MACHINE LEARNING TOOLS FOR SMARTER AUTOMATION AND DIAGNOSTICS IN THE DEVELOPMENT OF PERSONALIZED MEDICINE FROM SIZE-LIMITED DATASETS

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in

Department of Biomedical Engineering

Jennifer M Bone

Carnegie Mellon University

Pittsburgh, PA

May, 2021

Acknowledgments

This work was funded in part by the Center for Machine Learning and Health (CMLH) and UPMC Enterprises. I would like to thank CMLH for giving me the opportunity to do this work.

I would like to express my sincerest gratitude for the community of mentors, collaborators, colleagues, friends, and family that have guided me throughout the course of my PhD. It is my interaction with each individual that has allowed me to get this far in my academic career.

First, I would like to thank my co-advisers, Professor Newell Washburn and Professor Phil LeDuc for their unwavering support and guidance throughout this thesis endeavor that has helped me grow as a scientist.

Next, I would like to thank Professor Keith Cook for this support as both a collaborator and mentor. I would also like to thank Professor Adam Feinberg for his collaboration on 3D bioprinting. Professor Michael Lotze has been an incredible mentor and collaborator over the last 8-9 months. It is with Professor Lotze that I was able to develop a new method for looking at TCR and BCR data. I foresee continuing to develop new models and technology together in the future. I feel incredibly fortunate to be part of your collaboration. Furthermore, I'd like to thank Pranav Murthy for sharing his insight into the clinical data and his continued support for this project.

I would also like to thank my colleagues in Professor Washburn's lab, including Dr. Chris Childs, Dr. Kedar Perkins, Dr. Aditya Menon, Joe Pugar, and Calvin Gang. It was our everyday interactions that always encouraged me to keep pushing forward. Our stories of the day-to-day life together as students formed the basis of my fond experience during this PhD. Chris, thank you for collaborating with me, reading my papers, helping me with machine learning, being a

great TA, and being a fantastic and supportive friend. I would also like to thank my colleagues in Professor LeDuc's lab who always appeared when I needed them for help. Our memorable experiences at conferences and lab trips will never be forgotten.

I would like thank Santiago Carrasquilla for his unwaivering support and friendship throughout my entire PhD. Thank you for reading my papers, helping me with paper figures, helping me get resources to do computational work, and listening to me rant about code at all hours of the night. Thank you for joining buggy with me, starting Tartan Salsa with me, teaching me to dance, and providing loving support throughout this entire endeavor.

I would also like to thank my BME colleagues who helped me throughout the last few years. I'd like to thank Dr. Andrew Lee, and Andrew Hudson for their intelligent feedback and support throughout the 3D bioprinting project. Finally, I would like to thank Yuxin Guo for being a fantastic undergraduate student and for collaborating with me on both the 3D printing and cancer immunology projects.

I would like to thank Tartan Salsa for playing a large role in my extracurricular happiness. Thank you for teaching me how to dance, and for providing me with an inclusive community outside the lab.

I'd like to thank my parents, Jeannie and Chris Bone, for their belief in me since before day 1. It is your hard work and tenacity that I emulate, and which got me through some of the most difficult times. I could not have done this without you. I'd also like to thank my family, Emily and Andrew Yowler, for supporting me and always knowing how to make me laugh. Someday, we will collectively discover what Quantum of Solids really is. I'd also like to thank Jeff and Paula Bone who are talented engineers and whom I respect greatly. Finally, I'd like to thank

Uncle Dan and Aunt Betsy for connecting with me and making me feel loved throughout my time in Pittsburgh. I feel so proud to call you family.

I'd like to thank my Grandma Jan who always encouraged me to be a strong woman. She was a talented writer, knitter, musician, and loving grandmother. Throughout my life she has been such an inspiration to me, and it is because of her and Grandpa Dave that I wanted to be and felt I could someday be a graduate-level scientist. Grandma Jan encouraged me to pursue a PhD. I wish I could have shown her this thesis. I'd also like to thank my Grandma Gerry and Grandpa Dub who believed in my education from the beginning and set me up for success.

I would like to thank my lifelong friends, Don and Scarlett Hibner for being a part of my support circle since before I can remember. I admire you greatly and hope to make you proud. Your passion for intellectual pursuits motivated me to go after pursuits of my own. I would also like to thank my lifelong friend, Ellen Narver, for mentoring me in both horsemanship and life. Our intellectual conversations encouraged me and motivated me to deeply invest in education.

I would like to thank the mentors who helped me get to where I am today. Dr. Tao Kwan, thank you for giving me my start as my first research mentor. It is because of you that I found the confidence to pursue a graduate degree and began learning the skills I would need to make it happen. I would also like to thank Mrs. Lawrence, who was my AP biology teacher and who encouraged me to love studying biology. I still have the picture of the first electrophoresis I ever ran in our class hanging on my bookshelf. I'd also like to thank Mr. Steve Cooperman who encouraged me to love studying physics no matter how hard or difficult it became. Finally, I'd like to thank Mrs. Savage, my first ever science teacher. You are the reason I fell in love with

science. I still tell people about my experiences in your class. In this thesis, I hope I have finally made good on my promise to write the Great American Novel.

Finally, I would like to thank my esteemed thesis committee:

Dr. Newell R. Washburn (Chair)

Dr. Philip R. LeDuc, PhD

Dr. Keith Cook, PhD

Russell Schwartz, PhD

Amir Barati Farimani, PhD

Abstract

With medical datasets becoming more readily available and standardized, machine learning (ML) has revolutionized healthcare through improved analysis of multi-variable clinical data, discovery of causal relationships or hidden states, and the generalization of predictive models to new and unseen patient data. Current ML architectures such as deep learning and canonical neural networks, rely on large datasets in order to make accurate models. However, variations in patient response due to heterogeneity in populations such as genomic, environmental, and physiological factors and processes suggest the need to tailor medical solutions to the unique features possessed by individual patients. As healthcare becomes more patient-specific, so too does the need to balance an ever-increasing feature-space (model complexity) with smaller numbers of patients. Thus, inherent in the applications of ML for smarter diagnostics and automation in patient-specific solutions is the drive to leverage biomedical datasets that are rich in information but limited in sample size. This work seeks to adapt ML techniques for feature-importance and predictions from size-limited data in the context of automating 3D-bioprinting for patient-specific implants and transplants, and early diagnosis of renal cancer progression for clinical decision support.

Additive manufacturing (AM) of biologically and physiologically active materials such as hydrogels, cell scaffold proteins, and cells is a promising avenue towards developing patient-specific implants and organ transplants using rapid fabrication and flexible design. However, the “plug-and-play” vision of bio-printed cell scaffolds and organs remains elusive due to the variability of biological materials. The heterogeneity of material response to the same physical process settings results in a complex feature-space that is difficult to optimize. As a

result, Hierarchical Machine Learning (HML) is used to embed domain knowledge into a statistical inference framework to reduce the experimental data necessary to model error bias in process design choices. HML-optimized predictors were shown to produce high-fidelity bio-printed constructs that deviate from expected dimensions by less than 10%. Furthermore, the use of a supervised physical middle layer that connects predictors to the quality of print response is shown to aid in transfer learning to new print materials suggesting a method for rapid optimization of parallel 3D bioprinting systems.

Disease diagnosis can also benefit from small experimental or phase 1 clinical data. An innovative Markov model is developed to perform early classification of patient response to hydroxychloroquine/Aldesleukin (IL-2) treatment for progressive renal cancer. The model reduces the high-dimensional (10^{15} – 10^{25}) feature-space of T-cell receptor (TCR) and B-cell receptor (TCR) systems biology to an intermediate-dimensional space of 400 descriptors, revealing the causal features responsible for predicting the final state of 30 patients after 15 days of treatment with 95% classification accuracy. Through quantitative monitoring of amino acid motifs in the primary structure of TCRs and BCRS over 3 treatment points, a mechanistic understanding of the orchestration of TCRs and BCRs towards patient recovery is discussed. These results suggest that this Markov model could be a powerful diagnostic tool for leveraging phase 1 clinical data towards early patient diagnosis, informing an early and individualized medical response.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Machine learning (ML) is a tool that is revolutionizing healthcare and biomedical research. The number of publications that have utilized ML for advances in healthcare and biomedical sciences has increased almost exponentially in the last two decades (Figure 1). With large numbers of datasets more available and standardized, ML is becoming a robust tool for multiple disciplines from improved patient care, to diagnostic predictions and biomedical research.[1] Some major ways in which ML has been useful in these areas are (1) feature selection (2) unveiling of causal relationships, and (3) optimization of objectives.

Feature selection has been a powerful tool for decomposing large healthcare datasets into the dominant underlying features that describe the data. A notable use of feature selection in healthcare datasets is the modeling of large genomic data from next generation sequencing (NGS). By harnessing disease signatures, such as mutations or up-/down- regulation of mRNA expression, high-level observations of diseased phenotypes can be elucidated for disease-risk predictions.[2] For example, NGS data has been used to correlate genetic scores to predict the outcome of acute myeloid leukemia.[3] Causal relationships between variables can also be useful for classification and advanced mechanistic understanding between seemingly disparate variables. One example is in the discovery of the correlation of ejection fraction (percentage of blood in the left ventricle after each heart contraction) with predictions for patient survival from heart failure.[4] Finally, an example of machine learning for the optimization of a target objective can be seen in a recent study involving codon optimization for high performance synthetic genes.[5]

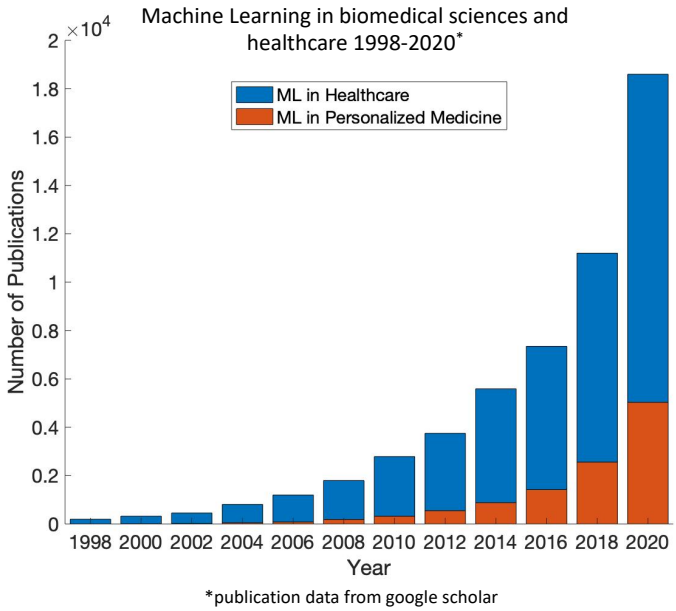

Figure 1.1. Publications using ML for biomedical sciences and healthcare from 1998 – 2020 show a nearly exponential growth trend (blue) over approximately two decades. ML for personalized medicine has only grown in the last decade and is still relatively nascent. Publication data was gathered using google scholar with the search terms [“machine learning” or “artificial intelligence”] and [“Bio” or “biomedical” or “healthcare]. Publication data for ML in personal medicine was gathered using the search terms [“machine learning” or “artificial intelligence”] and [“personal” or “patient-specific”] and [“medicine” or “healthcare”]. The analysis is meant to show a general trend in the scientific community.

Large datasets clearly provide promising avenues for predictions and pattern discovery that might not be realizable with traditional statistical techniques. However, many real-world datasets lack the sample size necessary to sift through redundant or irrelevant features, impeding model performance, generalizability, and interpretation.[6] In fact, it is often difficult to achieve large sample sizes in real-world settings particularly if data acquisition is expensive, or the data being studied are rare.[7] As a result, the potential for utilizing ML strategies to gain improved predictive power or mechanistic understanding from size-limited data remains, many times, untapped. In addition to resource challenges, size limitations in data arise from natural

heterogeneity in populations. Nearly 2,500 years ago, the Greek physician Hippocrates postulated the "individuality of disease," noting that not all solutions are created equal for every individual.[8] Known as the "Father of Western Medicine," Hippocrates is perhaps the oldest proponent of patient-specific healthcare.

Over two millennia later, a grand goal of future medicine still lies in patient-specific healthcare. The intersection of physics, biology, and computational power of the last decade have provided new tools to make the forward-thinking advice of an ancient Greek physician a reality. However, despite the tools and age-old motivation, the number of publications for ML in patient-specific (personalized) healthcare has been less prolific compared to the use of ML in healthcare and biomedical sciences (Figure 1.1). One reason for this may be that models that are truly patient-specific must ultimately wrestle with population heterogeneity, both in features that describe patients, and patient responses to their environment. As a result, patient-specific ML models experience a curse of dimensionality: higher patient specificity splits seemingly homogeneous populations into more features, which in turn can have a negative effect on the sample size required for a model to properly learn and answer questions. After all, how can a model perform well on the whole population and also be highly specific to an individual? The sensitivity-specificity relationship is a well-known, albeit challenging, tradeoff in many disciplines. ML is not a cure. Instead, the question becomes to what limits can ML help push this tradeoff in favor of highly sensitive and highly generalizable models?

To begin to answer this question, the relationship between patient specificity, model complexity, sample size, and theoretical model feasibility is theorized in Figure 1.2. Initially, model feasibility scales with increasing complexity and data size (region-II). Moving along the sample-

size axis, tall arrays (in which samples outnumber features) are typically well-suited for canonical ML algorithms as the sample size is sufficient for data-driven learning (region-IV). In the limit that there are large features combined with large data (region-III), model performance can eventually be limited by computational feasibility. Moving along the complexity axis, as models seek to become more patient-specific (yellow), higher model complexity is expected (region-I). However, the data available per feature decreases. Compared to region-IV, the region-I regime changes the shape of the data, which can cause difficulty with canonical ML architectures. Since Region I represents the data regime in which we still desire good model performance, ML strategies must be adapted.

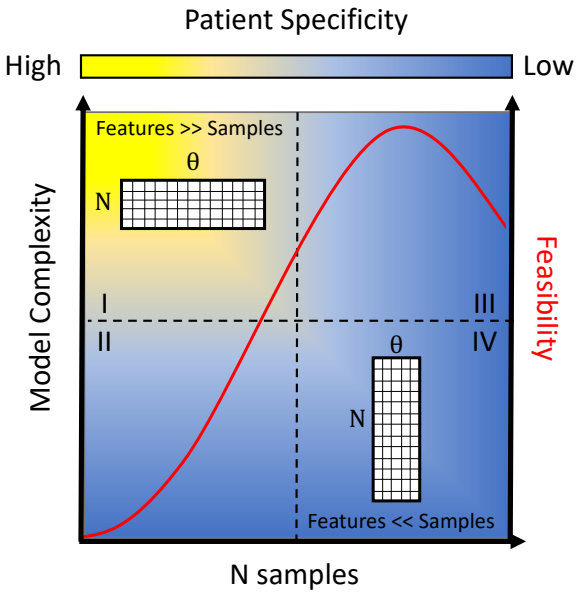

Figure 1.2. The intersection of machine learning with patient-specific models in terms of model feasibility. (I). The limit of high patient specificity (low sample number) and high model complexity. (II) Model feasibility scales with low model complexity and low sample number. (III) The limit of high complexity, high sample number in which computational feasibility eventually reaches a limit. (IV) The regime of low-intermediate model complexity combined with high sample number in which most conventional ML architectures are typically utilized.

Previous work for adapting ML strategies towards good model performance from high features and low sample-size has illuminated the challenge. A comprehensive survey of multiple ML models on a small materials dataset (<100 samples) showed that, irrespective of model strategy (SVM, ordinary least-square regression, random forest, and LASSO to name a few listed), the relationship of training data size to degrees of freedom in the models (defined as non-zero features after regularization) was most predictive of model performance.[9] In summary, the relationship between data size and complexity was shown to be generally more important for good models than the strategy of the models themselves. Perhaps this result is somewhat expected as a natural consequence of the bias-variance tradeoff. However, the authors showed that providing low-quality estimations of the targeted property – essentially prior knowledge – consistently improved model performance for the same data size. Not surprisingly, Bayesian-inspired methods that encourage the use of prior and posterior learned distributions such as Gaussian Process[10] and Bayesian networks[11, 12], are frequently used to work with size-limited data with good model outcomes. A method that will be discussed in depth in this work is to embed known empirical relationships within statistical inference models to drive down the data necessary to discover variable relationships with the system response. By initially supervising potentially notable relationships between predictors, rather than allowing predictor relationships to be discovered by inference on the data alone, models have been shown to have improved accuracy for small data.[13] While physical laws are difficult to ascertain in healthcare, the success of Zhang and Ling *et al* from embedding crude measurements of target properties in their models implies a high likelihood that even loose estimates of prior knowledge from patient averages in healthcare literature could be both relevant and useful for improving model predictions.

This thesis will discuss strategies for learning from sample-limited datasets that are non-ideal candidates for "plug-and-play" into canonical ML architectures in the context of 3D bioprinting and early renal cancer diagnostics (Figure 1.3). 3D bioprinting is an exciting solution for fabricating organs that tailor to natural heterogeneities in the population due the ability to flexibly and rapidly prototype size variations of the same organ model. However, in order to meet the growing need for transplants at an industrial level, it will be crucial to validate that small variations in design can be printed reliably and with high fidelity given uncertainties in biological material composition and handling. In this thesis, we will argue that by leveraging physical knowledge of the 3D bioprinting system, we can predict high-fidelity prints from a capitalized experimental space. Furthermore, we will show how the empirical relationships that survive regularization improve model interpretability and create a knowledge bridge for predicting high-fidelity prints from parallel printing systems.

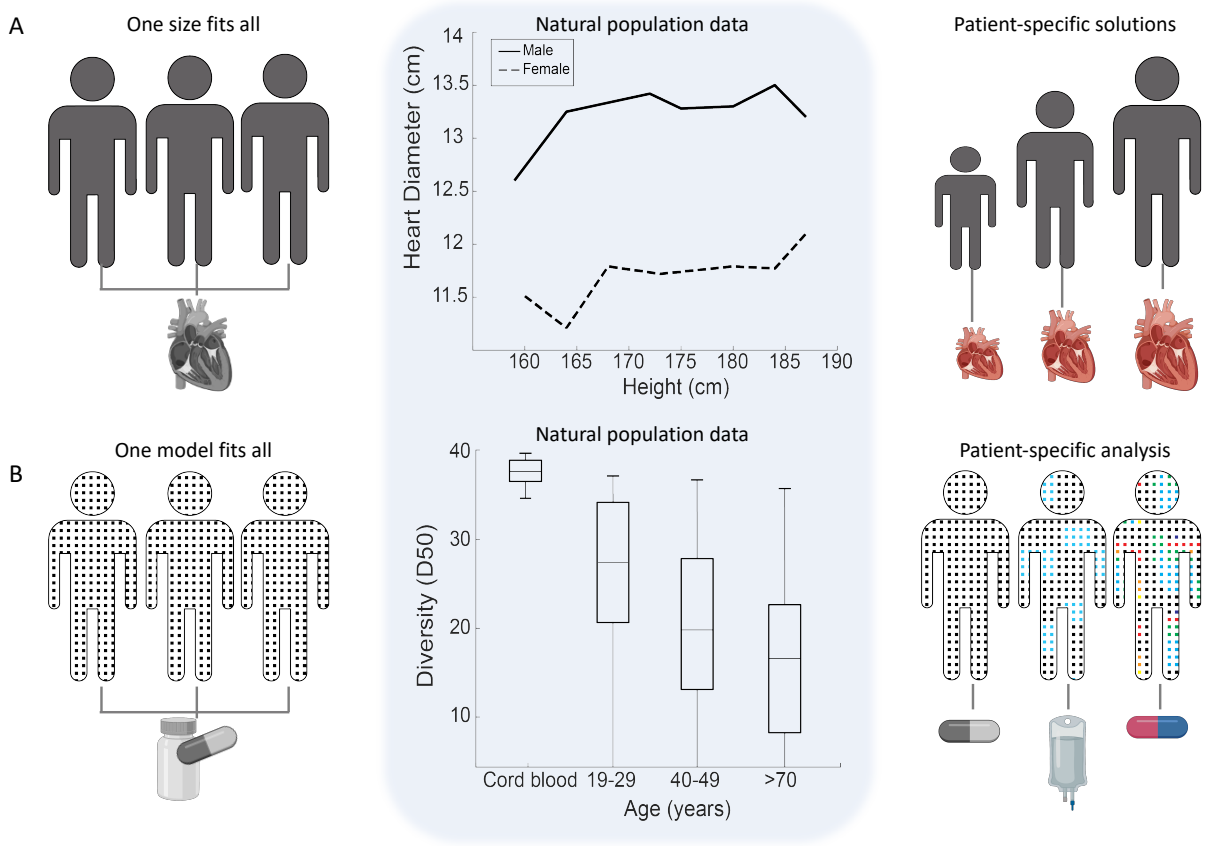

Figure 1.3. Patient-specific healthcare models and solutions are a grand future goal for machine learning and medicine. (A) shows the need for size variations in heart transplants. Data for this figure was adapted from Oberman and Karunas et al. Circulation, 1967.[14] (B) shows how diversity of T-cell receptors naturally varies according to age in years but can occlude cancer signatures. Data for D50 variation is unpublished from iRepertoire.[15]

Phase 1 clinical data encompasses the study of small numbers of patients (<50 patients) who are either healthy or are attempting experimental treatments due to a lack of response to more conventional treatments. Adaptome diversity (D50) has been shown to be a key indicator for disease.[16] Decreased diversity has been associated with poor immune health due to disease epitopes. However, with 10^{15} - 10^{25} possible features in the adaptome, natural variations in diversity due to age (shown in Figure 1.3B) and other environmental or genetic factors can

occlude causal disease features necessary for patient diagnosis and treatment. In this thesis, we will argue that a carefully selected feature space combined with probability estimations of feature importance for distinguishing disease phenotypes, can harness the expansive dimensionality of the adaptome allowing early prediction for renal cancer response to HCQ/IL-2 treatment in phase 1 clinical data.

This work shows how patient-specific healthcare can become more realizable as ML algorithms are strategized to make predictions from data structures that are atypical for conventional methods. In this work, traditional statistical inference methods are supplemented with supervised physical middle-layer variables to allow both data-specific learning and transfer learning to unseen material types. The result is a model that can predict up-front optimal process settings for 3D bioprints that can themselves respond to heterogeneous needs in patient populations. Furthermore, we will show that careful selection of a common feature space can leverage key predictive knowledge in a Phase 1 clinical trial of <50 patients. We intend for the impact of this work to trend towards more patient-centered, flexible care, from reliable personalized implants and transplants to early diagnosis in the clinic.

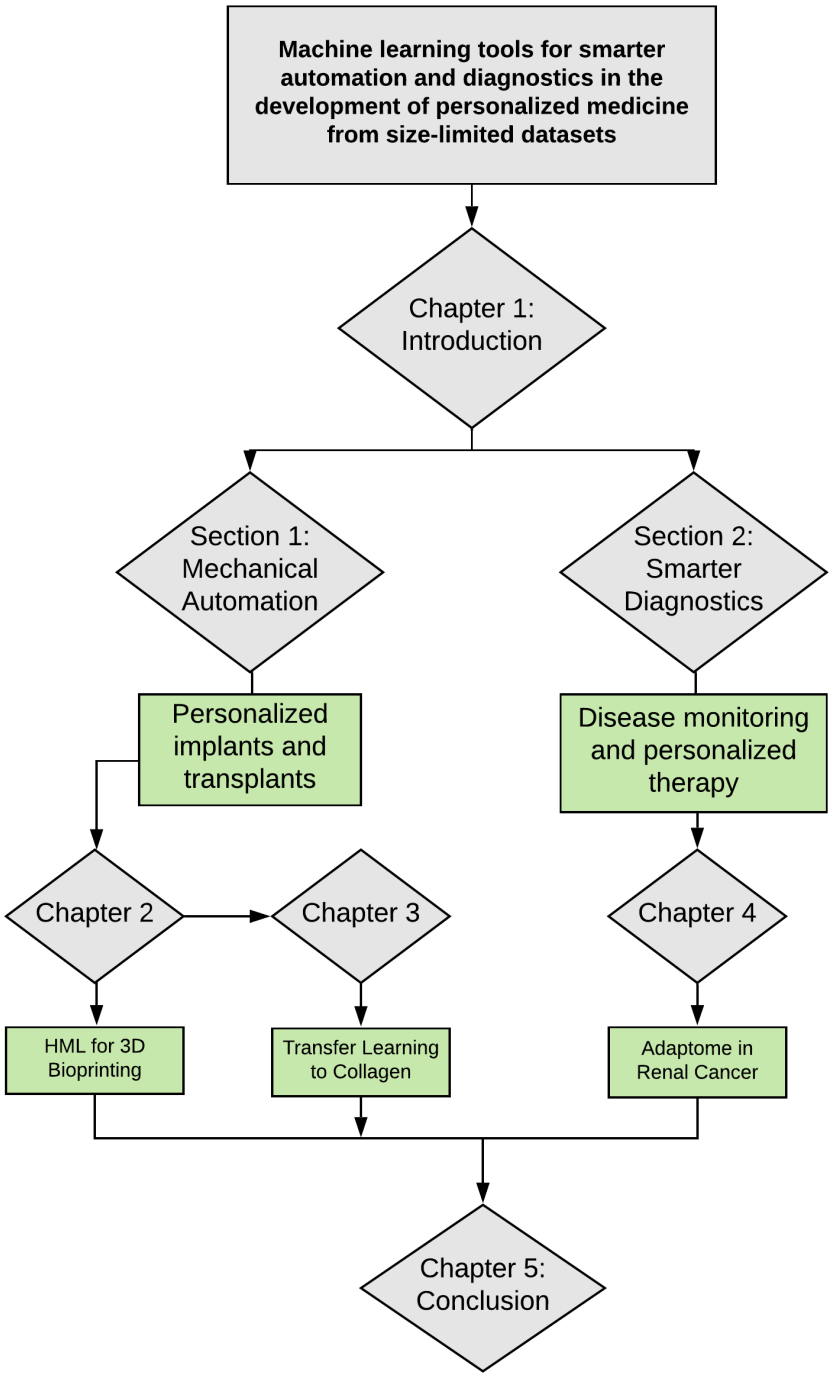

Figure 1.4. Outline and organizational flow of the thesis.

Section 1. Machine Learning for smarter automation from size-limited datasets in the context of optimizing 3D bio-printed constructs

Chapter 2. Hierarchical Machine Learning for High-Fidelity Bio-printed Constructs

2. 1. Introduction

Major structural and functional failure in the human body is often detrimental and requires rapid medical intervention via organ transplants, stents, hip or knee replacements, or prosthetics. At any given time, nearly 3,500 – 4,000 people are waiting for a heart or heart-lung transplant[17], and every ten minutes a new person is added to the national transplant waiting list.[18] Moreover, the prevalence of rejection and immunosuppression currently impact the success of transplantation and highlight the need for patient-specific transplants that will increase the rate of survival. Due to this high demand, the market for patient-specific implants is projected to reach more than $10B by 2021.[19] 3D printing is an emerging technology that could strongly impact the future of research, translation, and industry. Three major areas of current impact are: transplantation, drug testing, and desktop print technology as shown in Figure 2.1.

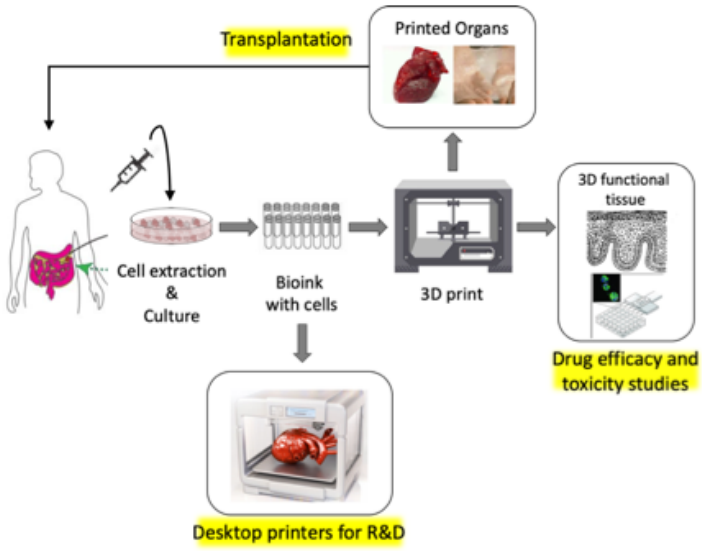

Figure 2.1. The diverse and multi-variate space of 3D printing breaking into healthcare.

Transplantation of functional tissue constructs leaped from benchtop to industry when 3-D Bioprinting Solutions printed a vascularized, functional thyroid, which they successfully implanted in mice in 2015.[20] Additionally, 3D printed human tissue found in the heart, liver, and kidneys has been used to assess drug efficacy and toxicity testing. 3D printing company Organovo has generated revenue via functional printed liver organoid tissue for drug toxicity studies.[21] Finally, there is a growing market for desktop bioprinters that can be sold for in-house applications and research, such as BioBots who make and sell alginate-based hydrogel inks laden with various biological substrates, including cells. It is clear that 3D printing is moving from prototype to industry with applications spanning a multivariate space composed of eclectic inks and printing systems.[22] However, while the concept of diverse applications is clearly a key element to the success of 3D printing in healthcare on an industry level, this laudable flexibility to print with many different biological materials comes at a price.

Workflow for additive manufacturing (AM) typically requires copious iterative testing. Using a full Design of Experiments (DOE) to find the "golden batch" of parameters that produce the highest-quality prints is costly and time-intensive. Furthermore, limited domain knowledge of the system is gained with this optimization method, making predictions for different ink-printer systems – even those composed of the same system but with a slightly altered design – challenging. 3D printing as a manufacturing method to meet patient-specific implant or transplant needs on an industry level is weakened when design changes require iterative testing for every new design.

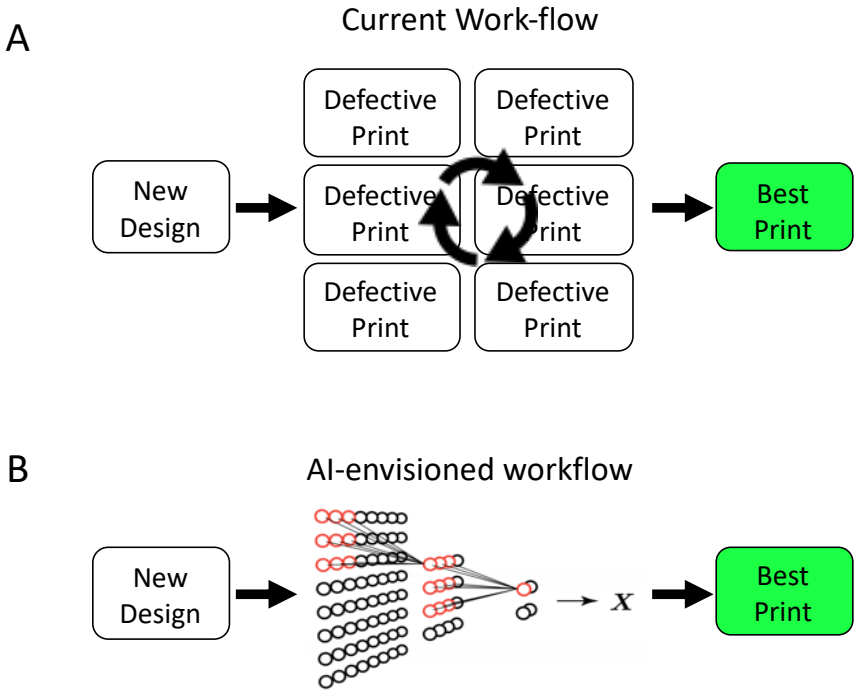

Figure 2.2. Envisioned AI-driven workflow compared to current methods. (A) 3D printing is not "plug-and-play." In reality, 3D printing requires multiple iterations of various process settings in order to achieve an optimal print. Small changes in design often requires a re-collection of experimental data. (B) The AI-envisioned workflow strives to efficiently use the experimental space necessary to predict high-fidelity prints. Moreover, the predictive model could extend to new designs without the need to collect additional data.

An ideal optimization system would be composed of a small training dataset over which physical knowledge of the system could be leveraged to make targeted predictions for new materials and designs as shown in Figure 2.2.

While hard materials such as ceramics and metal printing have started to incorporate AI in their work-flow to reduce error,[23] state-of-the-art methods for optimizing 3D-printed soft materials, such as silicone polymers, still require large datasets even for a more streamlined hill climb optimization.[24] Recently, Menon *et al* showed that small datasets of silicone 3D prints could be optimized via hierarchical machine learning (HML).[25] Process predictors, such as ink viscosity and print speed, were related to physical equations and a measurable print score (output). LASSO regression parameterized the complex rheological space into its fundamental physical interactions. Armed with domain knowledge, optimal print parameters were elucidated. Print speed and defects were optimized and mitigated, respectively for a small dataset of 38 prints. In contrast to silicone printing, biomaterials, such as hydrogels, cells, and proteins, offer a unique challenge due to their non-Newtonian behavior and intrinsic variability in feedstocks. The complexity of the system makes 3D printing of biomaterials a good candidate for an HML framework since the presence of rheo-physical equations in the middle layer can better correlate experimental parameters to the final printed outcome.

To date, there is limited technology that integrates machine learning with 3D biomaterial printing. The essence of soft-material additive manufacturing, which consists of polymer and non-Newtonian fluid flow and gelation through a small capillary, is an age-old and well-studied phenomenon[26] that has gained broader use in recent years due to the accessibility and promise of 3D printing as an impactful manufacturing method. Integrating multiple engineering disciplines,

many successful bioprint prototypes utilizing a plethora of tested inks and print settings have impacted the literature;[27] however, often the methods for reliably conferring system parameters with high-fidelity prints is 'guess-and-check' with limited fundamental knowledge of the system or materials. The time is right for rich datasets to be parameterized by physical modeling that can combine both the global laws of non-Newtonian soft-material flow with constraints specific to the desired printing system and design.

A Hierarchical Machine Learning (HML) framework is presented that uses a small dataset to learn and predict the dominant build parameters necessary to print high fidelity 3D features of alginate hydrogels. We examine the 3D printing of soft hydrogel forms printed with the Freeform Reversible Embedding of Suspended Hydrogels (FRESH) method based on a CAD file that isolated the single-strand diameter and shape fidelity of printed alginate. Combinations of system variables ranging from print speed, flow rate, ink concentration, and nozzle diameter were systematically varied to generate a small dataset of 48 prints. Prints were imaged and scored according to their dimensional similarity to the CAD file, and high print fidelity was defined as prints with less than 10% error from the CAD file. As part of the HML framework, statistical inference was performed, here using the Least Absolute Shrinkage and Selection Operator (LASSO) to find the dominant variables that drive error in the final prints. Model fit between system parameters and print score was elucidated and improved by a parameterized middle layer of variable relationships which showed good performance between predicted and observed data ($R^2 = 0.643$). Optimization allowed for the prediction of build parameters that gave rise to high-fidelity prints of the measured features. A trade-off was identified when optimizing for the fidelity of different features printed within the same construct, showing the need for complex predictive

design tools. A combination of known and discovered relationships was used to generate process maps for the 3D bio-printing designer that show error minimums based on chosen input variables. Our approach offers a promising pathway towards scaling 3D bioprinting by optimizing print fidelity via learned build parameters that reduces the need for iterative testing.

Furthermore, applications of HML to parallel 3D printing systems will be discussed. Specifically, transfer learning for the optimization of target printing systems will be shown for the first time. We will show that the supervised middle layer can be leveraged for predicting FRESH-printed collagen features. We will also discuss the use of HML for rapid optimization of new biomaterials with FRESH, including an optically transparent bath solution and cell-laden inks.

Machine learning (ML) is a collection of statistical tools that are used to discover relationships in data, allowing for modeling and optimization of complex systems that have several underlying and intertwined laws. [28, 13] ML tools are particularly powerful for biomedical research due to the highly interdisciplinary nature of the field that often does not result in simple analytical solutions. Consequently, ML methods to study biological processes, process engineering, and healthcare have grown significantly in number in the last decade.[29] The increasing availability of large datasets has enabled the use of ML across multiple disciplines from diagnostics and patient care[1] to bacterial genome analysis and antibiotic resistance prediction.[30] However, conventional ML methods typically rely on large datasets and statistical inference alone, with predictions decreasing in accuracy in small data domains (such as those that have less than 100 instances per feature). Thus, the full utility of ML tools in many biomedical applications is still largely unexplored.

To extend the benefits of ML to a wider range of biomedical problems, it is essential to develop modeling techniques that can cope with systems defined by several complex relationships but

which are limited in data size or completeness. Clymer *et al* showed that the prediction of labral tear severity for a small medical dataset of 34 patients could be achieved using transfer learning from larger image databases.[31] However, a large repository of similar data to train models is not always available, such as in the case of tissue engineering or analysis, wherein bottlenecks in experimentation time lead to small datasets. Shaikhina *et al* predicted risk of bone fracture for severe osteoarthritis from a dataset of only 35 femora bone tissue specimens using a multiple-run strategy to find the best-performing neural network from a small set of predictors.[32] One major question that remains is how to perform feature selection in complex, collinear spaces to enable accurate predictions of new combinations of variables while still working from the small data domain. Such is the case with 3D bio-printing where an especially complex mixture of observed forces interact in ways that are difficult to predict and contribute to print error. A tool which could illuminate the driving physical laws behind inputs and desired outcomes would be useful for directing bio-printing.

Physical and chemical insight (domain knowledge) can be leveraged to reduce the size of the dataset needed to discover complex predictor-response relationships.[13] Hierarchical Machine Learning (HML) is a hybrid physical-statistical machine learning methodology developed on small experimental datasets in which predictors are connected to the system response by a middle layer whose variables are parameterized by known physiochemical relationships from domain knowledge pertaining to the system.[33] Conventional regression techniques can then be used to connect the middle layer to the system response for prediction and optimization. In this work, we demonstrate that HML can be used to model and optimize Freeform Reversible Embedding of

Suspended Hydrogels (FRESH)[34], a printing technique rich in physiochemical relationships but plagued by limited data for given conditions.

FRESH is a rheochemical process that allows for the creation of three-dimensional, functional structures through layer-by-layer precise spatial control of biological building blocks such as hydrogels[34], cells and proteins.[35] This ability has advanced 3D bio-printing as a promising manufacturing method to meet the growing need for patient-specific medicine through the development of advanced cell scaffolds for tissue engineering,[36] tissue organoids for high throughput drug testing,[21] and ultimately fully functional printed organs.[37] In this work, we focus on hydrogel bio-inks, which are widely utilized in bio-printing to make flexible, biocompatible cell scaffolds or cell-laden constructs.

A significant challenge in soft-material additive manufacturing is the ability to print biological materials with high fidelity. We define high-fidelity prints as those in which the physical dimensionality of the final print deviates from the CAD design file by less than 10%. The challenge is to determine the optimal process settings that maximize print fidelity from a massive variable range and space covering materials selection, materials formulation, and process parameters. For biological materials, these variables are strongly coupled: Inks based on solutions of biomacromolecules and cells have rheological behavior that is highly dependent on concentration and shear rate,[38] creating an enormous design challenge.

ML has shown good predictive capabilities in aspects of additive manufacturing, such as gaussian process regression and Bayesian analysis for prediction of porosity defects in metal 3D prints[23] and development of power-velocity process maps to aid print designers.[39] State-of-the-art methods for optimizing 3D-printed soft materials, such as silicone polymers, still require large

datasets even for a more streamlined hill-climb optimization .[24] However, Menon et al showed that small datasets of silicone 3D prints could be optimized via HML and optimal print parameters were elucidated.[25] Additional methods such as neural networks have been discussed for improving print fidelity, predicting optimal process parameters, and error. As Yu and Jiang note, there is limited use of machine learning in 3D bio-printing processes due to the need for sufficient data to make predictions for a highly complex bio-fabrication system.[40] Figure 2.3 describes how HML methodologies can be advantageous for addressing small 3D bio-printing datasets and draws a comparison with traditional neural networks.

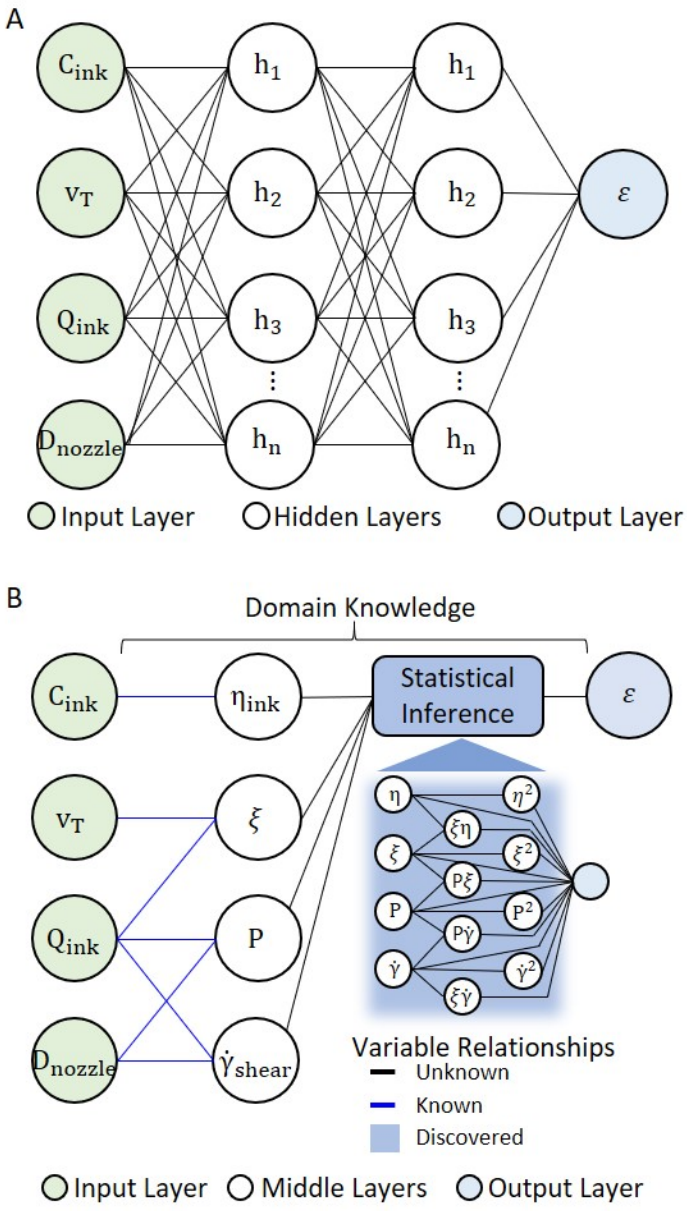

Figure 2.3. (A) The methodology of a conventional neural network wherein variable relationships are discovered and represented by hidden layers. (B) HML provides a methodology to leverage experimenter knowledge and experience to reduce the data-driven burden of variable relationship discovery. Domain knowledge inputs known, general physical relationships into the model via a middle layer of physical variables parameterized by the input layer. Statistical inference and cross-validation discover more complex, system-specific relationships and evaluate the ability of the middle layer to describe the system response.

In this work, we propose a HML model for predicting and optimizing the print fidelity of hydrogel 3D prints in terms of linewidth and shape fidelity. To begin, we generate a dataset of both high and low-fidelity alginate prints by systematically varying print input parameters and assessing the resulting prints in terms of dimensional similarity to the original CAD designs. To address the problem of a small dataset, an HML algorithm was constructed wherein the structure of the middle layer leverages known physical relationships in the flow-gelation process of alginate. The model fit was assessed by cross-validation and a comparison of R^2 values across multiple fitting strategies. We compare the performance of HML to a conventional neural network. We further describe the improved fit of including a middle layer over direct modeling of bottom-layer variables to output. We then show an optimization of the HML model in which we minimize print error to generate a new set of optimized input variables predicted to generate high-fidelity prints with an error of less than 10% in dimensionality from the original CAD specification. We test these predictions experimentally by evaluating error in prints created from HML-predicted optimal print parameters. Finally, we leverage discovered variable relationships to generate process maps for the 3D bio-printing designer that foreshadow the success and pitfalls of choosing input variable combinations on the expected feature size.

2.2 Experimental Methods

A 3D bio-printing dataset was generated consisting of a series of prints manufactured on a pilot printing system by optimizing the FRESH printing method.[35] In FRESH, a biomaterial ink is extruded into a sacrificial support fluid that cures and holds the ink in place. Once cured, the sacrificial solution is removed, leaving behind the printed form. The use of a sacrificial support fluid for the printing system alleviates errors due to post-processing damage caused from the

removal of cured extraneous support material that normally plagues finished prints. As a result, individual, free-floating printed fibers can be printed and imaged that allow for isolated study of individual process parameters.

2.2.1. CAD model design

The fundamental building blocks in layer-by-layer fabrication, namely the ability to print lines and corners with accuracy, play an important role in print fidelity. Many prints fail due to early errors in print features that are then carried throughout the entire print. In order to better isolate the physical characteristics of ink as it is extruded from the nozzle and cured, it is important to analyze free-floating printed strands that are not fused to an underlying layer. As a result, the CAD model was designed as a 10 mm x 10 mm x 10 mm cube containing a window modifier in which alginate strands are bridged from one end to the other thus allowing for the isolation of flow-gelation properties. Furthermore, corners were analyzed under the same variable combinations, providing implications for the estimation of path planning with material feedback.

2.2.2. Ink material selection

Sodium alginate was chosen due to its ubiquity in acellular bio-printing, and for its usage in cell scaffold printing[41] and cell-laden inks.[42] This approach has the potential to be generalized to other biomaterials such as collagen, chitosan, methacrylated hyaluronic acid, gelatin and other curable soft materials.

2.2.3. FRESH Process

Gelatin as a sacrificial support material in FRESH has multiple advantages. It is a cost-effective yield-stress material with well-studied biocompatibility.[43] Furthermore, it undergoes a sol-gel transition at physiologically compatible temperatures, making it an excellent candidate for work

with cells-laden inks. The hydrophilicity of gelatin also has a compatible chemistry with hydrophilic alginate ink. This is important because print errors are likely to be prevalent in systems with highly mismatched surface energies. Alginate ink gelation readily occurs when its aqueous form is extruded into a medium that contains metallic divalent cations such as calcium. Thus, a gelatin support bath of packed microparticles containing calcium chloride can be used to print complex alginate constructs that can be released upon adding heat to the system.

Alginate inks were prepared by dissolving 30 mg, 40 mg, and 50 mg of alginate separately into 10 mL of DI-water to make 3%, 4%, and 5% w/v inks, respectively. The range was limited to 5% w/v due to the decreased flow of alginate at higher viscosities. Gelatin sacrificial supports displaying Bingham plastic rheology were prepared by dissolving Gelatin Type B (Sigma) into 200-proof ethanol: DI-water heated to 60 °C. The pH was adjusted to 6.0 and then the solution was allowed to cool to room temperature overnight (or a minimum of 6 hours) while stirring. Gelatin was then processed to make the support material. Aliquots of gelatin solution were collected into centrifuge tubes and centrifuged at 500 rpm. The supernatant was discarded and replaced to the same volume with DI-water. Tubes were vortexed until the gelatin was resuspended, and centrifuged at 2000 rpm for 2 min. The supernatant was discarded and replaced with 11 mM calcium chloride (Sigma) for support material to be used with the 152 μm nozzle diameter and with 8 mM calcium chloride for baths to be used with the 80 μm nozzle diameter. The particles were washed and resuspended 3 times using calcium chloride through centrifugation at 2000 rpm. The final centrifugation for concentrating the gelatin particles was conducted at 3000 rpm. Concentrated gelatin was then distributed into small petri dishes for printing.

All CAD files were processed using Slix3R. Layer height was kept constant at 0.06 mm for prints made using the 152 μm nozzle diameter, and at 0.04 mm for prints made using the 80 μm nozzle diameter. Nozzle length was kept constant for each nozzle diameter and is inputted in the rheological flow calculations in the HML middle layer. Translational nozzle speed describes the velocity at which the nozzle extruding the ink traverses across the 3-dimensional print bed and was the range of speeds was chosen to span up to one order of magnitude. Flow rate was set and varied by an extrusion modifier. We ensured that defaults in the program were manually overridden by examining the g-code.

2.2.4. Training set metrics

We will briefly discuss the chosen metric for print system response and the print variables chosen as predictors.

System Response Metric: Previous work in soft-material 3D printing has highlighted the importance of metrically quantifying the success of prints. We developed a method of standard metrics to quantify how closely a print dimensionally matches expectation based on the CAD file. Once printed and released, prints were visualized under a phase-contrast microscope. The diameter of the free-floating printed fibers, δ (μm), and corner radius, r (mm), were then measured using ImageJ, and print score was determined as the percent difference in observed dimension compared to the original CAD design. In Eq. 2.1, ε, error is the difference between the observed and expected dimensions. The size of printed strands is expected to match the inner-diameter of the extrusion nozzle ($\delta_{exp} = \delta_{CAD} = D_{nozzle}$). Corner estimation can also be measured at the same variable combinations and is expected to have a radius of 0 mm (corresponding to a 90° angle). We scored

prints according to the percent error from expected dimensions using the absolute value of error, although over- and under-estimation of error will be briefly discussed.

$$P_{score} = 1 - \frac{|\varepsilon|}{\delta_{CAD}} \times 100\% \qquad \text{Eq. 2.1}$$

2.2.5. Data Set and Parameter Space

A range of print variable predictors were chosen carefully to span the print design space shown in Figure 2.4 and summarized in Table 2.1.

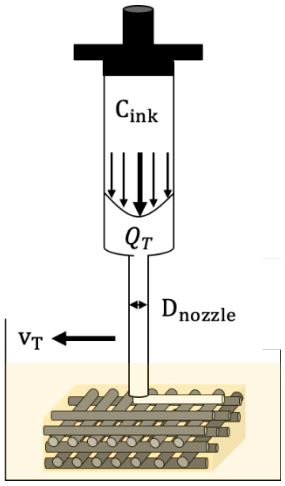

Figure 2.4. A schematic of bio-printing variables for the FRESH process in which $\mathbf{C_{ink}}$ refers to the concentration of alginate ink dissolved in DI water as a w/v%, $\mathbf{D_{nozzle}}$ reflects the inner-diameter of the print nozzle that extrudes alginate ink, and Q is a normalized flow rate, and $\mathbf{v_T}$ is the translational velocity of the nozzle speed. FRESH printing extrudes pre-crosslinked alginate into a gelatin sacrificial support bath (held constant for this analysis) where it is quickly crosslinked by calcium chloride in the support bath. Prints were heated to 35-37 °C for release and further characterization.

Table 2.1. The experimental predictor space and associated tested ranges. The printer flow rate was set to 0.1 **µ**L/s and altered in the g-code as percentage change from this value in the form of an extrusion multiplier (EM). For example, 0.4 EM represents 40% of 0.1 **µ**L/s which would give a flow rate of 0.04 **µ**L/s. While most variables are effectively continuous (the user can easily input fractional values or even alter variables during printing), the nozzle size remains discrete as it must be selected each time a dimensional design change is required. Optimization of predictor values to produce high fidelity features must ultimately be constrained to available nozzle sizes.

Predictor	Description	Tested Range	Continuity
Q	Normalized Flow Rate	$\underbrace{\text{EM}}$ i = [0.4, 0.6, 1.0, 1.5]× 0.1 µL/s	Continuous
v_T	Nozzle speed	j = [10, 20, 30, 50, 100] mm/s	Continuous
C_{ink}	Alginate concentration	k = [0.03,0.04, 0.05] w/v%	Continuous
D_{nozzle}	Nozzle diameter	l = [80, 152] µm	Discrete

A total of 48 prints with combinations of variable predictors covering the range in Table 2.1 were generated. Printed features were assessed as shown in Figure 2.5 and assigned print scores (Eq. 2.1) according to their similarity to the CAD design file.

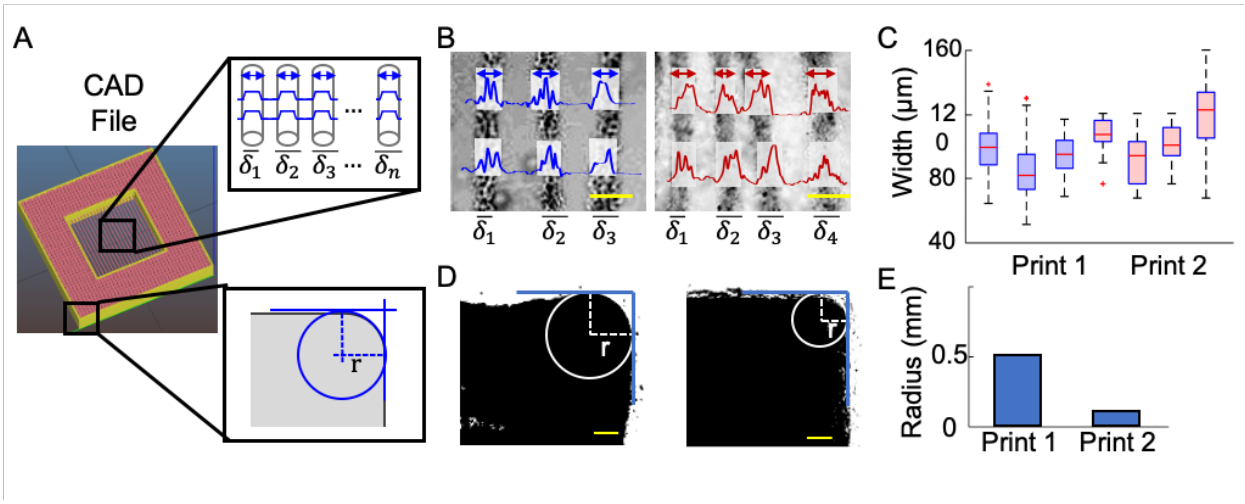

Figure 2.5. (A) CAD file design to be printed with alginate consists of a 10 mm x 10 mm x 10 mm box with 30% infill. Lines and corner features are extracted from each print. A window modifier was placed in the center and used to create freely floating printed strands which are analyzed as linewidths, $\boldsymbol{\delta}$ **(μm)**. Linewidths were free of support or infill material and were used to isolate print parameters related to the flow-gelation of alginate. Prints were evaluated by average linewidth similarity to CAD file, $\boldsymbol{\delta_{CAD}}$ **(μm)** and sharp edge fidelity (corner radius) to the CAD file, $\mathbf{r_{CAD}}$. (B) Images representing random regions of interest (ROI) are loaded into Matlab and background subtracted. Each row of the image is scanned for feature signal and $\boldsymbol{\delta}$ is defined as the measure of the linewidth for each line in the scan. The results are shown in (C) which uses a boxplot to show the variability in each linewidth throughout the full ROI, and the average linewidth, $\overline{\boldsymbol{\delta}}$ for each line. The variable $\overline{\boldsymbol{\delta}}_{\mathbf{avg}}$ will be used to describe the average of all linewidths in the ROI for a given set of predictors. (D) The corner radius **r (mm)** describes the roundness of corners and can be used as a metric to demonstrate the system's ability to estimate a rectangular edge. The corner radius was measured as the radius of the circle that is created if the corner arc is extended to form a complete circle. Corner radius $\mathbf{r\ (mm)} \rightarrow \mathbf{0\ (mm)}$ as the corner angle $\boldsymbol{\theta_{corner}} \rightarrow \mathbf{90°}$. (E) The relationship between two corner radii resulting from two different sets of predictors. Scale bars for (B) and (D) represent 250 **μm**.

Figure 2.6 shows the visible effects of printing the same CAD file but with varying combinations of process parameters. Increasing the flow rate produces thicker, vertically taller constructs while decreasing the print speed at constant flow seems to have a mitigating effect. This is reasonable considering the conservation of mass, which will be further explored in section 2.3 of the model. However, interesting effects were observed when print parameter changes were combined with composition alterations in the ink. A closer look in Figure 2.6D shows that increasing nozzle speed

at fixed flow rate produces higher-fidelity corners that were closer in shape to a right angle. Furthermore, the concentration of ink, which was used as a pseudo-viscosity modifier due to the shear-thinning qualities of alginate, can powerfully alter the expected dimensions of the printed filaments, referred to as linewidths as shown in Figure 2.6C. Thus, it is exceedingly important for the bio-print designer to be aware of the causal effects of these basal print parameters on the final output. We will show that domain knowledge of the print system in the form of parameterized physical variables will bridge experimental predictors to the print score on the small, printed dataset described in Figure 2.6. We will further demonstrate that these relationships can be predicted and optimized. To visualize the entire dataset, including additional features not discussed in this work, please see Appendix Figure 3.5.

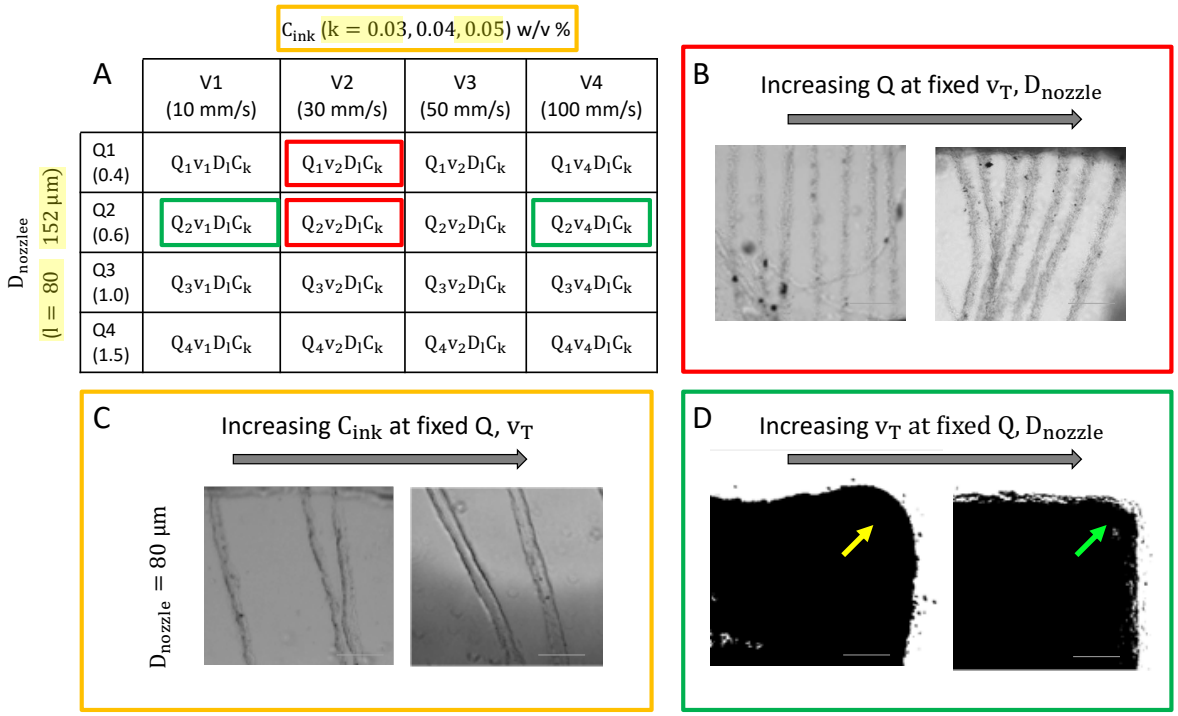

Figure 2.6. (A) The training set of prints was fabricated by combining the print predictors in Table 2.1 with their listed ranges for a total of 48 training prints, and examples of how variable combinations affect print outcomes are shown. Visual differences arise by combining the predictor inputs in different ways. (B) Fixed print speed and nozzle diameter, but at different flow rates, shows visibly different linewidths related to mass conservation. (C) Fixed print parameters but different w/v% alginate ink demonstrate the stark effect of ink composition characteristics on print outcomes related to shearing. (D) Estimation of how 90° edge changed with print speed. Scale bars represent 250 **µ**m.

2.3. Computational Methodology

2.3.1. HML Framework.

The goal of our HML model is to use system variables to predict and optimize a system response given a small training set. We developed an HML model of FRESH printing to predict and optimize for print fidelity using the small training set of print predictors described in 2.2, the experimental section. A summary of the HML model is shown in Figure 2.7 and the full model can be viewed in Appendix Table 3.3.

The HML model consists of a bottom layer (labeled as system variables in Figure 2.7), a middle layer, and a top layer (labeled system response in Figure 2.7). The bottom layer consists of a set of chosen predictors ranging from ink-related variables to machine settings that can be systematically varied in the laboratory as discussed in 2.2, the experimental section. The middle layer is a set of generalized physical equations, which along with the training set, were selected (here, using LASSO) and tuned via cross-validated as important driving forces for print fidelity. Finally, the top layer (complex system response) is the measured output (in this case print score) that is described by discovered relationships in the middle layer. Print score is a measure of the print fidelity based on dimensional errors identified in the training set prints. HML is important in this work because complex systems have identifiable predictors whose effect on the system response can be non-intuitive. One example of this is shown in Appendix Figure 3.7 in which changes in nozzle speed, v_T, result in observable differences in the corner shape fidelity for fixed flow and ink composition.

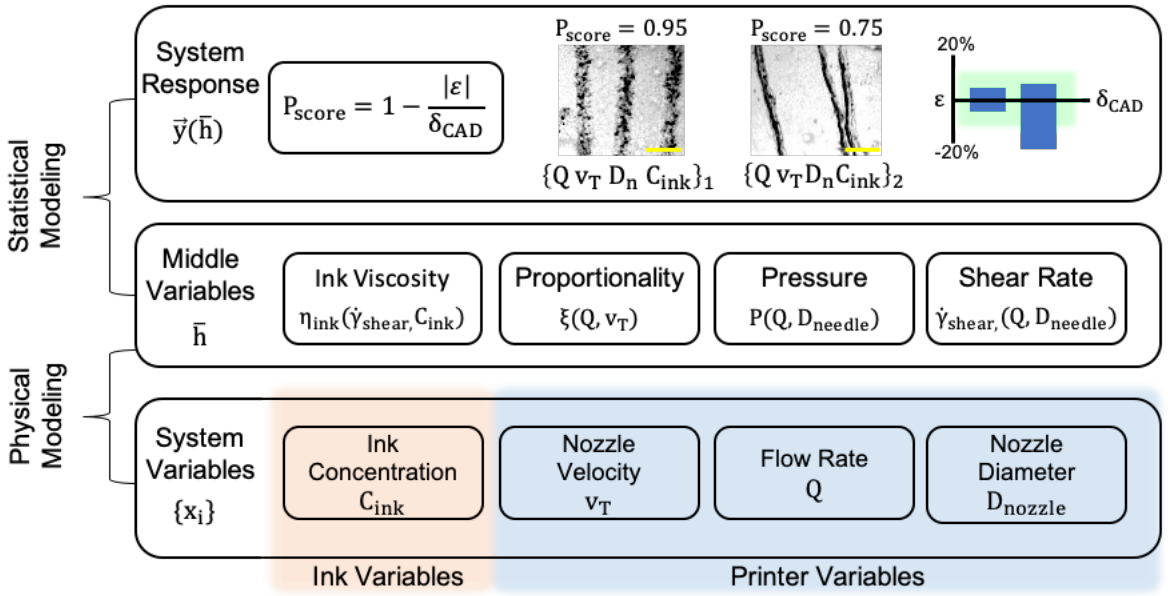

Figure 2.7. HML model of the FRESH printing process represented by a tiered structure. The bottom layer consists of system variables (predictors) that are directly controlled in the laboratory. The middle layer is a set of physical variables chosen to describe the print system and are parameterized by the bottom-layer predictors. Statistical inference in the form of LASSO is used to determine and determine the system response (print score for lines and corners). Print score is related to print fidelity by prints that have less than 10% error compared to the CAD file (or a score of at least 90%) in randomly assessed regions of interest on the print. Scale bars are 250 **µm**.

One perceived correlation between C_{ink} and the print response derives from shear-thinning fluids which display a power-law dependence of viscosity on concentration and shear rate, among other physical variables, that affect pressure and flow. The middle layer is important in this respect, and when combined with regression techniques discussed below, demonstrates how simple predictors can be bridged to the observable output via the parameterized equations in the middle layer. By leveraging domain knowledge, HML allows the small dataset to be modeled in lieu of a full design of experiments. In order to understand the effectiveness of the middle layer, we assessed the model fit error using only predictors compared to passing predictors first through the middle layer before modeling the print response. HML is a type of multi-level hierarchical statistical modeling in

which equations are used to connect pre-determined features and layers.[44] The difference between HML and conventional hierarchical approaches lies in feature discovery. While Yao and Moon[45] used a hybrid unsupervised hierarchical model combined with a SVM to discover latent relationships in the data, both the layers and features in HML are expert-guided. Furthermore, HML allows for additional complex coupling between variables that exist in different layers. Known physicochemical relationships provide interpretability of results to the user and may provide extendibility to new prints governed by the same physical principles

2.3.2. Designing the middle layer

The HML middle layer aids in correlating the underlying physical interactions and processes between experimental predictors and the system output. Towards this goal, we chose a series of rheo-physical parameters to describe the fundamental physical processes that underlie 3D printing of non-Newtonian, soft materials. It is important to note that it is not pertinent or necessary that these middle variables are immediately or expertly known. Dimensional reduction can be performed to assess the importance of the chosen variables in the middle layer towards affecting the desired system response. When the chosen middle layer variables correlate strongly to the system response, the R^2 score approaches unity, reflecting a more exact model. The following sections will discuss the chosen middle layer for the model FRESH 3D printing system, followed by a regularization and dimensional reduction that assess its accuracy in describing the system response.

2.3.2.1. Proportionality

In extrusion there is proportionality between the size of a printed filament, δ μm, with the average set flow rate, Q $\mu L/_S$ and translational nozzle speed, v_T (mm/s).[46] In principle, nozzle speeds and

flow rates are inversely related. For example, increasing the nozzle speed at a fixed flow rate will produce filaments of narrower diameter. The interplay of Q, v_T and the resulting shear forces play an important role in determining filament size, resolution, and precision when comparing printed constructs to the CAD file. This interplay is represented in Eq. 2.2.

$$\xi = \sqrt{\frac{4Q}{\pi v_T}} \qquad \text{Eq. 2.2}$$

2.3.2.2. Ink Viscosity

Alginate is a shear-thinning polymer and thus its viscosity, η_{ink} Pa · s, is non-Newtonian at high shear rates, $\dot{\gamma}\ {}^{1}/_{S}$, and is dependent on the concentration of the polymer, C_{ink}. At low shear rates ($\dot{\gamma} < \dot{\gamma}_{crit}$), alginate responds as a Newtonian fluid as the alginate polymer chain network reorganizes under flow. Above this critical shear rate, the viscosity decreases and is dependent on shear rate. The observed shear thinning of polymer solutions is caused by disentanglement of the polymer chains with increased orientation of the polymer coils in the direction of flow. The degree of disentanglement and orientation will depend on the concentration of dissolved polymer and the shear rate, which is captured in the filament size and morphology during the gelation process.[26] Comparatively high alginate concentrations require a longer chain relaxation time ($1/\dot{\gamma}_{crit}$) w hich is reflected in observed differences between 3% and 5% alginate in the training set.[47] In order to capture shear-thinning behavior, we tested the viscosity of alginate at specific concentrations of C_{ink} using a 40 mm cone-in-plate rheometer. A cross-model[21] was fitted to the data, and ultimately added to the HML middle layer as shown in Eq. 2.3. We further confirmed that the critical shear rate at which alginate begins to lose viscosity is concentration-dependent.

$$\eta_{ink} = \eta_{\infty} + \left(\frac{\eta_o - \eta_{\infty}}{1 + 0.02\gamma^n}\right) \quad \text{Eq. 2.3}$$

2.3.2.3. Effective shear rate

The effective shear rate depends on the nozzle size, D_{nozzle}, and flow rate, Q. While shear rates can vary at the nozzle walls of the extrusion nozzle, an effective shear rate was calculated for the total effective flow-rate measured in the system (Q ~0.1 μL/s). For a specified flow for small extrusion diameters relevant to the training set, 5% alginate is relatively more highly shear-thinned compared to other lower concentrations of alginate resulting in altered print features. A simple model for shear is added to the HML middle layer as shown in Eq. 2.4.

$$\gamma = \frac{Q/A}{D_{nozzle}/2} = \frac{8Q}{\pi {D_{nozzle}}^3} \quad \text{Eq. 2.4}$$

2.3.2.4. Pressure

The pressure of a shear-thinning fluid inside a small capillary tube is altered due to Hagen-Poiseuille flow for non-compressible fluids. Its form is dependent on the fluid viscosity, power coefficient from shear thinning, flow rate, and nozzle diameter size as shown in Eq. 2.5.[48]

$$P = 32\eta_{ink} L \left(\frac{3n+1}{4n}\right) \frac{4Q}{\pi {D_{nozzle}}^4} \quad \text{Eq. 2.5}$$

2.3.3. Model Assessment

As we envision the future of bio-printing to include costly and complex components such as cell-laden inks, the throughput of experimentation will continue to be challenging, limiting the dataset

in size and necessitating the development of improved ML strategies that can cope with small data. The effectiveness of multiple modeling strategies is assessed by comparing the residual sum of squares to the total sum of squares, or R^2 value, between model predictions and observed data. We compared HML to a shallow, two-layer feed-forward neural network (the method of which is described in Appendix Figure 3.8). For a comparison of model performances, see Appendix Figure 3.9.

The middle layer is mapped to print response by statistical inference in the form of Least Absolute Shrinkage and Selection Operator (LASSO). LASSO is a powerful tool for regularization that drives model variables with weak correlations to zero, reducing the complexity of an otherwise high-dimensional print-space and retaining only those variables that correlate strongly with system properties. The variables used for LASSO included the squared and cross-terms of the four middle layer variables. First, the range of tuning parameters was set and cross-validated to determine the correct hyperparameter to use for the regression. The response was a vector of coefficients that reflected a parameterization of the response surface (line and corner scores) in terms of the middle layer variables (Eq. 2.6 and Eq. 2.7). A simple neural network was run on the same set of predictors using a two-layer feed-forward network with a sigmoid activation function for hidden layer neurons and a linear function for output neurons. The network was trained with the Levenberg-Marquardt back-propagation algorithm following the Neural Fitting tool in Matlab. The R^2 values comparing the model performances are reported in Figure 2.8.

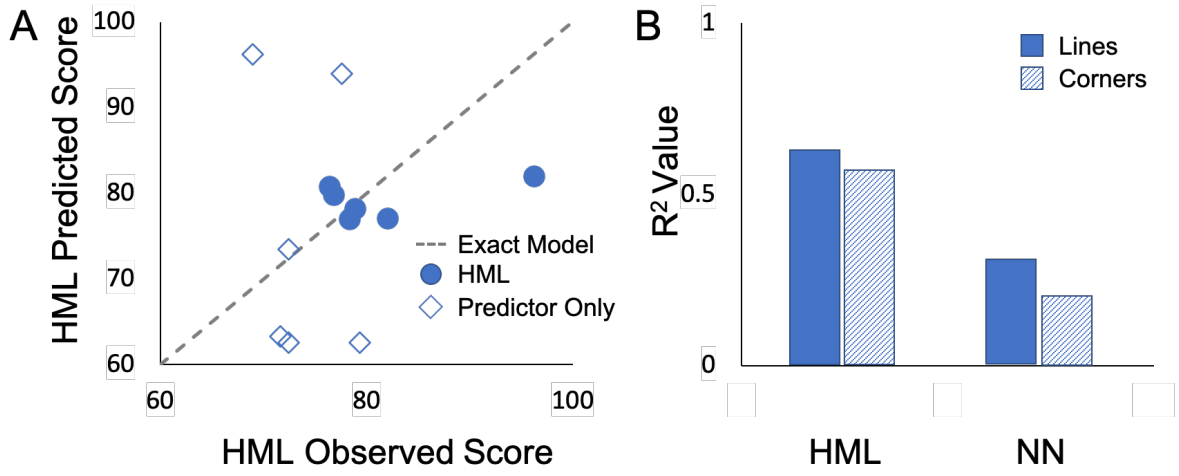

Figure 2.8. We performed an HML fit and compared the performance to both conventional statistical inference and to a simple neural network. (A) We demonstrate how the middle layer in HML improves model fit of linewidth scores compared to conventional statistical inference with predictor-only inputs. LASSO is used to model the linewidth scores using only the four print predictors and leave-one-out cross-validation. The resulting $\mathbf{R^2}$ is $\mathbf{-0.439}$, which is worse than fitting the mean to the data. Addition of the middle layer resulted in an improved test score, $\mathbf{R^2}$ of $\mathbf{0.643}$. The model accuracy is demonstrated by plotting the predicted score vs. the actual score where the 45º line represents error-free print fidelity. (B) HML is compared to a simple neural network with 10 neurons. The R^2 values for both models are reported for linewidths and for corner print features. In both cases HML out-performs a black-box network.

The goal here is not to assert that HML will always out-perform a carefully trained and cross-validated neural network that has sufficient data for model fitting and discovery of variable relationships. We have observed some fitting success with neural networks for predicting corner fidelity, and also with direct predictor-only fitting using non-linear methods such as random forest (Appendix Figure 3.8). In each analysis, however, the addition of the middle layer improved model fit. The aim of this analysis is to leverage physical knowledge to improve predictive power and to gain interpretability of the results for downstream analysis.

Statistical regression correlates variables to the output for the HML method. A LASSO regression from the parameterized middle layer to the top layer showed good model performance with an R^2

of 0.643. In contrast, when the experimental variables with squared and cross-terms included were modeled directly to the response, we observed poor model performance on the test set with $R^2 = 0.23$ while middle layer performance was generalizable to a score of 0.50 on the same test set. This demonstrates the importance of the HML framework, specifically in leveraging the middle layer for improved performance. Eq. 2.6 and Eq. 2.7 represent the print scores as linear combinations of the dominant middle layer variables, their cross-terms, and squared terms.

$$P_{\text{lines}} = 2.17\gamma + 1.21\eta_{ink} - 2.68P + 1.63\xi + 2.23\gamma P - 0.869\gamma\xi - 0.500\eta_{ink}\xi + 1.01P\xi - 2.23\gamma^2 - 0.47{\eta_{ink}}^2 + 0.25P^2 - 0.61\xi^2 \quad \text{Eq. 2.6}$$

$$P_{\text{corners}} = 2.17\gamma + 1.212\eta_{ink} - 2.678P + 1.63\xi + 2.232\gamma\xi - 0.870\eta_{ink}\xi - 0.498\gamma^2 + 1.010{\eta_{ink}}^2 - 2.232P^2 - 0.61\xi^2 \quad \text{Eq. 2.7}$$

2.3.5. Physical Interpretation

Feature importance is useful information for the designer to build physical intuition for the system. For HML, feature importance emerges from minimizing the L_1 norm and fitting coefficients to the middle layer variables. Eq. 2.6 and Eq. 2.7 show variables that met the regularization criteria. Ink viscosity played a significant role in linewidth, Figure 2.9A, which shows altered line dimensions resulting from a concentration difference in the ink, quantified in Figure 2.9B. Shear rate and ink viscosity interplay with ink concentration as the flow profile for shear-thinned alginate is different from a typical non-compressible fluid, Figure 2.9D. Noticeable changes in surface roughness can also be noted in the images, which may be attributed to the effects of polymer chain relaxation on the gelation process, shown in Figure 2.9C.[26] Finally, feature selection shows shear rate becomes important for nozzle directional changes which are necessary to print 90º angles (corresponding to

a corner radius of 0 mm). Visual comparison of corners at varying nozzle speeds is shown in Appendix Figure 3.7. The effects of acceleration and deceleration have important implications for reversing fluid flow, which can be challenging for highly viscous materials like hydrogels. Thus, higher w/v% concentration of alginates may be more suitable for designs that require smaller feature size.

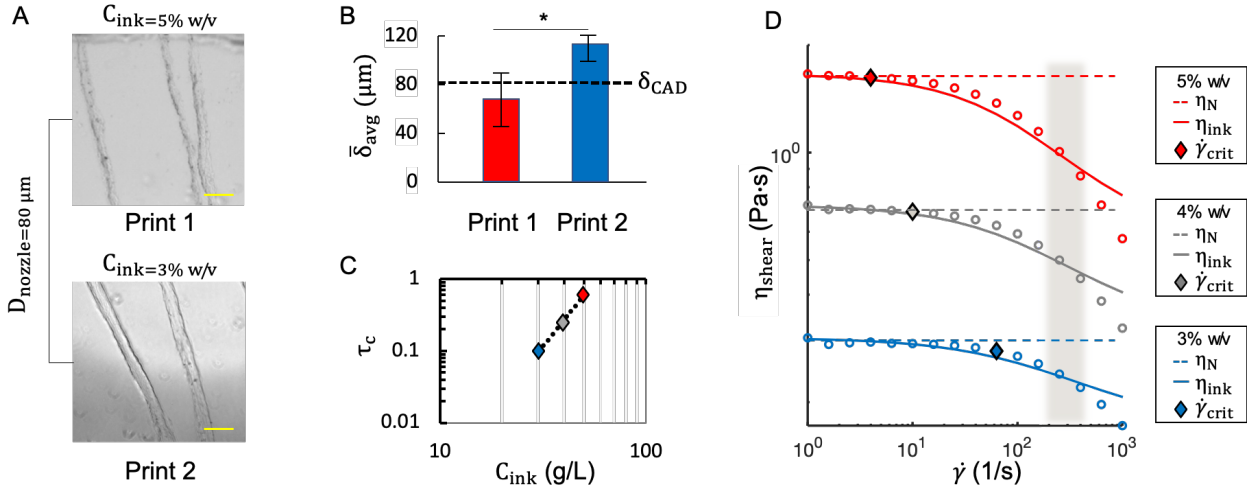

Figure 2.9. (A) Visual comparison of two prints at fixed *Q*, $\mathbf{v_T}$, $\mathbf{D_{nozzle}}$ but at different ink composition: print 1 at $\mathbf{C_{ink}}$ = 5% w/v and print 2 at $\mathbf{C_{ink}}$ = 3% w/v. (B) Quantification of average linewidth, $\bar{\boldsymbol{\delta}}_{\mathbf{avg}}$, shows an over- and under-estimation of the desired linewidth, $\boldsymbol{\delta_{CAD} = 80\ \mu m}$, from print 1 and print 2, respectively. Print 2 had a significantly higher linewidth compared to print 1 ($p < 0.001$; n=8). (C) Relaxation time, defined as $\boldsymbol{\tau_c = 1/\dot{\gamma}_{crit}}$, of alginate in DI water for 3% (blue), 4% (grey), and 5% (red) w/v showing a linear trend on a log-log plot with the slope $\sim \boldsymbol{C_{ink}^{3.6}}$ implying the blob overlap concentration regime for the alginate ink polymer.[47] 5% w/v alginate requires more time to relax to its original state compared to 3% w/v. (D) Shear viscosity of alginate solutions at concentrations of 3%, 4%, and 5% w/v as a function of shear rate measured with a 40 mm cone-and-plate rheometer. Newtonian viscosity, $\boldsymbol{\eta_N}$ and the concentration dependence of ink viscosity, $\boldsymbol{\eta_{ink}}$, after the critical shear rate (diamond). At $\boldsymbol{D_{nozzle} = 80\ \mu m}$, we hypothesize that the 5% w/v ink was in a regime where it is shear-thinned more heavily than for 3% w/v alginate (grey box).

Dominating coefficients in the HML equations suggest that proportionality is a driving feature for print score. Based on mass conservation for Newtonian inks, the volumetric flow rate Q scales

with linewidth δ and translational velocity v_T as $Q = \frac{\pi\delta^2}{4} v_T$. Appendix Figure 3.6 shows how the data from the training set scales linearly with proportionality. The cross-sectional areas of printed lines for 3% and 5% alginate in the training set are plotted as a function of the ratio of print parameters, δ^2 versus Q/v_T. The effect of ink concentration on proportionality is apparent in the constant offset of 3% w/v printed features compared to the theoretical relation, $\delta^2 \sim \frac{4}{\pi}\frac{Q}{v_T}$ shown in Appendix Figure 3.6(A). Data for 5% alginate w/v scales well for smaller diameters, Appendix Figure 3.6(B). We observe in the data that 5% alginate produced higher print scores for small linewidths but resulted in lower print scores at high flow rate where the ink may thinning due experiencing higher shear forces.

Finally, the model predicted that printing corners with smaller corner radii generally had a better performance at higher translational nozzle speeds. A clear trend can be seen in the training set as shown in Appendix Figure 3.7 in which a speed of 100 mm/s had more accurate corner scores compared to 10 mm/s for 4% alginate. Physical interpretations of the HML model of the most fundamental building blocks of construct design (lines and corners) can act as starting point for selecting the best parameters that will go on to build more complex shapes, and can act as a gateway to introduce constraints in future optimization.

2.3.6. Optimization

The model was optimized with differential evolution, an optimization method built in SciPy, used to minimize the print error.[49] Constraint ranges were applied to each of the bottom layer variables based on the range of the inputs. The HML equations given in Eq. 2.6 and Eq. 2.7 in combination with the physical equations in the middle layer were used to back-calculate the values of the

experimental predictors that give rise to HML-predicted high-fidelity prints for lines and corners, respectively. New prints were generated to test the optimization predictions, the results of which are summarized in Figure 2.10 and Table 2.2. As predicted, prints generated from optimal predictors reflected less than 10% deviation in dimensionality from the CAD file in randomly analyzed regions (and thus have a score of at least 90%), which we define as define high-fidelity. The HML model fit and optimization methodology are summarized and discussed in detail in Appendix Figure 3.9.

Table 2.2. Optimized print parameters from HML equations and predicted print score showing parameter settings that resulted in the highest print fidelity. New experiments were run to validate the HML predictions.

Feature	Desired Dimension	Optimized Printer and Material Settings				Predicted Print Score	Measured Print Score
		D_{nozzle} µm	C_{ink} w/v%	Q µL/s	v_T mm/s		
Linewidth	80 µm	80	0.05	0.06 (EM 0.6)	24.67	92%	95%
	152 µm	152	0.03	0.16 (EM 1.49)	11.09	95.0%	92.0%
Corner Radius	0 mm	80	0.04	0.06 (EM 0.6)	89.00	98%	98%
	0 mm	152	0.04	0.13 (EM 1.47)	92.06	96%	96%

Results from Table 2.2 show that a clear trend is established between the corner radius and faster translational nozzle speeds, predicting that 89 mm/s and ~92 mm/s will produce high-fidelity corners with 4% w/v alginate. Furthermore, higher concentrations of alginate were expected to produce lines with smaller diameters[50] which was observed in the HML model prediction,

showing 5% w/v alginate as optimal concentration to print 80 µm. Additionally, we noticed smaller nozzle diameters produced better print scores, and observed a trade-off between optimized features, as shown in Figure 2.10. Predictors for optimal linewidth scores gave sub-optimal corner radius estimation while optimized corner predictors traded linewidth for corner fidelity. The clear trade-off between optimizing for directional control and line control shows how 3D printing systems will require multi-objective optimization for complex printed constructs.

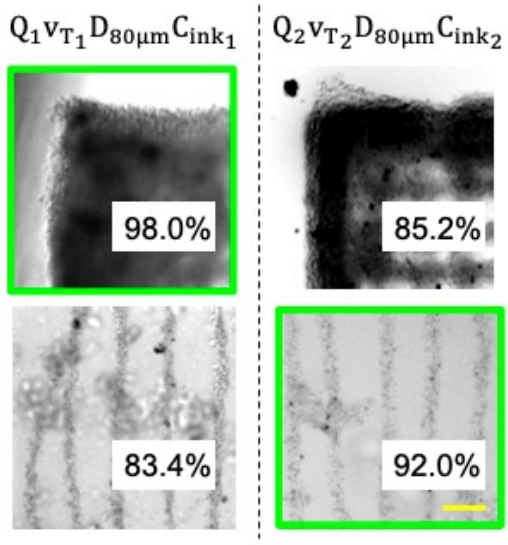

Figure 2.10. A tradeoff between optimized features, showing that conditions for optimizing corners were less well suited for lines (left), and conversely optimization of line morphology reduced the print fidelity for corners (right). In this experiment, $\mathbf{Q_1 = 0.6\ EM}$, $\mathbf{v_{T_1} = 24.7\ mm/s}$, $\mathbf{C_{ink_1} = 5\%\ w/v}$ and $\mathbf{Q_2 = 0.6\ EM}$, $\mathbf{v_{T_2} = 89\ mm/s}$, $\mathbf{C_{ink_2} = 4\%\ w/v}$. Scale bar represents 250 µm.

2.3.7. Process maps and printability: modeling error bias due to material feedback from the design space

The impact of the optimized model is realized by creating a generalized tool for the 3D bio-printing designer that shows the predicted success and pitfalls of variable combinations. Due to the proportionality of printer variables, namely that the cross-sectional area (A_x) of printed material scales with the ratio of flow rate and print speed, $A_x \propto {}^{Q}/_{v_T}$, the 3D printing designer theoretically has nearly infinite possibilities of ${}^{Q}/_{v_T}$ ratios that can produce a desired, fixed cross-sectional feature size. Figure 2.11A shows this relationship in more detail where (Q, v_T) can be varied to produce a constant slope, A_x. However, the non-Newtonian, shear-dependent nature of biopolymer ink extruded into a non-Newtonian, Bingham plastic support fluid means that not all combinations of (Q, v_T) are created equal and for large, costly prints, it is paramount to predict the best (Q, v_T) parameters that minimize error from the start.

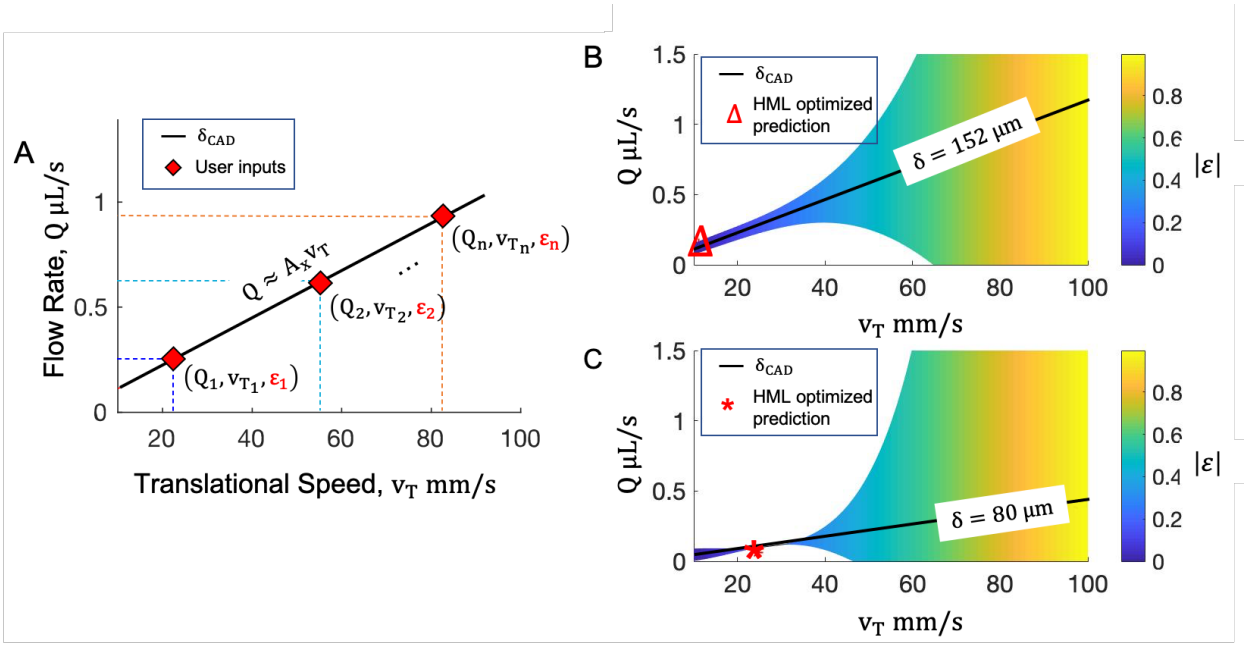

Figure 2.11. Printability plots for desired feature size representing the interplay of printer machine variables, flow rate $\mathbf{Q}$ and translational speed $\mathbf{v_T}$, on the predicted error in printing at a given nozzle diameter and ink concentration. The color scale and width of the curve show the HML predicted error for printing at different values of flow rate and translational print speed with optimal values identified in each. Furthermore, we updated the specificity of the plots by using a fitted equation to describe the dependence of $\boldsymbol{\delta}$ on $\mathbf{C_{ink}}$. Print data and corresponding fits, $\boldsymbol{\delta(C_{ink})}$ are shown in Appendix Figure 3.6A and B for $\boldsymbol{\delta = 152\ \mu m}$, and $\boldsymbol{\delta = 80\ \mu m}$ respectively. (A) For linewidths, under the assumption that printed material is freely extruded into the gelatin support bath from a flat, non- tapered nozzle, then $\mathbf{A_x}$ can be loosely approximated to linewidth as $\mathbf{A_x} \approx \boldsymbol{\pi/4\,\delta^2}$ giving rise to a slope that is proportional to linewidth (B) Printability plot for a 152 $\boldsymbol{\mu}$m nozzle with 3% w/v alginate and (C) 5% alginate from an 80 $\boldsymbol{\mu}$m nozzle showing a predicted minimum region of error for each. The color map represents the magnitude of HML predicted percent dimensional error from the CAD file. The width and shape of the error curves are mapped onto the chart as a visual tool to demonstrate to the user where the print error minimum lies given the optimal $\boldsymbol{(Q, v_T)}$.

Figure 2.11B and C shows basic process maps for printing 152 µm and 80 µm features for 3% w/v and 5% w/v alginate, respectively. The color map shows the magnitude of HML predicted percent error from the CAD file. An equation for the magnitude of the error (calculated from Appendix

Table 3.4 HML results) is then parameterized onto the chart to create a visual of the location of minimum error, $|\varepsilon|$ given the optimal user choice of (Q, v_T). The parameters from

Table 3.4 that gave rise optimal linewidths for 3% and 5% alginate are shown to have errors that lie within the minimized regions in both Figure 2.11B and C. We envision the development and use of process maps towards more complex bio-print design and shapes wherein constraints add new challenges to optimization that can be inputted into the HML middle layer. For example, printing with cell-laden inks may restrict predictors to those that offer safe shear forces for cells. Process maps such as those described in Figure 2.11 can help locate the minimum dimensionality error under such constraints.

The generalizability of the HML model to related systems is an area for future investigation. We propose using HML to optimize a FRESH printing system with new materials, variables, or printer settings, for example, with various shear-thinning bio-inks whose viscosity is outside the range of ink tested in the training set but follows the generalized cross law for viscosity described in Eq. 2.3 in the middle layer. Fixed physical relationships between the bottom and the middle layer provide a facile route for new materials or printer variables into the model-specific relationships from LASSO, such as the introduction of new nozzle sizes, or different dilutions of alginate that span broader viscosity regimes. Furthermore, additional predictors are expected to improve model fit, such as the inclusion of a predictor for monovalent salts in the ink. We foresee the use of the HML model and process described here as a future methodology for optimization of new print systems using a manageably sized training set from the laboratory, thereby allowing bio-printing

to be a good candidate for machine learning without the need for a large dataset. Furthermore, we believe this model has the potential to transfer to more complex printed constructs via the physical gateway in the middle layer, such as for printed constructs containing biological matter such as collagen or cell-laden inks with potential clinical relevance.

2.4. Conclusion

We have applied an HML methodology in order to use a small dataset to build predictions for high-fidelity bioprinting with alginate. The model utilized physical relationships to link build parameters to the print score, allowing improved accuracy for both training and test data. LASSO was utilized for tuning model coefficient and for feature selection of dominating causal physical forces that drive print error. Optimization allowed for the prediction of build parameters that give rise to high-fidelity prints, and a trade-off was elucidated between the most basic printing elements: printed lines and 90º corners. This approach could be used to guide printing constructs by selecting optimal material, formulation, and process variables for a given form. Known physical equations help expand the future generalizability of our method to incorporate a diversity of inks including biopolymers and cells, and these can be used to guide printing of complex constructs where figures of merit include both structural fidelity and biological function.

Chapter 3. Application of HML models: Proof of concept for leveraging the physical middle layer for rapid optimization of parallel 3D bioprinting systems

3.1. Introduction

It is a vision of 3D bioprinting that we can print with multiple or new materials. From a processing standpoint, despite the fine-tuning of parameters, the overall physical system remains largely defined by a small number of fundamental physical parameters. In theory, the appropriate middle layer developed from HML on FRESH-printed alginate is composed of the minimum physical features required to describe the system and can act as a knowledge source for unseen 3D bioprinting systems. The following discussion explores the generalizability of the middle layer to target bioprinting systems by updating the physical variables with user-defined estimations of the new materials used. To demonstrate the functionality of this approach, we task the source model with predicting print outcomes from estimations of the properties of unseen inks. Furthermore, we generally discuss how more complex target systems can be rapidly optimized by greedily leveraging similar data and updating variable weights.

Transfer learning can be imagined as an ultimate form of generalizability, not just to unseen data, but to new target systems. Transfer learning generalizes to new systems by leveraging knowledge of source material in order to predict new target systems that have intersecting mechanisms, features, or predictor distributions, all without the need for copious experimental data.[52] The generalizability of 3D bioprinting models is of great interest due to the plethora of available printing mechanisms, biological materials, and design choices that may be used in the clinic. Resource challenges for re-collecting data in order to conduct data-driven optimization of new printed materials via design of experiments lies in time-sensitive clinical needs for "plug-and-play" prototyping, lack of facile, quantitative shape/function assessment techniques, and high biological

material costs. The objective of this section is to extend the HML model framework developed for alginate FRESH printing (hereon called the HML-alginate model) to successfully predict the fabrication parameters for high-fidelity FRESH-printed constructs with new materials. Specifically, we will show a proof-of-concept of the rapid optimization of two different FRESH 3D bioprinting materials: collagen I ink and an optically transparent support bath via knowledge bridges constructed from latent variables in the HML embedded middle layer. Free-floating collagen I lines are predicted to within 10% of expected dimensions and a pilot dataset for rapid HML modeling and optimization of a Pluronic-based support bath is discussed. Finally, we discuss the future applications of HML in 3D FRESH bioprinting for cell-laden inks.

3.2. Proof of concept: leveraging the physical middle layer for predicting parallel 3D bioprinting systems

Existing approaches to statistical transfer learning (transferring information from source to target) can be very broadly categorized by the following methods: (1) instance-based transfer learning, (2) feature-based transfer learning and (3) parameter-based transfer learning.[53] Instance-based transfer learning can be used when the source and the target instances are generated from two different but closely related distributions, allowing the source data to be reused in the target task. Feature-based transfer approaches attempt to forge a bridge for knowledge transfer by learning common features or a common structure between source and target data. Parameter-based transfer learning assumes that source and target tasks share some common parameters, hyper-parameters, or prior distributions. Previous work involving transfer learning in 3D printing is sparse. One notable work seeks to address the challenge of comprehensively improving shape fidelity in FDM plastic printing in the face of a high-complexity error-space and limited training data.[54] In this

scheme, the dimensional error of various printed shapes is predicted in a parameter-based transfer learning approach. In particular, shape deviation in the training data from expected dimension is decomposed into two independent error models that differentiate shape-independent error from shape-specific error. Shared dominant parameters discovered from each statistical model inform one another, forging a bridge that can infer up-front error in new shapes.

The discussion of transfer learning in 3D bioprinting is still quite nascent as the similarities and differences between source and target printing systems that would inform the type of knowledge transfer necessary for good predictive models is still complex and elusive. Unlike plastic printing, bioprinting can experience additional constraints on the experimental space due rheological and functional complexity, motivating the powerful use of transfer learning for modeling in a reduced experimental space. To date, the extension of learned knowledge to predict relatable 3D bioprinting systems has not been attempted. We will show that a supervised middle physical layer allows for facile predictions of optimal linewidth for collagen I via the trained HML-alginate model without introducing additional experimental data. Next, we begin to optimize a clear Pluronic-based support bath solution to be used in place of gelatin support baths in FRESH. This analysis seeks to be proof-of-concept, and we will discuss future steps for more rigorous investigation.

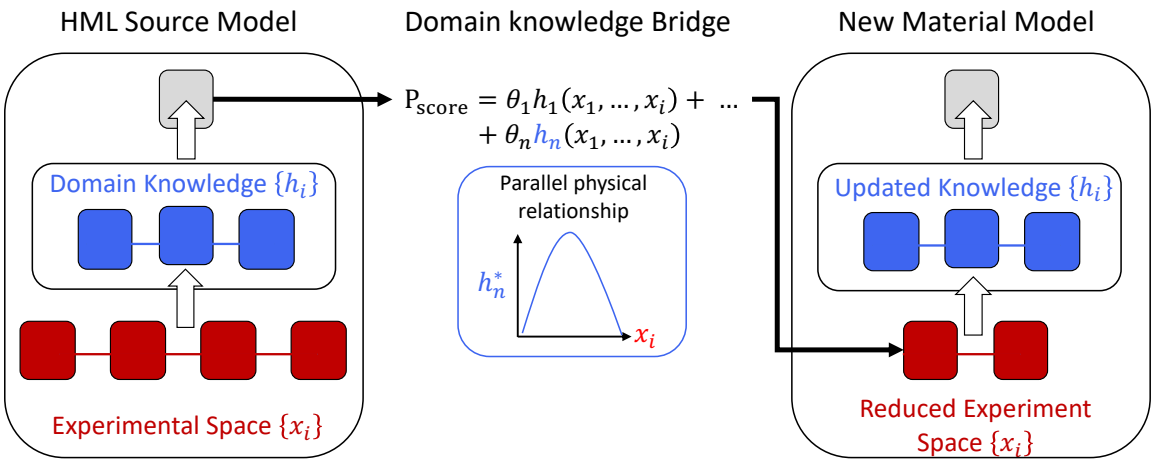

Figure 3.1. HML-optimized model for FRESH-printed alginate (discussed in Chapter 1) can be used for reducing the experimental space necessary to predict high-fidelity constructs from parallel printing systems. In this work, we define a parallel printing system as one that shares physical mechanisms captured in the physical middle layer of the source system. Shear-thinning inks such as alginate printed on the same FRESH printer and support bath, or novel print materials (such as a new sacrificial bath) that demonstrate Bingham plastic behavior are examples of new materials that can be bridged through the middle layer of the original HML-alginate model. The properties of new materials are bridged into a new model by updating the associated latent variable equations, $\boldsymbol{h_n(x_1 \dots x_i) \rightarrow h_n^*(x_i \dots x_i^*)}$. New variable weights from statistical inference between the middle and top layers are calculated in the event of new predictors, but with a reduced experimental space. In some cases, such as for Collagen I, the print space for alginate optimization can be re-used to predict high-fidelity printed lines. In the case of novel material optimization, such as a novel Bingham support bath printed with alginate ink, additional but reduced experiments are required.

3.2.1. Materials and Methods for Collagen I ink prediction

Collagen I lines were extruded into a gelatin support bath prepared as described previously (See FRESH printing in section 2.2.3). For collagen ink, 5 mL of collagen (Lifeink 200, Type 1 purified high concentration, Advanced BioMatrix) was diluted from 35 mg/mL to 23 mg/mL in 2.5mL of 0.24M acetic acid. An air-tight syringe was filled with 3 mL of diluted collagen ink and centrifuged for at 3000 G for 30s to get rid of bubbles. Once bubble-free, the collagen ink was loaded into the extruder for printing. The gelatin support bath was prepared according to section 1.2.3. However, instead of replacing ethanol with calcium chloride (as was done for alginate printing), the cross-

linker used was 0.5M Na-HEPES at a pH of 7.4 since change in pH is the appropriate cross-linker for collagen. For printing, a152 μm nozzle was used with the following print settings derived from the HML-alginate model ($(\mathrm{Q}, \mathrm{v_T}, \mathrm{D_{nozzle}}, \mathrm{C_{ink}}) = \left(0.1\frac{nL}{s}, 23\frac{mm}{s}, 152\ \mu\mathrm{m}, 23\ \mathrm{mg/mL}\right)$. In this system, $\mathrm{Q} = 0.1\frac{\mathrm{nL}}{\mathrm{s}}$ corresponds to an extrusion multiplier (EM) of 1. Furthermore, the layer height was kept the same as for alginate prints at 60 μm. The CAD file printed was kept the same as for alginate, and collagen free-floating lines were analyzed as in Figure 2.5. Printed collagen lines were assessed with a phase-contrast microscope with a 5x objective lens for deviations in dimension from the expected size of 152 μm. Collagen I viscosity was estimated and entered into Eq. 2.6. The measured print score for collagen was then compared to the HML-predicted score. Small deviations in collagen line straightness were considered to be a small uncertainty. These deviations are likely a result of natural properties of the polymer that are revealed due to extrusion and due to interaction with heterogeneous particles in the gelatin bath solution.

3.2.3. High-fidelity collagen I bioprinting from HML-alginate models: proof-of-concept

Free-floating collagen I lines were printed from expert-user process settings described in Lee *et al*. Science 2019.[35] To demonstrate transfer learning of the HML-alginate model to collagen, we tasked the HML-alginate model with matching the measured collagen print score (printed with expert-user settings) with an HML-predicted score based on updated, collagen-specific estimates of physical variables in the middle layer bridge. The HML-predicted score was calculated by updating η_{ink} in the HML-alginate model (Eq. 2.6) using an estimated value of collagen viscosity as a function of shear rate (see Figure **3.2**) that was gathered from the literature.[55] A profile of collagen I viscosity as a function of shear-rate can be measured with a 40 mm cone-in-plate

rheometer for better shear-dependent viscosity estimates. Nevertheless, collagen lines were predicted to within 10% of their measured score. These results show the HML model can transfer to differences in ink compositions, thus allowing the model to accurately predict printed feature outcomes with different inks.

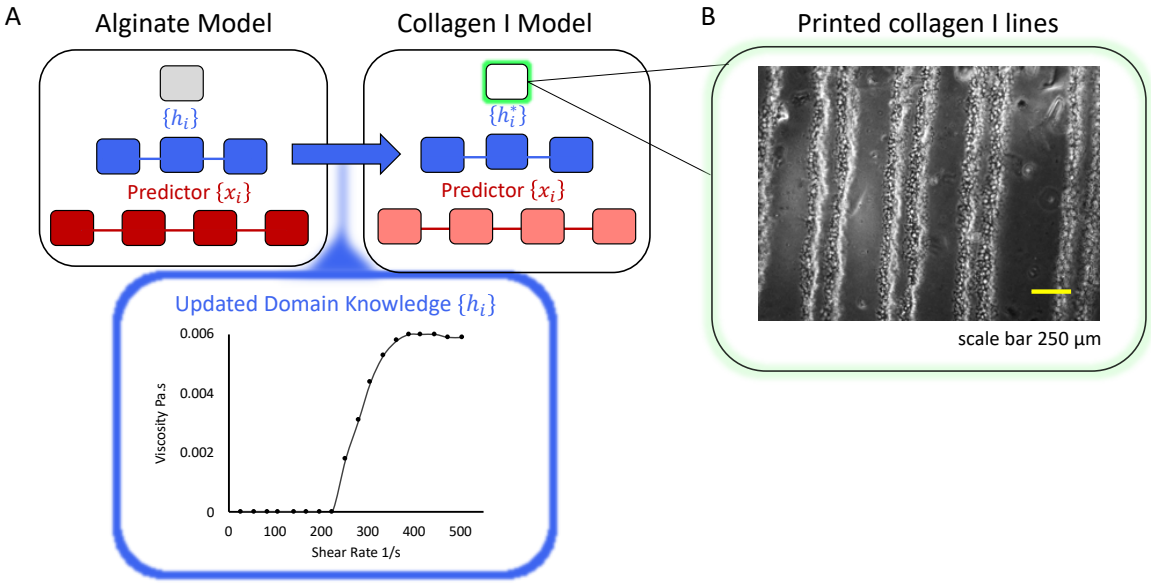

Figure 3.2. Proof-of-concept for adapting the HML model generated for alginate on collagen I, a parallel material ink. The HML-alginate model, which has learned FRESH printing over alginate features, is used to generate an equation of parameterized physical variables needed to predict high-fidelity prints in alginate. The middle layer then acts as a bridge to predict FRESH printing with collagen I ink. Native collagen was used as an estimate for collagen I viscosity in $\boldsymbol{Pa \cdot s}$. Better estimates can be made on the ink itself with a cone-in-plate rheometer. Nevertheless, the model was able to closely predict the lines for collagen extrusion based on updating the middle layer with collagen-specific estimates in Eq. 2.6 using the model weights cross-validated for alginate. As a result, the middle layer acts as a bridge for parallel printing systems via shared physical links, and it was able to translate to a collagen-ink space without running additional experiments. The analysis is meant to show a proof-of-concept for future endeavors that strive to make more complicated print predictions. The graph of viscosity curves is derived from Draipandy *et al*. J Mater Chem B, 2015.[55]

As shown in Table 3.1, the HML-predicted collagen prints (via the middle-layer parameterization) predicted lines with ~8% error compared to the measured values which measured approximately

10% error. We expect collagen to be less shear-dependent than alginate and thus more closely follow the inner-diameter nozzle dimensions. As the HML-alginate model would correctly predict a print score of 80% for alginate lines for the same print settings, we suspect that the feature importance relationships discovered from LASSO regularization in chapter 2 has a high generalizability regarding the causal physical parameters that drive the print response. These causal relationships are exploitable to unseen print materials, such as collagen, which undergo relatable physical forces in the printing system.

Table 3.1. Collagen I printed lines from expert process settings from HML predictions.

	Expert-user collagen print settings
$P_{score}(\text{measured})$	90%
$P_{score}(\text{HML} - \text{Predicted})$	92%

More complicated print predictions requiring an expanded feature space such as multiple ink chemistries, dual-extruders, support bath optimization, and complex shapes may necessitate the use of additional predictors not covered in the original HML-alginate model. In the next section, we will discuss how additional experiments can be used in combination with an updated middle physical layer for rapid optimization of new print materials in a reduced experimental space via the use of collected data and statistical inference to update variable weights. While random forest and simple neural networks are powerful for discovering predictor-response relationships on the training data, the use of a supervised middle layer provides the model interpretability necessary for transferring targeted knowledge to parallel 3D bioprinting systems.

3.3. Future Directions

The application of HML to more complex models will briefly be discussed in the context of designing an optimization strategy for a new optically transparent support bath solution to replace gelatin support bath in FRESH. Second, the application of HML to cell-laden inks will be discussed in the context of functional outputs (such as cellular viability), and process maps for multi-objective optimization.

3.3.1. Pilot data for rapid optimization of new support bath for FRESH printing

While we have shown that 3D printed inks can be optimized, the gelatin support bath solution has remained distinctly constant. Unlike silicone FRE printing in which the yield-stress of the Bingham PDMS support bath can be tuned by the polymer concentration and cross-linker, gelatin microparticles and their subsequent effect on yield stress are much more difficult to control. As a result, ink material and print predictors are tuned to the gelatin bath and not the other way around. The yield stress in gelatin-based support bath can be related to microparticle size, concentration of viscosity modifiers (such as gum arabic), packing density, and other factors.[56] Furthermore, gelatin microparticle support baths form an opaque hue due to large microparticle size that occludes real-time view of the construct during the printing process. While the index of refraction of prints could be tuned to match that of the gelatin support bath, allowing for real-time monitoring of the print process, an optically transparent support bath with tunable yield-stress and thermo-reversible properties within a biologically compatible temperature range may provide great benefit to the bio-print optimization process. Pluronic F-127 is an excellent candidate for a FRESH support bath as it is biocompatible, behaves as a Bingham plastic, and is thermo-reversible.[57] Furthermore, the

nature of micellular-packing gelation could mimic the gelatin microparticles previously used in FRESH printing.

In this section, the ability to extend HML methodologies to improve print optimization via tuning of the support bath will be briefly discussed. Additional experimental data will be necessary in this case to create a usable model since the original HML-alginate model did not include yield stress as a predictor (the gelatin support bath was kept constant). The use of, for example, storage and loss modulus data as middle layer variables could allow for additional tuning on the 3D bioprinting model for optimized prints. The dependence of the F-127 yield stress on temperature and concentration of dissolved polymer is a relationship that can be leveraged as a physical variable rather than a discovered relationship. The experimental space necessary to model a target printing system with this new bath solution in which there are numerous benefits to the printing process will briefly be discussed.

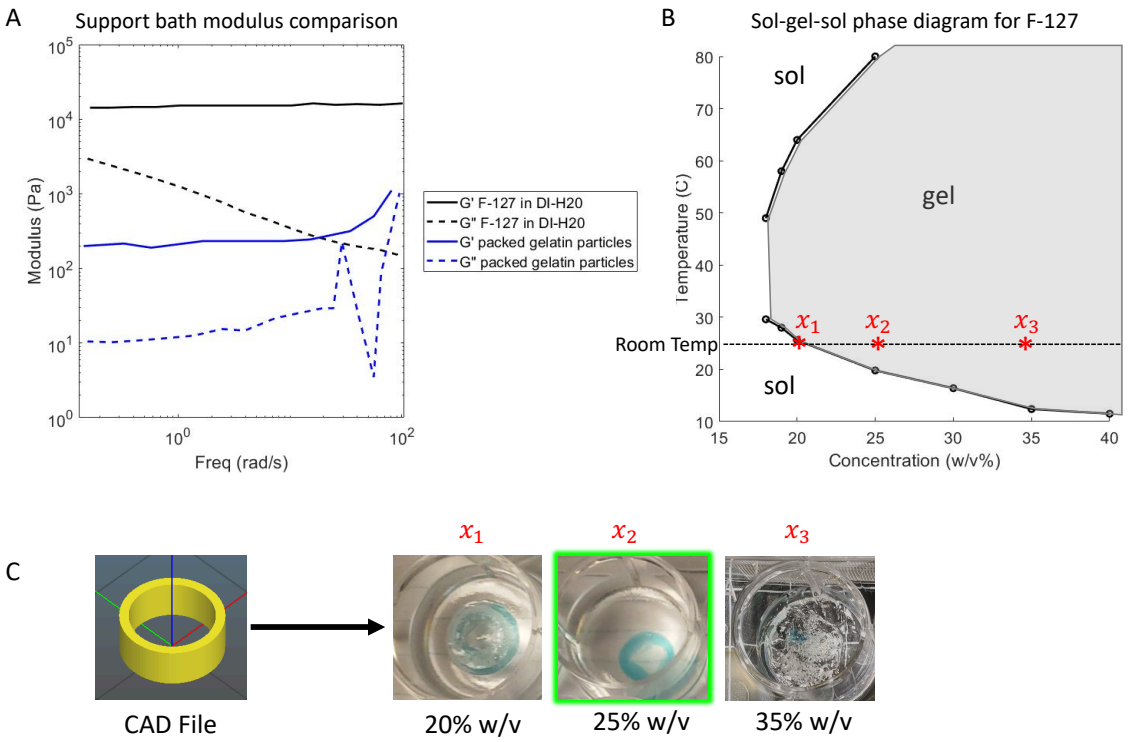

Figure 3.3. Early development of an optically transparent support bath for alginate printing. (A) The micelles formed when F-127 is dissolved in aqueous environments at 37°C causes the polymer to gel and act as a yield-stress material similar to packed gelatin particles used in the HML-alginate model in Chapter 1. One major difference between Pluronic bath and gelatin bath is in the magnitude of the storage and loss modulus: Pluronic is two orders of magnitude stronger than the gelatin bath. As a result, we should expect that the translational print speeds for alginate printed in Pluronic bath solution will be higher compared to printing in gelatin. The data for F-127 in DI-H20 is taken from Gioffredi *et al*. Procedia CIRP 2016[58] and the data for gelatin is taken from Hinton *et al*. Scientific Advances 2015.[34] (B) shows the thermo-responsive nature of the Pluronic polymer and its ability to release prints just below room temperature. As we seek to keep bio-printing at room temperatures, the variables $\boldsymbol{x_1, x_2,}$ and $\boldsymbol{x_3}$ represent the predictor-space for $\boldsymbol{C_{bath}}$ (concentration of dissolved Pluronic polymer) tested for purpose of optimization via the HML framework. Data for the phase diagram is taken from Gioffredi *et al*. Procedia CIRP 2016 (C) shows a CAD file of a circular tube and the resulting alginate prints at $\boldsymbol{C_{bath}} = \{\boldsymbol{20\%, 25\%, 30\%, 35\%}\}$ w/v F-127. Low $\boldsymbol{C_{bath}}$(x2) results in thermally unstable bath due to the proximity to the sol-gel transition. High $\boldsymbol{C_{bath}}$ (x3) is too stiff for realizable prints and is likely the product of increased yield stress due to higher concentration of dissolved polymer. $\boldsymbol{C_{bath}}$ is both a function of yield-stress and sol-gel transition.

3.3.1.1. Materials and methods for support bath optimization

Pluronic Bath Solution Preparation. First, 50 mL stocks of Pluronic F-127 (Sigma) at various w/v% concentrations were created. F-127 was dissolved in $CaCl_2$ (in DI-H20) at 4ºC at the following w/v% concentrations: 20% w/v, 25% w/v, 30% w/v and 35% w/v. Calcium chloride was varied at 1 mM, 5.5 mM, 8 mM, and 11 mM for each Pluronic concentration. Dissolved Pluronic stocks were made in a 4ºC cold room and left to dissolve overnight. Once the solutions were fully dissolved (about 12 hours), they were clear in appearance. Then, 5-mLs of each were then pipetted into a series of 6-well petri dishes to be used as support material for FRESH printing. The Pluronic bath solutions were then brought to room temperature which allowed them to gel. The baths can also be placed in a 35 ºC incubator for faster gelling transition. Once gelled, a FRESH 3D bioprinter was loaded with 4% w/v alginate in DI-H20. Alginate constructs were then printed into the gelled Pluronic bath containing calcium chloride cross-linker. Various print settings were tested on the different bath solutions, keeping the alginate ink constant. During printing, the entire system was kept at room temperature. Once the printing was complete, printed constructs were imaged and then released at 4 ºC. To improve the release, prints were diluted with cold $CaCl_2$ in DI-H20 (to help prevent alginate osmotic swelling). Released prints were then ready for additional assessment for dimensional similarity to the original CAD file.

3.3.1.2. Pilot results for utilization in HML modeling

A pilot dataset was created to model the optimization of Pluronic F-127 as an optically transparent bath solution for FRESH 3D bioprinting using HML. Multiple w/v% of Pluronic F-127 were tested in combination with variation in calcium chloride cross-linker, and variation of different printed shapes. Furthermore, a range of print-predictors (Q and v_T) were tested. We found that 20% w/v

Pluronic was too unstable to be used for prints, which is unsurprising given the bath concentration exists on the border of the sol-gel transition for F-127 at room temperature. Consequently, we found that 35% w/v baths were unable to maintain Bingham plastic behavior, resulting in poor-quality prints. We suspect that the yield-stress of the bath (which is also temperature dependent) is too high at 35% w/v for high-fidelity printing at the print speeds used. Higher print speeds are likely necessary for higher Pluronic bath concentrations. Finally, we noticed that 11 mM calcium chloride (used in alginate-FRESH printing) occluded the nozzle, preventing the print from emerging. This phenomenon was mitigated by printing at lower cross-linker concentrations compared to what was used in the HML-alginate model. Overall, we observed a complex interplay between print translational speed, flow rate, cross-linker, bath concentration, and bath yield stress that makes the system a great candidate for HML modeling and optimization.

3.3.1.3. Discussion of HML optimization of new support bath

The implementation of HML optimization for printing alginate in an optically clear Pluronic bath solution will now be discussed. The underlying assumption is that some predictors remain constant between the source model (HML-alginate) and the target model (HML-alginate + Pluronic). For example, we can use a consistent printing system or consistent ink between models. We believe it is a reasonable assumption, for example, that 4% alginate ink experiences shear-thinning and relaxation times independent of the bath. Consequently, the bath solution is expected to behave as a Bingham plastic similar to gelatin microparticles, a property that can be updated in the HML middle layer via the known relationship of storage and loss modulus. By harnessing known aspects of the physical space, statistical inference is allowed to do what it does best: discover the remaining connections in the data that drive the system response. It is up to the experimenter to interpret these

discovered relationships, which in theory (if over- or under-fitting is appropriately avoided through cross-validation) describe complex physical processes and material interactions. Unfortunately, due to recent events and laboratory shutdowns, we were unable to fully gather the experiments necessary to fully model the new Pluronic system. However, we are able to show promising pilot results of print-space necessary to model alginate FRESH printing in an optically clear bath solution. Completing the experimental space and testing the HML-optimized predictions is an exciting and motivating future step.

3.3.2. Applications to cell-laden inks

The system response in HML need not be dimensional fidelity but can instead contain a *functional* output. For example, a measurable response of 3D printed cells could be their viability[27] (calcein AM stain), alignment (orientation-parameter),[28] or attachment (E-cadherin),[29] to name a few. Thus, alginate inks laden with cells would be highly relevant as an extension of this work in which our known alginate system can be optimized alongside a functional output relating to the properties of cells. Furthermore, the introduction of constraints and visualization of key design-space areas from a process map, as has been done for metals,[30] could be useful for more complex printing processes that require the optimization of a multivariate space. Consider, for example, the pressing question whether 3D printing viable tissue constructs are feasible given high shear rates in the small, capillary-sized extrusion nozzles. High shear rates could negatively affect cells or other biomaterial inks during printing. It is well known that high shear forces can cause undesired coagulation,[31] reduce cell viability and differentiation, or even rupture cells. Functional cell outcomes add attainable constraints to the multi-objective optimization problem as it narrows the design-space for the designer as shown in Figure 3.4. Table 3.2 shows a case study regarding the

optimization of 3D printed small line widths using a cell-laden alginate ink requiring theoretical cell rupture constraints and mechanical failure conditions.[32]

Table 3.2. Case study constraints

Smallest feature size	At most 80 μm
Dimensional Precision	At least 95%
Critical shear stress	Less than 500 Pa

By mapping these constraints onto a Q-V space as in Figure 3.4, the user can direct machine-driven optimization parameters. In Figure 3.4, normalized user-defined print settings are plotted against measured effect on the cross-sectional area of the printed feature. Measured cross-sectional area from the data is shown. Print settings (Q and v_T) are plotted with measured cross-sectional area for three concentrations of C_{ink} printed with an 80 μm nozzle reflecting the ink material interaction with print settings. A red line is marked showing the critical shear stress beyond which cells printed into the system are expected to rupture. Furthermore, a line representing a cross-sectional area ~6400 μm^2 (derived from a nozzle diameter of 80 μm) is plotted with regions showing the location of 95% precision (or 5% deviation from 80 μm). Finally, the HML-alginate model is cross-referenced from Figure 2.11 to identify error in chosen print parameters. Taken together, the user can begin to identify the print settings (Q, v_T, C_{ink}) for $D_{nozzle} = 80\ \mu m$ that satisfy the design constraints from Table 2.2. Future work will involve updated error predictions from viscosity measurements of cell-laden alginate inks.

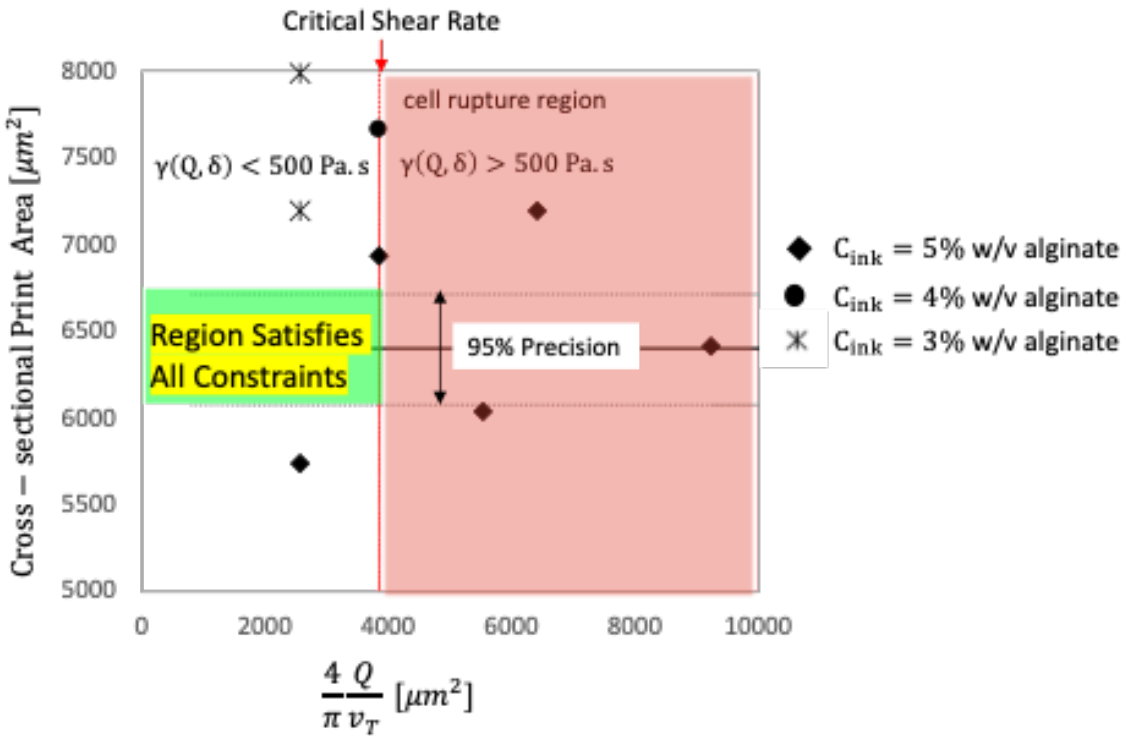

Figure 3.4. A pilot Q-V process map was generated using theoretical cell rupture conditions (72% strain, 8.7 **µ**N force, on a 30 **µ**m diameter cell). This shows the constrained design space for printed cell lines. From this map $\boldsymbol{Q}$, $\mathbf{v_T}$, and $\mathbf{C_{ink}}$ can be chosen for achieving dimensional accuracy with $\mathbf{D_{nozzle} = 80\ \mu m}$ based on material feedback from the data. Figure 2.9 can be cross-referenced to give estimates on HML-predicted dimensional error for the chosen print settings.

Preliminary work involving HML for modeling functional cellular outputs was conducted in which both shear stress from bioprinting and functionalization of alginate with RGB groups were to be correlated with cell viability and cell alignment on printed constructs. Alginate polymers align in the direction of shear and retain this alignment after cross-linking. Since it has been shown that some cells (such as skeletal cells) align to small micro-grooves in their environment[64], we aimed to devise an HML model in which shear rate in bioprinting nozzles could be used to map cell viability and alignment to the printed construct. Future work in this area is both exciting and necessary for the fidelity (both dimensionally and functionally) of future cell-laden prints.

3.4. Conclusion

The benefits of model interpretability from an embedded physical middle layer become evident when aspiring to greater model generalizability. Initially, an HML-alginate model was created from a streamlined experiment space of 48 prints in which the effect of print predictors on error was systematically sampled and modeled to predict high-fidelity shapes in alginate. While random forest and neural networks were powerful statistical inference methods for predicting print outcomes, the supervised middle layer combined with regularization revealed the dominant and interpretable physical variables necessary to describe high-fidelity printing. These dominant physical variables, namely combinations of pressure, viscosity, and shear-rates became integral for transferring the HML-alginate model to unseen bio-inks. By updating the HML-alginate model with collagen-specific latent variable estimations, we were able to show that variables in the embedded middle layer can be leveraged to predict parallel 3D printing systems with good accuracy. As a result, we showed a proof-of-concept of one of the first attempts at transfer learning in 3D bioprinting. The decomposition of the print-space into its dominant physical parameters also helped to identify modes of transfer to parallel systems for rapid optimization of new materials, such when choosing the experimental predictors for modeling a new, optically transparent support bath solution for FRESH. Finally, we can use HML-predicted error bias to create process maps that can help identify the trade-offs of choosing from a range of acceptable process parameters on the desired output. For example, the best shear-rates for printing with 80 μm dimensional precision may be detrimental to the survival or function of some cell types and an intermediate value must be chosen.

Section 1 Appendix

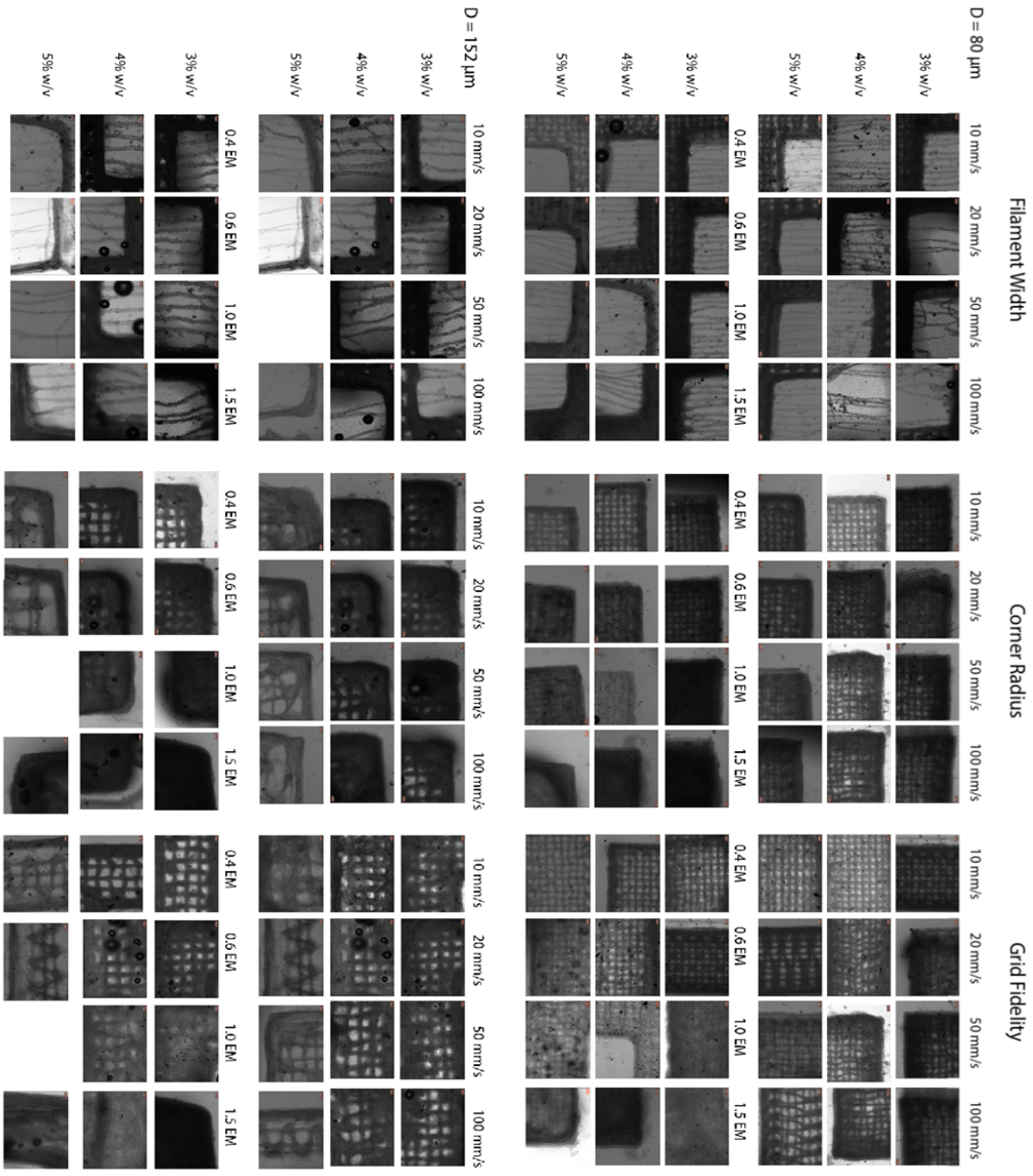

Figure 3.5. Above dataset that was generated to study the fidelity of prints in terms of the HML bottom layer variables: $\mathbf{D_{nozzle}}$ **at [80 µm, 152 µm]**, $\mathbf{v_T}$ [10 mm/s – 100 mm/s], $\boldsymbol{Q}$ [0.4 EM – 1.5 EM], and $\mathbf{C_{ink}}$ [3% w/v, 4%w/v, 5% w/v]. The EM is a multiplicative factor and the conversion from EM to flow rate is Q = 0.1 $\boldsymbol{\mu L/s}$ ×EM. Nozzle length could not be individually tested for this dataset and is represented in middle layer equations (see Table 3.3).

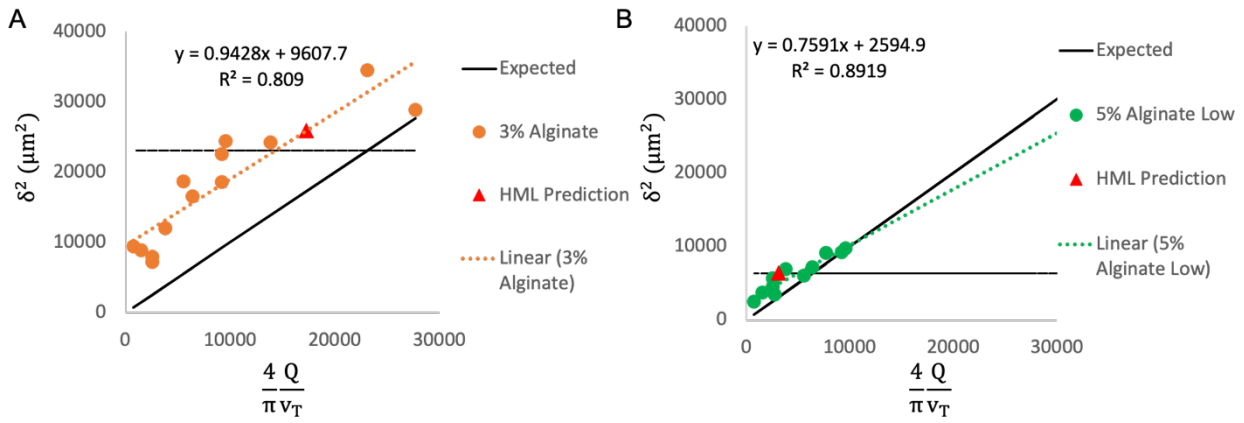

Figure 3.6. $\boldsymbol{\partial^2}$ vs $(\boldsymbol{Q}, \boldsymbol{v_T})$ plot shows the expected and measured relationship between cross-sectional area of printed material and print parameters. Cross-sectional area of each linewidth is plotted as a function of bottom-layer printer variables via the equation $\boldsymbol{A_x = {}^{4}/_{\pi}\, {}^{Q}/_{v_T} = \partial^2}$ where linewidths are assumed to be measured from cylindrical print fibers and thus are $\boldsymbol{\sqrt{A_x} = \delta}$. The location on the y-axis that corresponds to 152 $\boldsymbol{\mu m}$ has been shown for convenience. The QV space provides a way to look at the linewidth data measured from prints created from multiple print variables. (A) shows the dataset that attempts to make 152 **µm** diameter linewidths using 3% alginate. Dashed line shows the expected theoretical scaling of cross-sectional printed area with print parameters. However, the data points measured at various $(\boldsymbol{Q}, \boldsymbol{v_T})$ instead fall just above this expected line (orange dashed line). Red triangle represents the measured linewidth from print settings derived from the HML-random forest optimized prediction. (B) The dataset that attempts to make 80 $\boldsymbol{\mu m}$ diameter linewidths is shown in green compared to the expected relationship (black). The location on the y-axis that corresponds to 80 $\boldsymbol{\mu m}$ has been shown for convenience. The HML predictions for optimized print settings is shown to fall within 10% of 80 **µm**. The following variable relationships are used for printability maps for 3%, 152 **µm** error prediction, and 5%, 80 **µm** error prediction: $\boldsymbol{\delta^2_{152\ \mu m}(C_{ink} = 3\%\ w/v) = 0.9428\left({}^{4}/_{\pi}\, {}^{Q}/_{v_T}\right) + 9607.7}$ and $\boldsymbol{\delta^2_{80\ \mu m}(C_{ink} = 5\%\ w/v) = 0.7591\left({}^{4}/_{\pi}\, {}^{Q}/_{v_T}\right) + 2594.9}$.

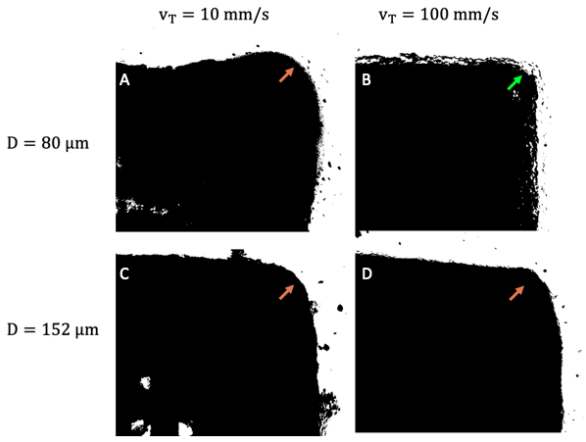

Figure 3.7. A macro-view of shape fidelity was analyzed by measuring the corner radius of the box window-frame at various printing conditions. At a fixed nozzle diameter of 80 **µm**, (A) and (B) the corner radius is shown to improve at higher printing speeds. Increasing print speed similarly improves the corner radius at a fixed, slightly higher nozzle diameter of 152 **µm** (C) and (D). In general, higher print speeds and smaller nozzles gave better control over alginate during directional changes in the nozzle.

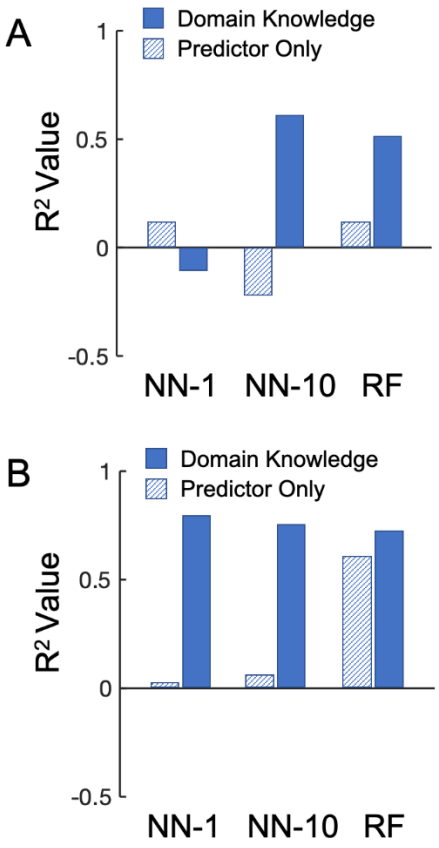

Figure 3.8. A schematic showing various modeling strategies with and without a middle layer for (A) linewidth features and (B) corner features. NN-1 represents a two-layer feed forward neural network with one neuron and NN-10 shows the same structure for 10 neurons. For each feature, we found modeling is improved with the addition of a middle layer. Random Forest (RF) and NN-10 also show promising model fits for linewidths. Neural networks and RF score reasonably well for corner predictions with the addition of the middle layer compared to HML in the text. HML was ultimately chosen in this work for its ability to perform well modeling both print outcomes. Furthermore, middle layer physical interpretation combined with the print scoring equations (Eq. 2.6 and Eq. 2.7) aid in the development of downstream analysis such as 3D bioprinter process maps for alginate and bridging to new material systems.

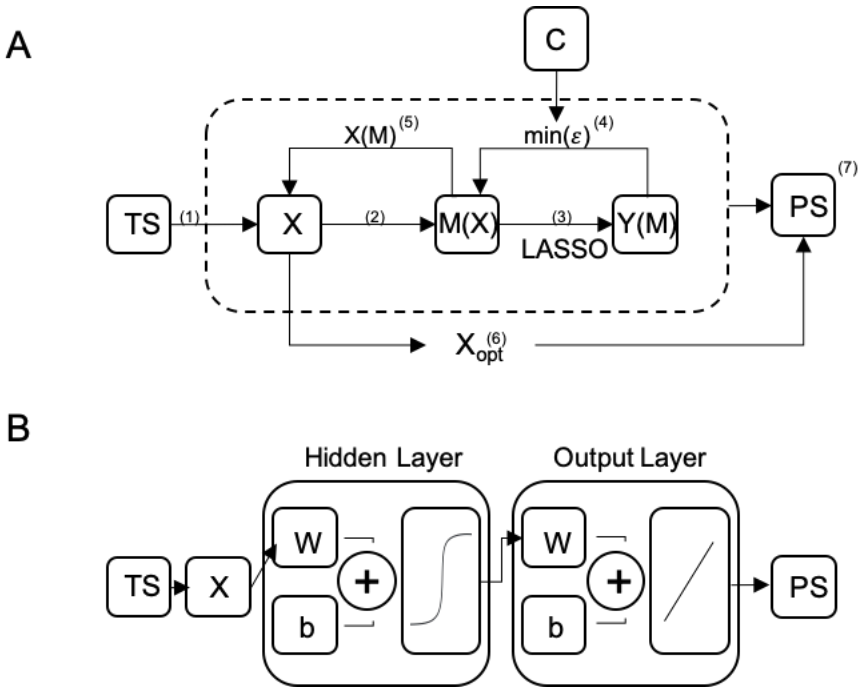

Figure 3.9. (A) HML fitting schematic and the optimization process. (1) TS = training set is generated using a combination of the set of X predictors. (2) X predictors are combined to parameterize a middle layer, M(X), consisting of known physical equations (see Table 3.3). (3) Statistical inference methods (LASSO in this analysis) discover combinations of middle layer variables that fit and further parameterize the system response Y(M) making it more specific to the system at hand. (4). Error, $\boldsymbol{\varepsilon}$ defined as $\boldsymbol{\varepsilon = \delta_{obs} - \delta_{CAD}}$ is minimized using constraints C. (5) Parameterization equations from part 2 are used to back-calculate the optimal set of predictors Xopt. (6) New experiments using the optimized predictors from Xopt are done to test the model predictions. (7) PS = predicted set which consists of optimized prints. (B) A schematic of a shallow neural network with a 2-layer feed-forward network. The input layer has four predictors and between 1 and 10 neurons in the hidden layer. The network is trained with a Levenberg-Marquardt backpropagation algorithm in accordance with the Matlab Neural Network Fitting tool.

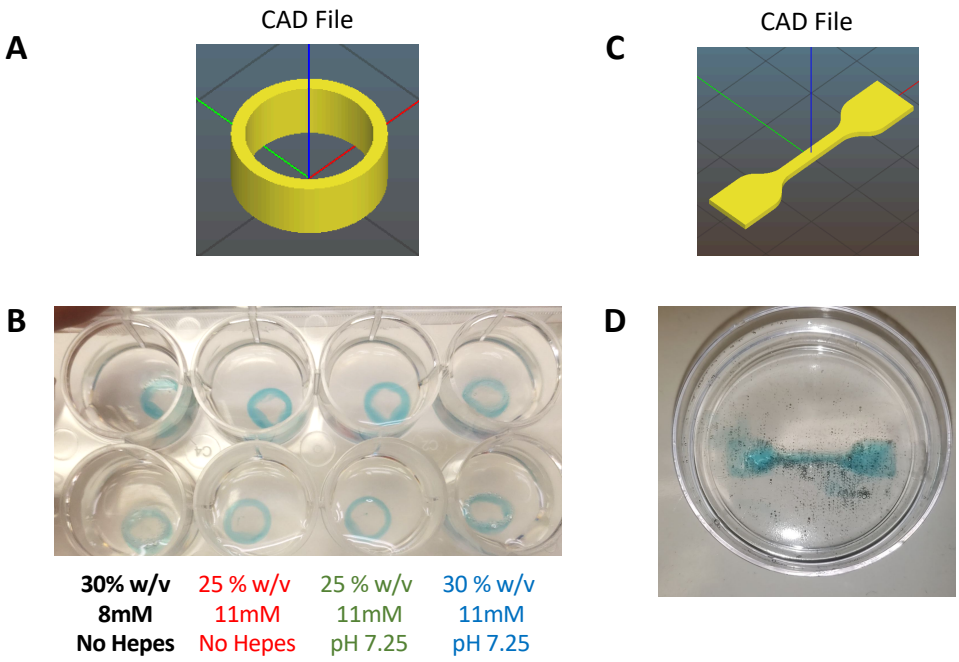

Figure 3.10. Additional shapes and experimental conditions explored for printing alginate with Pluronic F-127 bath solution. Since pH plays a large role in both micellular concentration and packing density and in the cross-linking of future bioinks (such as collagen), hepes (pH = 7.25) was tested as a diluent in comparison to pure DI-H20. For an HML model, each of these prints would be scored quantitatively according to the expected CAD dimensions. Once scored, the print score and subsequent process conditions that led to this score would be inputted into the model. A similar model would be used as HML-alginate, with the exception of new predictor variables for concentration of bath $\boldsymbol{C_{bath}}$ and concentration of cross-linker, $\boldsymbol{C_{cross}}$. The relevance of new middle layer variables such the relationship between $\boldsymbol{C_{cross}}$, bath yield stress, and $\boldsymbol{v_T}$ in predicting quality prints would be explored.

Table 3.3. Hierarchical machine learning middle layer equations.

	Variable	Parameter	Description	Equation or Range
Predictors	x_1	C_{ink}	Ink concentration	[0.03,0.04, 0.05] w/v% in DI water
	x_2	D_{nozzle}	Nozzle diameter	80 µm, 152 µm
	x_3	Q	Normalized Flow Rate (Extrusion Multiplier)	[0.4, 0.6, 1.0, 1.5]× 0.1 µL/s
	x_4	v_T	Nozzle speed	[10, 20, 30, 50, 100] mm/s
	x_5 *	C_{CaCl_2}	Alginate cross-linker concentration	[11 mM, 5.5 mM]*
Hidden Layer	h_1	$\eta_{ink}(C_{ink}, \dot{\gamma})$	Ink viscosity [Pa · s]	$\eta_{ink} = \eta_\infty + \left(\frac{\eta_o - \eta_\infty}{1 + 0.02\dot{\gamma}^n}\right)$
		$\eta_\infty(C_{ink})$	Ink viscosity (infinite shear)	$\eta_\infty = 20.11C_{ink} - 0.4483$
		$\eta_o(C_{ink})$	Ink viscosity (zero shear)	$\eta_o = 67.892C_{ink} - 1.8237$
		$n\ (C_{ink})$	Power law exponent	$n = 4.08C_{ink} + 0.5919$
	h_2	$\gamma(Q, D_{nozzle})$	Effective shear rate [1/s]	$\gamma = \frac{Q/A_x}{D_{nozzle}/2} = \frac{8Q}{\pi D_{nozzle}^3}$
		$A_x(D_{nozzle})$	Cross-sectional area [µm^2]	$A_x = \pi D_{needle}^2/4$
	h_3	$P(Q, D_{nozzle})$	Pressure [Pa]	$P = 32\eta_{ink}L\left(\frac{3n+1}{4n}\right)\frac{4Q}{\pi D_{nozzle}^4}$
		L	Needle Length [mm]	

				L = 25.5 mm for $D_{nozzle}(152\ \mu m)$ L = 6.5 mm for $D_{nozzle}\ (80\ \mu m)$
	h_4	$\xi\ (Q, v_T)$	Proportionality [μm]	$\xi = \sqrt{\frac{4Q}{\pi v_T}}$
Response	y_1	$P_{lines}(h_1, \dots, h_4)$	Print Score [0-100]% (measured line width [μm])	$P_{lines} = 1 - \frac{\lvert\varepsilon\rvert}{\delta_{exp}} \times 100\%$ $\varepsilon = \delta_{obs} - \delta_{exp}$
	y_2	$P_{radius}(h_1, \dots, h_4)$	Print Score [0-100]% (measured corner radius φ)	$P_{radius} = 1 - \frac{\lvert\varepsilon\rvert}{\varphi_{exp}} \times 100\%$ $\varepsilon = \varphi_{obs} - \varphi_{exp}$

Table 3.4. Coefficients from Eq. 2.6 and Eq. 2.7 in the text

Parameterization	Coefficients (Lines)	Coefficients (Corners)
γ	2.170	2.170
η_{ink}	1.210	1.212
P	-2.680	-2.678
ξ	1.630	1.631
γP	2.230	0
$\gamma\xi$	-0.869	2.232
$\eta_{ink}\xi$	-0.500	-0.870
$P\xi$	1.010	0
γ^2	-2.230	-0.498
${\eta_{ink}}^2$	-0.470	1.010
P^2	0.250	-2.232
ξ^2	-0.610	-0.614

Section 1 References

(1) Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J. T. Deep Learning for Healthcare: Review, Opportunities and Challenges. *Brief. Bioinform.* **2018**, *19* (6), 1236–1246. https://doi.org/10.1093/bib/bbx044.

(2) Xu, C.; Jackson, S. A. Machine Learning and Complex Biological Data. *Genome Biol.* **2019**, *20* (1), 76. https://doi.org/10.1186/s13059-019-1689-0.

(3) Patkar, N.; Shaikh, A. F.; Kakirde, C.; Nathany, S.; Ramesh, H.; Bhanshe, P.; Joshi, S.; Chaudhary, S.; Kannan, S.; Khizer, S. H.; Chatterjee, G.; Tembhare, P.; Shetty, D.; Gokarn, A.; Punatkar, S.; Bonda, A.; Nayak, L.; Jain, H.; Khattry, N.; Bagal, B.; Sengar, M.; Gujral, S.; Subramanian, P. A Novel Machine-Learning-Derived Genetic Score Correlates with Measurable Residual Disease and Is Highly Predictive of Outcome in Acute Myeloid Leukemia with Mutated NPM1. *Blood Cancer J.* **2019**, *9* (10), 1–4. https://doi.org/10.1038/s41408-019-0244-2.

(4) Chicco, D.; Jurman, G. Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone. *BMC Med. Inform. Decis. Mak.* **2020**, *20* (1), 16. https://doi.org/10.1186/s12911-020-1023-5.

(5) Chin, J. X.; Chung, B. K.-S.; Lee, D.-Y. Codon Optimization OnLine (COOL): A Web-Based Multi-Objective Optimization Platform for Synthetic Gene Design. *Bioinformatics* **2014**, *30* (15), 2210–2212. https://doi.org/10.1093/bioinformatics/btu192.

(6) Plant, D.; Barton, A. Machine Learning in Precision Medicine: Lessons to Learn. *Nat. Rev. Rheumatol.* **2021**, *17* (1), 5–6. https://doi.org/10.1038/s41584-020-00538-2.

(7) Wilkinson, J.; Arnold, K. F.; Murray, E. J.; Smeden, M. van; Carr, K.; Sippy, R.; Kamps, M. de; Beam, A.; Konigorski, S.; Lippert, C.; Gilthorpe, M. S.; Tennant, P. W. G. Time to Reality Check the Promises of Machine Learning-Powered Precision Medicine. *Lancet Digit. Health* **2020**, *2* (12), e677–e680. https://doi.org/10.1016/S2589-7500(20)30200-4.

(8) Katsanis, S. H.; Javitt, G.; Hudson, K. PUBLIC HEALTH: A Case Study of Personalized Medicine. *Science* **2008**, *320* (5872), 53–54. https://doi.org/10.1126/science.1156604.

(9) Zhang, Y.; Ling, C. A Strategy to Apply Machine Learning to Small Datasets in Materials Science. *Npj Comput. Mater.* **2018**, *4* (1), 1–8. https://doi.org/10.1038/s41524-018-0081-z.

(10) Knudde, N.; Raes, W.; Bruycker, J. D.; Dhaene, T.; Stevens, N. Data-Efficient Gaussian Process Regression for Accurate Visible Light Positioning. *IEEE Commun. Lett.* **2020**, *24* (8), 1705–1709. https://doi.org/10.1109/LCOMM.2020.2990950.

(11) Oniśko, A.; Druzdzel, M. J.; Wasyluk, H. Learning Bayesian Network Parameters from Small Data Sets: Application of Noisy-OR Gates. *Int. J. Approx. Reason.* **2001**, *27* (2), 165–182. https://doi.org/10.1016/S0888-613X(01)00039-1.

(12) Bromberg, F.; Margaritis, D. Improving the Reliability of Causal Discovery from Small Data Sets Using the Argumentation Framework. *Comput. Sci. Tech. Rep.* **2007**.

(13) Childs, C. M.; Washburn, N. R. Embedding Domain Knowledge for Machine Learning of Complex Material Systems. *MRS Commun.* **2019**, *9* (2). https://doi.org/10.1557/mrc.2019.90.

(14) Oberman Albert; Myers Allen R.; Karunas Thomas M.; Epstein Frederick H. Heart Size of Adults in a Natural Population-Tecumseh, Michigan. *Circulation* **1967**, *35* (4), 724–733. https://doi.org/10.1161/01.CIR.35.4.724.

(15) iRepertiore | Pioneers in Sequencing the Immune Adaptome https://irepertoire.com/ (accessed Feb 17, 2021).

(16) The Adaptome as Biomarker for Assessing Cancer Immunity and Immunotherapy | SpringerLink https://link.springer.com/protocol/10.1007/978-1-4939-9773-2_17 (accessed Feb 17, 2021).

(17) Stehlik, J.; Edwards, L. B.; Kucheryavaya, A. Y.; Benden, C.; Christie, J. D.; Dobbels, F.; Kirk, R.; Rahmel, A. O.; Hertz, M. I. The Registry of the International Society for Heart and Lung Transplantation: Twenty-Eighth Adult Heart Transplant Report—2011. *J. Heart Lung Transplant.* **2011**, *30* (10), 1078–1094. https://doi.org/10.1016/j.healun.2011.08.003.

(18) UNOS Data and Transplant Statistics | Organ Donation Data https://unos.org/data/ (accessed Apr 15, 2019).

(19) McCue, T. 3D Printing Stock Bubble? $10.8 Billion By 2021 https://www.forbes.com/sites/tjmccue/2013/12/30/3d-printing-stock-bubble-10-8-billion-by-2021/#6f1d4385bc06 (accessed Apr 15, 2019).

(20) Patel, P. The Path of Printed Body Parts. *ACS Cent. Sci.* **2016**, *2* (9), 581–583. https://doi.org/10.1021/acscentsci.6b00269.

(21) Nguyen, D. G.; Funk, J.; Robbins, J. B.; Crogan-Grundy, C.; Singer, T.; Roth, A. B. Bioprinted 3D Primary Liver Tissues Allow Assessment of Organ-Level Response to Clinical Drug Induced Toxicity In Vitro. *PLoS One* **2016**, *11* (7). https://doi.org/doi: 10.1371/journal.pone.0158674.

(22) Akmal, J. S.; Salmi, M.; Mäkitie, A.; Björkstrand, R.; Partanen, J. Implementation of Industrial Additive Manufacturing: Intelligent Implants and Drug Delivery Systems. *J. Funct. Biomater.* **2018**, *9* (3), 41. https://doi.org/10.3390/jfb9030041.

(23) Tapia, G.; Elwany, A. H.; Sang, H. Prediction of Porosity in Metal-Based Additive Manufacturing Using Spatial Gaussian Process Models. *Addit. Manuf.* **2016**, *12*, 282–290. https://doi.org/10.1016/j.addma.2016.05.009.

(24) Abdollahi, S.; Davis, A.; Miller, J. H.; Feinberg, A. W. Expert-Guided Optimization for 3D Printing of Soft and Liquid Materials. *PLoS One* **2018**, *13* (4). https://doi.org/10.1371/journal.pone.0194890.

(25) Menon, A.; Póczos, B.; Feinberg, A. W.; Washburn, N. R. Optimization of Silicone 3D Printing with Hierarchical Machine Learning. *3D Print. Addit. Manuf.* **2019**, *6* (4), 181–189. https://doi.org/10.1089/3dp.2018.0088.

(26) Yokoyama, F.; Achife, E. C.; Matsuoka, M.; Shimamura, K.; Yamashita, Y.; Monobe, K. Morphology of Oriented Calcium Alginate Gels Obtained by the Flow-Gelation Method. *Polymer* **1991**, *32* (16), 2911–2916. https://doi.org/10.1016/0032-3861(91)90186-M.

(27) Yan, Q.; Dong, H.; Su, J.; Han, J.; Song, B.; Wei, Q.; Shi, Y. A Review of 3D Printing Technology for Medical Applications. *Engineering* **2018**, *4* (5), 729–742. https://doi.org/10.1016/j.eng.2018.07.021.

(28) Goh, G. D.; Sing, S. L.; Yeong, W. Y. A Review on Machine Learning in 3D Printing: Applications, Potential, and Challenges. *Artif. Intell. Rev.* **2020**. https://doi.org/10.1007/s10462-020-09876-9.

(29) Chen, P.-H. C.; Liu, Y.; Peng, L. How to Develop Machine Learning Models for Healthcare. *Nat. Mater.* **2019**, *18* (5), 410–414. https://doi.org/10.1038/s41563-019-0345-0.

(30) Her, H.-L.; Wu, Y.-W. A Pan-Genome-Based Machine Learning Approach for Predicting Antimicrobial Resistance Activities of the Escherichia Coli Strains. *Bioinformatics* **2018**, *34* (13), i89–i95. https://doi.org/10.1093/bioinformatics/bty276.

(31) Clymer, D. R.; Long, J.; Latona, C.; Akhavan, S.; LeDuc, P.; Cagan, J. Applying Machine Learning Methods Toward Classification Based on Small Datasets: Application to Shoulder Labral Tears. *J. Eng. Sci. Med. Diagn. Ther.* **2020**, *3* (1), 011004. https://doi.org/10.1115/1.4044645.

(32) Shaikhina, T.; Lowe, D.; Daga, S.; Briggs, D.; Higgins, R.; Khovanova, N. Machine Learning for Predictive Modelling Based on Small Data in Biomedical Engineering. *IFAC-Pap.* **2015**, *48* (20), 469–474. https://doi.org/10.1016/j.ifacol.2015.10.185.

(33) Menon, A.; Gupta, C.; M. Perkins, K.; L. DeCost, B.; Budwal, N.; T. Rios, R.; Zhang, K.; Póczos, B.; R. Washburn, N. Elucidating Multi-Physics Interactions in Suspensions for the Design of Polymeric Dispersants: A Hierarchical Machine Learning Approach. *Mol. Syst. Des. Eng.* **2017**, *2* (3), 263–273. https://doi.org/10.1039/C7ME00027H.

(34) Hinton, T. J.; Jallerat, Q.; Palchesko, R. N.; Park, J. H.; Grodzicki, M. S.; Shue, H.-J.; Ramadan, M. H.; Hudson, A. R.; Feinberg, A. W. Three-Dimensional Printing of Complex Biological Structures by Freeform Reversible Embedding of Suspended Hydrogels. *Sci. Adv.* **2015**, *1* (9). https://doi.org/10.1126/sciadv.1500758.

(35) Lee, A.; Hudson, A. R.; Shiwarski, D. J.; Tashman, J. W.; Hinton, T. J.; Yerneni, S.; Bliley, J. M.; Campbell, P. G.; Feinberg, A. W. 3D Bioprinting of Collagen to Rebuild Components of the Human Heart. *Science* **2019**, *365* (6452), 482–487. https://doi.org/10.1126/science.aav9051.

(36) Farokhi, M.; Shariatzadeh, F. J.; Solouk, A.; Mirzadeh, H. Alginate Based Scaffolds for Cartilage Tissue Engineering: A Review. *Int. J. Polym. Mater. Polym. Biomater.* **2019**. https://doi.org/10.1080/00914037.2018.1562924.

(37) Murphy, S. V.; Atala, A. 3D Bioprinting of Tissues and Organs. *Nat. Biotechnol.* **2014**, *32* (8), 773–785. https://doi.org/10.1038/nbt.2958.

(38) Cross, M. M. Relation between Viscoelasticity and Shear-Thinning Behaviour in Liquids. *Rheol. Acta* **1979**, *18* (5), 609–614. https://doi.org/10.1007/BF01520357.

(39) Clymer, D. R.; Cagen, J.; Beuth, J. Power-Velocty Process Design Charts for Powder Bed Additive Manufacturing. *J. Mech. Des.* **2017**, *139* (10). https://doi.org/10.1115/1.4037302.

(40) Yu, C.; Jiang, J. A Perspective on Using Machine Learning in 3D Bioprinting. **2020**, *6*, 1–8. https://doi.org/10.18063/ijb.v6i1.253.

(41) An, J.; Teoh, J. E. M.; Suntornnond, R.; Chua, C. K. Design and 3D Printing of Scaffolds and Tissues. *Engineering* **2015**, *1* (2), 261–268. https://doi.org/10.15302/J-ENG-2015061.

(42) Tabriz, A. G.; Hermida, M. A.; Leslie, N. R.; Shu, W. Three-Dimensional Bioprinting of Complex Cell Laden Alginate Hydrogel Structures. *Biofabrication* **2015**, *7* (4), 045012. https://doi.org/10.1088/1758-5090/7/4/045012.

(43) Stevens, K. R.; Einerson, N. J.; Burmania, J. A.; Kao W. J. In Vivo Biocompatibility of Gelatin-Based Hydrogels and Interpenetrating Networks. *J Biomater Sci Polym Ed* **2002**, *12* (12), 1353–1366.

(44) Stryhn, H.; Christensen, J. The Analysis—Hierarchical Models: Past, Present and Future. *Prev. Vet. Med.* **2014**, *113* (3), 304–312. https://doi.org/10.1016/j.prevetmed.2013.10.001.

(45) Yao, X.; Moon, S. K.; Bi, G. A Hybrid Machine Learning Approach for Additive Manufacturing Design Feature Recommendation. *Rapid Prototyp. J.* **2017**, *23* (6), 983–997. https://doi.org/10.1108/RPJ-03-2016-0041.

(46) O'Bryan, C. S.; Bhattacharjee, T.; Niemi, S. R.; Balachandar, S.; Baldwin, N.; Ellison, S. T.; Taylor, C. R.; Sawyer, W. G.; Angelini, T. E. Three-Dimensional Printing with Sacrificial Materials for Soft Matter Manufacturing. *MRS Bull.* **2017**, *42* (8), 571–577. https://doi.org/10.1557/mrs.2017.167.

(47) Roger, S.; Sang, Y. Y. C.; Bee, A.; Perzynski, R.; Di Meglio, J. M.; Ponton, A. Structural and Multi-Scale Rheophysical Investigation of Diphasic Magneto-Sensitive Materials Based on Biopolymers. *Eur. Phys. J. E* **2015**, *38* (8). https://doi.org/10.1140/epje/i2015-15088-1.

(48) Suntornnond, R.; Tan, E.; An, J.; Chua, C. A Mathematical Model on the Resolution of Extrusion Bioprinting for the Development of New Bioinks. *Materials* **2016**, *9* (9), 756. https://doi.org/10.3390/ma9090756.

(49) Oliphant, J. E.; Peterson, P. *SciPy: Open Source Scientific Tools for Python*; 2001.

(50) Li, H.; Liu, S.; Li, L. Rheological Study on 3D Printability of Alginate Hydrogel and Effect of Graphene Oxide. *Int. J. Bioprinting* **2016**, *2* (2), 54–66. https://doi.org/10.18063/IJB.2016.02.007.

(51) Brzezińska, M.; Szparaga, G. The Effect Of Sodium Alginate Concentration On The Rheological Parameters Of Spinning Solutions. *Autex Res. J.* **2015**, *15* (2), 123–126. https://doi.org/10.2478/aut-2014-0044.

(52) Torrey, L.; Shavlik, J. Transfer Learning www.igi-global.com/chapter/transfer-learning/36988 (accessed Feb 16, 2021). https://doi.org/10.4018/978-1-60566-766-9.ch011.

(53) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A Survey of Transfer Learning. *J. Big Data* **2016**, *3* (1), 9. https://doi.org/10.1186/s40537-016-0043-6.

(54) Cheng, L.; Tsung, F.; Wang, A. A Statistical Transfer Learning Perspective for Modeling Shape Deviations in Additive Manufacturing. *IEEE Robot. Autom. Lett.* **2017**, *2* (4), 1988–1993. https://doi.org/10.1109/LRA.2017.2713238.

(55) Duraipandy, N.; Lakra, R.; Srivatsan, K. V.; Ramamoorthy, U.; Korrapati, P. S.; Kiran, M. S. Plumbagin Caged Silver Nanoparticle Stabilized Collagen Scaffold for Wound Dressing. *J. Mater. Chem. B* **2015**, *3* (7), 1415–1425. https://doi.org/10.1039/C4TB01791A.

(56) Djaković, L. J.; Sovilj, V.; Milošević, S. Rheological Behaviour of Thixotropic Starch and Gelatin Gels. *Starch - Stärke* **1990**, *42* (10), 380–385. https://doi.org/10.1002/star.19900421004.

(57) Suntornnond, R.; Tan, E. Y. S.; An, J.; Chua, C. K. A Highly Printable and Biocompatible Hydrogel Composite for Direct Printing of Soft and Perfusable Vasculature-like Structures. *Sci. Rep.* **2017**, *7* (1), 16902. https://doi.org/10.1038/s41598-017-17198-0.

(58) Pluronic F127 Hydrogel Characterization and Biofabrication in Cellularized Constructs for Tissue Engineering Applications | Elsevier Enhanced Reader https://reader.elsevier.com/reader/sd/pii/S2212827115010628?token=978DF10A8BEE4A4B38B4D62B4D6641DC6CA8A0857CA03ADDDE1408156A877D839175FF29F67547D9014B0B63D90DABF0 (accessed Aug 20, 2019). https://doi.org/10.1016/j.procir.2015.11.001.

(59) Chang, R.; Nam, J.; Sun, W. Effects of Dispensing Pressure and Nozzle Diameter on Cell Survival from Solid Freeform Fabrication–Based Direct Cell Writing. *Tissue Eng. Part A* **2008**, *14* (1), 41–48. https://doi.org/10.1089/ten.a.2007.0004.

(60) Davidson, P.; Bigerelle, M.; Bounichane, B.; Giazzon, M.; Anselme, K. Definition of a Simple Statistical Parameter for the Quantification of Orientation in Two Dimensions: Application to Cells on Grooves of Nanometric Depths. *Acta Biomater.* **2010**, *6* (7), 2590–2598. https://doi.org/10.1016/j.actbio.2010.01.038.

(61) Panorchan, P.; Thompson, M. S.; Davis, K. J.; Tseng, Y.; Konstantopoulos, K.; Wirtz, D. Single-Molecule Analysis of Cadherin-Mediated Cell-Cell Adhesion. *J Cell Sci* **2006**, *119* (1), 66–74. https://doi.org/10.1242/jcs.02719.

(62) Alexander, D. E. *Nature's Machines: An Introduction to Organismal Biomechanics*; Elsevier Science & Technology: Saint Louis, UNITED STATES, 2017.

(63) Peeters, E. A. G.; Oomens, C. W. J.; Bouten, C. V. C.; Bader, D. L.; Baaijens, F. P. T. Mechanical and Failure Properties of Single Attached Cells under Compression. *J. Biomech.* **2005**, *38* (8), 1685–1693. https://doi.org/10.1016/j.jbiomech.2004.07.018.

(64) Wang, P.-Y.; Yu, H.-T.; Tsai, W.-B. Modulation of Alignment and Differentiation of Skeletal Myoblasts by Submicron Ridges/Grooves Surface Structure. *Biotechnol. Bioeng.* **2010**, *106* (2), 285–294. https://doi.org/10.1002/bit.22697.

Section 2: Smarter Diagnostics- Merging systems biology with molecular biology: ML methods for leveraging Phase 1 clinical data.

Chapter 4. A Markov model for early prediction of renal cancer response to HCQ/IL-2 treatment and disease monitoring from Phase 1 clinical data

4.1. Introduction

The advent of multiplex PCR and next generation sequencing recently has enabled robust high throughput analysis of the immune repertoire in response to disease and treatment.[1 2] It has been well established that the diversity of the T cell repertoire gradually decreases with age correlating with an increased incidence of cancer.[3 4] In a 30-year longitudinal study of 6 healthy individuals, 10-year sequential TCR Vβ sequencing identified a stable CD4 repertoire, but a reduction in CD8 diversity.[3] This progressive increase in circulating repertoire clonality was validated in a large cohort of pan-cancer patients (n=218) and age matched healthy controls (n=95), which identified an age-specific reduction in repertoire richness (unique clones) and evenness (clone frequency) across both healthy individuals and cancer patients older than 40 years.[4] Intriguingly, despite administration of cytotoxic, lymphodepleting chemotherapies, patients with hematologic and solid tumors experienced an age specific rebound in TCR diversity that was attributed to thymic rebound rather than peripheral TCR clonal expansion.[4] These results demonstrate the dynamic nature of the adaptive immune repertoire, and the necessity to monitor such changes during disease development and treatment.

With some of the earliest reports of cancer immunotherapy response identified in patients with metastatic clear cell renal carcinoma treated with high dose interleukin-2 (IL-2),[5 6 7] several studies have since dissected the adaptive immune response in such cancers and other tumor types. Early reports established the tumor specific clonal expansion associated with tumor

infiltrating lymphocytes (TIL), whose oligoclonal population and longer CDR3 length drastically differed from circulating T cells in the blood.[8] [9] [10] Modern high dimensional multiomics analysis of localized RCC TIL, adjacent normal tissue, and peripheral blood lymphocytes (PBL) from 40 patients classified tumors into immune regulated, immune activated, and immune silent subtypes. Although an increased infiltration of T cells was identified in both immune regulated and activated subtypes, an enrichment of ICOS+, PD-1+, LAG3+, CTLA4+ CD4 and CD8 cells in the immune regulated subtype corresponded with reduced TCR clonality, higher pathologic grade, larger tumors, and poorer overall survival.[11]

Given immunologic control of health and cancer progression, the advent of multimodal cancer therapy, including chemotherapy, radiation, targeted therapy, immunotherapy, and surgery has warranted the need to assess treatment effects on the adaptive immune system. TCR sequencing of PBL before and after nephrectomy in 45 patients with localized RCC found that patients with higher baseline diversity had a reduced neutrophil to lymphocyte ratio (negative prognostic indicator), lower clinical stage, and longer overall survival. Trauma and acute inflammation following surgery resulted in a mobilization of naïve T cells and a reduction in T cells with an exhausted phenotype that corresponded to an increase in peripheral TCR diversity.[12] In a small cohort of patients with RCC treated with and without neoadjuvant stereotactic body radiation (SBRT) prior to nephrectomy, TCR Vβ sequencing of the resected tumor and pre and post SBRT peripheral blood identified an intratumoral and peripheral clonal expansion in SBRT treated patients. This clonal expansion was accompanied by an increase in convergent TCR nucleotides coding for the same amino acid sequence and was present within two weeks of SBRT. This was slightly reduced compared to baseline by the time of nephrectomy (4 weeks)[13] in two

independent cohorts of metastatic RCC, patients with high baseline TCR repertoire diversity had greater survival and response following combined high dose IL-2[14] and autophagy inhibition or anti-VEGF therapy.[15] This strongly suggests a central role for receptor diversity in robust responses to cancer.

Immune checkpoint blockade (ICB), targeting CTLA-4, PD-1, and PD-L1 has transformed oncology practice and now serves as a first line therapy for patients with advanced clear cell RCC.[16] [17] [18] Of course, patient heterogeneity, including tumor and T cell intrinsic factors, limit anti-tumor responses and leave significant patient populations without clinical benefit. In a recent report, 25 patients with metastatic RCC treated with anti-PD-1 (nivolumab) ICB, serial TCR Vα and Vβ sequencing helped to retrospectively identify patient response. Unlike the other metastatic cohorts, baseline TCR diversity was not associated with treatment response. Patients displaying an increase in peripheral TCR clonality at 3 and 6 months on therapy were more likely to respond to therapy and have greater overall survival. Clonal expansion of circulating T cells in response to presumed tumor specific antigens resulted in an increase in the number of shared tumor and peripheral blood TCRs and elevated CTL gene signatures (granzyme B, perforin, CD39, and PD-1) associated with response.[18]

The vast majority of studies evaluating the immune repertoire in relation to patient outcome or treatment response often utilize broad metrics (diversity, clonality, etc) that reduce the high dimensionality of heterogenous immune repertoires. Although repertoire diversity is an essential summary statistic that reflects the potential breadth of antigen specificities and degree of clonal expansion, it is highly susceptible to biased calculation with under sampled TCR repertoires[19] and does not verifiably capture subtleties associated with antigen recognition.

Advances in machine learning algorithms have enabled superior understanding and utilization of immune repertoires in the monitoring and classification of disease outcomes. In a cohort of 90 patients with gastric cancer, a convolutional neural network outperformed support vector machine and random forest models in classifying cancer vs normal tissue based on BCR sequencing that utilized amino acid features coded into matrices with Kidera factors, CDR3 length, and V/J frame, and number of somatic hypermutations.[20] Utilizing the public TCGA database of 4200 RNA sequenced solid tumors, researchers at the University of Texas Southwestern developed a neural network, DeepCAT for *de novo* prediction of cancer associated TCRs that can provide a highly sensitive noninvasive cancer detection platform. Following removal of public CDR3s, 2x2 amino acid sequences are plotted and separated by CDR3 length and correlated to provide a Cancer score that was shown to be a robust predictor of cancer in several early and late-stage epithelial tumors of non-viral origin compared to non-cancer PBMCs. The Cancer score was prospectively validated in untreated RCC, ovarian, and PDAC cohorts with high sensitivity that could classify patients by tumor stage.[21] Following γδ TCR sequencing of 54 FFPE tissue specimens from patients with coeliac (CED) and non-coeliac disease, a machine learning algorithm that considered hypervariable CDR3 regions, short overlapping segments CDR3 sequences (kmers), and kmer sequence position (start, middle, end) utilized principal component analysis and hierarchical clustering with clinical indications to predict CED diagnosis that was accurate and independent of patient gluten sensitivity.[22] While machine learning algorithms utilizing immune repertoire sequencing data can greatly improve early cancer detection and diagnosis, they have not yet been utilized to predict treatment response and outcome. Patients with cancer receiving novel therapeutic agents often must wait several weeks if not months to receive potential clinical benefit, and in many cases

experience repertoire fitness and performance status decline and eventually succumb to their disease. A ML algorithm that can stratify treatments based on patient baseline or immediate immune repertoire response could dramatically improve precision medicine and clinical outcomes.

In this work, ML methodologies for early prediction of treatment efficacy based on patient repertoires are developed from patients treated with hydroxychloroquine (HCQ) and Adlesleukin (IL-2) for progressive renal cancer. Two main assumptions drive this work forward. The first is that the adaptome plays a role in renal cancer progression, remission, and disease equilibrium.[7] IL-2 is geared towards activating the immune system, and the original study shows that prolonged IL-2 treatment (with the help of HCQ) results in 20 of 29 patients reaching at least stable disease. The second assumption is that this role (and its mechanisms) is visible in the vast repertoires of patients, which can be observed as clonotypes that circulate the blood. Both the clonotypes that are observed and those that are not observed (due to down-regulation or migration to the tumor site) in intravenous blood could carry important information about the cancer progression/regression battlefield. Thus, detecting clonotype signatures related to disease phenotypes via intravenous blood could be a non-invasive and rapid way to elucidate an earlier diagnosis of renal cancer disease state compared to current standards.

A leading approach for CDR3 clonotypes-mediated prediction cancer states is to collapse a space of 10^{15}-10^{25} possible sequences into a low-dimensional space of a few descriptors such as whole sample sequence diversity and entropy. However large-scale descriptors may mask the more sensitive information encoded in partial-sequence (or within-sequence) information content of individual clonotypes that is more specific to disease outcomes. In this work, we

present a methodology for modeling renal cancer response to IL-2 based on information entropy and first-order Markov chains. A probabilistic analysis on observed amino acid pair motifs in whole receptor sequences was conducted for three purposes: (1) intermediate-dimensionality representations of CDR3 repertoires as compact 20x20 matrices; (2) early predictive classification of final patient states following 15 days of immunotherapy with an accuracy of 90%; (3) quantitative monitoring of changes in patient state via the information content of the adaptome. In comparison with t-SNE analysis of four standard diversity metrics, first order Markov chains of clonotypes elucidated entropic bias responses during IL-2 treatment mediated primarily by T cells or B cells. These results show the utility of Markov models for leveraging within-sequence information content as a powerful tool for advances in diagnostics in the clinic.

4.1.1. Cohort of Renal Cancer Patients

In this study, 29 patients with progressive renal cancer were treated with HCQ/IL-2 over the course of 15 days (Figure 4.1). At specified treatment points, namely day -14 (also called pre-IL-2 treatment in this work), day 1 (first day of IL-2 dosage), and day 15 (second dosage of IL-2), intravenous blood samples are extracted from each patient and sent to iRepertoire for analysis. A combination of NGS with damPCR and a proprietary bioinformatic pipeline were then used to extract clonotypes, after which the nucleotide and amino acid primary structure was reported. Examples of the extracted raw clonotype data used for this analysis, namely the CDR3 primary amino acid sequence and corresponding observed frequency are shown in Table 4.1 and Table 4.2 for a responder and non-responder from three different treatment points.

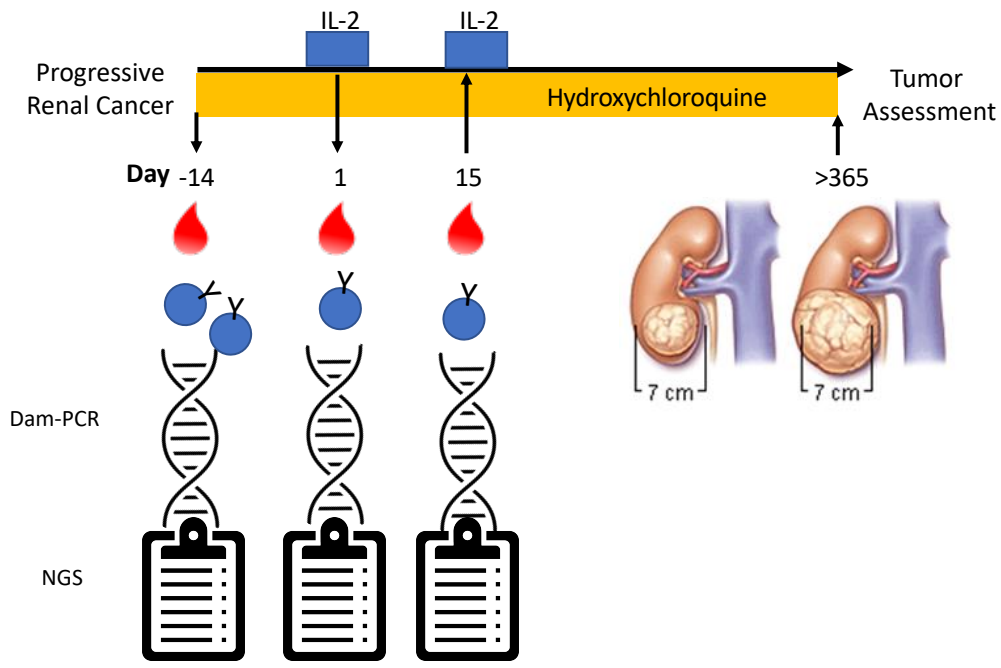

Figure 4.1. The role of the adaptive immune response in fighting renal cancer was studied by sequence-specific profiling of 29 progressive disease patients over the course of hydroxychloroquine (HCQ) and Adlesleukin (IL-2) treatment. All 29 patients received HCQ treatment at day -14 (pre-IL-2 treatment). HCQ/IL-2 was given two weeks later at day 1, and again at day 15. Blood samples for immune profiling were collected at all three time points. Out of the 29 patients studied, 20 patients had responses evaluated at day 15. RNA from blood samples was analyzed using a combination of dam-PCR and next generation sequencing (NGS). The iRweb (from iRepertoire) bio-informatics pipeline was used to assemble profiles of CDR3 variable region clonotypes for all seven TCR and BCR chains used in this analysis. Patient response to HCQ-IL-2 treatment was determined up to and even exceeding 100 weeks post-treatment. In our analysis, we intend to use patient CDR3s from the adaptive immune profiling described above to make early-stage predictions of patient outcomes by day 15.

It is important to note that the length of each individual sequence varies (10-25 amino acids on average) as does the size of each full sample list (10^2-10^5 total number of clonotypes observed in each sample, shown at the bottom of each list in Table 4.1 and Table 4.2). Furthermore, each patient sample was analyzed for all seven chains, resulting in 21 clonotype lists per patient. Since any combination of observed sequences from one of seven chains, frequency of clonotype

observation, or sample diversity could be a causal feature of patient response to treatment, a large and highly variable feature-space is obtained. Predictive models must carefully navigate over- and under- fitting the patient response. Figure 1.2 of this thesis reflects on the shape of data in which features drastically outnumber the patients studied. The challenge of good model feasibility in this regime lies in sufficient data to supplement features for more canonical data-driven learning. Acquiring "bigger data" in this case boils down to measuring a greater number of patients, which would exclude the analysis of data in Phase 1 trials. The most effective model strategies will adapt to balance high numbers of features per patient while leveraging the small patient cohort.

Table 4.1. Example of raw responder clonotype data (from IGH clonotypes)

Patient 0321 – Responder

Sample 1 Day -14 (Pre)		Sample 2 Day 1		Sample 3 Day 15	
Observed Sequences	Freq	Observed Sequences	Freq	Observed Sequences	Freq
ARDGVGATHFDH	5350	ARIPPMIEVVYYGMDV	26097	ARGVHNSYDPAGYDN	18446
ARGTTLVSRAEYFQD	4651	AKEHSSTSKGSFDI	18493	AHRRTYSSGWYFDY	5297
VTDPSWDILTGYTFDY	4618	AKEISGISSGSFDY	15747	ARMGRMDV	3595
AKYRIENMVHSGFDY	3563	AKEYSSVSKGSFDV	11227	ARTKRGVGGNFLYYFDY	3076
⋮		⋮		⋮	
~19.9k clonotypes		~31.5k clonotypes		~ 35.5k clonotypes	

Table 4.2. Example of raw non-responder clonotype data (from IGH clonotypes)

Patient 0627 – Non-Responder

Sample 4 Day -14 (Pre)		Sample 5 Day 1		Sample 6 Day 15	
Observed Sequences	Freq	Observed Sequences	Freq	Observed Sequences	Freq
ARTSAGRPGDY	13950	GSFTDS	14368	ARIYDSSVSYTGQDTFDI	14830
ARRYCEGGVCYDDRG	10757	ARDRLTTMTSLVLDH	13040	ARDSVTTYFDY	3138
AKDDDFWSGYYTFDD	8248	AKDRSYTIFGVFDY	10492	AKVPQNYGDSNLEY	2537
ARNLPPDY	7356	ARIKLKATDALAY	9634	VKEDDYHRSGRLDA	2452
⋮		⋮		⋮	
~45.9k clonotypes		~77.4k clonotypes		~ 105.4k clonotypes	

As previously stated, each patient sample is analyzed for all seven chains (TRD, TRG, TRA, TRB and IGH, IGK, IGL). Tree maps highlighting the diversity of observed sequences in each sample for any of the seven chains can be generated to highlight the overarching clonotype diversity for each patient sample. An example of a tree plot for Day 15 comparing Patient 0321 (Responder) to Patient 0627 (Non-Responder) is shown in Figure 4.2. Each colored, rounded rectangle represents a unique CDR3 that can be mapped back to a specific V- and J- gene alignment. The size of the rectangles represents the observed frequency of each unique CDR3 in the sample, reflecting the repertoire bias. As the adaptome has many functions other than tumor suppression, the tree plots exemplify the "needle in the haystack" challenge of locating the dominant CDR3s that are signatures of renal cancer disease within a myriad of possibilities. While the plots visualize diversity, they cannot give estimates on the clonal expansion of any specific CDR3s, which would theoretically come from the up- and down- regulation of relevant disease-fighting receptors. The next step will be to analyze the repertoire bias using quantitative diversity metrics as discussed in the following sections.

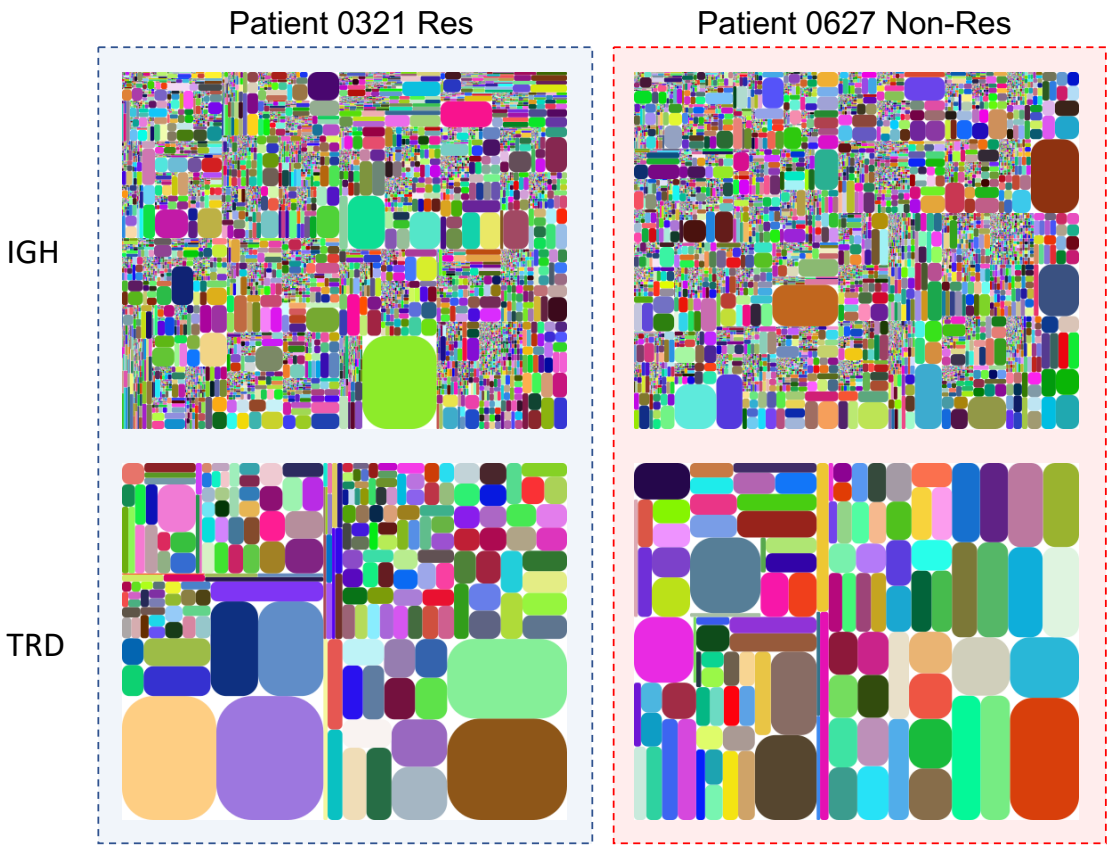

Figure 4.2. The tree plot shows chain diversity for two patient samples comparing BCR Heavy (IGH) and TCR Delta (TRD) chains for Day 15 from Table 4.1 and Table 4.2. Each rounded rectangle represents a unique CDR3 (uCDR3) entry based on V- and J- gene usage, gathered during alignment in the bio-informatic pipeline. The entire plot area is divided in to sub-divided according to V usage, which is subdivided according to J usage. The size of the rounded rectangle represents the relative frequency of observation for the uCDR3. We found that for all patients, BCR chains tended to demonstrate consistently greater diversity compared to TCR chains, as shown for the TRD and IGH above.

4.2. Results

4.2.1. Analysis of Standard Diversity Metrics

Standard diversity metrics from seven-chain sequence lists shown pictorially in Figure 4.2 were calculated from patient samples for Day 15. A detailed explanation of diversity calculations can be found in Appendix Table 4.5. Four standard diversity metrics were used: D50, Diversity Index (DI), Shannon Entropy, and Unique CDR3 count (uCDR3). After calculation, patient

samples at Day 15 were separated by known labels responders and non-responders, and the diversity metrics were individually analyzed for the null hypothesis that the patients come from a single distribution. Figure 4.3A shows the results of using diversity metrics on each of the seven chains to significantly distinguish two distinct (responder versus non-responder) patient distributions. In summary, neither TCR nor BCR diversity metrics showed significant distinction between responders and non-responders.

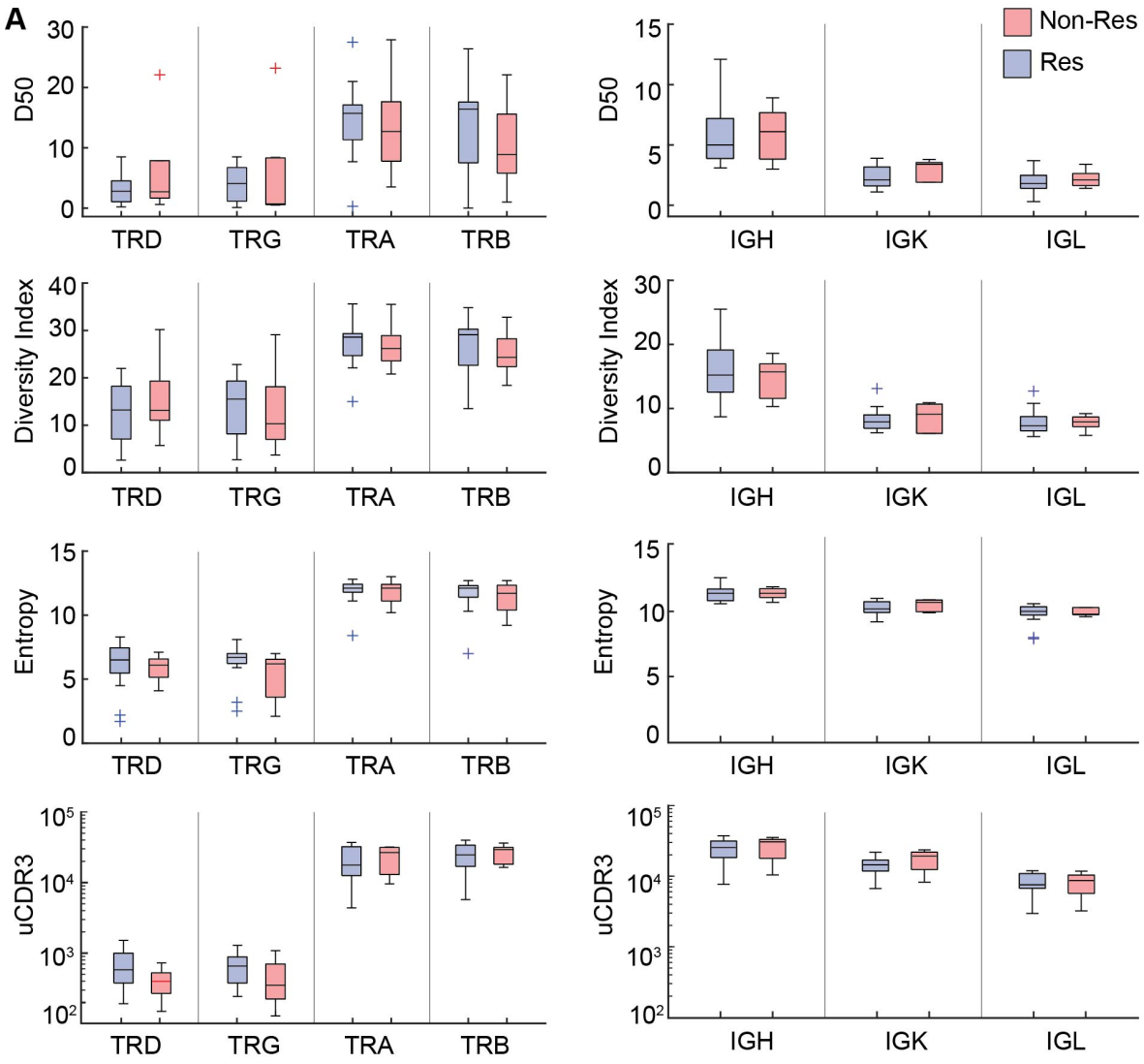

Figure 4.3. Four diversity metrics, D50, Diversity Index (DI), Entropy, and Unique CDR3 (uCDR3) analyzed for Day 15. For each chain, patient data in the corresponding diversity metric was separated into known labels to determine if patients could accurately be split into responder (blue) and non-responder (red) labels. Boxplots show the distributions of responder and non-responder patients for each metric. The significance of the data split by true its labels (shown) did not outperform splitting the data randomly in any of the four metrics calculated. We did notice that TRD and TRG chains had collectively lower diversity and lower unique CDR3s compared to other chains.

To better understand the role of diversity metrics in the data, and to rule out higher-order combinations of diversity factors as important splitting features, a high-power clustering

algorithm, t-distributed Stochastic Neighbor Embedding (t-SNE), was run to visually capture non-linear patterns. Data from all seven chains were aggregated for each of the four diversity metrics aforementioned, creating an intermediate dataset [4 diversity features x 105 samples] in size. The t-SNE algorithm was then tuned at various perplexity hyperparameters to project the data onto a 2D axis of principle vectors. Datapoints were then annotated by patient response and by chain to help interpret clusters. We found that patient clusters were not observable by patient response to treatment. However, we did find that non-linear transformation of diversity metrics via t-SNE tended to cluster based on TCR or BCR chain. We found distinct three distinct clusters (shown for Day 15 in Figure 4.4) that were invariant to the treatment point studied. Cluster I consisted of the TRD/TRG chains. Cluster II consisted of TRA/TRB chains, and cluster III contained IGH, IGK, IGL chains. This is an interesting result, both because it reflects the natural association of α/β and γ/δ subunits, but also suggests that the mechanism itself is more strongly observable in diversity metrics than disease signatures at Day 15. This may be due to non-specific expansion of the clonotypes as a result of IL-2 treatment.

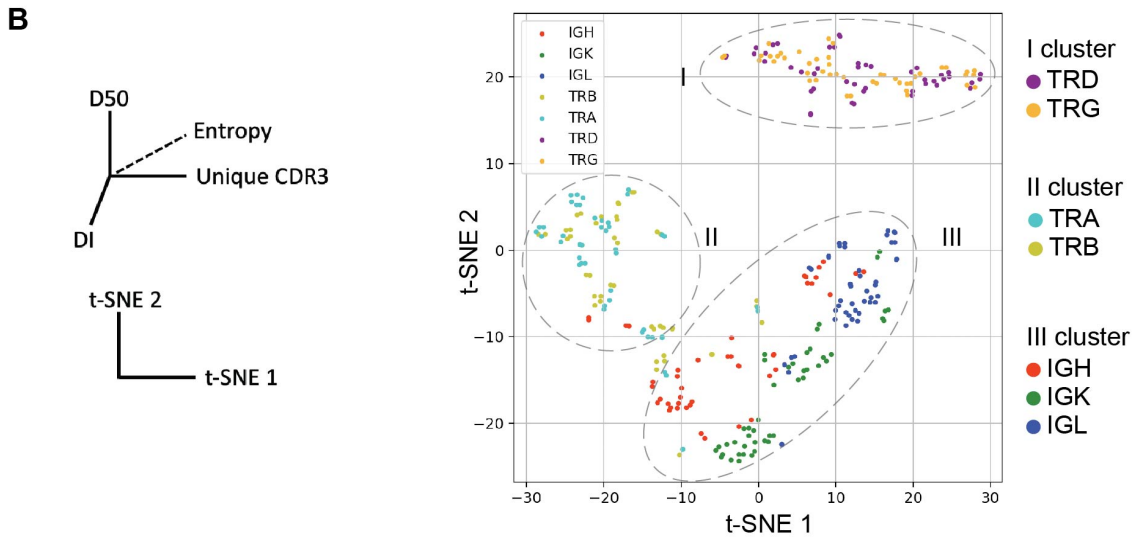

Figure 4.4. Unsupervised clustering by diversity metrics. Four diversity metrics, D50, Entropy, Unique CDR3s, and Diversity Index (DI) were reduced to two principle t-SNE components to illuminate non-linear patterns in the data. While t-SNE representation of the data did not pull out any latent groups, diversity metrics clustered most readily by chain with distinct clusters forming for (I) TCR delta/gamma chains, (II) TCR alpha/beta chains, and (III) light, heavy and kappa chains. Clustering of TCR and BCR chains by diversity reflects a conserved underlying mechanism of adaptome function.

Analysis of diversity metrics reveal the flaws with low dimensionality representations for predicting patient response. None of the 4 diversity metrics used were able to significantly split the patients into response and non-response. It is likely that the answer is embedded deeper in the data and may require more fine-tuned analysis of the individual clonotypes to uncover disease signatures. Consequently, we will conduct a series of full-length and partial-length analyses of clonotypes patterns to uncover the casual features necessary to predict patient response.

4.2.2. Markov Methods for classification from partial-length analysis

Analysis of diversity metrics imply the possibility that by day 15, the effect of IL-2 treatment is a large-scale fortification of entire repertoires, rather than the targeted generation of clonotypes specific to disease epitopes. Furthermore, unsupervised clustering indicated that full-sample diversity metrics, while an excellent description of the overall chain contributions and mechanisms, are too global to accurately classify patients.

Amino acids are often functionally interchangeable in terms of properties such as hydrophobicity, polarity, molecular weight, and acidity among other secondary structure attributes that affect binding to targets. In reality, the "lock and key" of perfect molecular recognition involves competition with complimentary constructs that have similar binding affinities but non-exact amino acid codes. Given the vast diversity and individuality of public repertoires due to random V(D)J recombination, full-length sequences likely contain too much individuality to be detected as shared, distinguishing features in responder or non-responder patient cohorts. Rather than exact sequences, the down-stream selection and expression of TCRs and BCRs in patient cohorts is far more likely to converge on similar, but non-exact clonotypes with comparable properties that arise between individuals with common selection pressures.

Not surprisingly, previous success has been achieved in using partial-length analysis of clonotypes for identifying causal motifs that give good classification of disease diagnosis. Using 4-6 length snippets, statistical classifiers for using immune repertoires to diagnose multiple sclerosis with passable accuracy have been created.[23] In this work, Ostmeyer, *et al*. further discuss the utility of including Achtley vectors, or hot-encoded amino acid property

representations,[24] in the model design, raising classification accuracy from 65% to 87% for a cohort of 125 patients. Additional success has been achieved through k-mer analysis combined with Achtley vectors, achieving up to 80% accuracy for classification of small patient cohorts. However, k-mer analysis of clonotypes quickly becomes computationally expensive as features will scale as 20^k where k is the number of amino acids in the sliding window in each sequence. For example, k=3 triplet analysis of clonotypes results in 20^3 = 8,000 features, nearing the size (or in the case of TRD/TRG increasing the size) of the features in the original dataset. Algorithms like PCA which has been used for feature selection from up to k=7-mer analysis, and the multitude of layers required for deep learning on amino acid pairs,[21] may continue to battle model complexity at low patient numbers. In this way, the field is in constant risk of model non-convergence or else over-fitting the data and losing generalization to unseen patients.

Inspired by the use of Markov models for the classification of DNA snippets into CpG islands,[25] we suggest a simple classification model built from measuring conditional probabilities of amino acid pairs (k=2-met analysis) to reduce the vast cardinality of full-length sequence lists into an intermediate-dimensional space of 400 features stored in transition probability maps calculated for each sample. Without the need for additional property features, we will show that the embedding of condition probability with amino acid pair motifs can classify patients with up to 90% accuracy regardless of the BCR or TCR chain analyzed at Day 15 of IL-2 treatment. First, patient samples will be transformed into a common matrix structure representing the normalized amino acid usage bias. Then each k=2-mer pair will be tested for success in splitting the patient data by true labels versus random splits, resulting in a Significant Feature Filter that keeps only the causal pairs. Leave-one-out patient cross-validation will assign

a feature importance score to the features in the filter. Finally, the Hadamard product of the filter with raw patient data, reveals distinct amino acid abundance in the clonotypes of non-responders versus responders. After filtering, patients are given a scalar classification score. A significance test is then run to test the null hypothesis that responders and non-responder scores arise from the same distribution. The generalizability of this method with respect to features that survive cross-validation and receive high significance weights will be discussed.

4.3. Model Building

A visualization of the model pipeline for the transformation of patient samples into an intermediate feature space and subsequent feature significance tests from patient samples at Day 15 is shown in Figure 4.5. First, whole clonotypes from the list are chopped into pairs. An example of k=2 pair analysis is as follows: GSFTDGS will be represented as GS | SF | FT | TD | DG | GS with a count of 1 for each pair except GS which receives a count of 2. Pairs in each sequence are counted and normalized in probability maps based on the frequency of observing a given amino acid with its previous neighbor. The probability of observing an amino acid pair in a single CDR3 sequence is $P_{obs}(X_i|X_{i-1}) = {}^{N_{X_i|X_{i-1}}}\!/_{N_{X_i|X_N}}$ where X_i and X_{i-1} are two neighboring amino acid pairs. The final vector $\langle X_i, X_{i-1}, P_{obs}\rangle$ encodes the location and magnitude of each observation in the probability map. As there are 20 possible amino acids, the feature-space of the pair-wise probability map scales with 20^k = 400 possible entries. Each clonotype in the list observed at day 15 for a given patient is analyzed with k=2-tuplet analysis and added to the map, resulting in a single probability map of 400 features for each patient. The rows of observations in the map are normalized to 1. Probability maps are generated for each

patient at day 15, ${}^{\text{patient}}\mathcal{M}_{day15_{chain}}(X)$. Each vector in the map, $\mathcal{M}(\langle X_i, X_{i-1}, P_{obs}\rangle)$ is considered an independent feature.

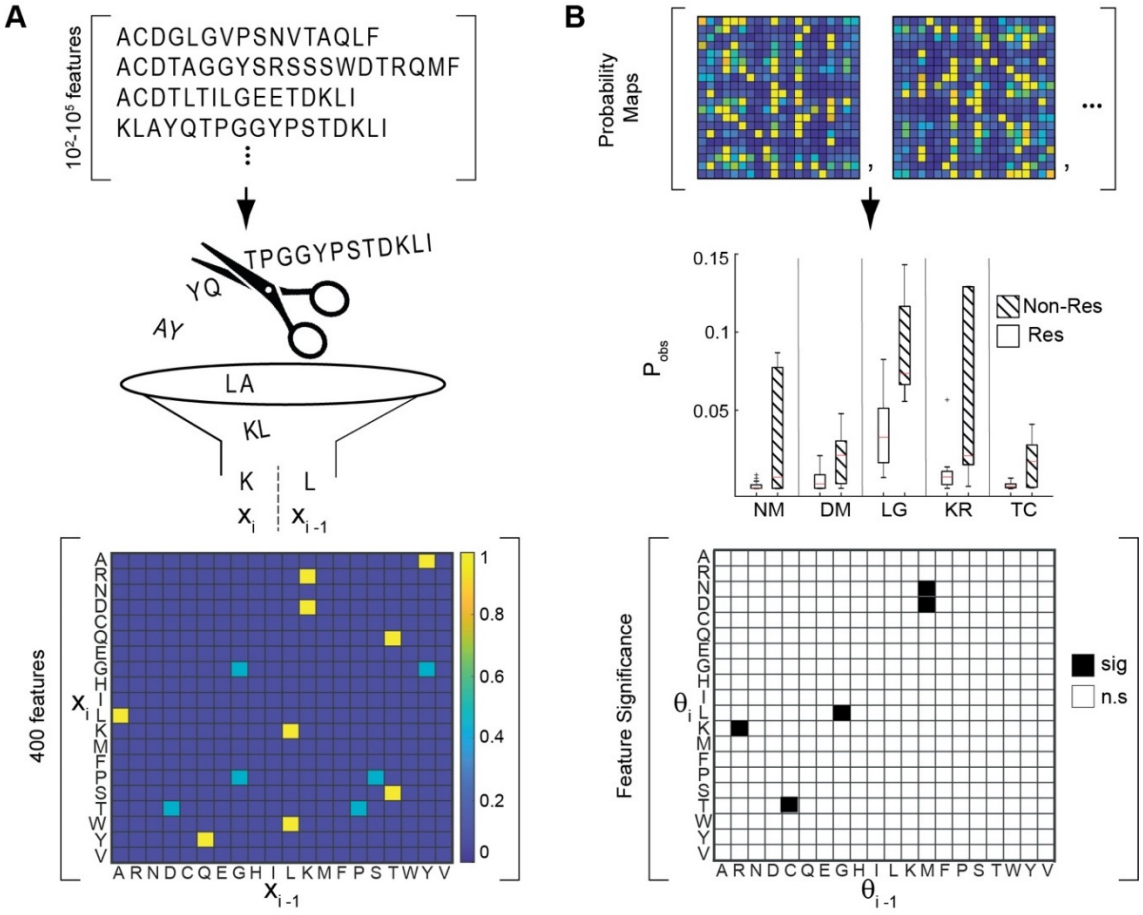

Figure 4.5. Initial classification pipeline starting from clonotype lists from patient samples at Day 15. (A) Individual CDR3 sequences from the cardinal list elucidated from a patient sample are chopped into dimers, counted, and normalized in a conditional probability map of $20^{k=2}$ = 400 features. (B) Probability maps from (A) are elucidated for each sample and compared for responder versus non-responders cohort for each k=2 pair feature. Significance is assigned to a feature if the observed probability distributions from the data can split the data more successfully for true patient-response labels than for random splits. The feature in the probability map is then replaced with a value of 1 or 0 depending on the significance test for describing patient outcomes, resulting in a Feature Significance matrix the keeps only the relevant features.

A supervised approach to finding the features driving patient response or non-response to treatment is conducted. Patients are divided into "responders" or "non-responders" and each feature (amino acid pair) in the map is assessed for this split individually. A feature is considered significant if splitting the data according to the true patient response performs better than splitting the data randomly. In the example of performing the analysis for patients at day 15 using the TRD chain, identified significant pairs are $\{\langle X_i|X_{i-1}\rangle\} = \{NM, DM, LG, KR, TC\}$. Significant features are assigned a value of 1 and placed into a new matrix, $\mathcal{F}(\theta_{ij})$, using the transformation $\langle X, X_{i-1}, P_{obs}\rangle \rightarrow \langle \theta_i, \theta_j, p_{sig}\rangle$. Significant features $\langle \theta_i, \theta_j, p_{sig=1}\rangle$ are represented as black squares. Insignificant features are assigned 0 using the vector $\langle \theta_i, \theta_j, p_{sig=0}\rangle$ and represented as white squares. The result is a transformed feature-space $\mathcal{M}(X) \rightarrow \mathcal{F}(\theta)$ in which suspected high-performing features in the training analysis are identified and stored.

Leave-one-out cross-validation (LOOCV) shown in Figure 4.6 is then conducted in which the analysis of significant features shown in Figure 4.5 is performed on n-1 patients with one sample removed each time. Given a cohort of 15 patients who gave Day 15 samples, 15 CV iterations were run. Each of the resulting significance matrices gave slightly different feature importance entries. Thus, the resulting feature selection matrix ensembles are then averaged so that features which survive leave-one-out cross validation for all data folds were given higher weights than those which only described a few iterations. The result is a final "Feature Significance Filter" $\mathcal{F}(\overline{\theta_{i,j}})$ in which values for entries, $\theta_{i,j}$ are a spectrum ranging from 0 (not-significant) to 1 (significant) wherein a weight of 0.8 would imply that the feature survives in only 80% of the

cross-validation data folds. We suspect that features that are able to elucidate higher significance scores will be more generalizable to unseen patient cohorts.

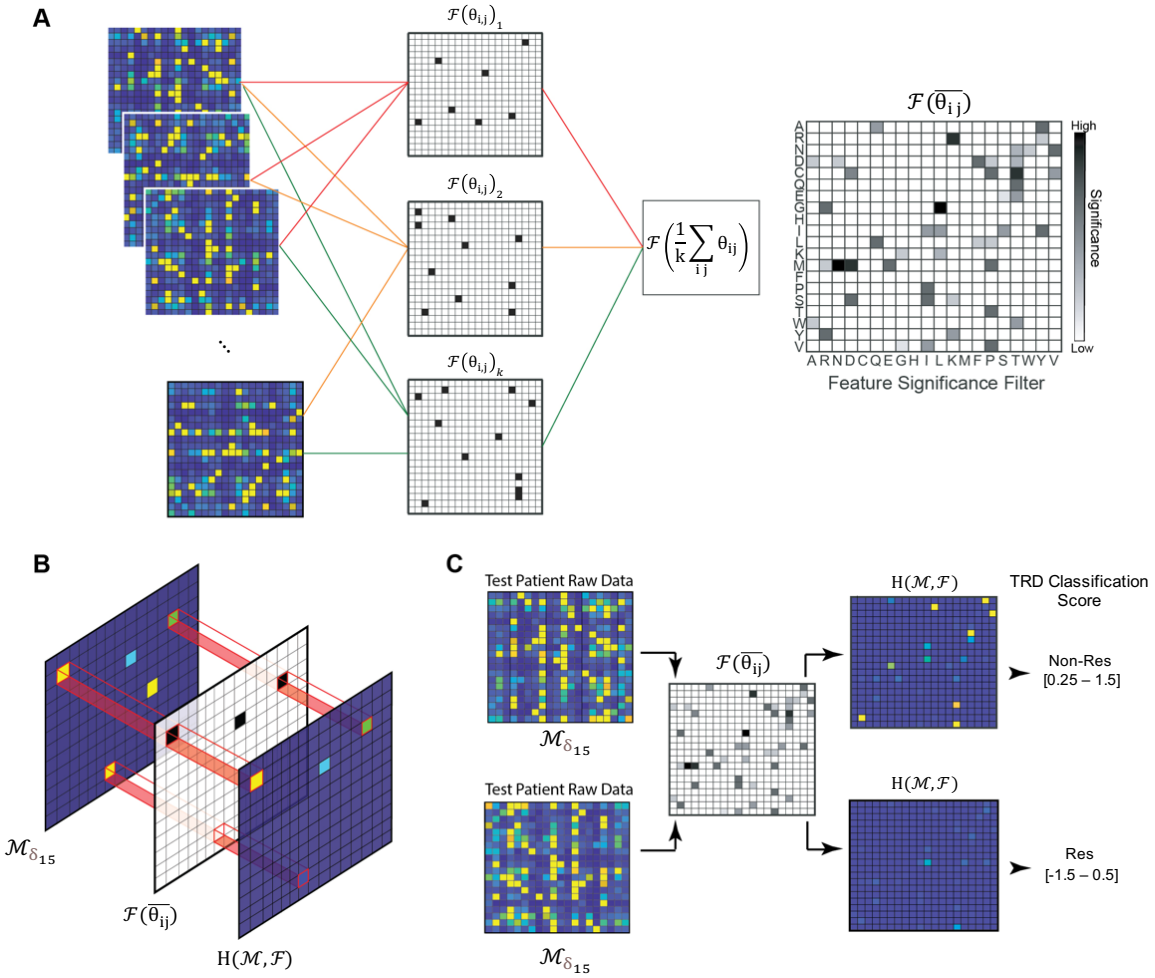

Figure 4.6. Ensembles of probability maps are used to evaluate the significance features in separating responders from non-responders. The features are cross validated using leave-one-out cross-validation. The filter from each cross-validation iteration is then averaged element-wise to make the final Feature Significance Filter wherein features that survived multiple data splits in the cross-validation receive a higher weight. The result is a pseudo-regularizer that retains only the features expected to successfully split responders from non-responders, sending all other features to zero. (B) coefficients in the filter, theta, represent the weights of the features. (C) Use of selection filter on raw patient data. The idea here is to show that the selection filter helps pick out the dominant motifs that separate out the patient groups. The algorithm is run independently for seven-chain analysis.

Armed with a cross-validated Feature Significance Filter, and inspired by image analysis techniques, the Hadamard product is calculated on raw patient data as shown in Figure 4.6B. The result is a new matrix $H_{i,j} = \left(\mathcal{M}_{15_{day}} \circ \mathcal{F}\left(\overline{\theta_{i,j}}\right)\right)$ which keeps the frequency values $\mathcal{M}_{\mathrm{i,j}}$ from the raw patient sample data that are most significant. Features determined from LOOCV analysis to be most significant ($\overline{\theta_{\mathrm{i,j}}} = 1$) retain the raw observed frequency values from $\mathcal{M}_{\mathrm{i,j}}$. Like a cloudy window, as $\overline{\theta_{\mathrm{i,j}}}$ trends to zero ($\overline{\theta_{\mathrm{i,j}}} \to 0$), its product with the corresponding $\mathcal{M}_{\mathrm{i,j}}$ raw data observation attenuates the raw values down to zero (no significance). Figure 4.6C shows the result of the Hadamard product between the Feature Significance Filter and two raw patient samples from Day 15 (TRD chain). Finally, the rows and columns of $H_{i,j}$ for each patient are summed to give a scaler score from significant features and corresponding weights.

$$score = \sum_{i}\sum_{j}\left(\mathcal{M}_{\mathrm{raw}} \circ \mathcal{F}\left(\overline{\theta_{i,j}}\right)\right) \qquad \text{Eq. 4.1}$$

$$score_{TRD} = \begin{cases} \text{range}\,[0.25\,,1.5] & \text{NonRes} \\ \text{range}[-1.5\,,0.5] & \text{Res} \end{cases} \qquad \text{Eq. 4.2}$$

From the feature significance tests, we observed that non-responder patients typically had a higher observed abundance of significant amino acids neighbors compared to non-responders (and example for this in TRD chain is shown in Figure 4.5B). In some cases, there were significant features which had higher observed abundance in responders. In order to consistently analyze attributes, features were assigned a value for $\overline{\theta_{\mathrm{i,j}}} = -1$ in the event that responders had higher abundance after the significance test. For simplicity, the model is discussed above in

terms of non-responder abundance. In reality, the range of model weights $\overline{\theta_{1,J}} = [-1,1]$ reflecting the possibility of responder abundance is used.

A summary representing the full analysis for all seven chains from the classification model is shown in Figure 4.7. Within each chain, we were able to discover distinguishing features that differentiated patients into responders and non-responders, resulting in good separation of patients by their true labels.

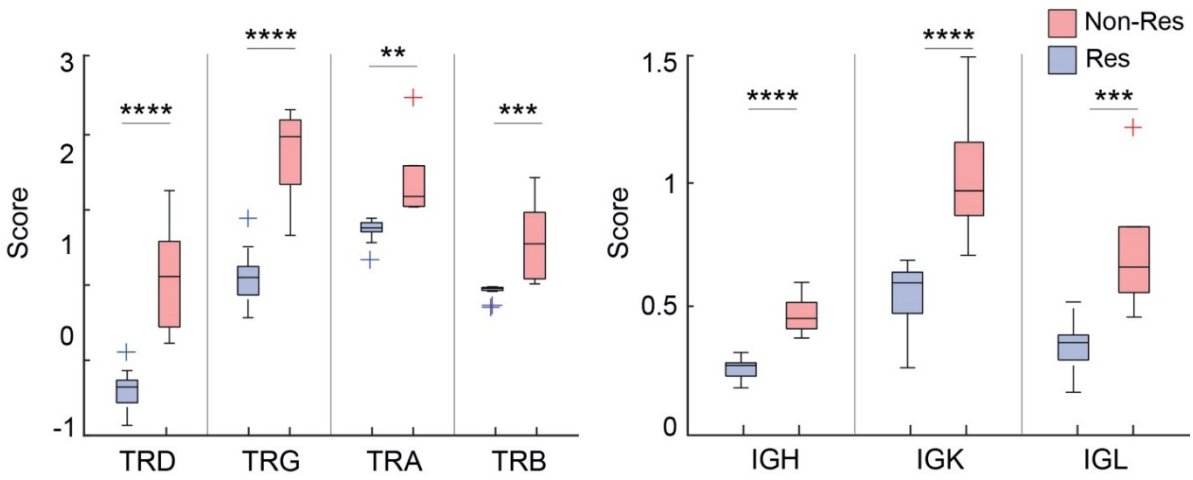

Figure 4.7. Classification of patients based on the scoring from the Feature Selection Filter and analysis. Each chain was able to pull out significant features in clonotypes that then distinguished patients by their true labels. A Shapiro-Wilk test was run to identify the presence of normal distributions in responder and non-responder cohorts, and ANOVA was run to identify the significance of the distributions by classification score. The significance values are **** for $p<0.0001$, ** for $p<0.001$, and ** for $p<0.01$.

More data will allow us to draw distinct, scalar cut-off thresholds for classification scores for each chain. For the purposes of this work, classification accuracy can be loosely identified as the number of patients who exist in the overlap between responder and non-responder score

distributions, a metric for confidence based on the degree of data separation. Specifically, Accuracy is defined as N(patients in overlapping regions)/N(total patients studied). The following table summarizes each chain's performance on classification with respect to true labels.

Table 4.3. Classification certainty by chain

Chain	Classification accuracy
TRD	18/20 = 90%
TRG	18/20 = 90%
TRA	20/20 = 100%
TRB	19/20 = 95%
IGH	20/20 = 100%
IGK	20/20 = 100%
IGL	18/20 = 90%

All seven chains (TCR and BCR) significantly distinguished the patients by their true labels with a significance of at least $p<0.01$. Furthermore, classification certainty for each chain was at the lowest only 90% for TRD, TRD, and IGL and at most 100% for TRA, IGH, and IGK. Overfitting the data is a concern given that the low number of patients meant we could not run a true test on unseen data. However, we hypothesize that features that earned higher weights in the model are more likely to generalize to large populations. Ultimately, the ability to glean good classification accuracy as early as 15 days post treatment could have large benefits for patients in the clinic who may be able to quickly alter treatment to get a more efficient response.

4.4. Disease monitoring: Markov models for within-sequence information entropy

The success of each chain in the Markov classification model in predicting early (Day 15) patient outcomes via observed motifs in clonotypes has the profound implication that all seven chains play an orchestral role in patient equilibrium with renal cancer. To further investigate this role, we will extend the Markov model to analyze individual patients with respect to themselves over three temporal treatment points. A benefit of monitoring individual patients against themselves is that with 400 features and 3 temporal treatment points, the shape of the data is well-suited to data-driven learning. Furthermore, we seek to gain an understanding of how each of the seven chains contributes to patient response over the course of treatment, ultimately allowing us to observe how the ensemble of chains function together and possibly obtaining additional mechanistic insight into immunotherapies.

Table 4.1 shows how the cardinal number of each set of clonotypes extracted from patient samples changes over treatment. For patient 0321, the set of IGH sequences measures from 31.5k in size at day 1 to 35.5k at day 15. Likely an effect of the IL-2-driven expansion, changes in the cardinal number of each sample show how CDR3s are generated (or sometimes removed) over the course of treatment as selection pressures up- and down- regulate TCRs and BCRs. A lack of constant features discourages the use of volcano plot strategies for measuring the significance versus expression ratio over treatment conditions. The goal of this section is to monitor the significant fold-changes of clonotypes for an individual patient over the treatment. Fortunately, we have an excellent method for transforming samples into a consistent feature space using the probability maps described in section 4.3. We will use probability maps

combined with first order Markov approximation of full-length CDR3s to study the magnitude and bias of clonotype expression as the distribution of receptors shifts for each patient over time.

4.4.1. Clonotype analysis for disease monitoring

As classification greedily hunts for significant motifs for early diagnosis, the model loses interpretability. However, motifs alone cannot encompass the whole story. Motivated by the need for targeted T-cell therapies,[26] there is additional interest in the full-length sequences of clonotypes with the highest therapeutic value. Furthermore, the data in this study provide a rare opportunity to observe the effect of high doses of IL-2 on repertoires, made possible from the mitigation of side-effects provided by HCQ. Thus, the effect of IL-2 on wholistic changes in the repertoire can be elucidated. Data-driven learning over clonotype expression during treatment could provide better mechanistic understanding of patient response to IL-2 and renal cancer pressures. We hypothesize that wholistic changes in repertoire expression can be studied via the probabilistic information carried by amino acid motifs within full-length clonotypes. By studying changes in k-mer clonotype patterns at a baseline (Day-14) compared to post IL-2 treatment at Day 1 and Day15, we will show the presence of entropic bias shifts for TCR and BCR each chain in response to IL-2.

In the literature, Shannon entropy is typically calculated as a diversity metric for whole samples. The equation $S = -\sum_i^{\mathrm{N}} p_i log2(p_i)$, where p_i is the frequency of observing the ith CDR3 in a sample of N sequences, measures the overall disorder of a sample. Higher values for entropy imply greater disorder in the observed list of full-length sequences. However, this representation neglects the information content contained within the primary structure of individual sequences. Within the primary structure of full-length CDR3 sequences is the combination of multiple

probabilistic V(D)J events that lead to the observance of specific amino acid motifs. Examples of such events include, VD or DJ insertions, P-nucleotide deletions, random recombination.[27] To begin to reconcile random V(D)J events from down-stream selection of dominant CDR3s, it will be important to intersect diversity metrics of full-length sequences with the information content from within-sequence patterns that arise due to variation in patient environments. In this work, we will show a novel application of how k-mer analysis of CDR3s combined with first-order Markov chains allows a wholistic analysis of samples over time, giving broad insight into repertoire chain dynamics, stemming from the most basic sequence patterns.

The abundance of within-sequence amino acid motifs can be represented in the probability maps of samples at each treatment point: $\{\mathcal{M}_{\mathrm{pre}}, \mathcal{M}_{\mathrm{day1}}, \text{and } \mathcal{M}_{\mathrm{day15}}\}$. Patient samples for three temporal points, Day -14 (pre IL-2), Day 1 (IL-2 dose 1), and Day 15 (IL-2 dose 2) were analyzed according to section 4.3 (summarized in Figure 4.3). Probability maps were generated for each treatment point for individual patients. For example, ${}^{0321}\{\mathcal{M}_{\mathrm{pre}}, \mathcal{M}_{\mathrm{day1}}, \text{and } \mathcal{M}_{\mathrm{day15}}\}_{\partial,\gamma,\alpha,\beta,\mathrm{IGH,IGK,IGL}}$ represents the set of three probability maps for patient 0321 for samples obtained at pre IL-2, Day 1, and Day 15 for each chain. A total of 21 maps per patient are created for the 3-timepoint analysis on all seven chains.

Baseline pre-treatment sequences (from Day-14) were calculated as first-order Markov chains using the $\mathcal{M}_{\mathrm{pre}}, \mathcal{M}_{\mathrm{day1}}, \text{and } \mathcal{M}_{\mathrm{day15}}$ data for conditional probability estimations. A given CDR3 sequence can be expressed as the product of conditional probabilities of its k-mer parts. The probability of observing a sequence of specific dimer motifs in a given sample list can be calculated according to Eq. 4.3 below:

$$P(X \ldots X_i X_{i-1}) = \prod P_X P(X_i | X_{i-1}) \quad \text{Eq. 4.3}$$

where P_X is the nonuniform composition of individual amino acids in proteins and can be estimated as ~4.18 bits/amino acid (but will cancel out in subsequent calculations of this analysis). The values for $P(X_i|X_{i-1})$ are found from the address coordinates $\langle X, X_{i-1}, P_{obs} \rangle$ given by probability maps $\mathcal{M}_{\mathrm{i=x,\ j=x-1}}$ measured from data.

For example, if the theoretical sequence $\mathrm{seq} = \mathrm{DRA}$ is evaluated as a first-order Markov chain using k=2-mer analysis and using the Day 1 probability map, then the dimers AR | RD can be represented as $P(X_i|X_{i-1}) \sim \mathrm{P(A)} * \mathrm{P(A}\,|R) * P(R|D)$ where $\mathrm{P(A}\,|R) = \mathcal{M}_{1_{i=A,j=R}} = \langle A, R, P_{obs_1} \rangle_{day1}$ and $\mathrm{P}(R\,|D) = \mathcal{M}_{1_{i=R,j=D}} = \langle R, D, P_{obs_1} \rangle_{\mathrm{day1}}$. If instead the sequence were evaluated at Day 15 probability map, then $\mathrm{P}(A\,|R) = \mathcal{M}_{15_{i=A,j=R}}$ and $\mathrm{P(R}\,|D) = \mathcal{M}_{15_{i=R,j=D}}$ resulting in $\langle A, R, P_{obs_{15}} \rangle_{\mathrm{day15}}$ and $\langle R, D, P_{obs_{15}} \rangle_{\mathrm{day15}}$, respectively. In general, $P(X \ldots X_i X_{i-1})|\mathcal{M}_1$ implies the first-order Markov chain for a CDR3 sequence evaluated with the $\mathcal{M}_1$ probability map, $P(X \ldots X_i X_{i-1})|\mathcal{M}_{15}$ is the first-order Markov chain for the same sequence calculated using the $\mathcal{M}_{15}$ map, and $P(X \ldots X_i X_{i-1})|\mathcal{M}_{\mathrm{pre}}$ is a baseline.

We can then take the log ratio $P(X \ldots X_i X_{i-1})|\mathcal{M}_{\mathrm{day}} \Big/ P(X \ldots X_i X_{i-1})|\mathcal{M}_{\mathrm{pre}}$, which allows a logistic comparison of how Day-14 sequences compare to the sample distributions from other treatment points. Since $log_2\left(A/B\right) = log2(A) - log2(B)$, we get a description of the baseline distribution similarity to either Day1 or Day15. If $P(X \ldots X_i X_{i-1})|\mathcal{M}_{\mathrm{day}} >> (X \ldots X_i X_{i-1})|\mathcal{M}_{\mathrm{pre}}$, then the score for $log_2\left(A/B\right)$ will be large and

positive. If $P(X \ldots X_i X_{i-1})|\mathcal{M}_{\text{day}} << (X \ldots X_i X_{i-1})|\mathcal{M}_{\text{pre}}$ then the score will be negative. If $P(X \ldots X_i X_{i-1})|\mathcal{M}_{\text{day}} \approx (X \ldots X_i X_{i-1})|\mathcal{M}_{\text{pre}}$, then the score becomes zero. Day-14 sequences can be used as a baseline to compare changes of within-sequence entropy over treatment, representing frequency of pattern expression for Day 1 and Day 15 using the following equations:

$$\Delta_{\frac{\text{day1}}{\text{pre}}} = \log_2 \left(\frac{\{\text{seq}\}_{\text{pre}}|\mathcal{M}_1}{\{\text{seq}\}_{\text{pre}}|\mathcal{M}_{\text{pre}}} \right) \quad \text{Eq. 4.4}$$

$$\Delta_{\frac{\text{day15}}{\text{pre}}} = \log_2 \left(\frac{\{\text{seq}\}_{\text{pre}}|\mathcal{M}_{15}}{\{\text{seq}\}_{\text{pre}}|\mathcal{M}_{\text{pre}}} \right) \quad \text{Eq. 4.5}$$

Eq. 4.4 and Eq. 4.5 indicate how pre-treatment sequences are analyzed, with $\{\text{seq}\}_{\text{pre}}|\mathcal{M}_1$ implying that Day-14 (pre) sequences were analyzed with the conditional probabilities observed in Day 1 observations, and $\{\text{seq}\}_{\text{pre}}|\mathcal{M}_{15}$ indicating the same set of Day-14 sequences was calculated with Day15 observations. The expression, $\{\text{seq}\}_{\text{pre}}|\mathcal{M}_{\text{pre}}$, acts as the baseline. In this way, the baseline (Day-14) sequences are calculated as first order Markov chains using two different probabilities distributions, and the log-ratio acts as a fold-change metric normalized to a baseline.

Each clonotype in Day-14 samples, $\{\text{seq}\}_{\text{pre}}$, is analyzed, resulting in transformed distributions of the sample to reflect similarities and differences between the next analyzed treatment points. The treatment points compared from Eq. 4.4 and Eq. 4.5 are shown in Figure 4.8A.

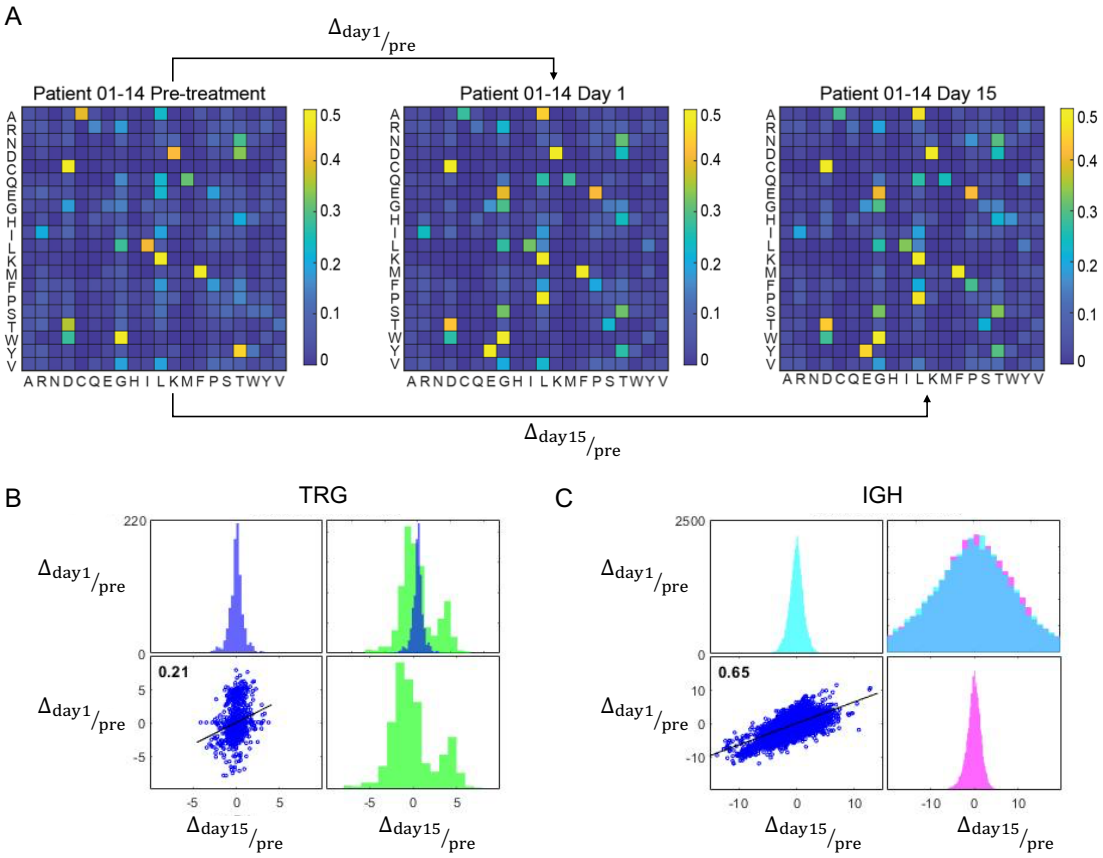

Figure 4.8. Analysis of patient repertoires over three treatment points. (A) First probability maps are generated showing normalized frequency of amino acid pair motifs in each sample. The fold-change pattern analysis is described by $\Delta_{\boldsymbol{day1}/\boldsymbol{pre}}$ (change from Day-14 to Day 1), and $\Delta_{\boldsymbol{day15}/\boldsymbol{pre}}$ (change from Day-14 to Day15). (B) Shows histograms of Day-14 sequences binned by their evolution or conservation of sequence patterns compared to respective treatment points for TRG and IGH. Diagonal plots show individual the individual Day-14 distributions as they are calculated with the conditional probabilities from Day1 and Day15 timepoints. Off-diagonal element (upper-left corners) shows the overlay of these two distributions, the entropic bias of which is quantified in the lower-off-diagonal Pearson correlation coefficient.

Histograms representing Day-14 sequences binned by the ratio of Markov probability estimates from equations 4 and 5 show the evolution of the repertoire over treatment based on within-

sequence patterns. Positive bins on the x-axis indicate new sequence motifs in Day1 (Eq. 4.4, Figure 4.8B top left), and Day15 (Eq. 4.5, Figure 4.8B bottom right), respectively. The more positive the bin, the greater the evolution of amino acid motifs from baseline. Consequently, negative bins show a conservation of patterns from Day-14. Finally, sequences that did not change in expression over treatment are binned at zero. Entire distributions overlaid (Figure 4.8, upper right), give a wholistic estimate for the entropic change in the receptors between Day1 and Day15 for the chain studied. Changes in the means of the distributions, for example, represent a paradigm shift in amino acid pattern bias (as demonstrated in the TRG chain histogram overlay in Figure 4.8B, upper right). As one could not presume to guess outright the shape of the resulting bias distributions, Pearson linear correlation coefficients, ρ, were used to compare the distributions from Eq. 4.4 and Eq. 4.5 (Figure 4.8B, bottom left) so as not to make assumptions about the distribution shapes. When significant change between Day1 and Day15 occurred, ρ trends towards 0 (shown for TRG in Figure 4.8B). When no change occurs, the internal entropy is frozen and ρ trends towards 1 (shown for IGH in Figure 4.8B).

4.4.2. Entropic bias shifts in patient repertoires

To get a sense for how all seven chains contribute to entropic changes in the overall repertoire in response to IL-2, the Pearson linear correlation coefficients were collected for each of the seven chains and viewed for each patient. Interestingly, we noticed four mechanistic patterns within responding and non-responding patients: BCR bias, TCR bias, mixed bias (shown in Figure 4.9), and no bias. A subset of responding patients showed a dichotomy between BCR and TCR contributions to repertoire bias in response to IL-2. For these patients, we noticed that either the T-cells or the B-cells had small Pearson coefficients, but not both, and the difference

between either BCR or TCR populations was significant, $p<0.05$. Furthermore, all non-responding patients showed either a mixed bias in which all seven chains had large bias shifts in the CDR3 receptors or no bias in which the entire repertoire was frozen. A summary of the repertoire bias responses from monitoring over the course of treatment is shown in Figure 4.9.

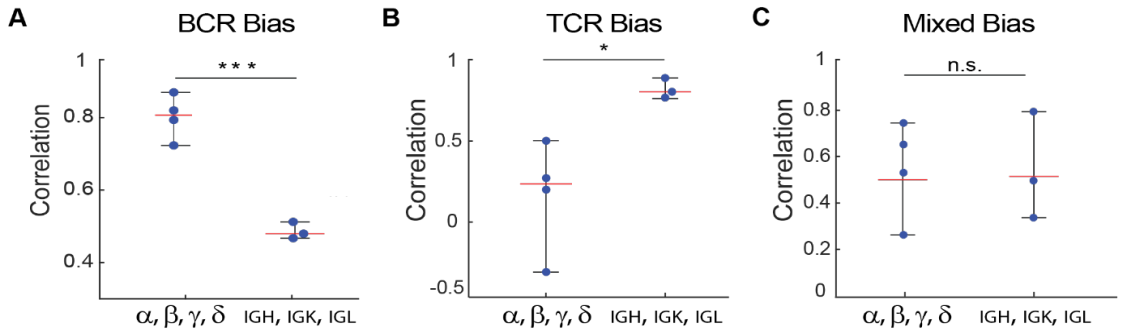

Figure 4.9. Examples of 3 patients with different measured repertoire biases as a result of within-sequence pattern monitoring over three treatment points. (A) An example of BCR bias in which the B-cells take on large entropic changes in receptor sequences while the T-cells conserve patterns from before treatment. (B) An example of TCR bias in which TCRs demonstrate evolution over treatment while BCRS conserve pre-treatment pattern information. (C) An example of mixed bias in which very little pre-treatment patterns are conserved in the sequences of BCRs or TCRs and significant entropic shifting is observed for all chains. Not shown is an example of no bias, in which one patient, who responded as PD to IL-2 treatment, experienced no entropic change in sequences at all.

A summary of measured repertoire bias as a function of patient outcomes to IL-2 treatment in shown in Table 4.4 below. Notably, only responders showed a clear orchestration of BCR and TCRs, allowing either one or the other to take on entropic shifts. In contrast, only non-responders demonstrated a mixed bias, or lack of clear orchestration between chains. Nearly half the responders also demonstrated mixed bias, and future work will lie in the investigation of how chain entropy orchestrations correlate to patient outcomes.

Table 4.4. Patients with TCR, BCR, or mixed bias.

	TCR Bias	BCR Bias	Mixed Bias
Responders	8	2	5
Non-Responders	0	0	5

4.5. Discussion

In this work, we designed a computational framework to analyze amino acid abundance in TCR and BCR clonotypes for early diagnosis of renal cancer response to IL-2, and to gain deeper mechanistic understanding of repertoire dynamics during treatment. We found that classic diversity metrics of patient repertoires occluded cancer signals at Day15, motivating a more in-depth analysis of within-sequence amino acid biases.

The initial challenge of in-depth clonotype analysis stemmed from a lack of a consistent feature space for making comparisons between patients, as the cardinal number of sequences observed in samples varied by individual patients and chains studied. Clonotype sequences from repertoires, analyzed by chain, were thus decomposed into amino acid strings and the frequency of observation of these strings was collected into an intermediate-dimensional space for comparison across patient cohorts, which we called probability maps. Since the feature-space of probability maps scales with the k-mer analysis as 20^k, and given that the number of observed clonotypes can span 10^2 for TRD/TRG to 10^5 for BCRS, k=2 representations were selected resulting in a space of 400 common features. This feature reduction by within-sequence amino acid neighboring pairs formed the basis by which early classification and in-depth entropic response monitoring over treatment could be conducted.

Surprisingly, we found that all seven (TCR and BCR) chains were able to distinguish responders from non-responders by Day15 suggesting an orchestral role of the seven repertoire chains in cancer progression and recovery. This is counterintuitive given the generally accepted function of human IL-2, which is a growth factor inducing proliferation of activated T cells, [28] suggesting that TCRs would carry the clonotypes with detectable disease-related moieties. That BCRs also contained recognizable sequence patterns causal to disease response suggests the presence of B cell interactions with antigen-specific CD4+ helper T cells in response to IL-2.[29] Since the signaling pathway is fortified by IL-2-driven T lymphocyte stimulation with an antigen, it is likely that TCRs will be more effective, and perhaps more generalizable, when this analysis is applied to unseen IL-2-treated patient cohorts. However, since the specificity of T cell clones is static once V(D)J rearrangement occurs (meaning new T cells must be created for every new clonotype observed), the comparable flexibility of BCR variability may play a more integral role in early detection of pre-IL-2 treated patients. Consistent with this, pilot data (unpublished) has shown that pre-IL-2 (Day-14) diversity in BCR heavy chain improves the likelihood that patients will show positive response outcomes to IL-2.

The mechanism of TCR and BCR interactions with high IL-2 environments was further explored in the t-SNE plots generated from the canonical diversity metrics discussed in this work, namely D50, DI, unique CDR3, and Entropy. Greater than the ability to differentiate patient response by sample diversity, was the resilient (to treatment point or patient cohorts studied) higher-order clustering of the diversity metrics by their chains of origin. More specifically, TRA and TRB clustered distinctly from TRG and TRD, reflecting the well-accepted association in the literature of TCR α/β and γ/δ subunits.[30] That TCR α/β and BCR

light, heavy, and kappa chains formed distinct clusters from each other based on diversity of observed CDR3s reflects a distinction in the B- and T- cell types available in the peripheral blood.

Further investigation into the mechanisms of TCR and BCR contribution to patient response/non-response to IL-2 was studied via monitoring the expression of observed clonotypes by chain over three treatment points (pre-IL-2, Day1 IL-2 dose 1, and Day15, IL-2 dose 2). Rather than considering full-length clonotypes, the information content of amino acid bias within full-length clonotypes was studied. This was possible in a large part by the selection of the feature-space for classification which encoded the conditional probabilities for k=2mer amino acid patterns. The evolution of patient repertoires in terms of their information content, was elucidated from pre-IL-2 sequences (from Day-14) that were calculated as first-order Markov chains using the conditional probabilities from Day1 and Day15 samples.

We found a distinct and significant entropic bias in TCR versus BCR populations in patients who responded to IL-2. TCR bias was defined a patient whose TCR chains collectively orchestrated high entropic change over treatment compared to BCRs in which the collective information content of clonotypes remained mostly the same. In contrast, some responder patients showed the opposite phenotype: BCR chains demonstrated high information shifts over treatment while TCRs remained mostly the same. Since the T cell repertoire is under constant turnover driven by antigen presentation and clearance, TCR or BCR bias may be reflected in the regulated, concerted effort of recruiting TILs to the tumor site. In the case of BCR bias, the effective elimination of malignant cells stemming from effector and memory T cell tissue homing, would effectively remove poignant T cells from circulation and thus from our detection

in peripheral blood.[31] In contrast, patients with TCR bias may be in early stages of T cell generation and differentiation, before recruitment. When the immune system and cancer reach an equilibrium state (or worse, an immune escape state), cancer-associated T cells have been shown to accumulate in the blood[32] wherein they may mix with both perfunctory and disease-poignant receptors resulting in a mixed bias response. We speculate that this is the stage in which renal cancer is most diagnosable, and responsible for the mis-fired amino acid abundance detected in the early classified non-responders. The validation of the role of entropic bias in TCR and BCR chains will need to be studied by comparing clonotypes from peripheral blood from those found near tumor tissue.

Section 2 Appendix

The t-SNE algorithm was calculated using sklearn in python.33 Perplexity for determining the number of nearest neighbors was tuned manually until visible clusters were established. The Markov model for classification and disease monitoring from amino acid motifs was generated from scratch using Matlab 2019. Example calculations are provided throughout the text. Descriptions of the computational functions and codes are provided in an appendix section 2.

Table 4.5. Model equations

Symbol /Equation	Variable/Operation
${}^{\text{patient}}\{\mathcal{M}_{day}\}_{chain}$ Example: ${}^{0321}\{\mathcal{M}_{\text{pre}}, \mathcal{M}_{\text{day1}}, \text{and } \mathcal{M}_{\text{day15}}\}_{\partial,\gamma,\alpha,\beta,\text{IGH,IGK,IGL}}$	Probability maps representing stored conditional probability of observing k=2mer amino acid pairs in a sample. This calculation is done for all seven chains in an individual patient sample.
$\mathcal{F}(\theta_{i,j})_k$	Feature matrix with entries, i, j that survive the significance test at each CV iteration where k signifies the data fold.
$\mathcal{F}(\overline{\theta_{i,j}})$	Feature significance filter consisting of feature weights derived from cross-validation. Range [-1, 1]. Features determined to be significant Features determined to be significant in
$\theta_{i,j}$	Feature weights derived from CV with range [-1,1] non-continuous. Keeps significant features and sends non-significant features to zero. (supplement: case where theta = -1
$H_{15_\delta} = H\left(\mathcal{M}_{day_{chain}} \circ \mathcal{F}(\overline{\theta_{i,j}})\right)$	Hadamard product of individual patient probability maps with significance filter to retrieve transformed probability maps that retain significant k-tuplet (k=2) features.
uCDR3	Unique CDR3. The number of unique peptide CDR3s observed within a sample.

$D50 = \frac{N_{uCDR3}^{50\%} * 100}{10{,}000}$	D50 for uCDR3 frequency is in the top 10,000 where $N_{uCDR3}^{50\%}$is the number of uCDR3s that make up 50% of the reads of the to 10,000 uCDR3s. The D50 is the percent of dominant and unique T or B cell clones that account for the cumulative 50% of the total CDR3s counted in the sample. The more diverse a library, the closer the value will be to 50
$(S)Entropy = -\sum_{i}^{10,000} p_i log2(p_i)$	Shannon entropy for a full patient sample where p_i is the frequency of observing the ith CDR3 within the top 10,000 CDR3. The CDR3s with frequencies below the top 10,000 were not included in this calculation
$DI = \frac{\sum_{i=1}^{k} f_{iCDR3}}{\sum_{i=1}^{n} f_{iCDR3}}$	Diversity index where r_i is the frequency of the i^{th} CDR3 and n is the total number of uCDR3s
$\Delta_{\frac{\text{day1}}{\text{pre}}} = \{seq\}_{\Delta_{\frac{\text{day1}}{\text{pre}}}}$ $\Delta_{\frac{\text{day1}}{\text{pre}}} = \log_2\left(\frac{\{\text{seq}\}_{\text{pre}}\|\mathcal{M}_1}{\{\text{seq}\}_{\text{pre}}\|\mathcal{M}_{\text{pre}}}\right)$	Fold change Day-14 versus Day1
$\Delta_{\frac{\text{day15}}{\text{pre}}} = \{seq\}_{\Delta_{\frac{\text{day15}}{\text{pre}}}}$ $\Delta_{\frac{\text{day15}}{\text{pre}}} = \log_2\left(\frac{\{\text{seq}\}_{\text{pre}}\|\mathcal{M}_{15}}{\{\text{seq}\}_{\text{pre}}\|\mathcal{M}_{\text{pre}}}\right)$	Fold change Day-14 versus Day 15

$$\rho = \frac{\mathrm{cov}\left(\Delta_{\frac{\mathrm{day15}}{\mathrm{pre}}}, \Delta_{\frac{\mathrm{day1}}{\mathrm{pre}}}\right)}{\sigma_{\Delta_{\frac{\mathrm{day1}}{\mathrm{pre}}}} \sigma_{\Delta_{\frac{\mathrm{day1}}{\mathrm{pre}}}}}$$ $$\rho = \frac{\sum\left(\Delta_{\frac{\mathrm{d15}}{\mathrm{pre}_i}} - \overline{\Delta_{\frac{\mathrm{d15}}{\mathrm{pre}}}}\right)\left(\Delta_{\frac{\mathrm{d1}}{\mathrm{pre}_i}} - \overline{\Delta_{\frac{\mathrm{d1}}{\mathrm{pre}}}}\right)}{\sigma_{\Delta_{\frac{\mathrm{d15}}{\mathrm{pre}}}} \sigma_{\Delta_{\frac{\mathrm{d1}}{\mathrm{pre}}}} (N-1)}$$	Pearson correlation coefficient used to determine the correlation between the fold change (day1/day14) to the fold change (day15/day1). The range is [-1 to 1] with ρ = 1 is complete correlation and ρ = -1 is anti-correlation (and signifies a shifting of the distribution towards day15).

Section 2 References

1. Han, J. & Lotze, M.T Adaptive Immunity and the Tumor Microenvironment. *Cancer Treat. Res.* **180**, 111–147 (2020).
2. Han, J. & Lotze, M. T. The Adaptome as Biomarker for Assessing Cancer Immunity and Immunotherapy. in *Biomarkers for Immunotherapy of Cancer: Methods and Protocols* (eds. Thurin, M., Cesano, A. & Marincola, F. M.) 369–397 (Springer, 2020). doi:10.1007/978-1-4939-9773-2_17.
3. Yoshida, K. *et al.* Aging-related changes in human T-cell repertoire over 20years delineated by deep sequencing of peripheral T-cell receptors. *Exp. Gerontol.* **96**, 29–37 (2017).
4. Simnica, D. *et al.* T cell receptor next-generation sequencing reveals cancer-associated repertoire metrics and reconstitution after chemotherapy in patients with hematological and solid tumors. *Oncoimmunology* **8**, e1644110 (2019).
5. Lotze, M. T. *et al.* High-Dose Recombinant Interleukin 2 in the Treatment of Patients With Disseminated Cancer: Responses, Treatment-Related Morbidity, and Histologic Findings. *JAMA* **256**, 3117–3124 (1986).
6. Rosenberg, S. A. Interleukin 2 for patients with renal cancer. *Nat. Clin. Pract. Oncol.* **4**, 497 (2007).
7. Rosenberg, S. A. IL-2: The First Effective Immunotherapy for Human Cancer. *J. Immunol. Baltim. Md 1950* **192**, 5451–5458 (2014).
8. Puisieux, I. *et al.* Restriction of the T-cell repertoire in tumor-infiltrating lymphocytes from nine patients with renal-cell carcinoma. Relevance of the CDR3 length analysis for the identification of in situ clonal T-cell expansions. *Int. J. Cancer* **66**, 201–208 (1996).
9. Massa, C. *et al.* Identification of patient-specific and tumor-shared T cell receptor sequences in renal cell carcinoma patients. *Oncotarget* **8**, 21212–21228 (2017).
10. Gerlinger, M. *et al.* Ultra-deep T cell receptor sequencing reveals the complexity and intratumour heterogeneity of T cell clones in renal cell carcinomas. *J. Pathol.* **231**, 424–432 (2013).
11. Giraldo, N. A. *et al.* Tumor-Infiltrating and Peripheral Blood T-cell Immunophenotypes Predict Early Relapse in Localized Clear Cell Renal Cell Carcinoma. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **23**, 4416–4428 (2017).
12. Guo, L. *et al.* Characteristics, dynamic changes, and prognostic significance of TCR repertoire profiling in patients with renal cell carcinoma. *J. Pathol.* **251**, 26–37 (2020).
13. Chow, J. *et al.* Radiation induces dynamic changes to the T cell repertoire in renal cell carcinoma patients. *Proc. Natl. Acad. Sci.* **117**, 23721–23729 (2020).
14. Lotze, M. T. *et al.* Full adaptome repertoire analysis of immunotherapy to predict responsiveness and correlation with CD8-LAG3, sLAG3, and hepatocyte growth factor levels in patients with renal cancer. *J. Clin. Oncol.* **37**, e16116–e16116 (2019).

15. Ho, T. H. *et al.* T-cell receptor (TCR) repertoire in metastatic renal cell carcinoma (RCC) patients treated with first-line vascular endothelial growth factor receptor blockade. *J. Clin. Oncol.* **34**, 501–501 (2016).

16. Motzer, R. J. *et al.* Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. *N. Engl. J. Med.* **373**, 1803–1813 (2015).

17. Topalian, S. L. *et al.* Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer. *N. Engl. J. Med.* **366**, 2443–2454 (2012).

18. Kato, T. *et al.* Peripheral T cell receptor repertoire features predict durable responses to anti-PD-1 inhibitor monotherapy in advanced renal cell carcinoma. *Oncoimmunology* **10**,.

19. Bortone, D. S., Woodcock, M. G., Parker, J. S. & Vincent, B. G. Improved T-cell Receptor Diversity Estimates Associate with Survival and Response to Anti–PD-1 Therapy. *Cancer Immunol. Res.* **9**, 103–112 (2021).

20. Konishi, H. *et al.* Capturing the differences between humoral immunity in the normal and tumor environments from repertoire-seq of B-cell receptors using supervised machine learning. *BMC Bioinformatics* **20**, 267 (2019).

21. Beshnova, D. *et al.* De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. *Sci. Transl. Med.* **12**, (2020).

22. Foers, A. D. *et al.* Classification of intestinal T-cell receptor repertoires using machine learning methods can identify patients with coeliac disease regardless of dietary gluten status. *J. Pathol.* **253**, 279–291 (2021).

23. Ostmeyer, J. *et al.* Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. *BMC Bioinformatics* **18**, (2017).

24. Atchley, W. R., Zhao, J., Fernandes, A. D. & Drüke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci.* **102**, 6395–6400 (2005).

25. Wu, H., Caffo, B., Jaffee, H. A., Irizarry, R. A. & Feinberg, A. P. Redefining CpG islands using hidden Markov models. *Biostatistics* **11**, 499–514 (2010).

26. Posadas, E. M., Limvorasak, S. & Figlin, R. A. Targeted therapies for renal cell carcinoma. *Nat. Rev. Nephrol.* **13**, 496–511 (2017).

27. Murugan, A., Mora, T., Walczak, A. M. & Callan, C. G. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci.* **109**, 16161–16166 (2012).

28. Mingari, M. C. *et al.* Human interleukin-2 promotes proliferation of activated B cells via surface receptors similar to those of activated T cells. *Nature* **312**, 641–643 (1984).

29. Hipp, N. *et al.* IL-2 imprints human naive B cell fate towards plasma cell through ERK/ELK1-mediated BACH2 repression. *Nat. Commun.* **8**, 1443 (2017).

30. Mahe, E., Pugh, T. & Kamel-Reid, S. T cell clonality assessment: past, present and future. *J. Clin. Pathol.* **71**, 195–200 (2018).

31. Obar, J. J. & Lefrançois, L. Memory CD8+ T cell differentiation. *Ann. N. Y. Acad. Sci.* **1183**, 251–266 (2010).

32. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion. *Science* **331**, 1565–1570 (2011).

33. Buitinck, L. et al., 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. pp. 108–122.

Chapter 5. Conclusion

This work envisions a patient-specific future in medicine driven by ML in which medical diagnostics and interventions can flexibly and accurately cater to individual patient needs. For example, 3D bioprinting is an exciting technology for healthcare as the fabrication method can respond to natural heterogeneities in the population by rapidly prototyping size variants of the same implant or transplant design. Additionally, small patient datasets such as Phase 1 clinical data can be leveraged for advanced understanding of patient-response to experimental drugs or procedures. The linchpin for learning from and for the benefit of better patient-specific healthcare is in the bias-variance tradeoff as higher patient-specificity comes at the price of high model complexity. The limit of large features and small data-size necessitates the adaptation of ML techniques to fit atypical data shapes for good and generalizable model predictions.

As previously mentioned, 3D bioprinting is a promising tool both for filling the gap of patients on the organ transplant waitlist and for tailoring custom organs and transplants to natural heterogeneities in the population. Despite the ability to rapidly prototype, the experimental data for optimizing 3D bioprinting is limited by bottlenecks in biological material handling and methods for rapid quantitative assessment of 3D-printed constructs, which retain the same high-absorbing biological matter that pose difficulties for imaging in conventional tissues. In this work, adapting ML techniques for 3D bioprinting looks like embedding domain knowledge of the bioprinting system into a statistical inference model to drive down the sample-size necessary for data-driven learning on the system. By inserting an intermediate layer of supervised relationships within input predictors, the data necessary to perform LASSO for simultaneous feature selection and inference was reduced. The resulting hierarchical model (HML) predicted print outcomes for just 48 prints with an R^2 of ~0.6 in test prints. Furthermore,

the optimal process settings for high-fidelity printed features in alginate were elucidated from process printability maps, which narrowed the vast user design-space to process parameters that minimized HML-predicted error.

The transfer of learning from one print system to the next is crucial for the intricate tissue features of functional biological tissue. Despite dual extrusion and multi-material print systems, transfer learning between parallel physical systems within bioprinting has not been readily attempted to date. In this work, the physical intermediate layer, tuned by cross-validation with model weights refined by LASSO, reflected the dominating physical relationships necessary to predict the print outcomes. As a result, a knowledge bridge was created to transfer learning to parallel printing systems, showing accurate predictions of FRESH-printed collagen ink. Future work involves the accelerated optimization of new materials in a FRESH printing system, including optically transparent support solutions, and multi-material or cell-laden inks.

The impact of ML for predicting high-fidelity prints also lies in the regulatory pathway towards determining the "readiness" of bioprinting technology for customizable medical products such as cell-scaffolds with personalized geometries and constituent cells. The regulation of industrial medical-grade 3D printed products is already extensive.[1] It is imperative that the data requirements for quality control of small batches of 3D bio-printed tissues do not discourage innovation through even higher regulation barriers. ML can aid in rapid regulation and approval of small sample-sizes tailored to patient geometries by predicting how changes in design affect the fidelity of the print.

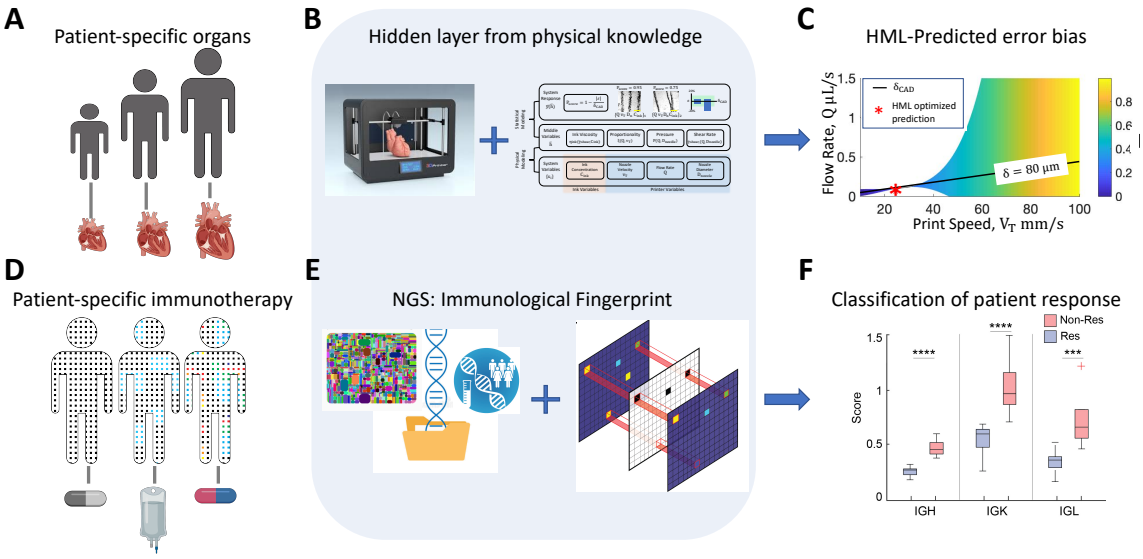

Figure 5.1. Reasoning, methodology, and end-result of ML for small biomedical datasets. (A) There is a need for organs that fit specific patients and not a "one-size-fits-all." (B) A hidden layer is generated from knowing the physical properties of the system. (C) Using HML, we can predict the error bias and optimized printing settings for a given filament diameter. (D) There is a need to understand heterogeneities in individuals for better immunotherapies. (E) Our ML approach creates an immunological disease fingerprint of the patient from their T and B cells. (F) Each patient can then be classified between responders and non-responders to a given therapy.

The Herculean task of merging the vast and ever-growing data from systems biology with the targeted, theoretical knowledge from molecular biology is a task well-suited for the field of ML. Data-driven learning from systems biology, such as big 'omics' data, can excel at forging relationships, correlations, and patterns that can then be interpreted as either relevant or circumstantial by supervised knowledge from related fields. Difficulties arise when attempting to conserve the benefits of Big Data learning towards size-limited datasets that are skewed with large features and limited patient numbers – a common data regime for patient-specific datasets (Figure 5.1B). As a result, ML algorithms are less "plug-and-play" for patient specific models. By adapting ML strategies to Phase 1 clinical data, we show successful model predictions and advanced discovery of feature relationships from a cohort of <50 patients. Early, Day 15

classification of repertoire response to HCQ/IL-2 treatment for 29 patients with progressive renal cancer was achieved to at least 90% accuracy for all seven TCR and BCR chains. Disease fingerprints were discovered by transforming the vast numbers of TCR and BCR predictors into a series of intermediate feature-spaces, which we called probability maps, based on the information entropy within neighboring amino acids in TRC and BCR sequences. Significant neighboring motifs, cross-validated for their ability to split the data into true labels better than random, were ascertained and used as a filter for comparing raw patient data of responder and non-responder patients (Figure 5.1E). As a result, the classification model from this study for predicting patient outcomes from TCR and BCRs currently out-performs recent models which rely on large feature spaces[2] or large numbers of model layers.[3]

Furthermore, data-driven learning can both validate long-standing theories and postulate new variable relationships. Within the classification endeavor, we encountered interesting relationships along the way. We began the search for metastatic recovery signatures from standard diversity metrics and were surprised to find that while TCR and BCR diversity metrics were unable to accurately classify patients, combined together the metrics preserved a mechanistic undercurrent of association between the chains. Clusters were identified for TCR γ/δ and TCR α/β chains, supporting the literature that γ/δ and α/β receptors interact as sub-units.[4] Additional mechanisms were elucidated from the data that will involve future study. Notably, disease monitoring in which first-order Markov models of clonotypes were calculated over three temporal treatment points, revealed an entropy bias for TCR and BCR chains. The bias was defined as a clearly measured entropic shift in either BCR populations or TCR populations, indicating a concerted orchestration of the chains. This orchestration may be

evidence of the antigen-recognition and clearing battleground of metastatic cells, parts of which can be viewed from peripheral blood. While some responders had clear bias, all non-responders showed a mixed bias in which no clear orchestration was present over treatment. We suspect that this repertoire "mis-firing" in which all TCRs and BCRs are circulating in the blood due to lack of recruitment to the tumor cite. Future research is on-going in which this disease monitoring analysis will be compared to clonotypes from tissue samples near the tumor site.

Conclusion References

1. Hourd, P., Medcalf, N., Segal, J. & Williams, D. J. A 3D bioprinting exemplar of the consequences of the regulatory requirements on customized processes. *Regen. Med.* **10**, 863–883 (2015).
2. Foers, A. D. *et al.* Classification of intestinal T-cell receptor repertoires using machine learning methods can identify patients with coeliac disease regardless of dietary gluten status. *J. Pathol.* **253**, 279–291 (2021).
3. Beshnova, D. *et al.* De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. *Sci. Transl. Med.* **12**, (2020).
4. Van Neerven, J., Coligan, J. E. & Koning, F. Structural comparison of alpha/beta and gamma/delta T cell receptor-CD3 complexes reveals identical subunit interactions but distinct cross-linking patterns of T cell receptor chains. *Eur. J. Immunol.* **20**, 2105–2111 (1990).