaked, with a mean of 1.98.

been observed in the linear and second harmonic phase advances, at the location following the increased sextupole. Further, the linear relationship between the amplitudes of the second harmonic spectral lines when the baseline has been removed, and the change in the field gradient has been observed. Further spectral analysis and use of the simple fitting model will continue, in order to develop a robust method for locating and measuring sextupole field gradients at CESR.

CHAPTER 5

MACHINE LEARNING-BASED SURROGATE MODELING IN THE PHYSICAL SCIENCES

5.1 What is Machine Learning?

Since the 1980's and 1990's, the rise of statistical and numerical methods led to the expansion of machine learning (ML) as a computational science and field of research. In particular, the study of how a computer could learn to perform certain tasks, make decisions, or recognize patterns. Popular examples include the DeepMind algorithm AlphaGo, which learned and successfully beat the professional Go player, Han Fui in 2008 (37), or self driving cars. Many self-driving car companies rely on ML methods to train an artificial intelligence (AI) how to accept high-frequency sensor data and use it to make decisions on the road. In both cases, a technique for using information from a highly complex and fast-changing system was learned, and then successfully used to make decisions or complete a task. These successes have spurred innovation and continued development in the field. Now, with improved hardware for completing long computations, as well as software that makes ML methods accessible, the interest in using ML tools for scientific applications has also grown.

As with any new technology it is prudent to be critical in its application and use. Carleo mentions the "suspicious eye," referring to those scientists who are wary of ML methods being peddled as the research equivalent of snake oil. However, they fail to mention those who are willing to try ML methods with a critical eye; optimistic about its uses but practical about its limitations.

Carleo et. al. state, quite eloquently, why in particular ML and physics are closely related, saying (38):

> ... both ML and physics share some of their methods as well as goals. The two disciplines are both concerned about the process of gathering and analyzing data to design models that can predict the behaviour of complex systems. However,

> the fields prominently differ in the way their fundamental goals are realized. On the one hand, physicists want to understand the mechanisms of Nature, and are proud of using their own knowledge, intelligence and intuition to inform their models. On the other hand, machine learning mostly does the opposite: models are agnostic and the machine provides the "intelligence" by extracting it from data. Although often powerful, the resulting models are notoriously known to be as opaque to our understanding as the data patterns themselves. Machine learning tools in physics are therefore welcomed enthusiastically by some, while being eyed with suspicions by others. What is difficult to deny is that they produce surprisingly good results in some cases.

In some ways, machine learning models are receptacles for knowledge that can be gained through observation. The training methods are analogous to the observation methods. By varying the types of models used, and the training strategies employed, the goal of machine learning for the physical sciences is to gain information about the relationships between observations. Unlike a calculation or function, a set of rules takes the inputs to outputs. However, machine learning methods aim to learn the rules when the inputs and outputs are known.

In this chapter, I will introduce focus on describing neural networks, and how they can be applied for surrogate modelling.

5.2 Training Paradigms

Before discussing the types of models, and in particular neural networks, the abstract idea of "training" a model should be described. In machine learning, the learning or training process refers to an optimization procedure that adjust numeric parameters to achieve the best result given a constraint. Thus, supervised learning occurs when a model is given labeled data; inputs and associated outputs which act as targets. The goal of the training process is then for the model to be able to map inputs to outputs, and reduce the overall error in predicting outputs. Conversely, unsupervised learning occurs when the training data is unlabeled. The success of the model is determined by extremizing a cost function.

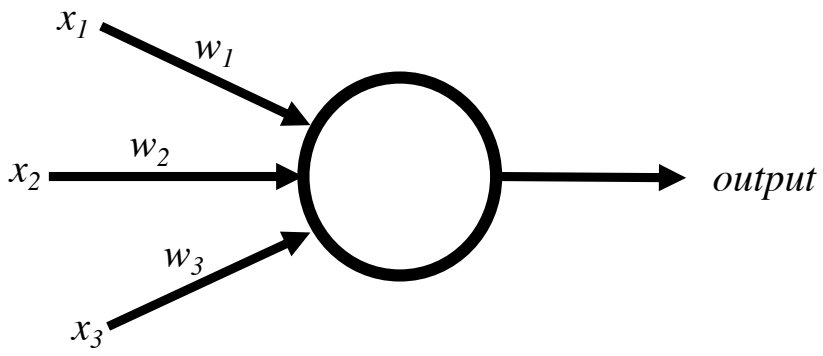


Figure 5.1: A simple schematic of a single perceptron, or neuron, with three inputs, and the associated weights, which can be used to calculate a single output.

5.3 Neural Networks

The foundation of a neural network is the artificial neuron, which emulates the neurons found in animal brains. This model accepts an input, and based on a nonlinear activation, can be "on" or "off" if a threshold is met.

Take for example a single neuron. Shown in fig 5.1 is the neuron with three inputs, x_1, x_2, x_3, and one output. The neuron can compute an output based on the following rule. The inputs are weighted by real numbers w_1, w_2, w_3. The output is then determined by a weighted sum $\Sigma_j w_j x_j$ compared against a threshold value. If the output is above the threshold, the neuron outputs 1, meaning it has turned "on", or 0, meaning it is "off".

By stacking neurons to accept inputs or pass information to subsequent layers of neurons, a single perceptron becomes a multi-layer perceptron, and then a neural network.

However, it is the activation function and associated threshold of each neuron which makes a neural network more than the matrix calculation it appears to be. A variety of common activation functions are shown in fig. 5.3. The activation function is what allows a neural network to model nonlinear relationships.

With these basic building blocks, a neural network can be built to approximate any nonlinear function (). Academic and industry research and development often builds on

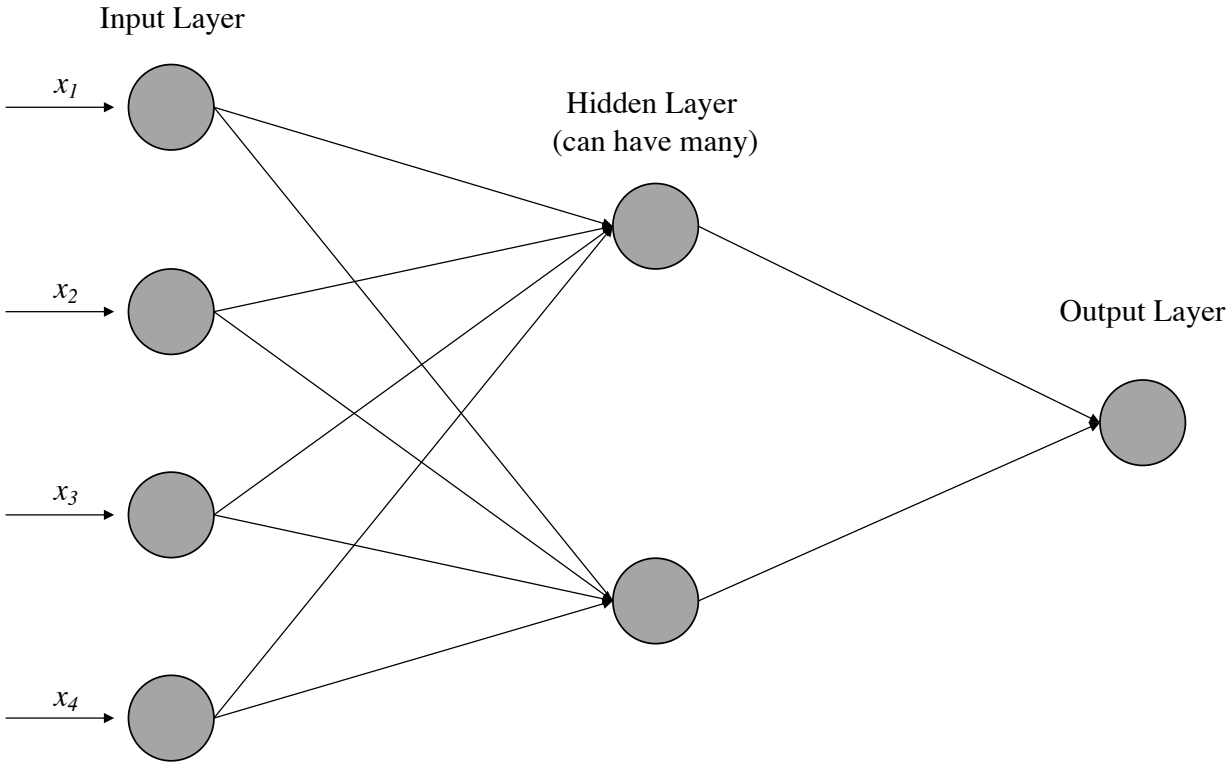


Figure 5.2: A simple example of linked neuron forming layers, which can then be connected to further layers of neurons. This is a very basic structure for a fully connected (all neurons pass information to each neuron in the next layer), but many variations exist based on different applications.

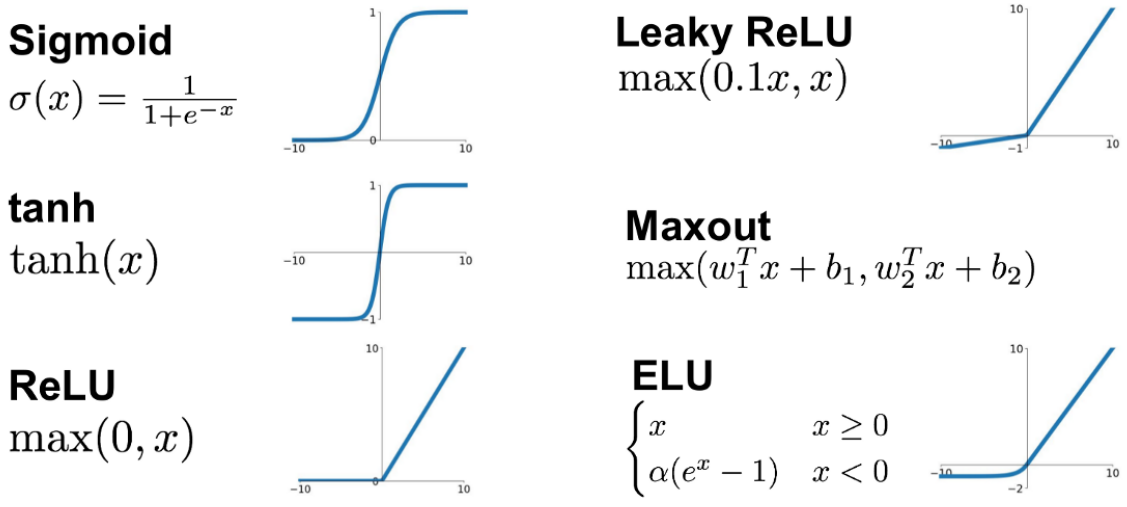


Figure 5.3: Shown are several common activation functions used for building neural networks. The activation function used often depends on the application.

these basic concepts to produce new configurations of networks, training algorithms, and methodologies. The physical sciences can benefit from this field of research, as classes of computational intractable problems become feasible with ML methods.

5.3.1 Convolutional Neural Networks

Convolutional neural networks, or CNNs, are very similar to NNs. They consist of neurons that are arranged to take inputs and produce outputs, and the training process also allows the weights between neurons to optimize given training data. The primary difference between CNNs and simpler NNs is that CNNs are mostly used for pattern recognition within images.

The reason that NNs are less effective for learning models from image-based data is due to the computational complexity required. A canonical example, training a model to recognize handwritten digits, can illuminate this further. The MNIST (39) data set consists of several thousand black-and-white images of handwritten digits and are labelled with the corresponding digit in a computer-readable format. The images are 28×28 pixels, thus unspooling these pixels into a layer results in 784 ($28 \times 28 \times 1$) neurons. If the pictures were any larger, or in color (RBG color images consist of 2 layers), then this process results in very large layers. The danger in having large models where single layers can be thousands of neurons is overfitting the model. Overfitting generally means that the model is unable to learn effectively. The result of overfitting is the model lacks the ability to find general features, and tends to recognize the specific features that it saw while training. Such a model would not make good predictions on general inputs, because it is too specialized. To avoid this, the CNN has built-in methods which can preserve spatial information while limiting overfitting.

CNNs consist of convolutional layers, pooling layers, and fully-connected layers. These layers are stacked and arranged as needed, which is referred to as the architecture of the model. After an input layer which can accept the image input, a convolutional layer is applied. The convolutional layer consists of filters which are applied to the small regions

of the image, and slid over the whole image. The filters typically result in finding edges or features, which can then be visualized. After convolving the image with a filter, the resulting image is condensed, with the features represented within the condensed image. Further condensing is done to continue to minimise the number of parameters within the model, by using pooling layers. These layers tend to identify the most important information within the condensed image, and replace several pixels with a single pixel. Common choices are max-pooling, meaning the maximum pixel value in some $n \times n$ region is used to represent the entire $n \times n$ region. Successive applications of these layers can results in preserving local information, but reducing the problem dimensions such that optimizing the weights is computationally feasible. In the case of classification problems, such as training a model to recognize the content of an image, the output layer often consists of predicting a single label. However, CNNs can be used to find patterns in output images as well. This can be done by using a deconvolution process to construct an image based on the heat map created by the convolution process. Because of this, CNNs are very powerful for building models to relate image data.

5.4 Machine Learning for Particle Accelerators

A key challenge for studying beam dynamics or designing components at particle accelerator facilities, is the ability to simulate the accelerator. Obtaining machine time for the sole purpose of characterization can be rare, and often certain measurements cannot be done simultaneously or taken without disrupting operation. Therefore simulations are vitally important in design and experiment planning. Simulation software are often developed due to the need for high-fidelity numerical calculations of difficult-to-model phenomena. However, simulations introduce other challenges such as missing physics due to model simplifications, and may require significant computational resources. This is particularly relevant for the injector for an accelerator, where space charge is a dominating effect. Therefore, it is a priority to produce a fast and reliable model which can be used for online tuning while

supplying beam to users, or for offline use in experiment planning and research.

There is significant effort currently (11; 12; 13; 14) towards using model-based control methods in real-time machine operation and tuning with the goal to make operation and tuning more efficient. Machine learning methods may help to automate tasks such as switching between standard operating schemes, or correcting small deviations that result in poor beam quality. A possible solution for model-based control methods could be the use of simulations, which are ubiquitous in the field. However, simulations are often very slow, taking minutes to hours to run on high-performance computing resources. This impedes the use of most commonly used simulation software as candidates for model-based control/automation.

Therefore, it is paramount to work towards developing fast, reliable, and realistic models of machines. Machine learning based surrogate models are computationally intensive for their inference (or training) phase, but engender rapid responses during predictions. This enables such machine learning based models to proffer orders of magnitude speed ups over classical simulation based schemes. Amongst different machine learning based algorithms, Neural network based surrogate models are being widely applied to addressing the issue of speed (12), and can therefore be used for very fast optimization. Surrogate models can be very powerful, especially when trained using data that spans the operational range of the physical inputs. While surrogate models clearly are fast enough for use in real-time operation, questions about their reliability and how realistic they are remain. This is due largely in part to the following issue: data-driven surrogate models often require large corpora of training data, which cannot always be collected via measurements. Consequently, archived simulation data is often used as a substitute while model training. In this scenario, the machine learning model learns machine behavior as established by the simulation. In this light, the performance of the machine learning model is determined by the fidelity with which the simulation model represents the machine. In most cases, large discrepancies are present due to collective effects or simplifications assumed in the simulations. The lattice or machine representation in a simulation is often a static representation of the designed

machine, and not an evolving representation of the physical machine. Simulations tend to represent ideal conditions, and therefore ideal beam dynamics. Unfortunately, this is not the case in actual operation of most machines. This is one of the major reasons that simulations are not used in the control room; the simulated lattice and predictions demonstrate the trends in the machine, but with many components in play, large absolute discrepancies often go unresolved. By using simulated data for training a surrogate model, static discrepancies and the beam conditions that result from them are learned by the surrogate model, and reproduced when predictions are made.

One approach is to attempt training surrogate models on measured data, but often there is not enough to do so. Due to limited machine studies or machine development beam time available at user facilities, it may be challenging to collect sufficient training data.

Many of these issues are addressed by demonstrating various approaches for making surrogate models more representative of the physical machines and components. The demonstrations shown in this thesis address both: the need for realistic simulated training data, as well as how to incorporate measured data into model training. While the demonstration is specific to the Linac Coherent Light Source (LCLS) superconducting injector (LCLS-II), during early injector commissioning (EIC), the methodology is largely agnostic to the component or machine. The hope is that these method will be applied widely, and used to make surrogate models for use in model-based control feasible.

CHAPTER 6

LINAC COHERENT LIGHT SOURCE II PHOTOINJECTOR CHARACTERIZATION AND SURROGATE MODELING

As soon as the LCLS achieved lasing at 8keV in 2009, design of the upgrade, LCLS-II, began. The primary goals for LCLS-II was to produce photon pulses at higher repetition rates, in a broader range of photon energies. Further, "the LCLS-II Project will position SLAC to retain its status as the world's preeminent Free-Electron Laser (FEL) research center, even as other FEL facilities begin operation around the world" (40). This includes a larger photon spectral range, the ability to control the photon polarization, as well as decreased pulse lengths for use in "pump then probe" imaging techniques (40).

The LCLS-II injector will produce the electron beam for the LCLS-II superconducting linac. repetition rate will be increased to 1MHz, which is several orders of magnitude above the current repition rate of 120 Hz. Not only would this allow for more "pump-probe" experiments to be performed, high-repetition rates can provide more data in less time, increasing statistics for experimenters. If experiments can be completed in less time, more users will have the opportunity to access the facility.

Further, high brightness machines require low-emittance, energetic electron bunches. In order to achieve this, producing high-quality bunches from the photoinjector is important. In a linear machine, the beam quality does not equilibrate in a storage ring. Degradation during the acceleration and transport of the electron bunch to the FEL can limit the photon quality. Thus, the brightness of photons produced at an FEL depend significantly on the injector performance.

Further, the technology used to produce electron bunches for lower repetition rate normal-conducting injectors cannot be scaled up for MHz repetition rates. Thus, the Advanced Photoinjector Experiment (APEX) (41) at Lawrence Berkeley National Lab determined how to successfully achieve an injector with the necessary capabilities.

Basic Principles of Photocathode Injectors

The basic principle of a photocathode injector is the photoelectric effect. By bombarding a cathode material with photons, the energy of the photons can be absorbed. If the energy supplied is higher than the binding energy of the electrons to the nuclei, some electrons will be emitted from the material. The science and technology of photocathode injectors was developed on this principle, to produce compact, high-charge electron bunches for use in particle accelerators.

The beam emittance tends to degrade after the injector due to particle energy loss, focusing errors, machine calibration or misalignments. Therefore it is the goal of injector design to produce bunches with the smallest possible beam emittance. However, this goal is in contest with another major goal: achieving high bunch charge. The metrics used to evaluate cathode materials is the quantum efficiency (QE). The metric QE defines the proportion of the absorbed photons to the electrons produced. Both the emittance and the QE can be estimated theoretically, and measured experimentally. The intrinsic emittance depends on the intrinsic transverse momentum distribution of the electrons in the cathode material, and is thus characteristic of the material (42). The QE depends largely on the energy of the incoming photon. The decision of which cathode material to use therefore may depend on the type of laser that will be used, the available photon energies, the desired beam emittance, and desired beam current.

Another important consideration is the laser which is used to produce the electron bunches. The laser spot size determines the physical transverse size of the electron bunch. Fluctuations in the laser intensity can result in aberrations that degrade the bunch quality. Even after accounting for laser stability, these fluctuations may play an important role in the resulting beam quality. This will be studied in detail, in order to inform how to create more robust surrogate models.

Table 6.1: Expected operating parameters for the LCLS-II injector, achieved after all construction and commissioning is complete. During this study, the injector was still in early commissioning, and had limited operational ability.

Parameter	**Value**	Unit
Charges	100	pC
Laser FWHM	20	ps
Laser Radius	1	mm
Field on Cathode	20	MV/m
Repetition Rate	1	MHz

6.0.1 The LCLS-II Injector

The injector, shown in Fig. 6.1, will be used for the LCLS-II project. The expected operating parameters for the injector are shown in Table 6.1. When the data was taken for this thesis, the injector was in early commissioning.

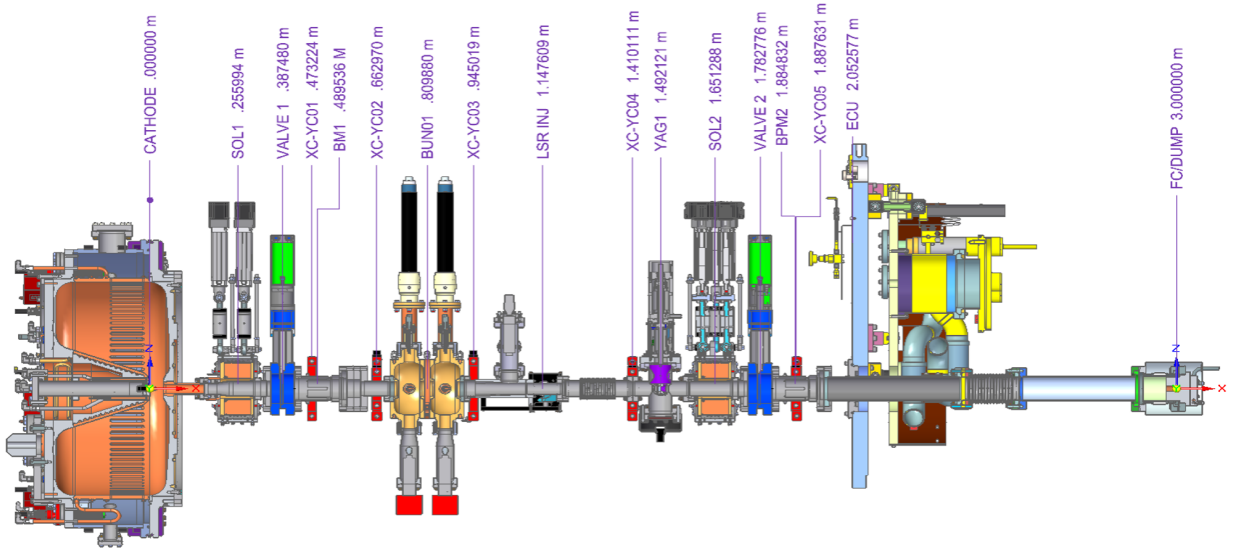


Figure 6.1: A schematic of the LCLS-II injector, showing each component and its position along the beam line (41).

6.1 Injector Surrogate Modelling

Training data was generated in two ways. For scalar model development, the standard particle generator in Astra was used by supplying laser radii as inputs. A large random

sample of the laser radius, cathode cavity RF phase, beam charge and solenoid strength was then simulated to create a data set which assumed standard incoming beam parameters. The beam quantities of interest are calculated, in simulation, by integrating the trajectories of thousands of macroparticles that represent the full charge of the bunch. There are several options for simulation tools to make data sets, which rely on sophisticated space charge calculations. In this case, the simulation software, Astra, developed at DESY in Hamburg, Germany was used (43), and particle generation was done using the Astra particle generator, as well as distgen (44). The SLAC-developed Python wrapper LUME-Astra (45) was used to create, set-up, and process simulated data into data sets. A sample in this data set consists of the scalar input values, and the associated bulk beam values such as emittances and beam sizes. This data was used for scalar model training.

To begin creating a surrogate model of the LCLS-II injector, training data was generated in simulation. The data was comprised of scalar injector parameters which are adjusted or tuned during operation, such as the bunch charge and the solenoid strength, and the resulting electron beam quantities, such as emittances and beam sizes. The beam quantities of interest are calculated, in simulation, by integrating the trajectories of thousands of macroparticles that represent the full charge of the bunch. There are several options for simulation tools to make data sets, which rely on sophisticated space charge calculations. In this case, the simulation software, Astra, developed at DESY in Hamburg, Germany was used (43), and particle generation was done using the Astra particle generator, as well as distgen (44). The SLAC-developed Python wrapper LUME-Astra (45) was used to create, set-up, and process simulated data.

Further training data was generated by running measured VCC laser distributions and idealized SG laser distributions through LUME-Astra while randomly sampling injector input settings. For each unique laser profile, 2,000 randomly-sampled points in the input space were generated. The predicted output includes the electron beam distribution as it might be measured at a YAG screen, along with bulk statistical quantities such as normalized

emittance and beam sizes. Two simulated data sets consisting of approximately 60,000 samples from SG particle distributions and 70,000 from VCC measurements were generated. Thus, a sample in this data set consisted of the scalar inputs including the dimensions of the input laser distribution and also the 50 $\times$ 50 binned input distribution, and the associated scalar output values and 50 $\times$ 50 binned output electron distribution.

6.1.1 Scalar Neural Network Model and Data Creation

The scalar surrogate model was trained primarily for two reasons, to aid offline experiment planning and start-to-end optimization by providing quick, non-invasive predictions of the beam behavior given injector settings. Fig. 6.2 shows the basic inputs and outputs. The neural network architecture itself consisted of 8 layers (6 hidden layers), each using a hyperbolic tangent activation function. The hidden layers each had 20 nodes; the input layer had 4 nodes corresponding with each input. The four inputs were: laser radius, RF cavity phase, solenoid magnet strength, and bunch charge. The model output 16 scalar prediction. The outputs were: RMS bunch positions (x-, y-, and s- directions), RMS bunch momenta (x-, y-, and s- directions), normalized emittances, beam kinetic energy and number of particles remaining.

These output parameters are relevant to optimization studies. Maximizing the bunch charge while minimizing the emittance is one optimization that may be done when designing an injector. This model uses the radius of uniform laser distribution as an input rather than a realistic laser distributions. This matches the typical practice of using a laser radius with a static profile for multi-objective optimization studies on injectors (including the LCLS-II injector).

The training data was generated using the Cori supercomputer at the National Energy Research Scientific Computing Center (NERSC) (46). The simulations were run in parallel on 128 CPUs, producing a data set of 21,600 samples. Creating this data took approximately 20 hours to complete. If it was not possible to run these simulations in parallel, and each

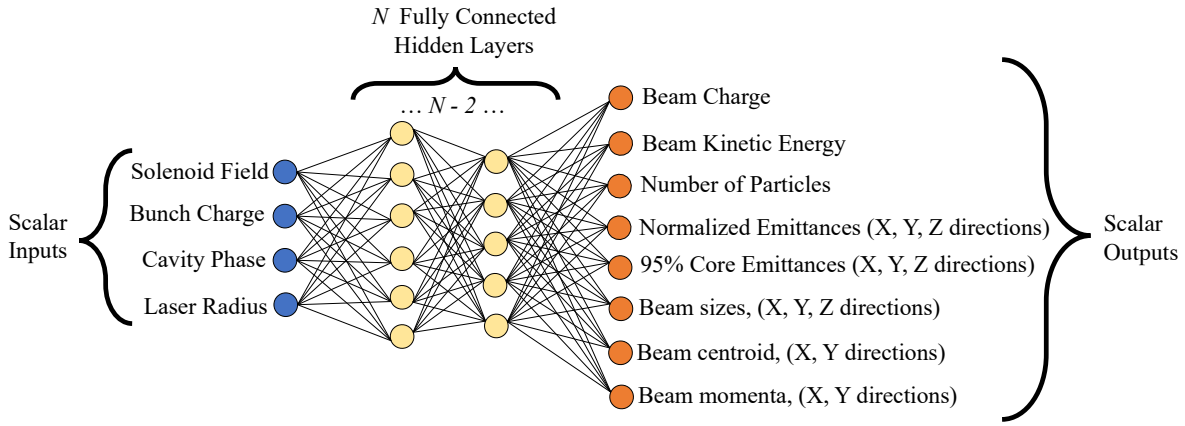


Figure 6.2: Schematic for the feed-forward, fully-connected neural network architecture used for the scalar-to-scalar surrogate model. The four scalar inputs are: laser radius, RF cavity phase, solenoid magnet strength, and bunch charge. The outputs were: RMS bunch positions (x-, y-, and s- directions), RMS bunch momenta (x-, y-, and s- directions), normalized emittances, beam kinetic energy and number of particles remaining.

simulation was completed sequentially, the total simulation time would have been at least 2500 hours (over 100 days).

To make the data, the input space was randomly sampled, to span the full parameter ranges. This is important for making a representative surrogate model. Depending on how many input dimensions there are, and the resolution needed, a grid could be used to systematically loop through input values. In this case, where a large volume of samples in a low-dimensional input space is taken, a random sample is sufficient. Thus, the surrogate model will be able to learn about the electron dynamics throughout the input-dimension volume. This is shown in Fig. 6.3.

The data was split into training, validation, and test samples. During any training process, the training samples are used for fitting the model. The validation loss is calculated and monitored during training to avoid over fitting, but is not included directly in the weight updates. All testing samples are withheld from the training process entirely. The training process was completed using a single CPU, and took two to four hours to converge. Thus, when considering the amount of time spent preparing a surrogate model, the training time

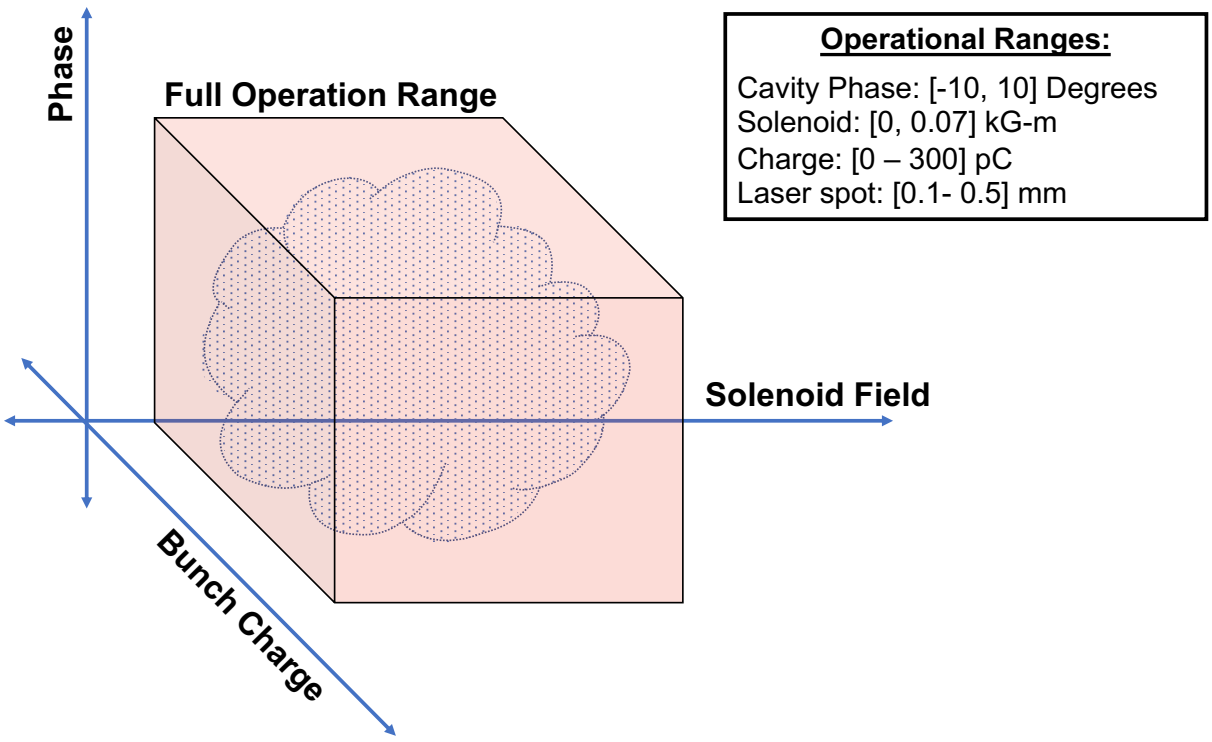


Figure 6.3: Shown is an example (projected to 3 dimensions) for how the input value sampling was done. For a given laser distribution or radius, the other three input values (beam charge, RF cavity phase, and solenoid strength) were varied randomly. The operational ranges define the range within which the input can be randomly sampled.

for a model with thousands of parameters is negligible relative to the time spent creating the training data.

6.1.2 Surrogate Model Performance

Shown in Fig. 6.4 is a small selection of the surrogate model performance on the test samples. The test set is sorted such that the target parameter that is being predicted by the surrogate model is presented in increasing order which makes it qualitatively clear if the surrogate model is performing well relative to the test samples.

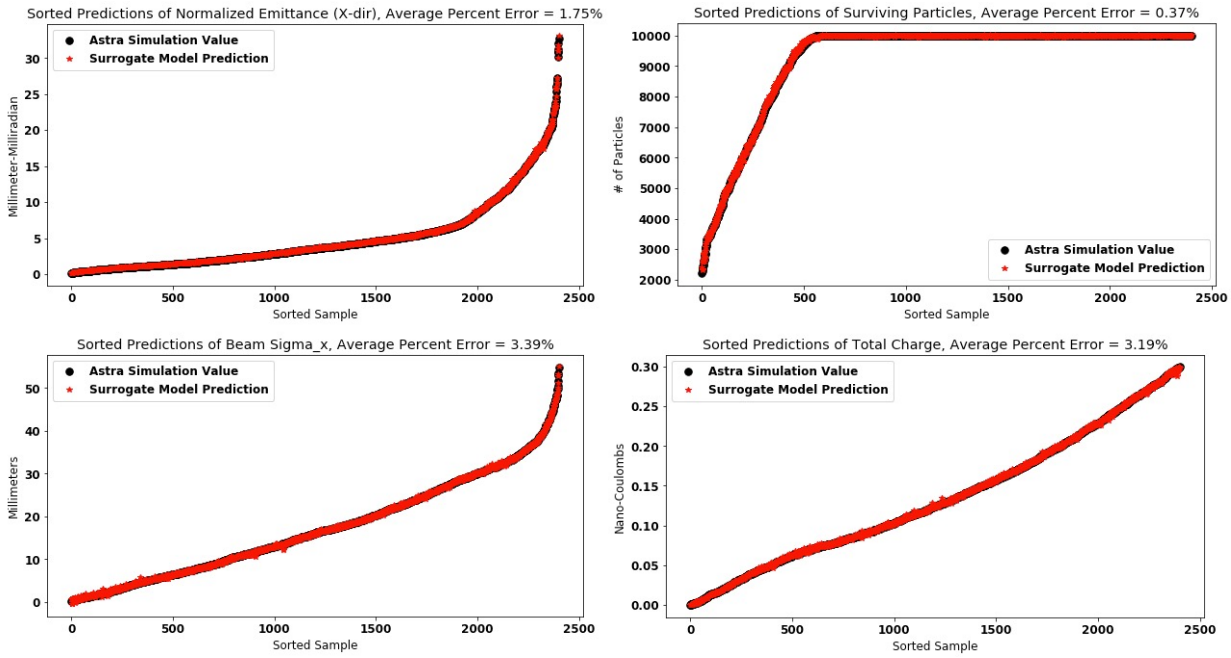


Figure 6.4: A selection of test set predictions made by the trained surrogate model, compared with the Astra simulated values. The average percent error for each quantity is provided.

Quantitatively, the average percent error for each parameter is calculated and provided in Tab. 6.2.

Table 6.2: The average percent error for key scalar output predictions from the surrogate model, relative to the simulated value.

Parameter	Average Percent Error
Total charge	3.19%
Beam size (x-dir.)	3.39%
Momentum spread (x-dir.)	5.10%
Normalized emittance (x-dir.)	1.75%
Kinetic energy	0.011%
Beam size (z-dir.)	0.65%
Normalized emittance (z-dir.)	3.78%
Particles lost	0.37%

6.2 Multi-Objective Genetic Algorithm Optimization Speed-Up Using a Surrogate Model

For accelerator design and parameter optimization, multi-objective genetic algorithms (MOGA) are a common and robust option. The algorithm can successfully sample from high-dimensional inputs spaces, to optimizing many outputs simultaneously, and can be used for constrained optimization with multiple objectives.

The algorithm, as suggested by the name, emulates evolution via "survival of the fittest." A population is randomly generated, where an individual in the population consists of input parameter values. The generation (all of the samples within the population) is evaluated via a fitness function, which is the is the rate-determining step in the algorithm. If the function that must be called requires significant time or computational resources to compute, the optimization will also take a long time. Once the fitness is determined, fit solutions are kept and unfit solutions are discarded. However, to ensure that the algorithm samples broadly, mutation can be used to alter fit solutions before they proceed to the next generation as a way to introduce variation.

The goal of optimization is to find the optimal solutions, which lie on a boundary called the Pareto optimal front. This optimality describes the plane of inputs resulting in outputs which cannot be further improved, without negatively impacting another competing output. This is shown in Fig. 6.5.

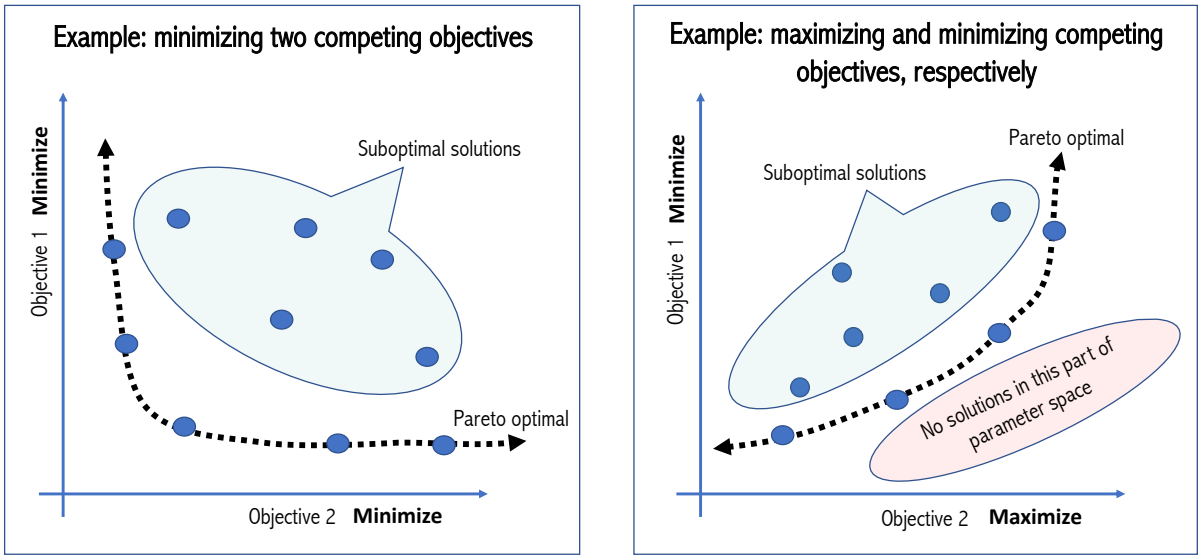


Figure 6.5: Two examples of Pareto optimal fronts are shown. Pareto optimally is the boundary along which one parameter cannot be improved further without detrimentally effecting another objective.

MOGA is favored among accelerator physicists because it can provide global optimization. Further, because accelerators are highly complex systems, multi-objective optimization with high-dimensional input spaces are ubiquitous. However, due to the computational expense as well as time needed to compute many simulations, it cannot be used during online operation. Even for offline design optimization, the resources needed can be prohibitive.

Thus, a well trained, robust surrogate model is a prime candidate for replacing a full simulation within a MOGA routine. This was demonstrated in (11). To demonstrate that the scalar surrogate model trained for the LCLS-II injector is similarly capable, the following optimization and comparison was done. The MOGA optimization was run for 200 generations, with 128 samples per population which was a similar set-up as done in (11). Figure 6.6 shows that the scalar surrogate model can successfully find Pareto optimal inputs which maximize the theoretically achievable beam charge, and minimize the beam emittance.

Shown in Tab. 6.3 are the timing comparisons. The optimization done by evaluating

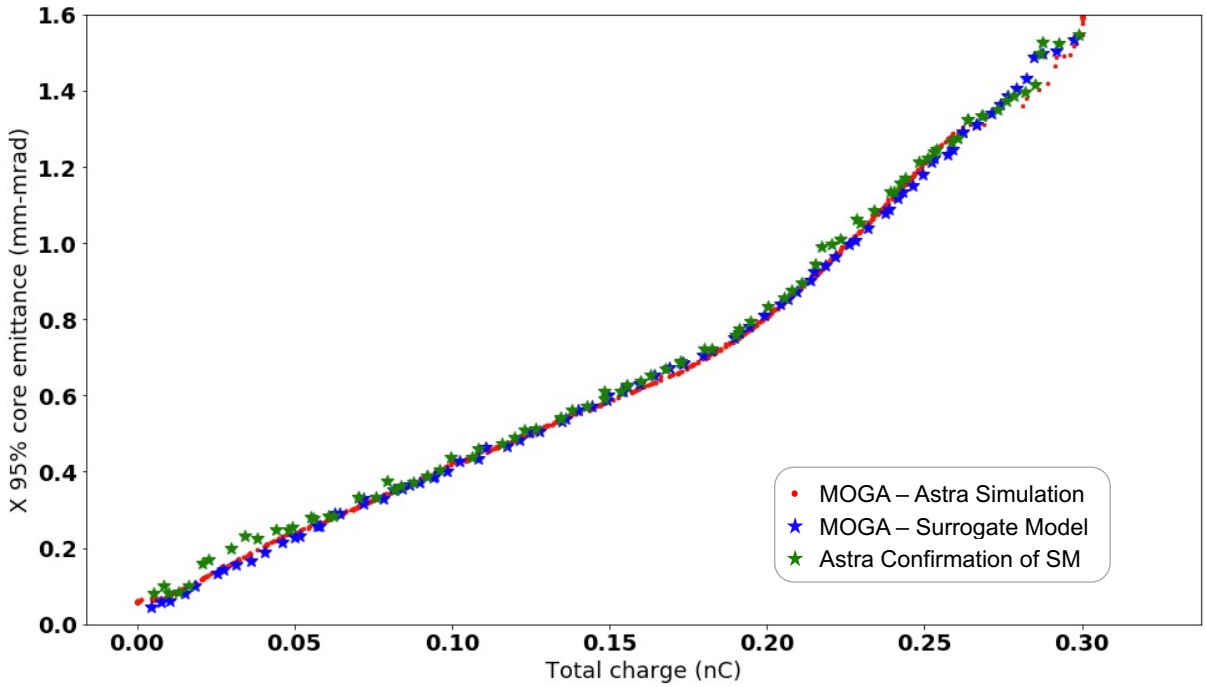


Figure 6.6: MOGA results determining optimal injector settings to minimize the beam emittance while maximizing bunch charge. The optimization was run using the Astra simulation, and compared to the optimization run using the surrogate model. To confirm if the surrogate model predictions were accurate, the Pareto optimal points were reproduced using Astra.

Astra simulations was run in parallel at NERSC, using 128 CPUs. By using the same population size as the number of CPUs available for simulation, each generation was done in parallel. Because each simulation took on average approximately 5 minutes, the full optimization took about 20 hours (though convergence may have been reached earlier, since only the number of generations was specified). Without the computational resources which allowed this parallelization, running sequential simulations would be untenable. To perform a similar process, a smaller population size would be required. However, because a trained neural network can execute a prediction in milliseconds, the surrogate model provides a significant speed up. In addition, Fig. 6.6 clearly shows that the surrogate model, if trained well, does not sacrifice on accuracy relative to the simulation. Thus, the surrogate model is a successful candidate for use in optimizations. The full simulation would likely still be used to confirm the results, or simulate optimal results for further investigation.

Table 6.3: Metrics for evaluating speed up of MOGA optimization to maximize bunch charge and minimize beam emittance, using the Astra simulation and the trained surrogate model.

Operation	**Astra Simulation**	**Surrogate Model**	**Speed Up**
Single function call	$\approx$ 5 minutes	$\approx$ milliseconds	$\approx 10^5$
Full optimization	$\approx$ 2500 hours (20 hours in parallel)	$\approx$ minutes	$\approx 10^5$

6.3 Characterization Studies of the LCLS-II Injector

In setting up a realistic surrogate model for the LCLS-II injector, it was important to assess how simulation results vary when using a realistic laser profile, as compared to an idealized Gaussian or uniform profile (as is assumed in most start-to-end injector optimizations). This determines whether it is necessary to include the full VCC image as an input to the model, or whether bulk metrics such as laser radius and an assumption of a Gaussian or uniform profile would be sufficient. We also characterized the LCLS-II injector with measured data scans and compared these to simulation data. This was done both to help improve the underlying physics simulation and assess both the need for and viability of using transfer learning for

this system to account for differences between the simulation and measurement domains.

Measurements were taken in Fall 2019. At the time, due to radiation safety limits, the injector would not be operated at the nominal operating parameters. In particular, the beam charge was limited. Measurements of laser input distributions and associated solenoid scans (where the electron beam size in the transverse direction was measured while the solenoid value was changed) were recorded. These scans were taken at several different beam charges, ranging from about 1 pC to 25 pC. Machine values such as magnet currents were also recorded.

6.4 Comparison Between Simulation and Measured Data

With such a high repetition rate, the cathode field gradient is lower than low-repetition rate photocathode injectors (41). Thus, the kinetic energy of the electrons as they are emitted and injected into the buncher is relatively low; up to 750 keV. At this energy, the dynamics are space-charged dominated. To study the dynamics in this regime and optimize the parameters for operation, particle-in-cell simulations are necessary. As such, these calculations can take several minutes to complete.

To evaluate how well the simulation data will compare to the measured data, for the purposes of surrogate modelling, the following was done. In Fig. 6.7, the beam sizes measured during two solenoid scans on the injector are compared to the predictions from Astra. The initial particle distribution for these comparison scans were generated by sampling Super Gaussian distributions in the transverse dimension, for 10,000 particles. The laser diameter was archived during the measurement and used for re-simulation. With first order matching of inputs (charge, radius, gradient, solenoid strength), there are clear discrepancies between simulation and measurement, as shown in Fig. 6.7. After investigating these errors through simulation scans, it is clear that the electric field gradient on the cathode has a significant impact on beam size and the location of the beam wais. There was no a spectrometer located in the beam line when this data was taken, therefore, the exact beam energy at the time of

measurement is not definitively known. Indirect measurements were attempted with a small corrector, but the results were inconclusive given they returned infeasible energy values (i.e. higher energies than possible given the amount of RF power supplied to the gun). This leaves the exact gun energy to be determined, and a probable cause of discrepancy in the measurement.

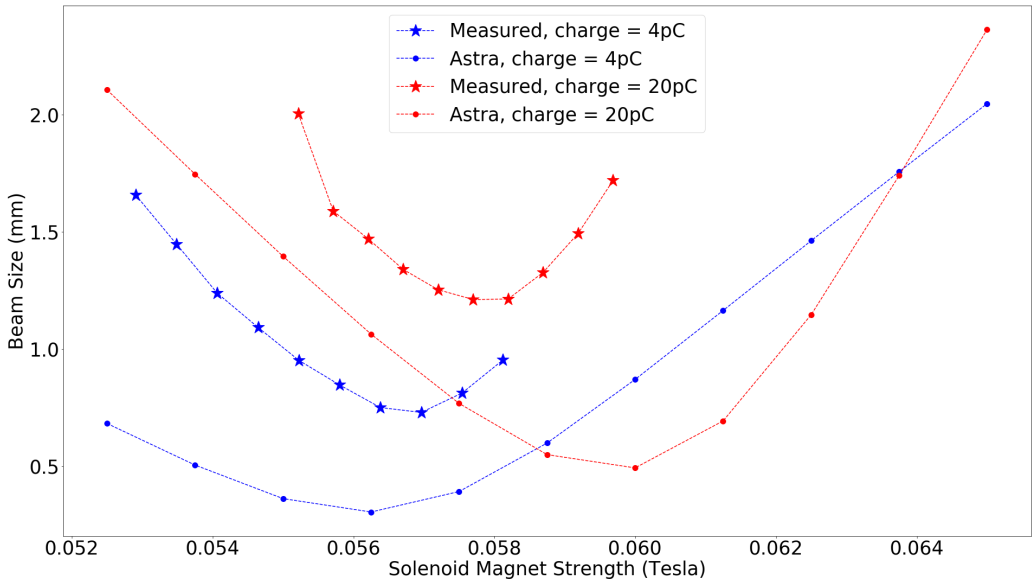


Figure 6.7: A comparison of simulated output values from Astra and the measurement values with the same machine parameters. All beam sizes are reported as the Gaussian standard deviation from fitting the particle distribution to a standard radial Gaussian distribution.

6.5 Simulation-based Sensitivity Studies

6.5.1 Generation of Electron Distributions from Laser Distributions

Using the measured laser distribution to sample particles for space charge simulation can minimize discrepancies between measurement and simulated output values (47). Therefore, in order to attempt to create more realistic simulated data, real laser profiles were used to generate particle distributions to be tracked in LUME-Astra. These laser profiles were

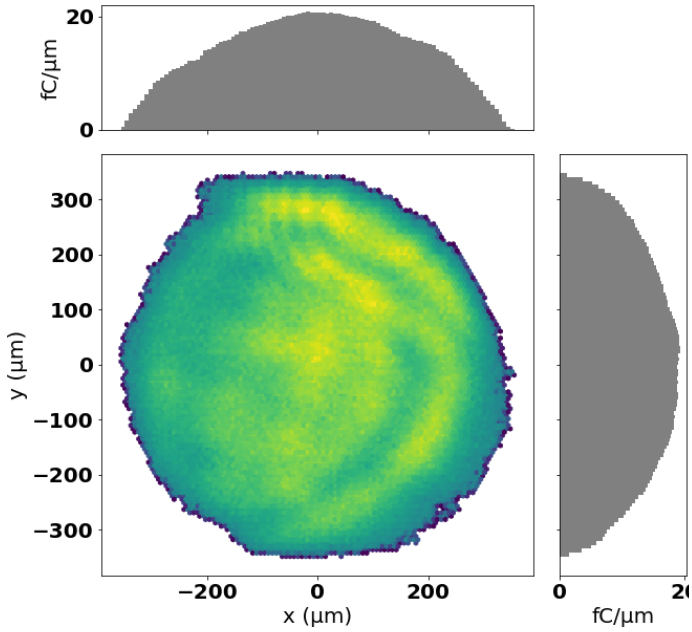
collected at the LCLS-II. The measurement is conducted as follows: the laser beam is passed through an optical splitter, such that approximately 5% of the beam intensity is directed towards a camera. The distance between the splitter and the camera is analogous to the distance between the splitter and the cathode, i.e. the transverse size of the laser at the camera location should be equal to the size at the cathode. The intensity at the camera is recorded, providing an image of how the laser intensity at the cathode appears.

We compared how the the beam sizes would differ when the initial particles were sampled from a realistic laser distribution and an ideal one. An idealized Super Gaussian (SG) density distribution is a common choice. The SG distribution, ρ, is given as function of radius r of the form:

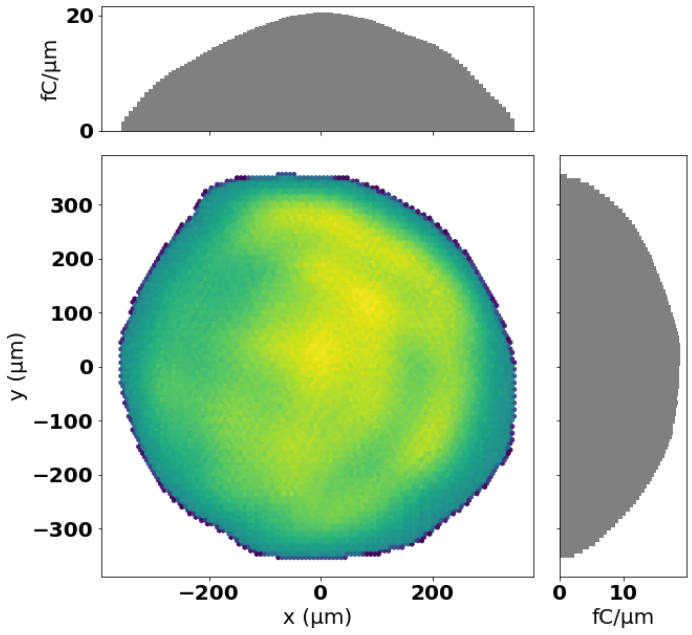
$$2\pi\rho(r) = \frac{1}{\Gamma(1+\frac{1}{p})\sigma^2}\exp[-(\frac{r^2}{2\sigma^2})^p] \tag{6.1}$$

where σ is the standard deviation, Γ denotes the Gamma-function, and p is the SG parameter. In the limit $p \to 1$, the SG is a standard Gaussian distribution. However, as $p \to \infty$, the SG distribution approaches a flat-top distribution. By parameterizing p by $p = 1/\alpha$, the α parameter is bounded by $[0, 1]$.

In order to make an idealized transverse SG distribution from measured VCC images, the following optimization procedure was completed. The measured laser profile was projected onto each transverse axis. A SG profile was then generated using an initial α value, and projected similarly into each transverse axis. An optimizer then iterated α via a Brent minimization algorithm (48), to minimize the residual between the projections. Many of the measured laser profiles are highly non-uniform or highly irregular edges. Therefore, results for which the per pixel percent error distribution standard deviation of less than 20% were selected. A nominal VCC and associated SG were chosen for the sensitivity analysis.

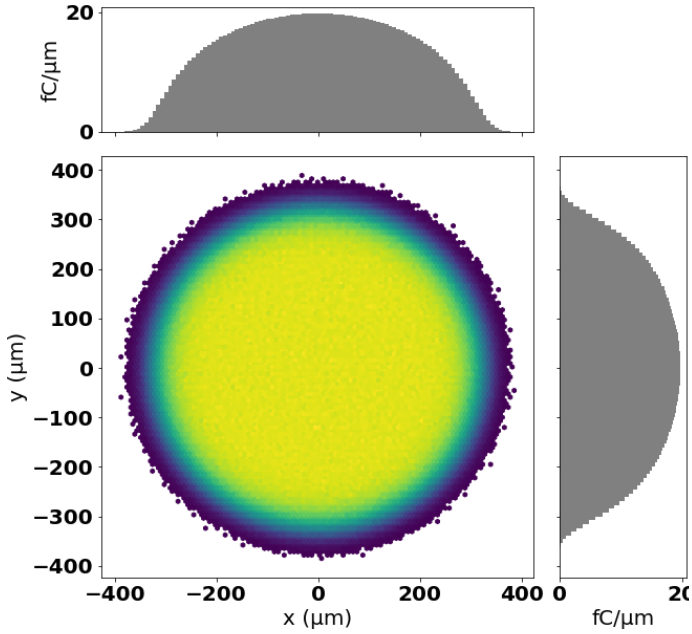


(a) VCC-generated initial electron distribution.

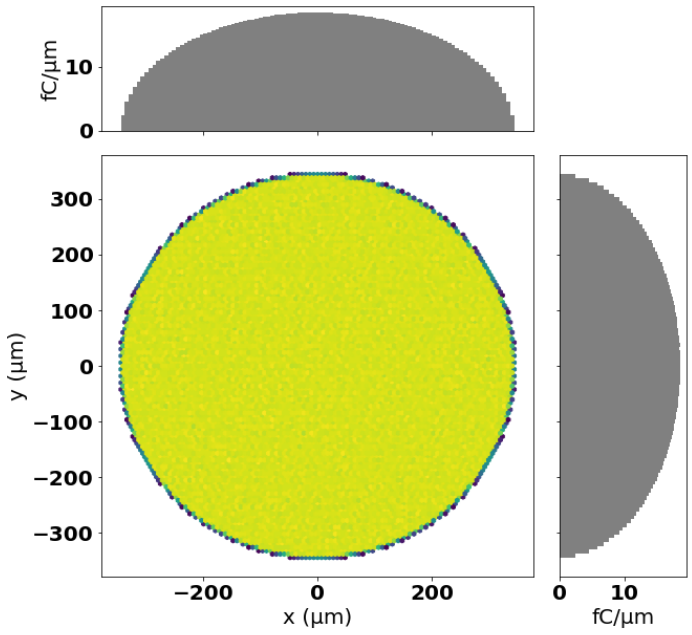


(b) Smoothed VCC-generated initial electron distribution.

Figure 6.8: Two of the five laser profiles used in the sensitivity study. The distributions are made from sampling 10,000 macro-particles from each laser profile. A simple convergence study confirmed 10,000 macro-particles was sufficient to calculate bulk beam parameters while reducing simulation run time.

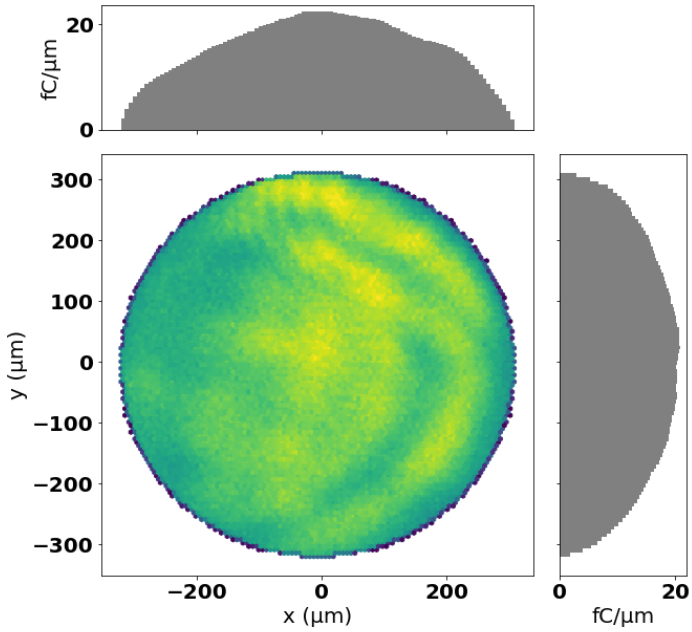


(c) Uniform, fixed radius initial electron distribution.



(d) SG generated laser distribution.

Figure 6.9: Another two of the five laser profiles used in the sensitivity study. The distributions are made from sampling 10,000 macro-particles from each laser profile. A simple convergence study confirmed 10,000 macro-particles was sufficient to calculate bulk beam parameters while reducing simulation run time.



(e) Circular-trimmed VCC-generated initial electron distribution

Figure 6.10: The last of the five laser profiles used in the sensitivity study. The distributions are made from sampling 10,000 macro-particles from each laser profile. A simple convergence study confirmed 10,000 macro-particles was sufficient to calculate bulk beam parameters while reducing simulation run time.

To determine how many particles would be sufficient to capture the dynamics, without increasing significantly the simulation time, the following was done. Several laser distributions were chosen and 10,000 particles, 30,000 particles, and 50,000 particles were sampled. The beam emittances and beam sizes at the end of the injector were recorded while scanning the solenoid strength and the beam charge separately. This was done for two VCC laser distributions, to determine if the shape of the input particles would contribute to needing more particles in order to maintain fidelity.

Two representative examples are shown in Fig. 6.11 and Fig. 6.12. It is clear from these examples that the only deviation between 10,000 particle runs and the higher particle runs, occurs at the wings of the solenoid distribution. Near the focal point of the range, the difference in the electron values is minimal. However, the time per simulation increases significantly as the number of particles increases. The increase is not linear; 30,000 particle

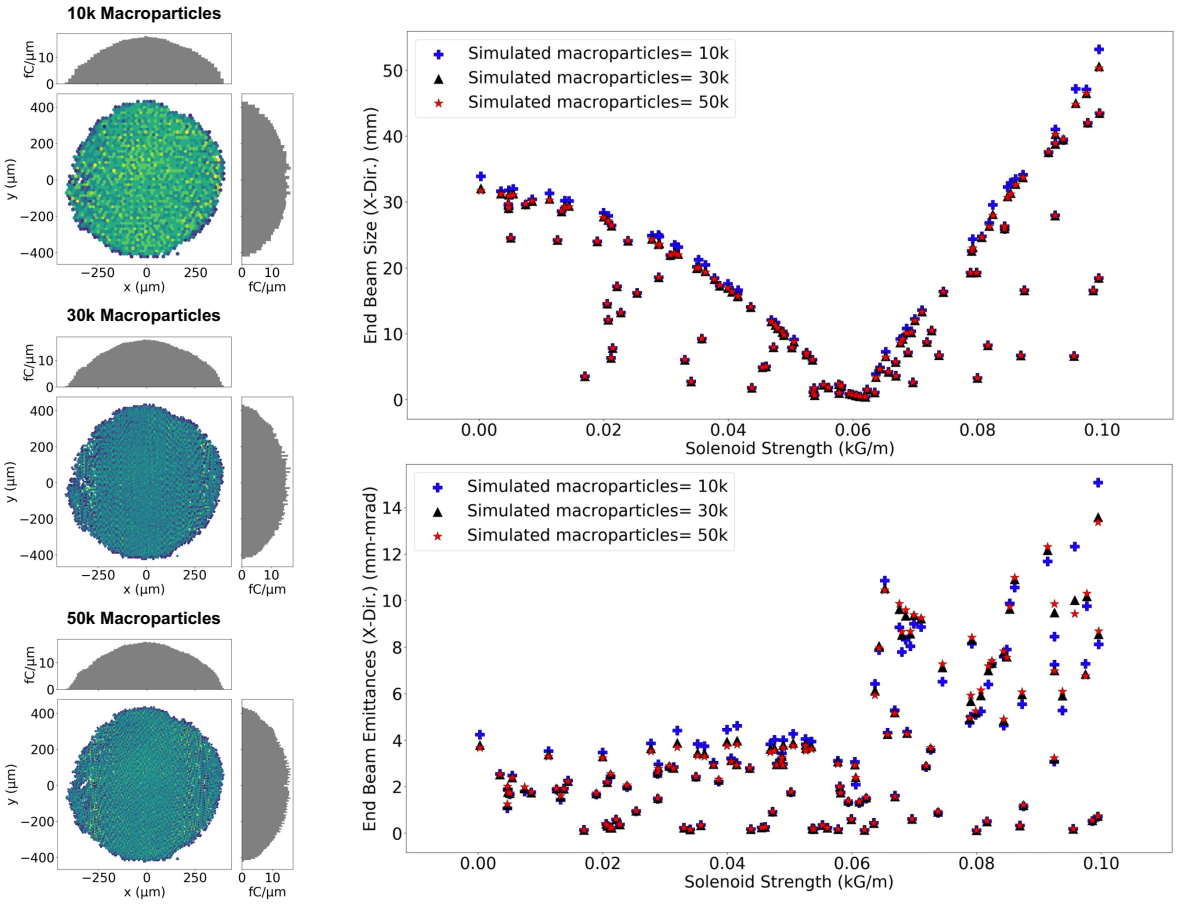


Figure 6.11: Shown are the laser distributions sampled 10,000 particles, 30,000 particles, and 50,000 particles for the same laser measurement. The simulated end beam emittance and end beam size are shown while scanning the solenoid magnet strength. The beam charge is varied between 1 - 300 pC.

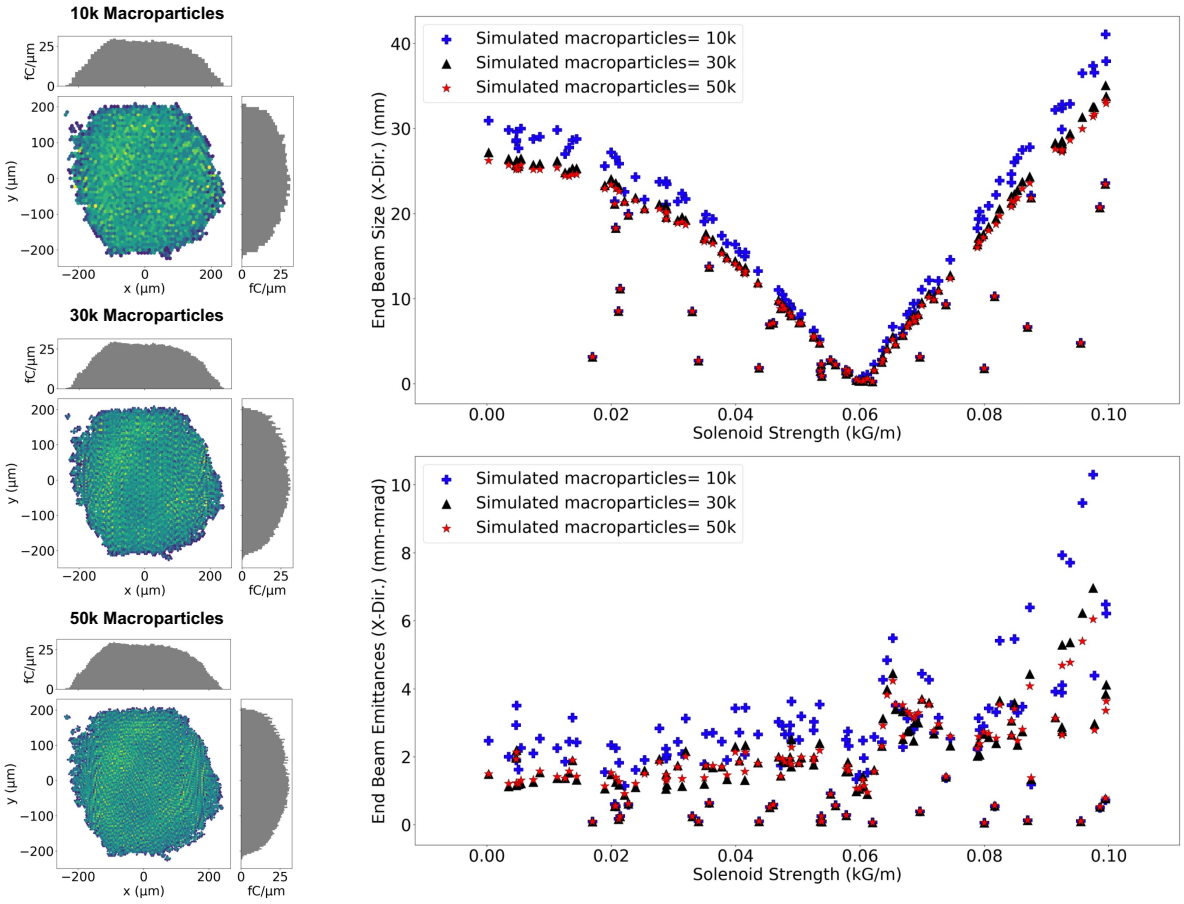


Figure 6.12: Shown are the laser distributions sampled 10,000 particles, 30,000 particles, and 50,000 particles for the same laser measurement. The simulated end beam emittance and end beam size are shown while scanning the solenoid magnet strength. The beam charge is varied between 1 - 300 pC.

simulations take about 7 times as long to complete, and 50,000 particle simulations take 16 times as long to complete, compared to the 10,000 particle simulations. In some cases, that resulted in simulation run times of one hour. Thus, there is very little improvement gained by increasing the number of macroparticles, but with a significant computation cost. Thus, 10,000 macroparticles is sufficient.

6.6 Sensitivity of LUME-Astra Predictions to Different Laser Distributions

The sensitivity of the simulation results on a realistic vs. idealized laser profile in simulation was also assessed. In this case, sensitivity is evaluated by whether the bulk properties of importance, normalized transverse emittance and transverse beam size, differ more than 10% from the values calculated from the uniform initial distribution. This threshold is close to the resolution of such measurements in the physical machine.

First, a candidate laser profile with features including rough edges and fluctuations within the bulk of the laser spot, was chosen. The transverse profiles compared include: uniform radial distribution, SG distribution, and the candidate laser profile with a Gaussian blur applied. The blurring was to determine to what extent the simulation is sensitive to non-uniformities. Shown in Fig. 6.8, Fig. 6.9 and Fig. 6.10 are each laser profile, with nominal charge of 10 pC total, for 10,000 sampled particles. The candidate VCC image is shown in Fig. 6.8 *a*, as well as the particles generated from a blurred version of the candidate VCC in Fig. 6.8 *b*. Having a similar, but slightly smoothed version of the candidate VCC will address whether the simulation is sensitive to the internal structure of the spot size, or just coarse features such as rough edges. This is further investigated by removing the edges of the candidate VCC images, and keeping internal features. This distribution and resulting particles are shown in Fig. 6.10 *e*.

Two highly uniform distributions were prepared for comparison as well. First, is a SG

distribution. The second, which is often used as the standard distribution for simulations, is a uniform density distribution with only a maximum radius specified. These distributions are shown in Fig. 6.9 *c* and *d*. The temporal structure for all of the particle distributions generated for this sensitivity study, as well as for the surrogate model training, was a Gaussian with standard deviation of 8.5 ps. The longitudinal time distribution was held constant for all simulations.

For each laser profile, a particle bunch with 10k particles was generated and tracked through the injector lattice in Astra to calculate various resulting beam outputs. It was determined that the bulk parameters in the simulation can be recovered with sufficient fidelity and speed using 10k particles. The primary quantities of interest in this study were the resulting normalized transverse emittances (95%, about two sigma, core emittance) and beam sizes. Astra simulations were completed for each laser profile at two charge settings (5 pC and 50 pC), with all other parameters, such as solenoid magnet gradient, held constant. The resulting transverse emittances and beam sizes at the YAG screen, 1.49 meters from the cathode, are shown in Fig. 6.13 and 6.14.

Figures 6.13 and 6.14 show that the emittance and beam size from Astra simulations are sensitive to the realistic beam distributions, relative to a uniform beam distribution. For the SG distribution, which emulates an ideal flat-top beam distribution, the emittance in each direction is the same, however there is a difference seen in emittances calculated from a VCC generated laser profile. Clearly the asymmetry of the VCC generated laser distributions can be captured by the simulation, as shown by the difference in beam size and emittances in each transverse direction, for a given VCC laser profile. These results suggest that using realistic laser profiles could result in simulated training data which is sufficiently different from that generated from idealized conditions.

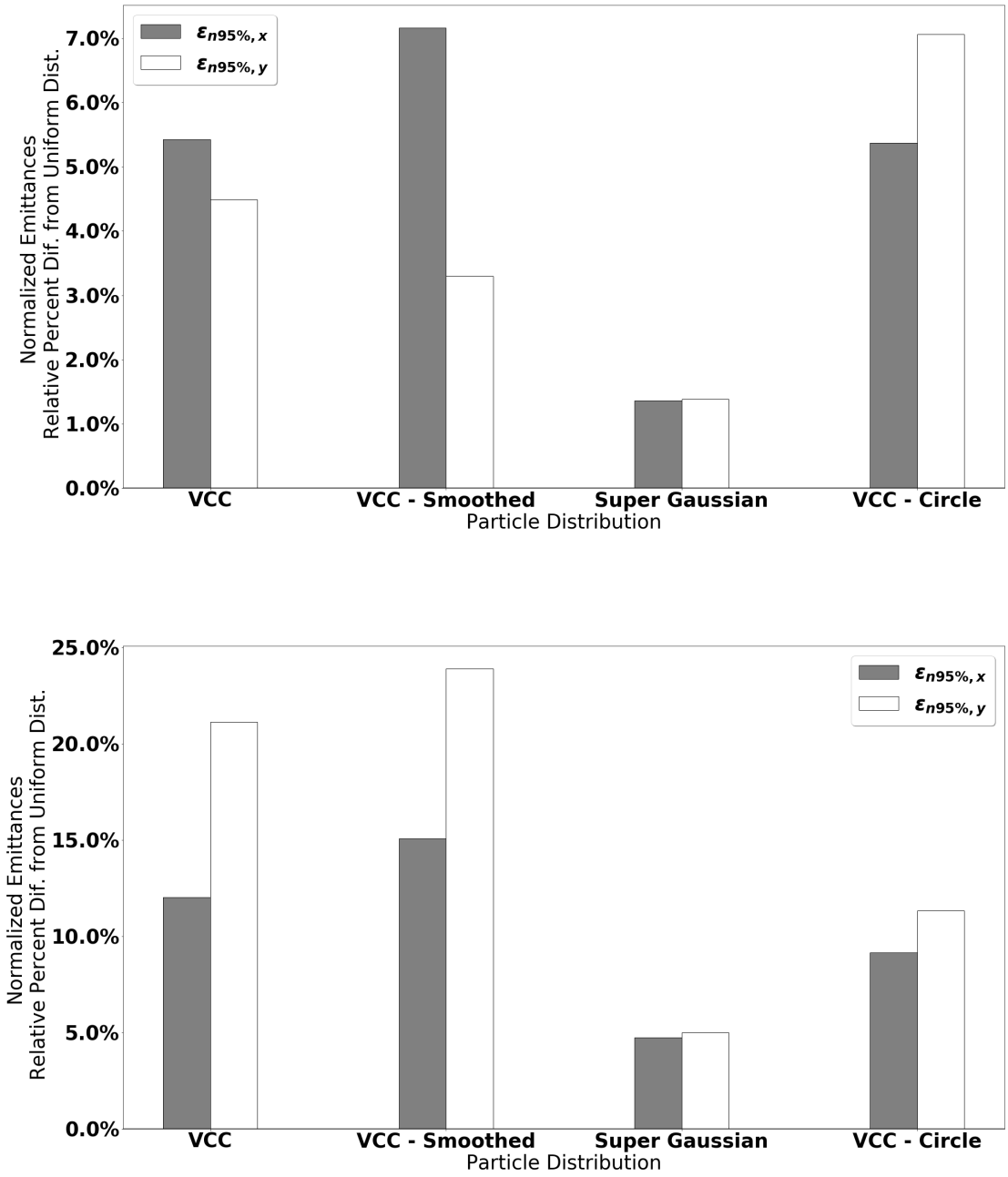


Figure 6.13: Shown are comparisons in the values of the end beam emittance simulated by LUME-Astra for each of the laser profiles shown in Fig. 6.8, Fig. 6.9 and Fig. 6.10, for two different bunch charges (top, 5 pc, bottom 50 pC). The percent difference is relative to the emittance as simulated from the uniform distributions (Fig. 6.9, d).

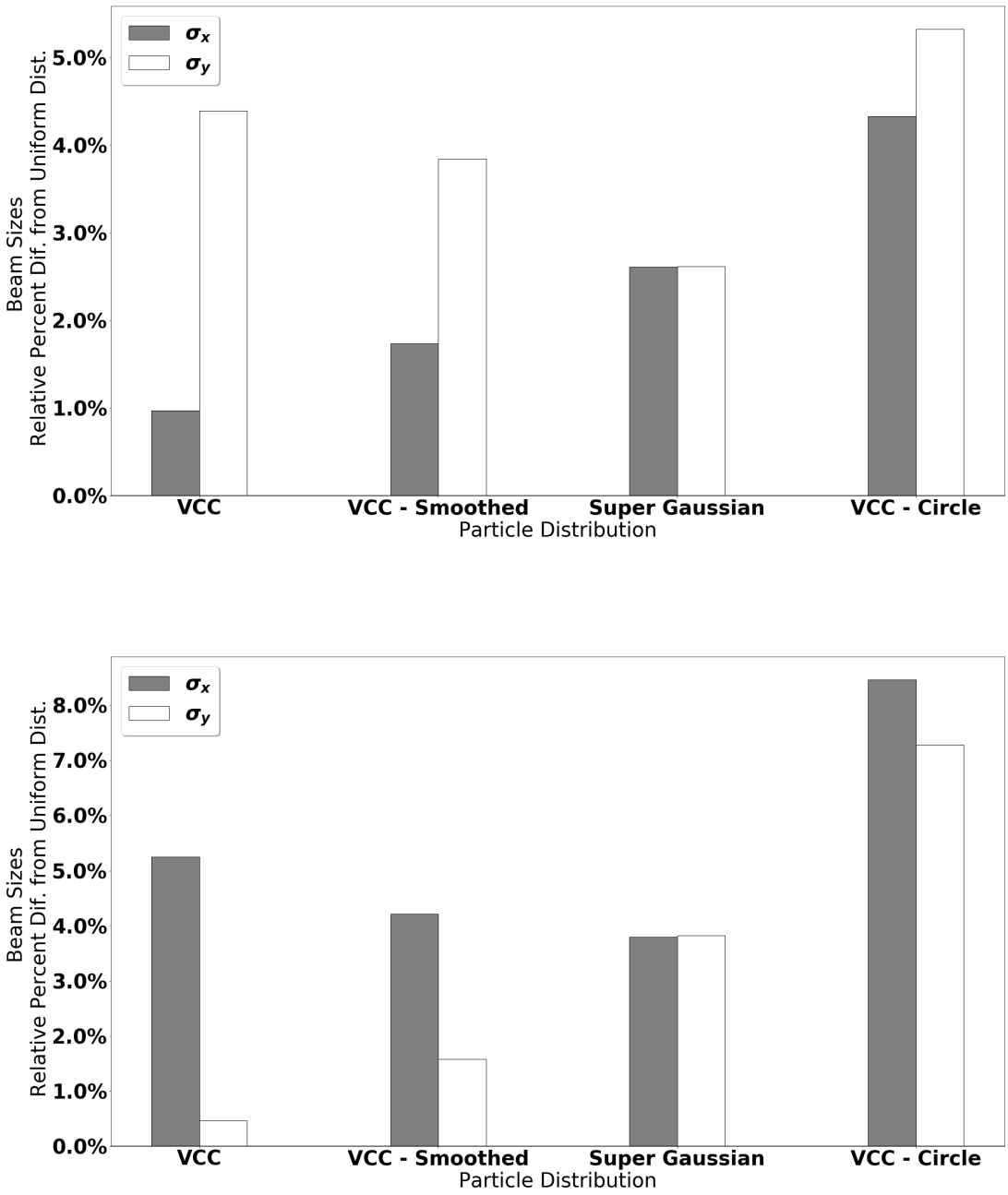


Figure 6.14: Comparisons in the values of the end beam emittance and beam sizes, simulated by LUME-Astra for each of the laser profiles shown in Fig. 6.8, Fig. 6.9 and Fig. 6.10, for two different bunch charges (top, 5 pC, bottom 50 pC). The relative difference, in percent, corresponds to the emittance or beam size as simulated from the uniform distribution (Fig. 6.9, c).

6.7 Convolutional Neural Network Surrogate Model

Having determined that the Astra simulation is sensitive to the nonuniform laser distributions, training data was generated by running measured VCC laser distributions and idealized SG laser distributions through LUME-Astra while randomly sampling injector input settings. For each unique laser profile, 2,000 randomly-sampled points in the input space were generated. The predicted output includes the electron beam distribution as it might be measured at a YAG screen, along with bulk statistical quantities such as normalized emittance and beam sizes. Two simulated data sets consisting of approximately 60,000 samples from SG particle distributions and 70,000 from VCC measurements were generated. Thus, a sample in this data set consisted of the scalar inputs including the dimensions of the input laser distribution and also the 50×50 binned input distribution, and the associated scalar output values and 50×50 binned output electron distribution.

The training was completed at NERSC, using a single GPU for training acceleration. This allowed the training of this model, which has 2,154,211 trainable parameters, to complete in hours. Some example results are shown in Fig. 6.16.

Further, the model can still successfully predict the bulk scalar electron beam parameters, such as beam sizes and emittances. This is shown in Fig. 6.19. In these cases, the average percent error is very low; lower than the measurement resolution for these quantities. Therefore, such predictions, once validated on the machine, could be very useful during operation or in offline use. Another important feature of these models is that all of the predictions can be calculated at the same time, due to the vectorized nature of the model evaluation. Thus, calculating thousands of predictions can be done simultaneously, and within seconds. Again, comparing this to the several minutes needed to run an Astra simulation, a well-trained model could quickly makes large parameter scans, online optimization, offline design work, and much more very accessible.

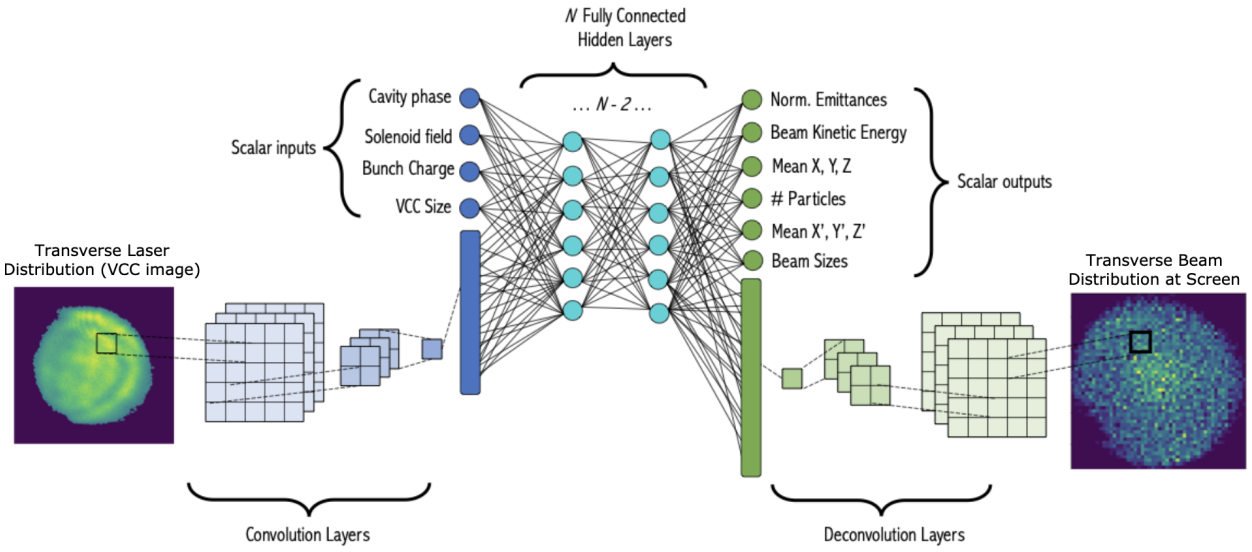


Figure 6.15: Encoder-Decoder CNN architecture used for prediction of beam transverse distributions and scalar beam parameters, with the VCC laser distribution as a variable input. To process the VCC images (binned into 50×50 pixels), the encoder consists of 3 convolutional layers with 10 filters each, alternating with max pooling layers for $2 \times$ downsampling. The scalar input settings are concatenated into the first of 4 fully-connected layers in between the encoder and decoder. The scalar outputs are obtained from the last of these layers. Finally the decoder CNN consists of 3 convolutional layers alternating with $2 \times$ upsampling layers, resulting in an output transverse beam prediction image with 50×50 bins.

Surrogate Model Generalization

Another critical consideration for surrogate modelling efforts is determining how well the model can stay relevant. Because the laser fluctuations and drift in the machine are ever-present and ever-growing, if the model is not trained on a wide parameter range including possible laser distributions, the model may only be relevant for a short period of time after it was prepared (if not retrained, discussed later).

To assess the ability of the model to generalize to laser distributions that were not used in training, and are significantly different from the training distributions, VCC images that had patches of intensity within the bulk of the laser spot missing, were held out of the training and validation sets. We find good agreement between simulation results and the surrogate model predictions shown in Fig. 6.18. This indicates that the model can be used online with

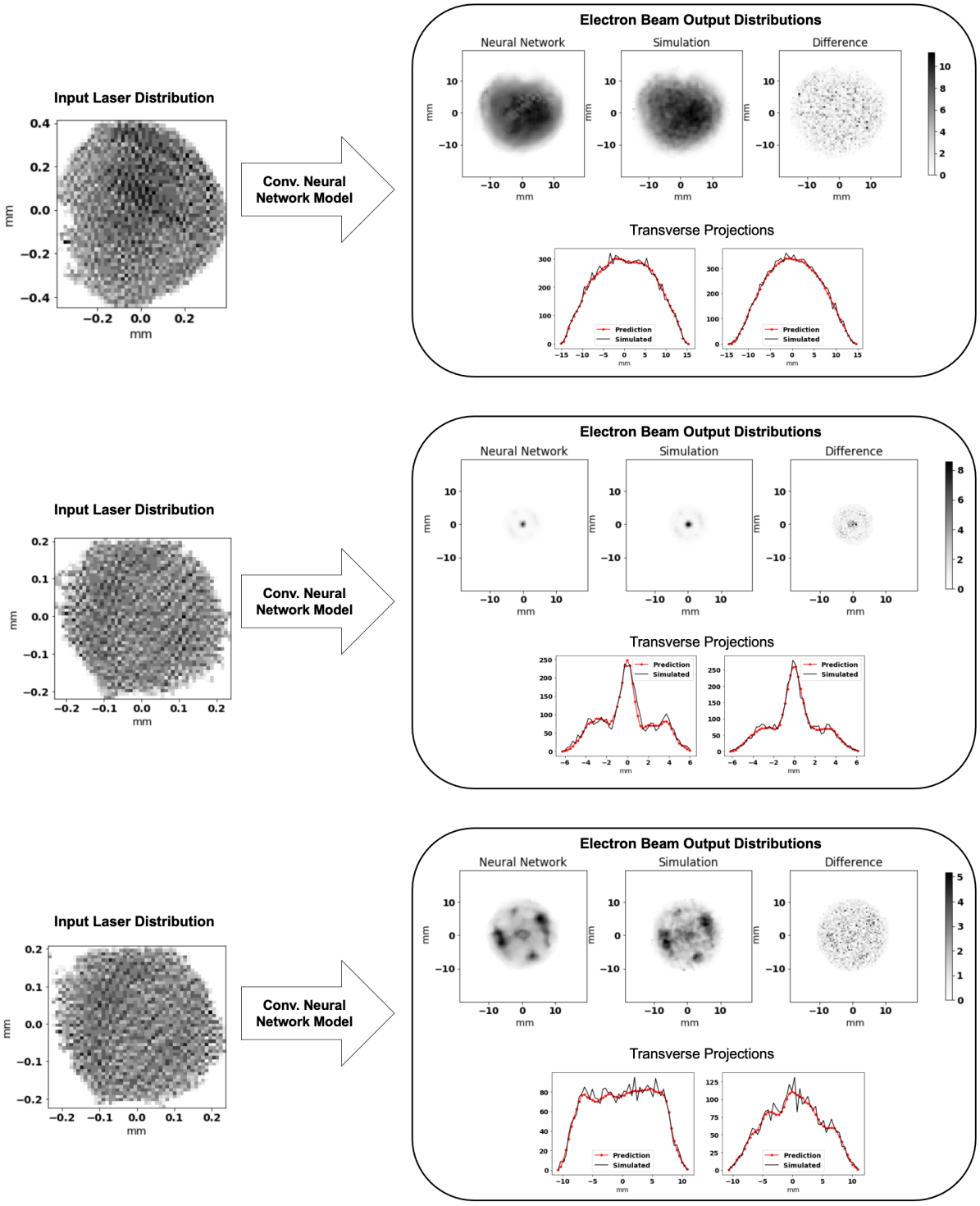


Figure 6.16: A nominal example of the CNN prediction, based on the shown laser distribution.

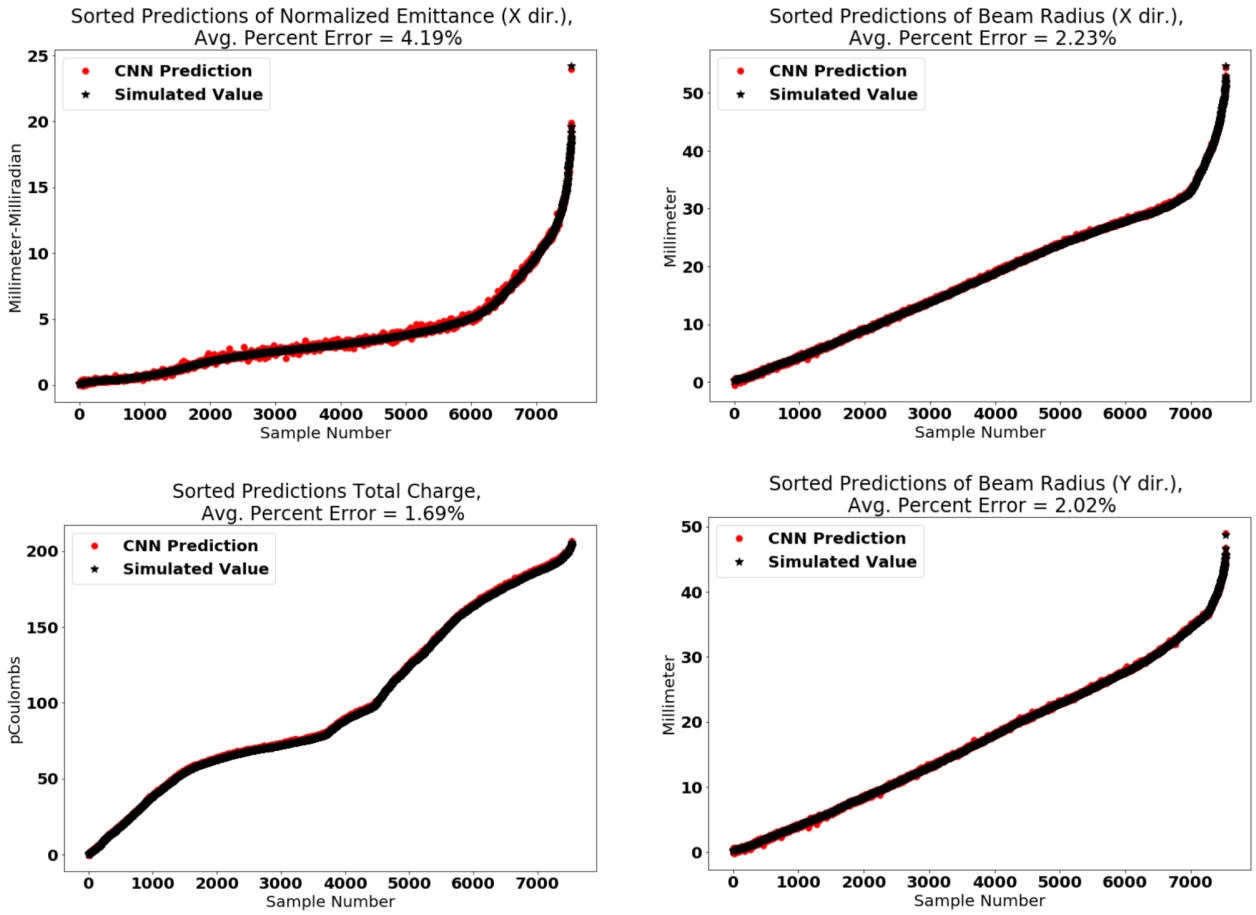


Figure 6.17: A selection of scalar electron beam bulk parameter predictions, sorted on magnitude in order to easily compare to the test set simulated value. These are representative examples of values that would be useful to predict rapidly, as well as the accuracy of this model.

the running accelerator to provide non-invasive estimates of the transverse beam profile (i.e. as both an online model of the injector and a virtual diagnostic), similar to how an online physics simulator could, but with much faster-to-execute predictions. The performance of the model on bulk scalar predictions is shown in Fig. 6.18.

6.8 Transfer Learning: from Simulation to Measured Training Data

Transfer learning encompasses a broad class of machine learning approaches wherein the performance of a model at a particular task or domain may be improved by transferring information from another related but different task or domain (49). In traditional approaches to machine learning, the distribution over feature space and the distribution in target space must be identical during training and deployment. If any such differences, termed distribution shifts, exist, the performance of the trained model is severely degraded (50). Transfer learning is thus one approach to handle distribution shifts between target domains (e.g. simulation to measurements, idealized laser beam shapes to non-idealized ones), and it has been successfully applied to diverse applications including image classification (51), anomaly detection (52), text sentiment analysis (53), etc.

We are able to compensate for the difference between the injector simulations and measurements by using transfer learning (54; 55), resulting in a surrogate model that is more representative of the real machine and can interpolate between VCC images more accurately than a model trained only on measured data.

Ample simulation data was generated, and a sparse sub-sample was used for surrogate model training, which can ensure the model is able to interpolate well, and minimize overfitting. Sub-sampling was also used to emulate small amounts of data for re-training (as one would typically have for measured data sets on an accelerator).

The model architecture described, and shown in Fig. 6.20, depicts the general architecture

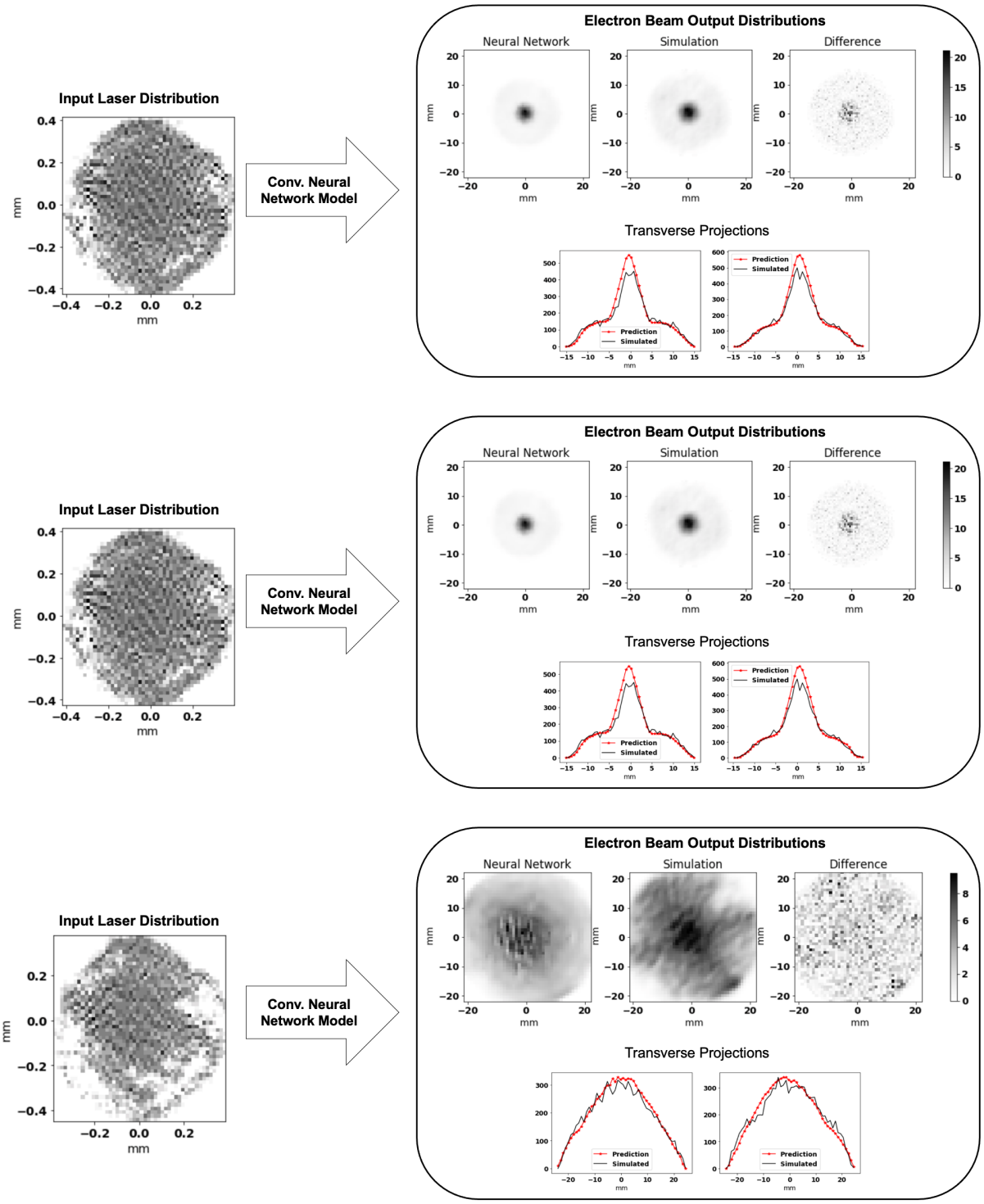


Figure 6.18: A nominal example of the CNN predictions for laser profiles which were significantly different than the laser samples used during the training process.

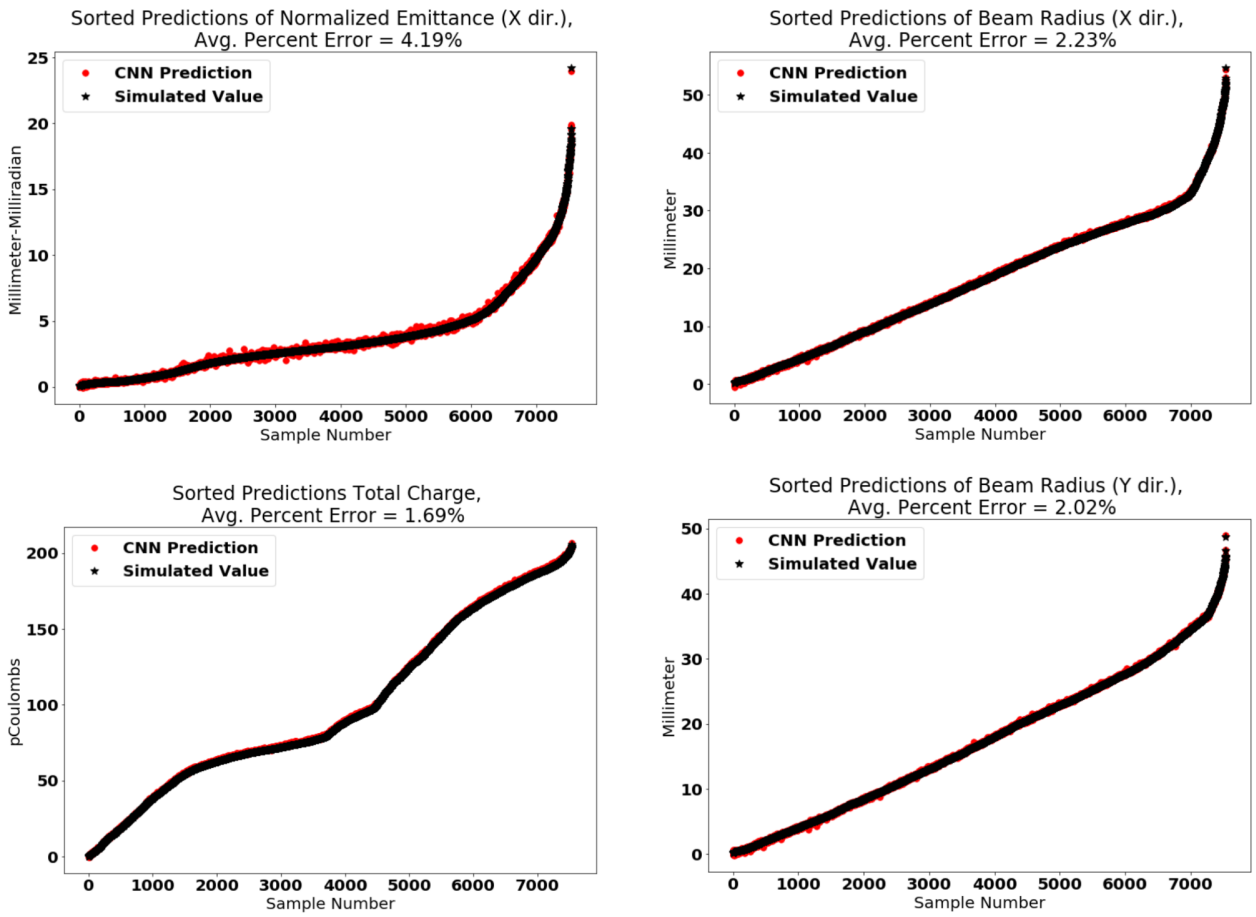


Figure 6.19: A selection of scalar electron beam bulk parameter predictions, sorted on magnitude in order to easily compare to the test set simulated value. These are representative examples of values that would be useful to predict rapidly, as well as the accuracy of this model.

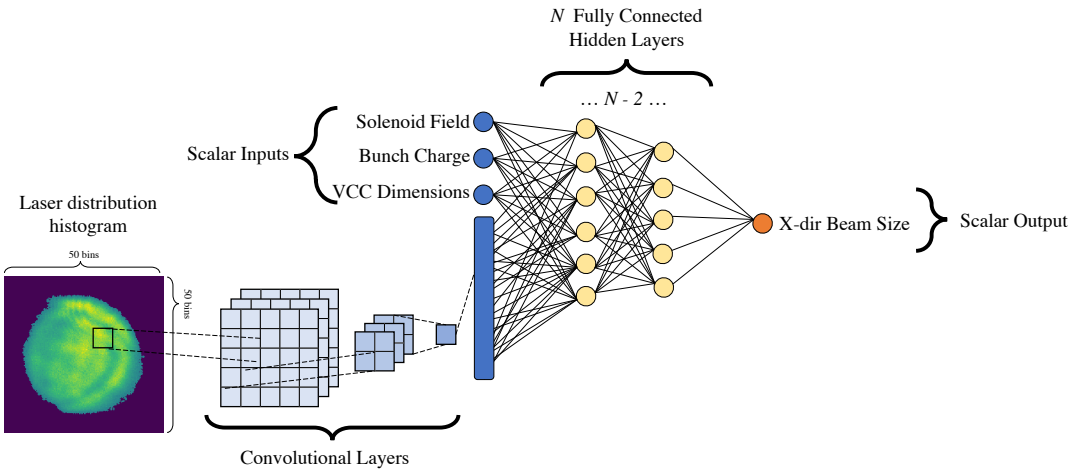


Figure 6.20: This schematic shows the surrogate model architecture used for transfer learning between the simulation and measurement domain. Scalar settings and a histogram of the laser distributions are used to predict scalar output values including emittances and beam sizes.

of the NN surrogate models. All models took scalar settings for the solenoid value and charge as inputs, along with the 2-dimensional histogram representation of the laser intensity on the cathode. The laser distributions were 50×50 bins. The size of the laser distribution were given as the horizontal and vertical extents of the histogram, relative to the center of the histogram. These six scalar values and the 50×50 bin laser distribution are considered inputs to the models. The binned images were input into convolutional layers. Three convolutional layers with 10 4×4 filters each are applied to the image inputs. The resulting nodes are then fed to densely-connected layers. The densely-connected part of the network consists of 6 hidden layers with 1024, 512, 256, 64, 32, 16, 6 neurons respectively. The output layer consists of one node predicting the transverse beam size in x.

As before, all neural network model development and training was done using the TensorFlow and Keras libraries (56). Each model was trained by minimizing a mean squared error loss (MSE) function, using the Adam optimization algorithm (57).

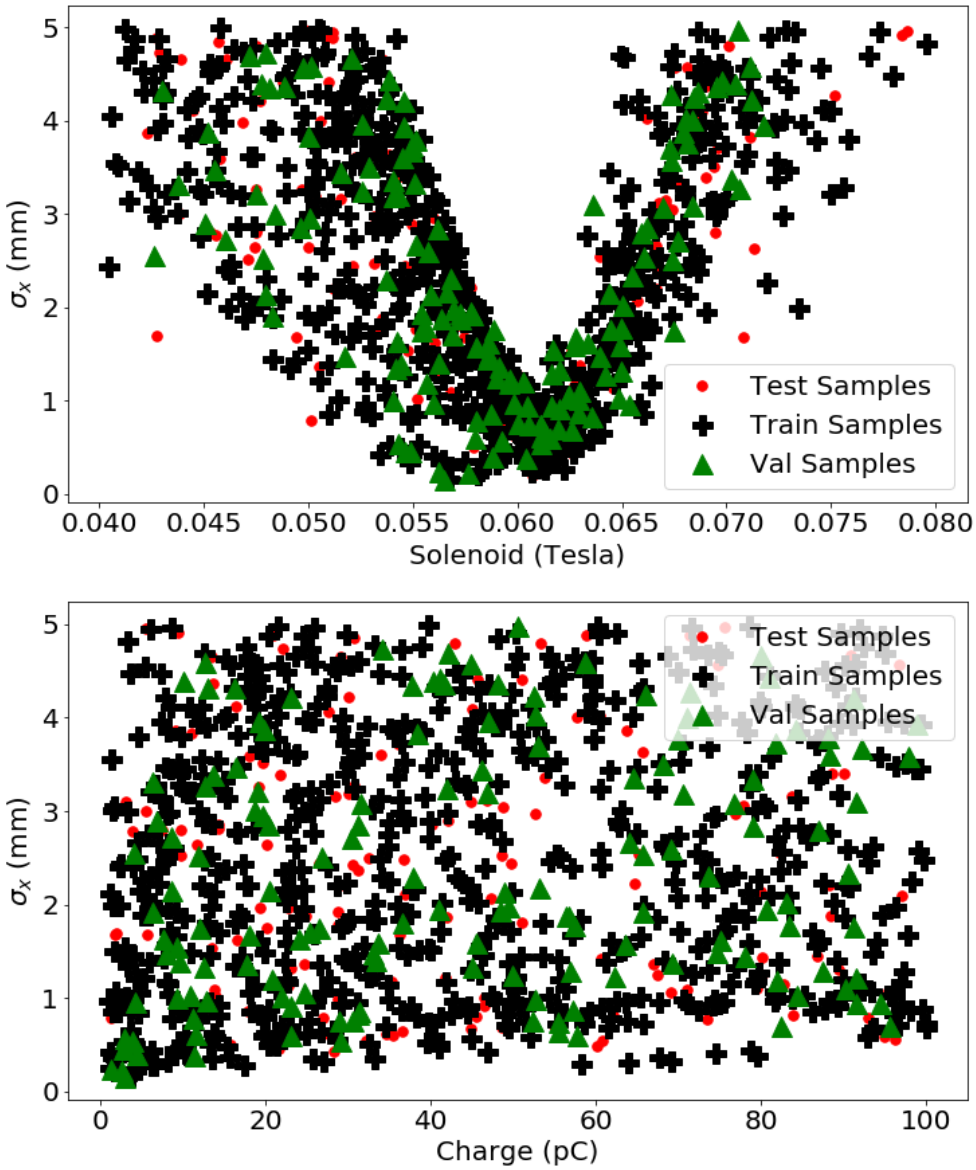


Figure 6.21: Shown are distribution of training, validation, and testing samples used to train the base simulation model. The data cover the scalar input parameter range, with a variety of SG laser distributions. There were 700 training samples, 150 validation samples, and 151 test samples (this is down-sampled data, ensuring the parameter space is not over-sampled).

6.8.1 Transfer Learning in Simulation: Idealized Laser Distributions to Measured Laser Distributions

The transfer learning approach was prototyped by training a model on SG laser distributions, then retraining the neural network model to predict VCC-based simulation results. Since this

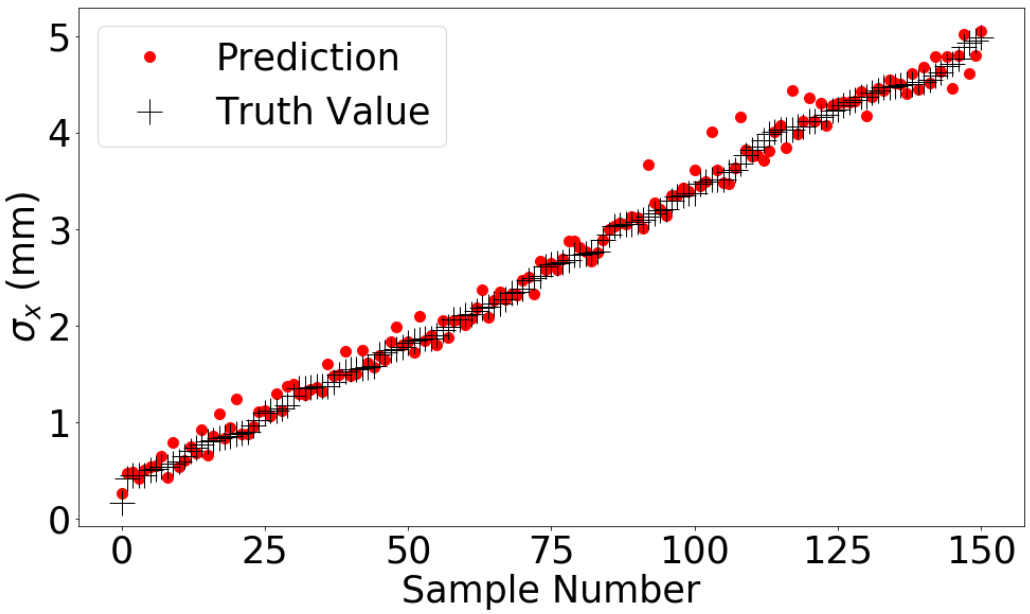


Figure 6.22: Predictions of the base model on test samples, which were withheld from training, sorted on magnitude. The MAPE for the test samples is 5.21%.

is a major potential source of disagreement between idealized simulations and the as-built injector, it enabled us to refine the approach prior to applying it to the measured data.

The SG generated data was used as the primary training data set, shown in Fig. 6.21. The data set was down-sampled to 700 training, 150 validation, and 151 test samples, which provided sufficiently sparse coverage to ensure we were not oversampling the parameter space. The resulting predictions are shown in Fig. 6.22. To evaluate the accuracy of the models, the mean absolute percent error (MAPE) was calculated as shown:

$$\text{MAPE} = \text{mean}\left(\frac{|y_{true} - y_{pred}|}{y_{true}}\right) \tag{6.2}$$

The final MAPE on the test values was 5.21%. Then, the VCC-based data set was down-sampled randomly to represent the small amount of measured data available. The down-sampled data is shown in Fig. 6.23.

In this case, after training the base model, we combine the training data sets such that the neural network is trained on both simulation data sets simultaneously (but, as described

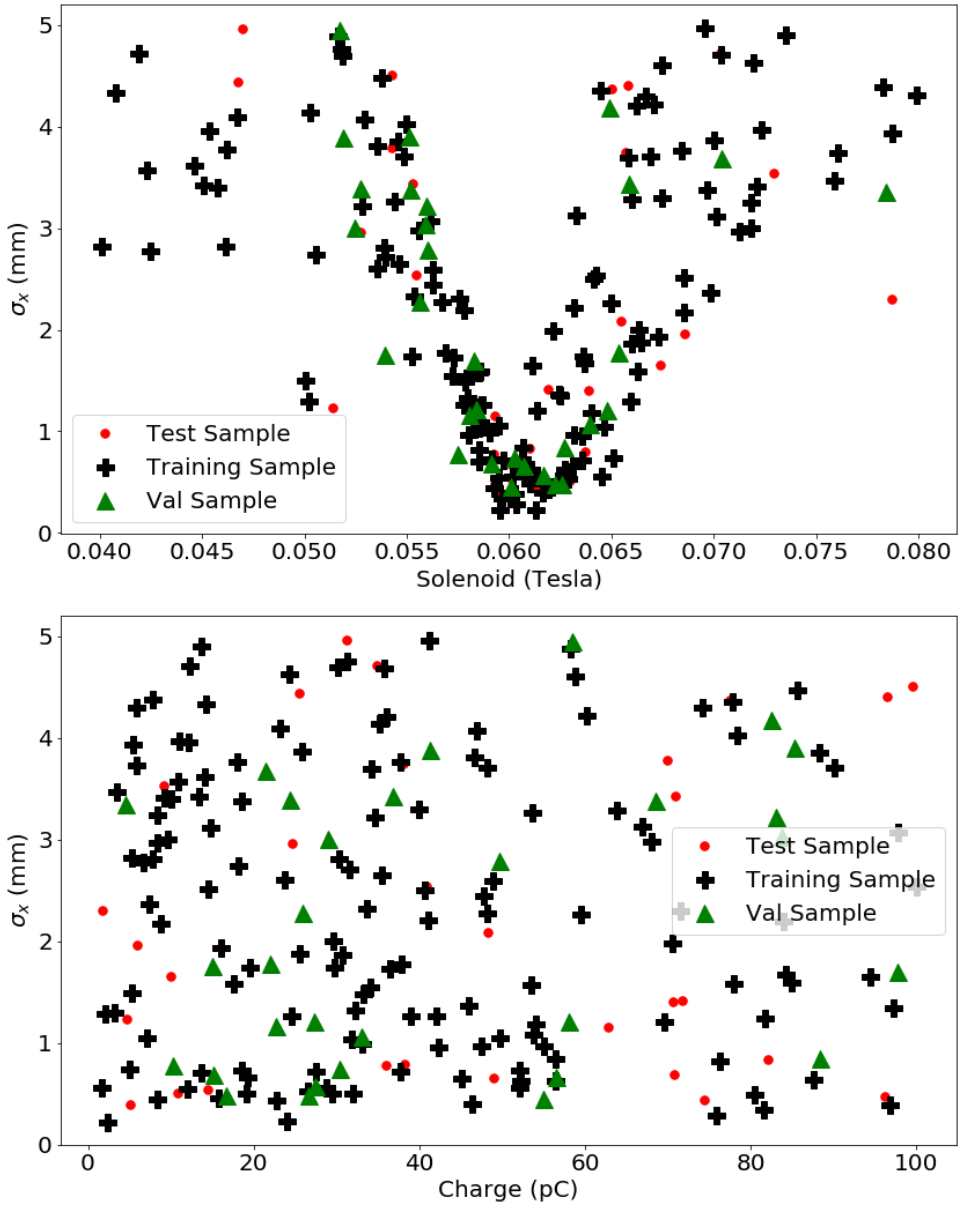


Figure 6.23: Shown are distribution of training, validation, and testing samples used to emulate a small, measured data set. These are simulation samples are generated with measured VCC laser distributions. There were 140 training samples, and 30 validation and test samples respectively.

earlier, with more limited adaption of the model allowed). Because the training data for the base model is 5 times larger, the smaller data set was repeated 5 times in order to create proportionally equal representation in the data set (a standard practice when dealing with imbalanced data sets). The performance on the combined test set (the test samples from the

SG data set, and the VCC-generated data) is shown in Fig. 6.24 and Fig. 6.25. In this case, we see that the model trained only on VCC data cannot predict the combined distribution as well as the model which underwent the transfer learning procedure. LCLS/figures/

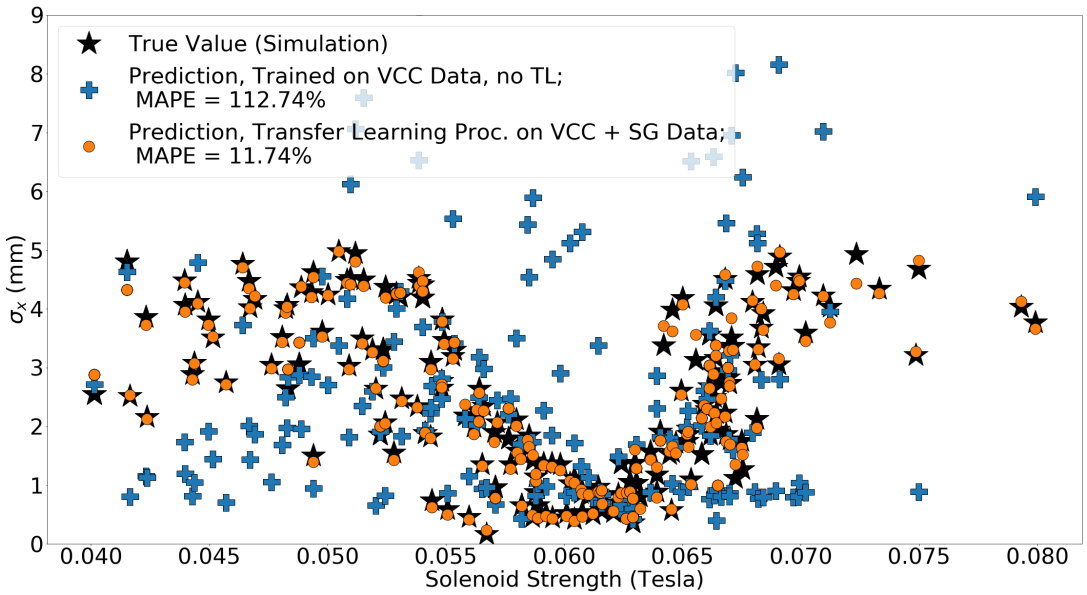


Figure 6.24: Transfer learning result in simulation, adapting from idealized laser distributions to measured distributions. Predictions of beam sizes are shown from a model trained only on measured VCC laser profiles without transfer learning, and from a model after transfer learning from idealized to measured profiles. The true values from simulation are sorted by the solenoid input value and represent the combined (idealized and measured) data.The performance of the model after transfer learning has better accuracy than a model trained soley on VCC-based data. This means the TL model can provide accurate predictions for a broader range of input parameters.

6.8.2 Transfer Learning to Measured Data

The results of the previous section show that a transfer learning procedure is necessary. The performance of the best model trained on Super Gaussian and VCC-generated data (such as the model shown in Fig. 6.24) does not predict the measured data sufficiently well, as shown in Fig. 6.26.

First, using the measured data collected, a simple surrogate model was trained for comparison with the transfer learning case. The training, validation and testing samples are

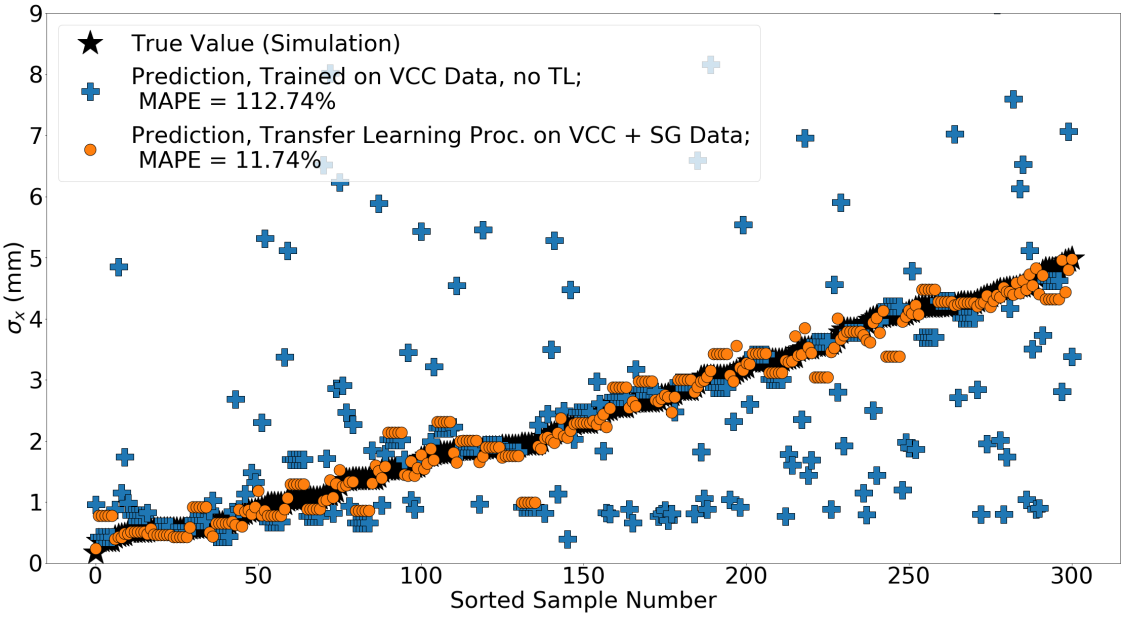


Figure 6.25: Transfer learning result in simulation, adapting from idealized laser distributions to measured distributions. Predictions of beam sizes from a model trained only on VCC laser profiles without transfer learning and a model after transfer learning with the combined data set was applied. The true values from simulation are sorted by the beam size magnitude.

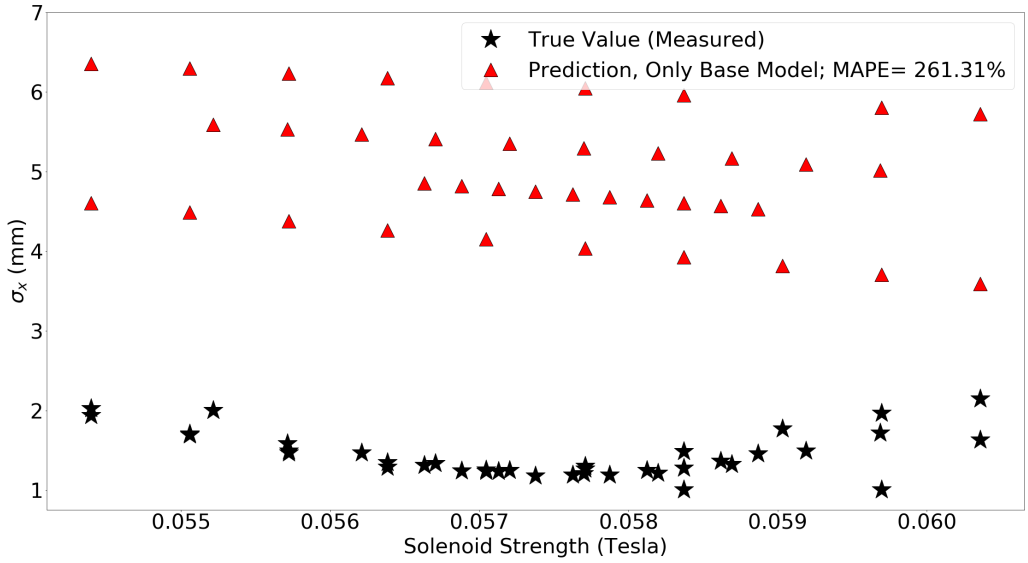


Figure 6.26: Predictions of measured data beam sizes from the simulation model (updated with measured VCC images). Despite the excellent performance on the simulated data, it is clear that the model trained only on simulation does not predict the measured beam sizes well.

shown in Fig. 6.27, where the test samples are selected from the same distribution as the training and validation data. The neural network model was able to learn from the limited data and successfully predict test samples with MAPE of 3.6%. The results are shown in Fig. 6.28. However, this measured data does not present the full range of operational values. Thus, if this model was used to predict beam size at a higher bunch charge, it would likely perform poorly relative to the true machine output. In addition, this data was already the result of generous beam time being provided for characterization (and is more than might be possible in many cases). In order to assess model performance when interpolating to new combinations of input settings we evaluated transfer learning from simulation in a case where the full operational range is present, to measurements where the model would need to interpolate to new setting ranges. This could be an effective method for producing a surrogate model, particularly when only limited measured data is available.

This scenario (missing ranges of parameter space) was emulated by withholding measurement values with charge between 20 pC and 22 pC, with the data distribution shown in Fig. 6.27. The previously prototyped procedure was attempted, but we found we needed to adjust the transfer learning procedure to accommodate the large systematic differences between the simulated and measure data. For the case with transfer learning to measured data, the main difference is the large systematic differences in the scalar output parameters, rather than in the types of input laser distributions seen during training. Thus, allowing more of the fully-connected layers to adapt to the new data was warranted.

The procedure was modified in the following ways. The base model (trained only on simulation data) for this procedure is the same as that produced previously during the simulation prototype. In this new case, the fully-connected portion (i.e. excluding the CNN layers of the neural network) were allowed to train with a reduced learning rate starting at 5e-5 and decreasing every 10 epochs for 2000 epochs. The final learning rate is then used while annealing the model (with all layers trained) for another 2000 epochs. The results are shown in Fig. 6.28.

Here, the transfer learning procedure resulted in a model that can predict the measured data as well as the model trained only on measured data, at 7.95% and 7.65%. The MAPE on the original simulated data sets were similarly comparable, at 100% and 119% respectively. There is a clear trade off between the accuracy reached on the target data set (measured data) versus the base data set (simulation data). Iteration on this procedure may further improve the agreement on both data sets as needed for experimental use. Further, the transfer learning model is able to predict on a broader range of laser input distributions, is expected to generalize better to new beam distributions.

6.9 Conclusions and Future Study

Surrogate models are a viable solution for many challenges faced while designing and operating particle accelerators. They can be used for real-time feedback in the form of virtual diagnostics, for offline experiment planning, and many other applications. In our study, we demonstrated novel methods for designing and training more comprehensive injector surrogate models.

First, a scalar surrogate model based on a wide range of simulated data was demonstrated, and we verified that it can be used for offline multi-objective optimization. Next, we showed that incorporating measured fluctuations in the initial laser distribution can improve surrogate model. Specifically, by including measured laser inputs during the training process, the model can more accurately predict beam outputs for out-of-distribution laser inputs. Previous injector surrogate models have not leveraged measured laser input fluctuations. Therefore, we showcase how important this inclusion is towards improving long term surrogate model viability during operation.

Then, to train a simulation-based surrogate model trained on idealized laser distributions to be more representative of the real machine, a simple data augmentation technique and a transfer learning procedure was able to successfully learn both output distributions (ideal and VCC-generated). As the LCLS-II injector is operated, additional measured VCC images

could be incorporated into the model using this approach. Other methods for data augmentation for improving disparity in sampling such as the Synthetic Minority Oversampling Technique (SMOTE) (58) could also be tried and compared.

Finally, we developed and applied a transfer learning procedure for transferring from simulation to measured data, which successfully reduced the model prediction error on a held out range of beam charges from 112.7% to 7.6%. Further iteration of the transfer learning process will likely improve the surrogate model training on both simulated and measured training data.

Further, the simulated data can be expanded to include more operational ranges such as gun gradient values, which may help resolve the shift in beam waist seen in the measured data. The ability of the surrogate model to successfully interpolate predictions within the range of possible input parameters (demonstrated in this case for previously unseen charges) can be very helpful for quickly estimating output parameters without needing experimental data. Our study shows that this is possible with a comprehensive machine-learning based surrogate model for the LCLS-II injector frontend. While we demonstrate this only for the LCLS-II Injector frontend, these approaches for improving online modeling of injector systems could be easily adapted to other accelerator facilities.

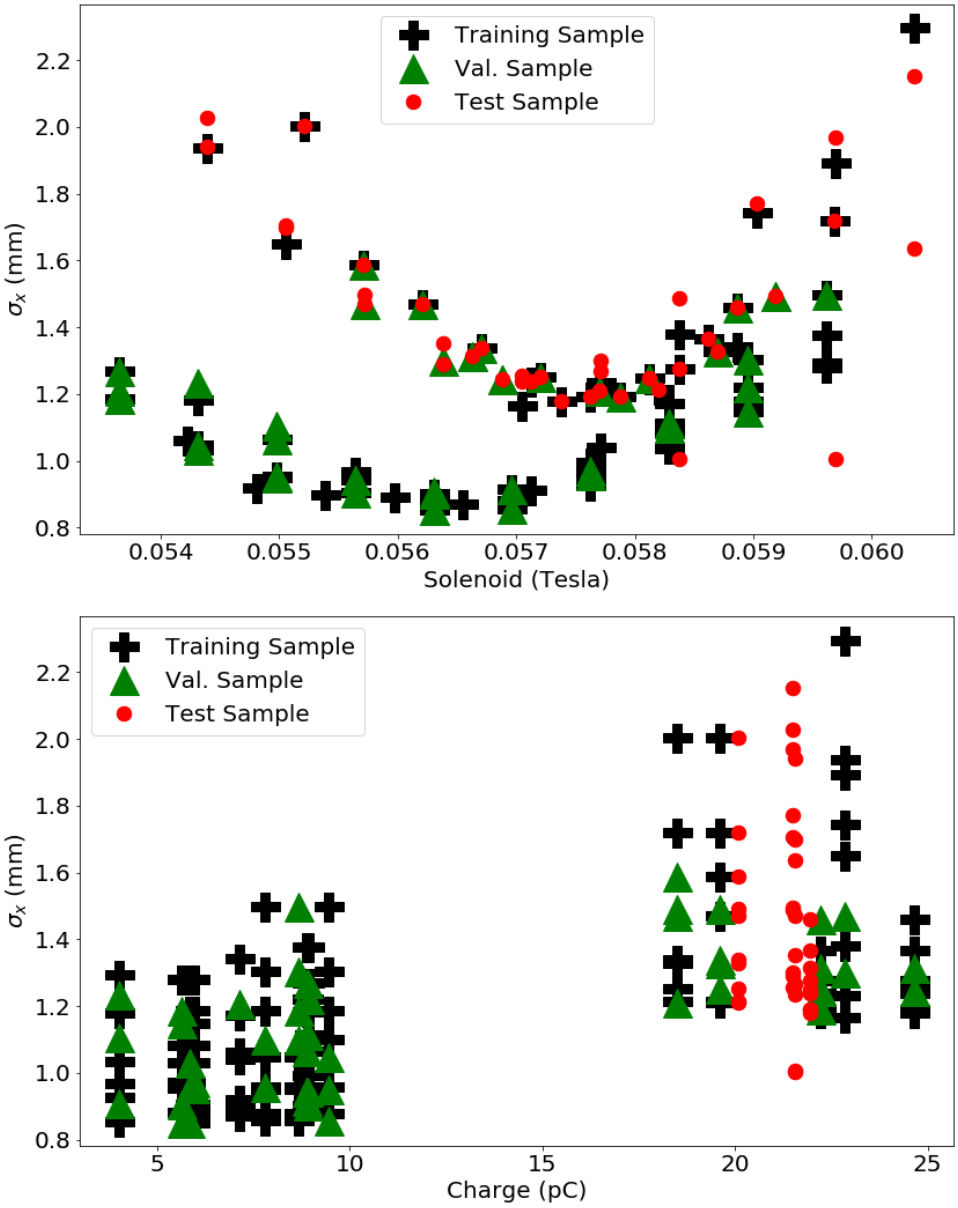


Figure 6.27: Training, validation, and testing samples for a surrogate model trained with only measured data, but with a large portion of measurements (in new beam charge ranges) withheld for test data. Shown are the 111 training samples, 48 validation samples, and 39 test samples.

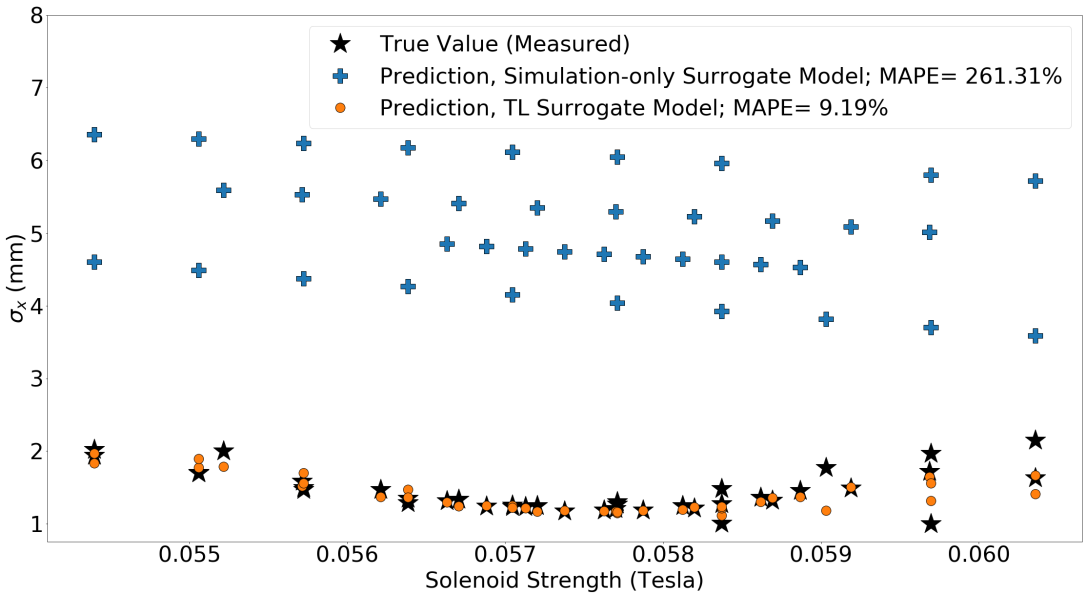


Figure 6.28: Prediction results for transfer learning between various models, predicting measured data. As shown previously, the transfer learning model trained only on simulation is still insufficient when predicting measurement. After the transfer learning procedure using, however, the model is able to successfully predict the measured data.

Appendices

APPENDIX A

DEEP LEARNING WITH QUANTIFIED UNCERTAINTY FOR A FREE ELECTRON LASER SCIENTIFIC FACILITY

This paper was submitted to the NeurIPS 2020 New in Machine Learning workshop.

A.1 Introduction and Motivation

Accelerator physicists and expert operators at dedicated light source facilities provide powerful X-rays to researchers to help advance their respective fields, while also researching how to improve accelerator technology. At the SLAC Linac Coherent Light Source (LCLS) (59), shown in Fig. A.1, electron beams are accelerated to speeds close to the speed of light, then made to radiate X-rays. This type of accelerator is called a Free Electron Laser (FEL). Due to their ability to create powerful, coherent X-ray pulses, access to FEL facilities for experiments is highly sought after. At present, the LCLS facilitates over a thousand scientific experiments per year, and many of these have fundamentally improved our understanding of important processes, such as photosynthesis (4), electron-phonon interactions that could aid the design of new materials (5; 6), and molecular interactions that assist drug discovery (**?**).

To accommodate as many experiments as possible, each visiting researcher may only have days or hours to conduct their experiment. With such profound time-pressure, it is very important that accelerator operators are able to accommodate requests for the custom beam parameters needed for the successful completion of each experiment (e.g. different x-ray pulse energies and durations) quickly, by adjusting many controllable variables on the accelerator. Therefore, time wasted during beam customization has a substantial negative impact on the scientific productivity at the facility, and is critical to address. Deep learning methods which can provide models with reliable and quantified uncertainty estimates are a viable solution to this problem.

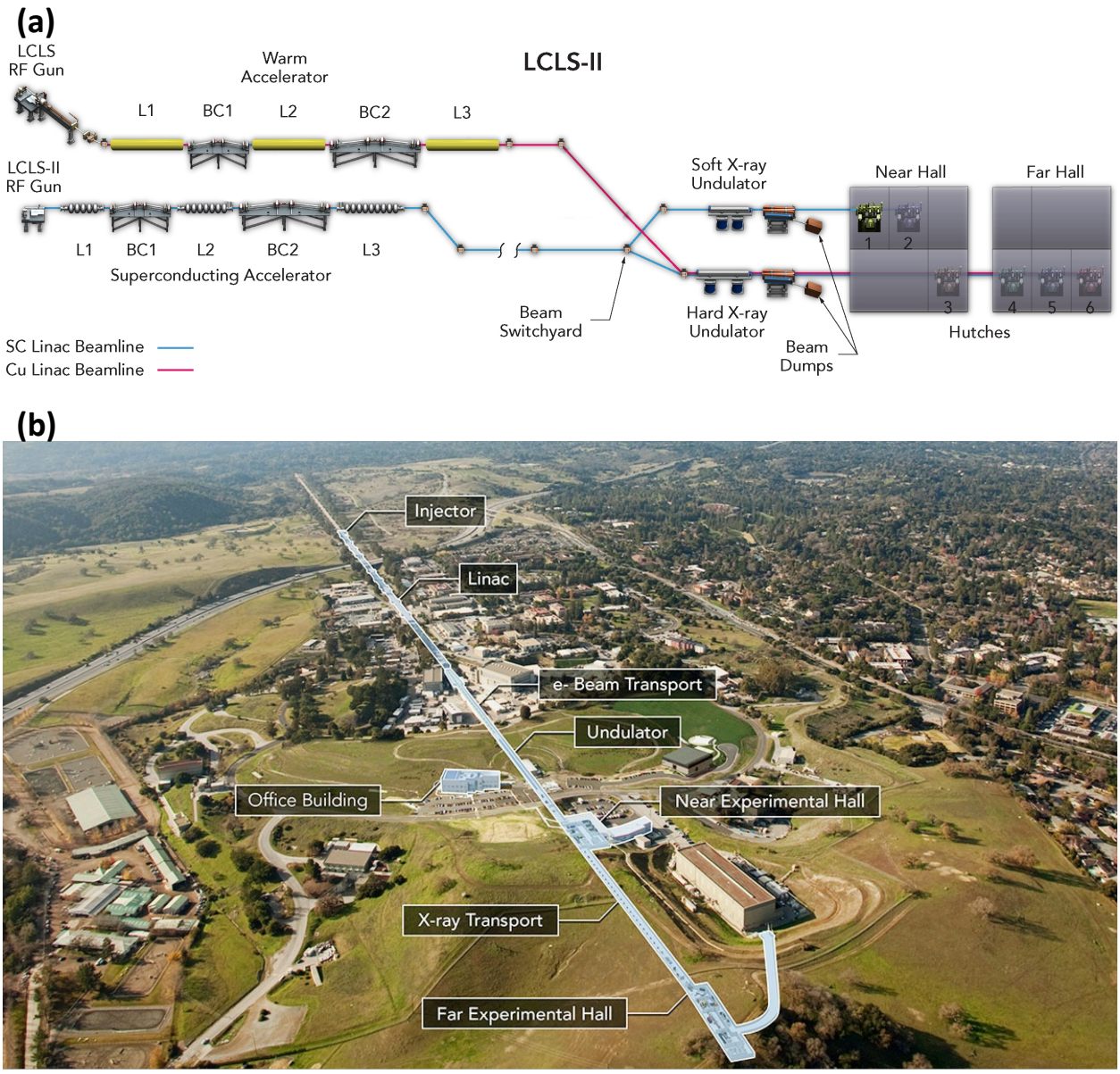


Figure A.1: The Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory. (a) Schematic of the LCLS-II accelerator complex. The Warm Accelerator, referred to henceforth as the LCLS, consists of three radio-frequency accelerating cavities: L1, L2, and L3. After the beam is accelerated, it is transported through an undulator, where it produces X-rays. These X-rays are sent to "hutches" that house experiments for visiting scientists. (b) Aerial view of the LCLS beamline and the surrounding area, in Menlo Park, California (**?**).

As accelerators have high-dimensional parameter spaces, searching for settings that will fulfill custom beam requests in a timely fashion is challenging. To help address this, one could use fast-executing models of the accelerator; however, accelerator systems are often difficult to model *a priori*. Thus, there is interest in using deep learning models to (1) speed up physics simulations of accelerators and (2) learn models directly from measured data. These models can execute very quickly to produce predictions of beam parameters given current machine settings. Predictions can also provide information about the beam that is not typically available during operation or easily measured. In addition, these models can be used to plan new experimental setups before trying them out on the accelerator.

Numerous sources of uncertainty (unknown misalignments in accelerator components, drift over time in unmeasured variables, intermittent anomalous behavior and erroneous signals) are present in the real accelerator system. In order to be useful in prediction and control tasks and avoid overly-confident predictions, these sources of uncertainty need to be taken into account in predictive models. Further, predictive uncertainties can be used directly to aid the search for optimal settings. For example, Gaussian processes and Bayesian optimization have been used to to optimize accelerator sub-sections (13; 14). However, Gaussian processes are computationally expensive for high dimensional data sets. This motivates the use of methods that are more capable of efficiently handling data with high-dimensional parameter spaces. To this end, here we investigate quantile regression neural networks (QRNN) and Bayesian neural networks (BNNs) as methods of providing quantified uncertainties for models of the X-ray pulse energy, given measured accelerator inputs. Here we assess model performance on noisy, high-dimensional data that covers a broad range of operating configurations, with the aim of obtaining an accurate model of the FEL pulse energy and associated uncertainty estimates, given a variety of accelerator settings. The results could inform future work on integrating deep learning into operation of the LCLS, including direct use of these models in online tuning of the accelerator and offline experiment planning.

A.2 Problem & Data

At many accelerator facilities, a large amount of data is stored automatically in an archive that is retained indefinitely. If one can learn an accurate system model just from this archived data, it removes the need to supplement the training data with invasive scans of accelerator parameters or simulation data. In this study, the data was curated from the LCLS archive, specifically targeting times when tuning was occurring. The full dataset includes 286,923 samples. Each sample consists of 76 scalar inputs, many of which are tunable and some of which characterize the initial beam at the start of the accelerator. The 76 inputs include focusing magnet strengths, accelerating cavity phases and amplitudes, and uncontrolled variables that can change over time. The output is a single scalar value: the photon pulse energy. Because the data spans several years of operation and different operating modes, there are several significant sources of irreducible uncertainty. One of these is "drift" in the system's response given input variables or unobserved hidden variables over time (due to, for example, equipment aging, part replacement, or other environmental changes). Another source of uncertainty is the sparse sampling of some of the inputs (e.g. due to some input variables being changed only very rarely).

Lastly, the pulse energy is a noisy measurement, due to both jitter in the beam parameters and noise in the detector response itself. Looking at periods of steady-state operation, the RMS noise in the measured data can be up to 0.3 mJ. The detector also has a sensitivity threshold below which the detector is not effective. In addition, the detector sometimes produces a positive signal when no beam is present. To assess model robustness, in this study, some of the models shown in section A.3 and section A.4 are trained without removing outliers. Models trained on data after removing erroneous or low-energy values are also shown, to provide a comparison of performance.

As the SLAC LCLS is a facility which provides custom photon pulses (X-rays) to experimenters, predicting this energy is very important. Thus, this dataset encompasses the primary service that is provided by the SLAC FEL to users, and is therefore important to

be able to model well.

A.3 Quantile Regression Neural Networks

Classical deterministic neural networks are an accurate and flexible approach to model a wide range of non-linear regression tasks. The training of such deterministic neural networks attempts to estimate the conditional mean of the target, via the minimization of a mean squared loss. However, in complex scientific applications, the data may have a significant noise component, and this noise process is often heteroscedastic. This is true for the dataset considered in this study. Additionally, the noise may correspond to an underlying distribution that is not approximated well by a Gaussian. For instance, the target data may be severely skewed, or strictly non-negative. In such cases, the classical neural networks may have limited benefit. It may be more advantageous to model the conditional distribution of the noise in the data, in addition to the conditional mean.

To this end, a parametric distribution may be pre-specified for the target, and different regression models are trained to estimate the parameters of this distribution, conditioned upon the data. Such applications are described as conditional density estimation networks, and provide estimates of the parameters of a pre-supposed family of distributions. A more robust alternative is to estimate the point predictions of different quantiles, using sets of quantile neural networks. For a random variable X with a cumulative distribution function, $F_X(x) = P(X \leq x)$, the t-quantile, q, is given by $qF_X(q) = t$. Such quantile regression approaches are more robust as they invoke no assumptions about the parametric form of the final distribution. Additionally, estimated quantiles are equivariant to monotonic transformations, thus the quantiles for the primary target quantity of interest can be used to predict quantiles of derived quantities of interest.

Such quantile regression approaches apply asymmetric weights to positive and negative

errors, and the error metric is often referred to as the tilted loss, given by:

$$\mathcal{L}(f(x_i)|\tau) = \begin{cases} \tau(y_i - f(x_i)), & \text{if} \quad y_i \geq f(x_i) \\ (\tau - 1)(y_i - f(x_i)) & \text{otherwise.} \end{cases} \tag{A.1}$$

Here we use QRNNs on the measured FEL data set. The design of the model architecture used to train the QRNNs was as follows. All model creation and training was done using the TensorFlow 2.0.0 library (60) and Keras API (56). The model consists of an input layer for all 76 scalar inputs, 8 fully connected hidden layers which decrease in the number of neurons from 80 to 10, by a factor of 10 neurons for each subsequent layer. The activation function for each layer was the hyperbolic tangent function. The final output layer of 1 scalar output provides the prediction of the photon energy. The median prediction (50% quantile), 97.5% quantile prediction, and 2.5% quantile prediction, were each fit using independent models. To create the median prediction, and the corresponding confidence intervals, the same training and validation data were used to train each independent model. A custom loss function was written based on Eq. A.1, and each model was optimized using Adam (61) for its given quantile. All QRNN models were trained with a batch size of 4096 samples for a maximum of 5000 training epochs. Early stopping was implemented, such that the loss on the validation set was monitored and training terminated when the loss showed no improvement for 500 epochs.

The basic QRNN model was trained on 80% of the full dataset, with the remaining 20% split equally into a validation and test set. No further hyperparameter optimization on the model or training parameters described above was conducted. The results on the test set are shown in Fig. A.2. Several individual predictions including the lower and upper quantile predictions are shown with the measurement value are shown in Fig. A.3. These samples are randomly chosen from the test set, and show how well the median prediction model performs.

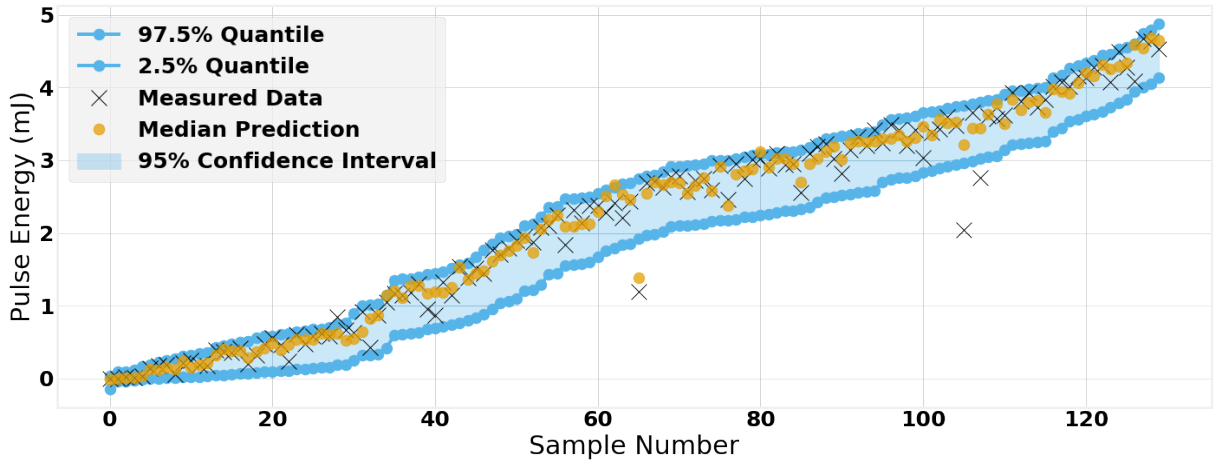


Figure A.2: The QRNN model results for the test set. The prediction coverage for these models is 92.84%. The test set mean absolute error (MAE) is 0.13 mJ.

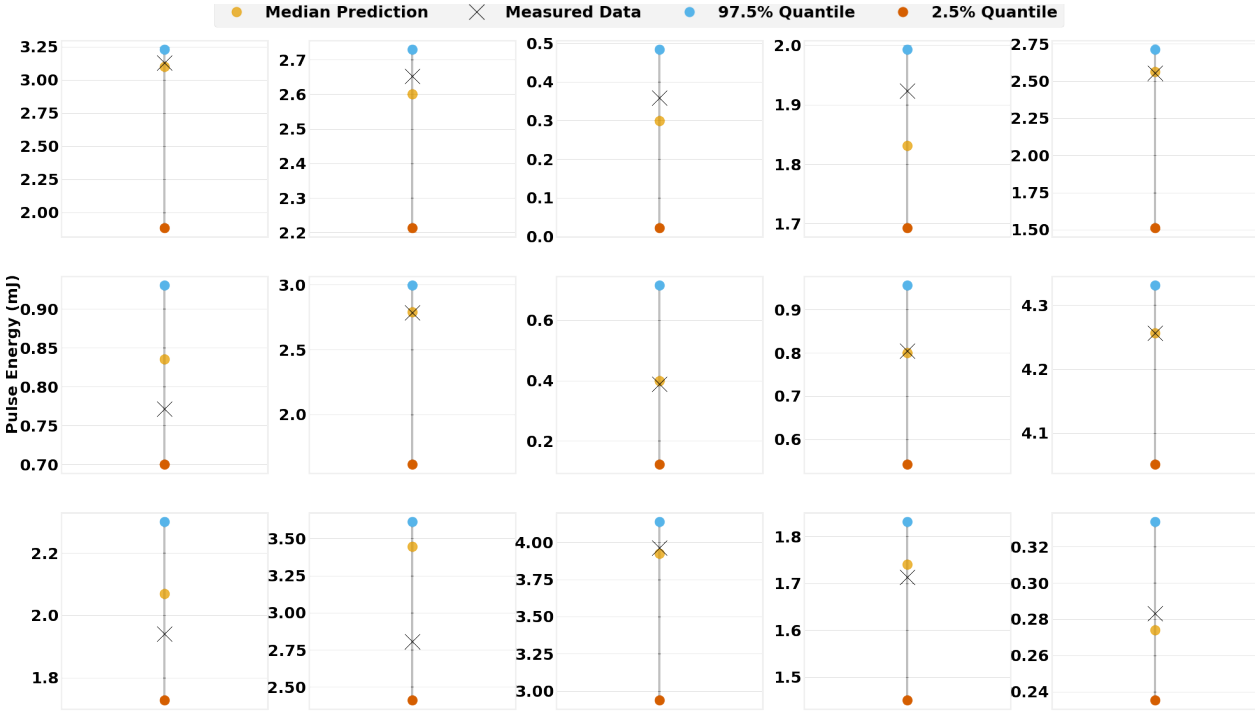


Figure A.3: QRNN prediction results on randomly selected test set samples, with median prediction and confidence interval shown along with the measurement value.

The coverage probability of measurements from the test set was 92.84%, which is close to the desired 95% which is the chosen confidence interval for uncertainty estimation. The mean absolute error (MAE) is 0.13 mJ. This performance was obtained in the presence of outliers

and measurements considered below the sensitivity threshold of the detector. By including these measurements, we can see that the QRNNs are very robust to these outliers and able to capture trends despite the presence of noise. These results suggest that QRNN models are good candidates for online prediction. Further, these models could be used for online model retraining as more data is collected, without any data filtering or manipulation required, and still provide reliable predictions of the FEL pulse energy with reasonable uncertainty estimates.

Because the aim is to obtain a predictive model of the FEL output, we also removed sections of time-ordered data as extra test sets. This is shown in Fig. A.4 and Fig. A.5. The data available for training is shown in black, and this was spilt into training and validation sets which were held constant for each set of QRNNs (the median prediction and confidence interval models). For the models shown in Fig. A.4, the output data spanned the full range of potential photon energies, 0 to 5 mJ. For the models shown in Fig. A.5, photon energies below 0.2 mJ were removed.

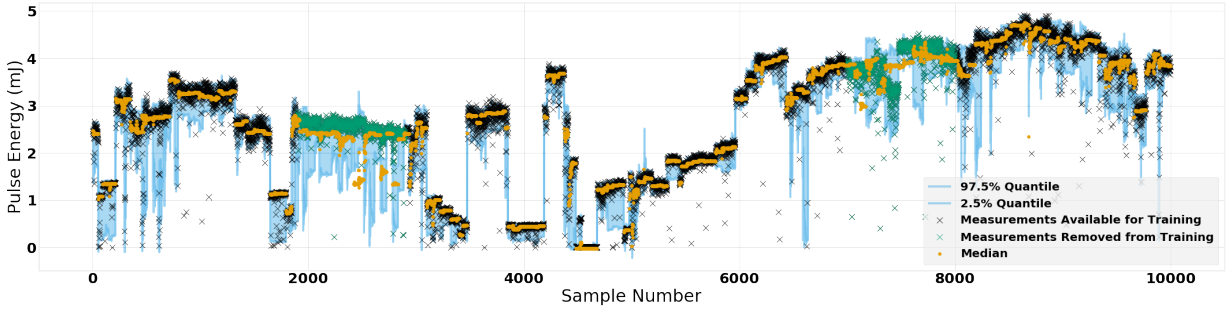


Figure A.4: QRNN results demonstrating predictions on regions of chronological data which were withheld from training. The coverage probability is 53.35%, and the MAE is 0.34 mJ. This model was trained on data spanning the full potential photon energy range (0 - 5 mJ).

First, a portion of chronological data in which the photon energy measurements are fairly stable throughout the segment was chosen. By observing the performance on this portion of the data, the ability for the model to predict on stable operating parameters can be observed. Another portion in which the measurement value changes significantly (more than 50%) was also chosen. Similarly, the ability for the model to predict different operational energies can

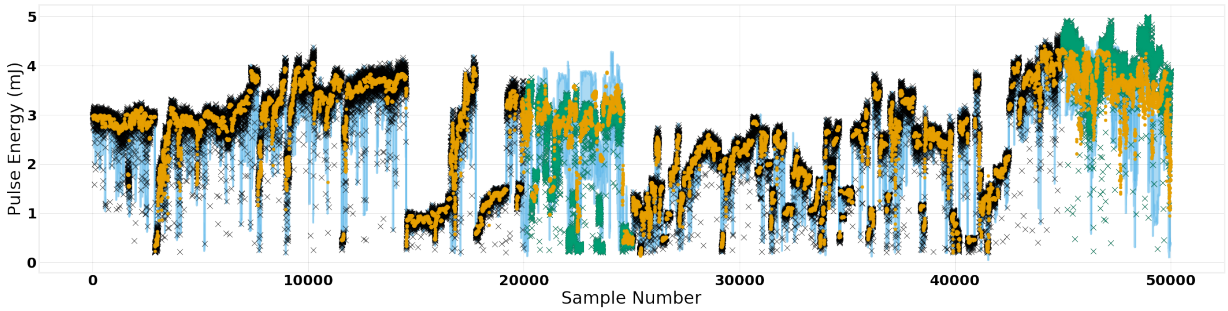


Figure A.5: QRNN results demonstrating predictions on regions of chronological data which were withheld from training, with legend provided in Fig. A.4. The coverage probability is 42.83%, and the MAE is 0.60. This model was trained on data spanning photon energies ranging from 0.2 - 5 mJ. The cut at 0.2 mJ was done to remove low-energy samples, where noise from the detector may dominate over the signal.

be observed in regions where the measured values change drastically.

The coverage probability was calculated on the regions of data removed from training. For Fig. A.4, the coverage probability for both removed portions of data was 53.33%, and for Fig. A.5 the coverage probability was 42.83%. While this demonstrates that the confidence interval is not capturing the full extent of possible measured values, the median prediction performs well, considering how challenging of a problem this is. For Fig. A.4, the MAE is 0.34 mJ, which is commensurate with the RMS noise in the measured data (and thus is about as good as we expect to be able to predict the output). Similarly for the models in Fig. A.5, MAE is 0.60 mJ. These results suggests that these models are able to capture the general trends present in the data, which is quite surprising considering how high-dimensional, noisy, and sparse the input space is. Large accelerators like the LCLS are notoriously difficult to model empirically, so this is a very encouraging result.

A.4 Bayesian Neural Networks

In contrast to deterministic neural networks, in Bayesian Neural Networks the weights and biases are assumed to be random variables with corresponding probability distributions. Deterministic weights and biases lead to deterministic predictions in the output of the neural

network. However, in the Bayesian Neural Network the weights and biases are random variables with probability distributions, leading to an output prediction that has an associated probability distribution as well. Given training data, Bayesian inference over these BNNs estimates the posterior probability distributions for these weights and biases. While querying from such a trained BNN, the mean of predictions is given via expectations over the posterior distributions: $P(y^*|x^*) = E_{P(W|D)}\left(P(y^*|x^*, W)\right)$.

Estimating the posterior distributions over the weights and biases for large networks was historically an intractable problem. Prior researchers had proposed searching instead for a variational approximation to the Bayesian posterior probability distribution on the biases and weights (**?**). Such variational approaches propose a family of parameterized distributions for the weights and biases, $q(W|\theta)$. Thence, the optimal value of the parameter, θ is approximated so as to minimize the Kullback-Leibler (KL) divergence between the parametrized distribution and the true Bayesian posterior probability distribution. Explicitly, this optimal parameter, θ^*, is given by

$$\begin{aligned}\theta^* =& \theta KL\big(q(W|\theta)||P(W|D)\big) \\ =& \theta \int q(W|\theta) \log \frac{q(W|\theta)}{P(W)P(D|W} dW \\ =& \theta \mathrm{KL}[q(W|\theta)||P(W)] - \mathbb{E}_{q(W|\theta)}[\log P(D|W)]\end{aligned} \tag{A.2}$$

Using this as a surrogate objective function, the problem of stochastic inference is reduced to one concerning optimization. The resultant loss is often referred to as the variational free energy, or the evidence lower bound (ELBO) and is given by

$$F(D, \theta) = \mathrm{KL}[q(W|\theta)||P(W)] - \mathbb{E}_{q(W|\theta)}[\log P(D|W)] \tag{A.3}$$

The first term of this objective is referred to as the complexity cost and depends on the priors used. The latter term is referred to as the likelihood cost and depends on the training data. In practice, approximations are utilized to simplify the expectation and the

minimization is carried out via steepest descent. In this investigation, we carry out this optimization using Monte Carlo sampling in conjunction with local parameterization (**?**). Here, the objective function is estimated via Monte Carlo sampling via

$$F(D,\theta) \approx \sum_{i=1}^{n} \log q(W^{(i)}|\theta) - \log P(W^{(i)}) - \log P(D|W^{(i)}), \tag{A.4}$$

and $W^{(i)}$ represents the i-th Monte Carlo sample from the variational posterior. Following best practice, we use a standard normal as the prior on the weights of the network. The joint prior over the entire set of weights and biases of the network is given by the product of the individual independent normal distributions over individual parameters, $P(W) = \prod_i \mathcal{N}(W_i|0,1)$. This approach adheres to the Bayes By Backprop algorithm. However, there are issues associated with the application of Bayesian Neural Networks to real world scientific problems. BNN inference algorithms often utilize simplifications and approximations to ensure scalability and computational tractability. As a consequence, This often leads to unreliable uncertainty estimates, due to the sensitivity to hyperparameters, the affect of outliers in the data, the verity of the assumptions made regarding the posterior distributions, etc.

For inference, approximate variational inference with the Bayes By Backprop algorithm (**?**) is used. The network architectures are selected using Bayesian Optimization (**?**). The final network architecture had 6 fully connected layers. To optimize the variational parameters, we utilize the Adam algorithm (61) wherein the rates are set using cross-validation. The activations across all neurons are Rectified Linear Units (ReLU), and all weights and biases are initialized with standard normal priors. The model was trained using measurement data in samples below the detector sensitivity threshold (0.2mJ) were removed.

Despite hyperparameter optimization, the BNN was not able to accurately or consistently predict the true value. Further, the confidence interval estimated from the BNN predictions at each point do not provide 95% coverage. While Fig. A.4 does display that the BNN is

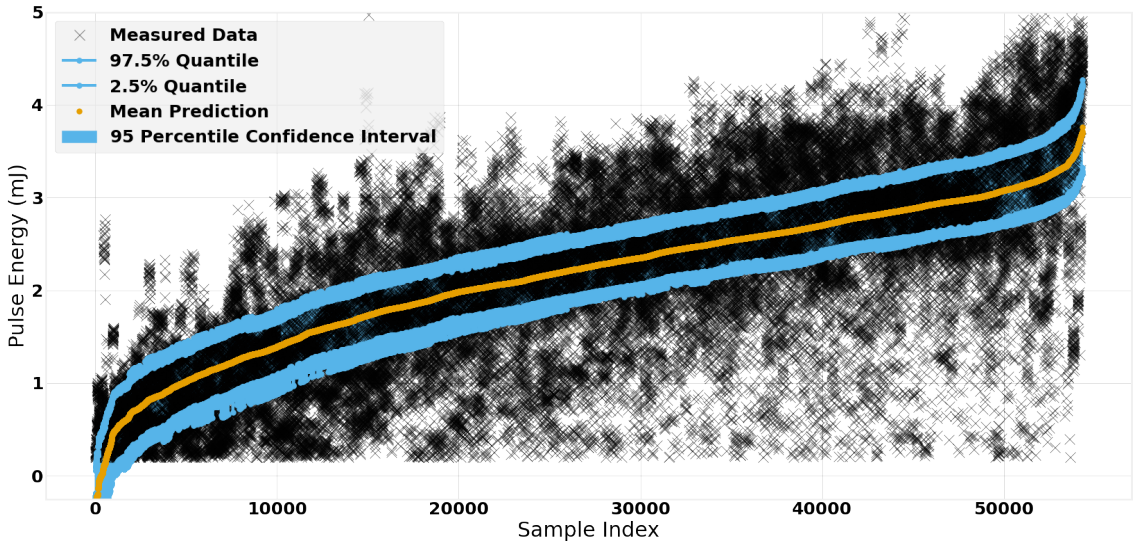


Figure A.6: BNN results on the test set, with coverage at 44.64%. The MAE is 0.56 mJ.

beginning to capture trends in the data, compared to the QRNN result shown in Fig. A.2, the BNN may require significantly more hyperparameter optimization.

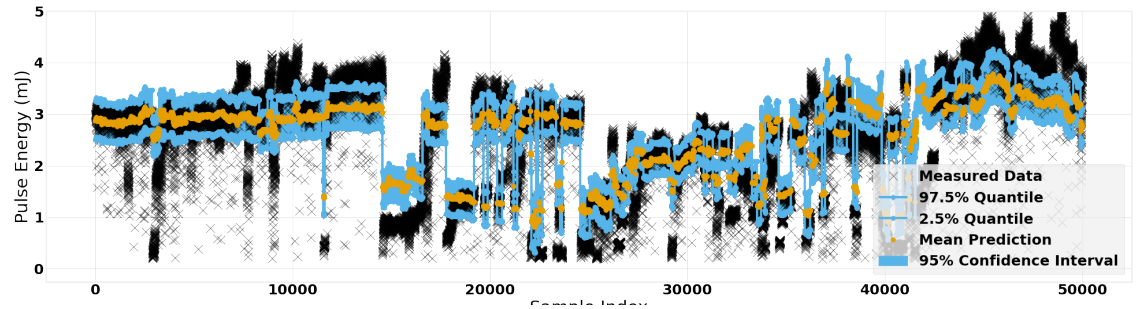


Figure A.7: A short range of data with the BNN median prediction and confidence interval. While the rough behavior of the FEL output is tracked, the coverage is not as good as the results obtained for QRNNs

Shown in Fig. A.4 is the same range of data as is shown in Fig. A.5. While portions of data were removed from training for the models shown in Fig. A.5, no portion of the data was explicitly excluded from being sampled in the train/test split to train the BNN. Despite all of the data being available for training, the BNN performance for the last 10,000 samples (which were removed from training for the models in A.5) was worse. The coverage for these

samples was 17.46%, and the MAE was 0.77mJ.

A.5 Conclusions and Future Outlook

To integrate deep learning models into particle accelerator control and operation, the models must provide predictions with reliable uncertainty estimates consistent with the measured data. We examined the ability of quantile regression and Bayesian neural networks to provide predictions and uncertainty estimates for the X-ray pulse energy of the LCLS, given historical data spanning several years, a wide variety of operating modes, and 76 measured input variables. Given the sparsity of the data and expected drift of system responses over time, it was not necessarily expected that a regression model could in principle be trained on this data to provide accurate predictions.

We find that QRNN models produce very reasonable median prediction and uncertainty estimates. The QRNN model shown in Fig. A.2 achieves 92.84% coverage, while the optimized BNN provides only 44.64% coverage. Further, the BNN result was obtained after removing outlier data, whereas the QRNN model performs well with outliers included in the training data. This highlights another advantage of the QRNN approach. We also find that the predictions track chronological sections of the data that were removed from training well. These sections may, for example, have non-i.i.d input relative to the training set. This will be investigated further in future studies, as developing methods for providing predictions without overestimating the models confidence in situations where the distribution of input data changes will be paramount for integration into machine control.

Because a goal for implementing model-based control of accelerators is to continuously retrain models as new data is collected, a model which is not sensitive to outliers would be a better candidate. Continuous retraining would address the irreducible uncertainty due to machine drift. An initial study examining this (by training on a sliding window of data and making predictions on future sections of the data) looks promising and will be continued in future work. Future work also includes applying feature selection on the inputs in this

dataset, to determine if the dimensionality can be reduced. We have started addressing this by evaluating the maximum information coefficient (62) for each input, and selecting only those with an MIC above a given threshold. This will be coupled with hyperparameter optimization on the BNN model and training procedure to see if more reliable BNN models for this task can be obtained. The eventual goal is to bring a refined version of either the QRNN or BNN model into operational use at the LCLS.

REFERENCES

[1] S. N. S.A. Antipov and A. Valishev, "Single-particle dynamics in a nonlinear accelerator lattice: attaining a large tune spread with octupoles in iota," *Journal of Instrumentation, Volume 12*, 2017.

[2] A. Seryi, *Unifying Physics of Accelerators, Lasers and Plasma*. CRC Press, 2016.

[3] K. Wille, *The Physics of Particle Accelerators: An Introduction, Translated by Jason McFall*. Oxford University Press, 2000.

[4] I. D. Young, M. Ibrahim, R. Chatterjee, S. Gul, F. D. Fuller, S. Koroidov, A. S. Brewster, R. Tran, R. Alonso-Mori, T. Kroll, T. Michels-Clark, H. Laksmono, R. G. Sierra, C. A. Stan, R. Hussein, M. Zhang, L. Douthit, M. Kubin, C. de Lichtenberg, L. Vo Pham, H. Nilsson, M. H. Cheah, D. Shevela, C. Saracini, M. A. Bean, I. Seuffert, D. Sokaras, T.-C. Weng, E. Pastor, C. Weninger, T. Fransson, L. Lassalle, P. Bräuer, P. Aller, P. T. Docker, B. Andi, A. M. Orville, J. M. Glownia, S. Nelson, M. Sikorski, D. Zhu, M. S. Hunter, T. J. Lane, A. Aquila, J. E. Koglin, J. Robinson, M. Liang, S. Boutet, A. Y. Lyubimov, M. Uervirojnangkoorn, N. W. Moriarty, D. Liebschner, P. V. Afonine, D. G. Waterman, G. Evans, P. Wernet, H. Dobbek, W. I. Weis, A. T. Brunger, P. H. Zwart, P. D. Adams, A. Zouni, J. Messinger, U. Bergmann, N. K. Sauter, J. Kern, V. K. Yachandra, and J. Yano, "Structure of photosystem ii and substrate binding at room temperature," 2016.

[5] M. P. Jiang, M. Trigo, I. Savić, S. Fahy, É. D. Murray, C. Bray, J. Clark, T. Henighan, M. Kozina, M. Chollet, J. M. Glownia, M. C. Hoffmann, D. Zhu, O. Delaire, A. F. May, B. C. Sales, A. M. Lindenberg, P. Zalden, T. Sato, R. Merlin, and D. A. Reis, "The origin of incipient ferroelectricity in lead telluride," 2016.

[6] A. Singer, S. K. K. Patel, R. Kukreja, V. Uhlíř, J. Wingert, S. Festersen, D. Zhu, J. M. Glownia, H. T. Lemke, S. Nelson, M. Kozina, K. Rossnagel, M. Bauer, B. M. Murphy,

O. M. Magnussen, E. E. Fullerton, and O. G. Shpyrko, "Photoinduced enhancement of the charge density wave amplitude," *Phys. Rev. Lett.*, vol. 117, p. 056401, Jul 2016.

[7] "Aps fact sheet."

[8] "Aps publication list."

[9] "Lclsii fact sheet."

[10] A. V. S. N. S. Webb, D. Bruhwiler and V. Danilov, ""chromatic and dispersive effects in nonlinear integrable optics","

[11] A. Edelen, N. Neveu, M. Frey, Y. Huber, C. Mayes, and A. Adelmann, "Machine learning for orders of magnitude speedup in multiobjective optimization of particle accelerator systems," *Phys. Rev. Accel. Beams*, vol. 23, p. 044601, Apr 2020.

[12] A. Scheinker, C. Emma, A. L. Edelen, and S. Gessner, "Advanced control methods for particle accelerators (acm4pa) 2019 workshop report," 2020.

[13] J. Duris, D. Kennedy, A. Hanuka, J. Shtalenkova, A. Edelen, P. Baxevanis, A. Egger, T. Cope, M. McIntire, S. Ermon, and D. Ratner, "Bayesian optimization of a free-electron laser," *Phys. Rev. Lett.*, vol. 124, p. 124801, Mar 2020.

[14] A. Hanuka, J. Duris, J. Shtalenkova, D. Kennedy, A. Edelen, D. Ratner, and X. Huang, "Online tuning and light source control using a physics-informed gaussian process adi," 2019.

[15] E. Courant and H. Snyder, "Theory of the alternating-gradient synchrotron," *Annals of Physics*, vol. 3, no. 1, pp. 1–48, 1958.

[16] R. Q. Twiss and N. H. Frank, "Orbital stability in a proton synchrotron," *Review of Scientific Instruments*, vol. 20, no. 1, pp. 1–17, 1949.

[17] A. Wolski, *Introduction to Beam Dynamics in High-Energy Electron Storage Rings*. Morgan Claypool Publishers; IOP Concise Physics, 2018.

[18] H. Wiedemann, *Particle Accelerator Physics II.* Berlin: Springer-Verlag, 1999.

[19] "Algorithm for chromatic sextupole optimization and dynamic aperture increase," 2006.

[20] "Improved step-by-step chroamticity compensation method for chromatic sextupole optimization.,"

[21] J. Bengtsson, "The sextupole scheme for the swiss light source (sls): An analytic approach," 1997.

[22] J. Bengtsson, I. P. S. Martin, J. H. Rowland, and R. Bartolini, "On-line control of the nonlinear dynamics for synchrotrons," *Phys. Rev. ST Accel. Beams*, vol. 18, p. 074002, Jul 2015.

[23] J. Ögren and V. Ziemann, "Optimum resonance control knobs for sextupoles," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 894, pp. 111–118, 2018.

[24] 2013.

[25] D. D. Caussyn, M. Ball, B. Brabson, J. Collins, S. A. Curtis, V. Derenchuck, D. DuPlantis, G. East, M. Ellison, T. Ellison, D. Friesel, B. Hamilton, W. P. Jones, W. Lamble, S. Y. Lee, D. Li, M. G. Minty, T. Sloan, G. Xu, A. W. Chao, K. Y. Ng, and S. Tepikian, "Experimental studies of nonlinear beam dynamics," *Phys. Rev. A*, vol. 46, pp. 7942–7952, Dec 1992.

[26] M. Giovannozzi, D. Quatraro, and G. Turchetti, "Generating unstable resonances for extraction schemes based on transverse splitting," *Phys. Rev. ST Accel. Beams*, vol. 12, p. 024003, Feb 2009.

[27] A. Franchi, L. Farvacque, F. Ewald, G. Le Bec, and K. B. Scheidt, "First simultaneous measurement of sextupolar and octupolar resonance driving terms in a circular accelerator from turn-by-turn beam position monitor data," *Phys. Rev. ST Accel. Beams*, vol. 17, p. 074001, Jul 2014.

[28] R. Tomás, "Normal form of particle motion under the influence of an ac dipole," *Phys. Rev. ST Accel. Beams*, vol. 5, p. 054001, May 2002.

[29] M. Billing, J. Dobbins, M. Forster, D. Kreinick, R. Meller, D. Peterson, G. Ramirez, M. Rendina, N. Rider, D. Sagan, J. Shanks, J. Sikora, M. Stedinger, C. Strohman, H. Williams, M. Palmer, R. Holtzapple, and J. Flanagan, "Instrumentation for the study of low emittance tuning and beam dynamics at CESR," *Journal of Instrumentation*, vol. 12, pp. T11006–T11006, nov 2017.

[30] D. Sagan, *The Tao Manual*. 2017.

[31] R. Storn and K. V. Price, "Minimizing the real function of the icec'96 contest by differential evolution," in *IEEE conf. on Evolutionary Computation, 842-844*, 1996.

[32] H. G. Hoffstaetter, "Comments on aberration correction in symmetric imaging energy filters," *Nuclear Instruments & Methods in Physics Research, 427, 275-281*, 1999.

[33] "First- and second-order charged particles optics. slac-pub, 68-70," 1984.

[34] H. J.L. and M. C. K., ""two-focal-length optical correlator,"," *Appl. Opt. 28, 5199-5201*, 1989.

[35] V. Danilov and S. Nagaitsev, "Nonlinear accelerator lattices with one and two analytic invariants," *Phys. Rev. ST Accel. Beams*, vol. 13, p. 084002, Aug 2010.

[36] M. Billing, W. Bergan, M. Forster, R. Meller, M. Rendina, N. Rider, D. Sagan, J. Shanks, J. Sikora, M. Stedinger, and et al., "Beam position monitoring system at cesr," *Journal of Instrumentation*, vol. 12, p. T09005–T09005, Sep 2017.

[37] "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[38] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, "Machine learning and the physical sciences," *Reviews of Modern Physics*, vol. 91, Dec 2019.

[39] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[40] J. Stohr, "Linac coherent light source ii (lcls-ii) conceptual design report,"

[41] F. Sannibale, D. Filippetto, C. F. Papadopoulos, J. Staples, R. Wells, B. Bailey, K. Baptiste, J. Corlett, C. Cork, S. De Santis, S. Dimaggio, L. Doolittle, J. Doyle, J. Feng, D. Garcia Quintas, G. Huang, H. Huang, T. Kramasz, S. Kwiatkowski, R. Lellinger, V. Moroz, W. E. Norum, H. Padmore, C. Pappas, G. Portmann, T. Vecchione, M. Vinco, M. Zolotorev, and F. Zucca, "Advanced photoinjector experiment photogun commissioning results," *Phys. Rev. ST Accel. Beams*, vol. 15, p. 103501, Oct 2012.

[42] T. Rao and D. H. Dowell, "An engineering guide to photoinjectors," 2014.

[43] K. Floettmann, "ASTRA: A Space Charge Tracking Algorithm," 1997-2021.

[44] C. Gulliford, "DistGen: Particle distribution generator," 2019.

[45] C. Mayes, "LUME-astra."

[46] "National energy research scientific computing center."

[47] A. Bartnik, C. Gulliford, I. Bazarov, L. Cultera, and B. Dunham, "Operational experience with nanocoulomb bunch charges in the cornell photoinjector," *Phys. Rev. ST Accel. Beams*, vol. 18, p. 083401, Aug 2015.

[48] R. P. Brent, *Algorithms for Minimization Without Derivatives*. Prentice-Hall, 1973.

[49] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[50] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.

[51] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *CVPR 2011*, pp. 1785–1792, IEEE, 2011.

[52] J. Nam, W. Fu, S. Kim, T. Menzies, and L. Tan, "Heterogeneous defect prediction," *IEEE Transactions on Software Engineering*, vol. 44, no. 9, pp. 874–896, 2017.

[53] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *IJCAI proceedings-international joint conference on artificial intelligence*, vol. 22, p. 1541, 2011.

[54] L. Y. Pratt, "Discriminability-based transfer between neural networks," *Advances in neural information processing systems*, pp. 204–204, 1993.

[55] S. Thrun and L. Pratt, *Learning to learn*. Springer Science & Business Media, 2012.

[56] F. Chollet *et al.*, "Keras." https://keras.io, 2015.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[58] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[59] C. Bostedt, J. D. Bozek, P. H. Bucksbaum, R. N. Coffee, J. B. Hastings, Z. Huang, R. W. Lee, S. Schorb, J. N. Corlett, P. Denes, P. Emma, R. W. Falcone, R. W. Schoenlein,

G. Doumy, E. P. Kanter, B. Kraessig, S. Southworth, L. Young, L. Fang, M. Hoener, N. Berrah, C. Roedig, and L. F. DiMauro, "Ultra-fast and ultra-intense x-ray sciences: first results from the linac coherent light source free-electron laser," *Journal of Physics B: Atomic, Molecular and Optical Physics*, vol. 46, p. 164003, Aug 2013.

[60] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.

[62] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science (New York, N.Y.)*, vol. 334, pp. 1518–1524, Dec 2011.