DESIGN AND OPTIMIZATION OF CEMENTITIOUS SYSTEMS WITH MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in

Department of Chemistry

Christopher M Childs

Carnegie Mellon University

Pittsburgh, PA

May, 2021

Acknowledgments

First, I would like to thank my advisor, Dr. Newell Washburn, for his guidance throughout my PhD studies. I appreciate the many unique opportunities he has provided for me over these years. I also am grateful for the understanding and compassion he has provided me beyond only academic issues, but also in my life. I look forward to the opportunity to work with him in the future.

To my PhD committee, I would like to thank: Dr. Stefanie Sydlik for her guidance and help with polymer synthesis and characterization techniques, Dr. David Yaron for his guidance and help in understanding machine learning as a lost first year graduate student, and Dr. Kimberly Kurtis for her guidance and collaboration for work on cementitious materials.

To my fellow Washburn lab members, I would like to thank you for not only the collaboration, but for the valued friendship. Dr. Kedar Perkins, Dr. Aditya Menon, Jennifer Bone, Joseph Pugar, and Calvin Gang, I appreciate the many work discussions, life discussions, and get togethers along the way. I would also like to thank the master's and undergraduate students who helped with my research along the way: Tia Kirby, Christine Huang, Cheng Zhang, and Jiangnan Zheng.

To my collaborators at Georgia Institute of Technology, I would like to thank you for the help you have provided and wonderful discussions. Particularly, I would like to thank Oğulcan Canbek and Renee Rios for their friendship and also for making me feel welcome in my weeks spent researching in Atlanta. I would also like to thank Aaron Miller for his collaboration on the UHPC project.

To the CMU chemistry department, I would like to thank the staff for their support along the way. I also thank Dr. Barnabás Poczós and Dr. Willie Neiswanger from the CMU Machine Learning Department for their collaboration. I would also like to thank the friends from the chemistry department I have made along the way, including: Stephen Schmidt, Michael Bautista, Site Li, and Dr. Meredith Schervish. I would also like to thank Dr. Lenny Vuocolo, who was incredible to TA for, along with our enjoyable excursions with Colby Yu to Penguins games and restaurants.

To my undergraduate professors: CPT. Bill Kowallis, PhD (USA), Dr. Ken Peters, Dr. Dave Gallaher, and Dr. Monique Hockman, thank you for your support and inspiration to continue my education and providing me a solid educational background to do so.

Before deciding on continuing a college education I would like to thank my former employers in the construction industry for providing me with a job and income. In particular I would like to thank Tony Bevec at A G Bevec Asphalt Maintenance and Paul McKenzie and Lou Casanova at CASMAC Line Striping. I would also like to thank a couple former coworkers who have supported me along the way, in particular Adam Wooten, Bob Hines, and Dorothy Church.

To my friends outside of academia, I thank you for your love and support over the years. Each of you hold a special place in my heart and I am proud to call you my friends. In particular, I would like to mention by name those friends who have been with me through a lot: Matt Ulrich, Sean Staltari, Radford Gunzenhauser, Devin Kelly, Evan Schardt, Captain Kevin S Tramaglini (USAF), Lawton Tellin, Patrick Quirk Jr, Derek Dunsey, Jordan Arroyo, Sean McClelland, Sarah Klos, and Alicia Koledin.

Finally, I would like to thank my family. The love and support you have provided through the years are beyond what any words can describe. I would like to thank my mom, Lisa Huggins, my dad and step-mom, Chris and Abby Childs, my grandparents Dan and Mary Mandus and

Charmaine Childs and Richard Childs and my sister, Caitlin Childs. I would also like to thank my Uncle, Geoff Childs, who played a major part in talking me into going back to college.

The space I need to properly thank all of those individuals who played a part in my life could be the length of a thesis itself. However, I will limit it to this and acknowledge that it is lacking in proper acknowledgment of all the love and support I have received throughout my life. I consider myself extremely blessed to have the friends and family that I have, and consider my greatest achievement as having the honor of knowing each and every one of you.

Abstract

Machine learning has revolutionized disciplines within materials science that have been able to generate sufficiently large datasets to utilize algorithms based on statistical inference, but for many important classes of materials the datasets remain small. After an introduction to various types of ML regression, Chapter 1 introduces a rapidly growing number of approaches to show how embedding domain knowledge of materials systems are reducing data requirements and allowing broader applications of machine learning. Furthermore, these hybrid approaches improve the interpretability of the predictions, allowing for greater physical insight into the factors that determine material properties. A survey of the modern utilization of machine learning for cementitious systems is discussed.

Chapter 2 discusses a background of cementitious systems along with a Hierarchical Machine Learning (HML) approach for improving their workability. The dispersion of cement paste induced by various hybrid polymers was explored. PEGylation of lignin derivatives has been shown to enhance emulsifying and dispersant activities. Here, the effects of anionic grafts were explored for dispersant activity within Portland cement. Kraft lignin and lignosulfonate are two important forms of purified lignin whose chemistries are characterized by low concentrations of carboxylate and high concentrations of sulfonate groups, respectively. The dispersion of cement paste by these hybrid polymers was compared with the PEGylated lignin analogues as well as a leading cement superplasticizer, poly (carboxylate ether) (PCE). Slump values significantly increased for both the PMAA-grafted lignin compared to the other analogues allowing for significant reductions in cement water content, with PMAA-grafted lignosulfonate approaching performance of the commercial PCE and suggesting that graft chemistry has a strong effect on dispersant function. Adsorption, zeta potential, and intrinsic viscosity were measured for the lignopolymer analogues to explore the interplay between lignin and graft chemistries in the mechanism of cement dispersion.

Blending metakaolin (MK), a calcined clay, into portland cement (PC) improves resulting concrete material properties, ranging from strength to durability, as well as reduces embodied CO_2 and energy. However, superplasticizers developed for PC can be inefficient or ineffective for improving the dispersion of PC-MK blends. Chapter 3 introduces a novel machine algorithm which was applied to tailor a superplasticizer to address poor flowability characteristic of 85/15 blends of MK-PC. A HML system was trained on a library of seven superplasticizers using a middle layer, which represents underlying physical interactions that determine system responses, based on polymer contributions to physicochemical forces in both the pore solution and particle surface. Synthesis of the algorithm prediction resulted in a water-soluble polymer with a high intrinsic viscosity and a resultant slump value in a cementitious paste that was comparable with leading poly(carboxylate ether) (PCE) superplasticizers. The results from this study demonstrate the importance of HML as a design tool for the molecular engineering of complex material systems.

Chapter 4 introduces alternative binder chemistries (ABC's) in the form of calcium sulfoaluminate (CSA) cements, which have lower embodied CO_2 compared to portland cement but set rapidly, often within 15 minutes, thus limiting their application. As such, set-retarding admixtures are added to increase the length of time before setting is achieved. These admixtures are typically small organic compounds with high anionic functionality. Retardation is achieved through a complex interplay of mechanisms which involve adsorption onto calcium in the clinker and subsequent prevention of clinker dissolution, complexation with dissociated calcium in the pore solution, and adsorption onto nucleated cement hydration products inhibiting further growth. A cheminformatics based machine learning methodology for the prediction and virtual screening of set retarders for these alternative binder chemistries. Discovery of such compounds is typically achieved through extensive iterative testing that does not ensure optimal solutions. Here, the use of cheminformatics, a data-driven approach used extensively in drug discovery, is demonstrated to identify new set retarders from small datasets for calcium sulfoaluminate (CSA) cements. Based on a sparse training set of 23 molecules containing polar and anionic functional groups, the cheminformatics approach was used to develop a predictive model relating chemical structure to the retarding capability. Then structures of 500,000 compounds were downloaded from a public database, and 365 were predicted to extend set time beyond 1 h. Among these, glyphosate is a commodity chemical that was found to impart a set time of 55 minutes. This cheminformatics approach could be used to develop structure-function relationships and perform rapid virtual screening of chemical admixtures to identify novel high-performance chemical admixtures.

Despite the growing body of work relating to the development of ultra-high performance concrete (UHPC) mixes, the process of designing a UHPC mix is still a highly iterative process. By aggregating previous work on UHPC, Chapter 5 introduces a machine learning model to predict and optimize mix designs based on materials not utilized in the training set. Cement blends are represented in terms of the latent variables of particle packing, water film thickness, and equivalent cement content in order to create a generalizable model. Two rounds of training and testing of were performed utilizing an uncertainty ensemble with a ridge regression while error analysis took place through comparing the Mean Squared Error, a prediction score which ignores Bayesian

probability and compares how well the mean values of the data fit to the best model; and Miscalibration area, a quantification of uncertainty in the model based on calibration techniques. The RMSE for the first iteration was 25 MPa and six blends predicted to obtain UHPC strength were tested. Two of the six blends were found to exceed 100 MPa in compressive strength, while the other four needed major blend modification in order to produce a blend of workable consistency. In the second round, three blends were selected with more tightly bound constraints to ensure mixture workability. All three blends exceeded 100 MPa, although model improvement and more informed feature selection is needed, it was shown that through the incorporation of latent variables, a generalizable model could be obtained to predict novel UHPC compositions with a disparate source of materials.

Designing Limestone Calcined Clay Cements (LC3) is a challenge because the factors that are correlated with strength are anticorrelated with workability. Chapter 6 presents a ML methodology for designing LC3 compositions for materials commonly found in North America subject to CO2 constraints. A hierarchical machine learning approach is performed to represent cement composition as a latent middle layer which can encode any arbitrary composition from a bottom, compositional, layer. Cement blends are represented in terms of their particle packing and water film thickness in both the prediction of workability and strength, while various parameters encoding particle-particle and pore solution forces for superplasticizers in the workability model. A random forest model was utilized in the prediction of workability returning an R^2=0.93 on the training set and R^2=0.81 on the test set. A gaussian process regression was utilized in the prediction of strength providing a final model training score with an R^2=1.0 and showing high generalizability to a test set with an R^2=0.97. Analysis of the effect of changes in the compositional variables was visualized through the gaussian process regression giving a posterior probability distribution for all predictions. Finally, these models were combined with a linear model capturing the CO_2 release for every compositional variable. A genetic algorithm was performed in order to predict a Pareto front corresponding to the points of maximum strength and workability, with minimized CO_2, predicting novel blends containing various ratios in combining different sizes of the supplementary cementitious materials.

Finally, Chapter 7 presents conclusions and future directions.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A_2	Second Osmotic Virial Coefficient
ABC	Alternative Binder Chemistry
AFt	Ettringite
AH_3	Gibbsite
AIBN	Azobisisobutyronitrile
ANN	Artificial Neural Network
ATMP	Nitrilotris(methylene)triphosphonate
BO	Bayesian Optimization
bpy	2,2’-bipyridine
C_2S	Dicalcium Silicate
C_3A	Tricalcium Aluminate
C_3S	Tricalcium Silicate
C_4AF	Tetracalcium Alumino Ferrite
CA	Citric Acid
CASH	Calcium aluminum silicate hydrate
CPM	Compressible Packing Model
CSA	Calcium Sulfoaluminate
CV	Cross Validation
D50	Average Particle Diameter
DFT	Density Functional Theory
DLVO	Derjaguin-Landau-Verway-Overbreek
DMF	Dimethylformamide
DOE	Design of Experiment
EBriB	Ethyl 2-bromoisobutyrate
ECFP	Extended-Connectivity Fingerprint
ECI	Effective Cluster Interactions
Glyphosine	N,N-bis(phosphonomethyl) glycine
GP	Gaussian Process
GWP	Global Warming Potential
HML	Hierarchical Machine Learning
HSP	Hansen Solubility Parameter
ITZ	Interfacial Transition Zone
KL	Kraft Lignin
Lasso	Least Absolute Shrinkage and Selection Operator
LC3	Limestone-calcined clay cement
LCA	Life Cycle Assessment
LOOCV	Leave-one-out-cross-validation
LS	Lignosulfonate
MAA	Methacrylic Acid
MD	Molecular Dynamics
MK	Metakaolin
MK-PC	Metakaolin-Portland Cement

ML	Machine Learning
MSE	Mean Squared Error
MSM	Dimethyl Sulfone
NaSS	Sodium Styrene Sulfonate
NSC	Normal Strength Concrete
OPC	Ordinary Portland Cement
PAA	Phosphonoacetic Acid
PBTCA	2-phosphonobutane-1,2,4-tricarboxylic Acid
PC	Portland Cement
PCE	Polycarboxylate Ether
PEGMA	Poly(ethylene glycol) Methacrylate
PIML	Physics Informed Machine Learning
PMAA	Poly(methacrylic acid)
PMDETA	N,N,N’,N’’,N’’-pentamethldiethylenetriamine
PMIDA	Phosphonomethyliminodiacetic Acid
PNS	Polynaphthalene Sulfonate
PSPMA	Poly(3-sulfopropyl methacrylate)
PSS	Poly(styrene sulfonate)
QC	Quantum Chemical
QSAR	Quantitative Structure Activity Relationship
RANS	Reynolds-averaged Navier Stokes
RED	Relative Energy Difference
RMSE	Root Mean Squared Error
s	Electrosteric Force
SCM	Supplementary Cementitious Material
SMILES	Simplified Molecular-input Line-entry System
SPMA	3-Sulfopropyl Methacrylate
SSA	Specific Surface Area
*t*BMA	Tert-Butyl Methacrylate
TFA	Trifluoroacetic Acid
TOC	Total Organic Carbon
UHPC	Ultra-high Performance Concrete
w/c	Water to Cement Ratio
w/cm	Water to Cementitious Material Ratio
α	Lasso L_1 Penalization Hyperparameter
ζ	Zeta Potential
η	Intrinsic Viscosity
θ	Fraction of Adsorbate on Adsorbent
λ	Lasso L_1 Penalization Hyperparameter

Chapter 1. Embedding Domain Knowledge for Machine Learning of Complex Material Systems[1]

1.1. Introduction

Many important materials are defined by a single underlying interaction or force, which allows modeling using analytical expressions having relatively few parameters. Examples include ferromagnets, where the magnetism is described by the exchange interactions between spins,[1] and elastomers, where the resistance to deformation is due to polymer chain entropy.[2] In contrast, the properties of complex materials are determined by multiple competing forces, the interplay of which lead to a rich diversity of physical properties and performance characteristics. Complex materials, such as complex fluids,[3] metal alloys,[4] and catalysts,[5] are ubiquitous, but predicting their properties remains a significant challenge.

Machine learning (ML) is a diverse collection of powerful techniques utilized to identify relationships in data, allowing for modeling and optimization of complex systems. With rapidly growing datasets available, ML has become a robust methodology applied across many materials disciplines and has been increasingly incorporated in conjunction with the Materials Genome Initiative.[6,7,8] However, the traditional methods of machine learning are based only on statistical inference, requiring large datasets to develop predictive models that connect composition and processing with properties. While some disciplines within materials science, such as metallurgy[9] or heterogeneous catalysis,[10] have developed methods for high-throughput experimentation to produce sufficiently large datasets, most disciplines still use traditional

[1] This chapter includes work that was published and reformatted:
Childs, C. M. & Washburn, N. R. Embedding domain knowledge for machine learning of complex material systems. *MRS Commun.* **9**, 806–820 (2019).

methods of materials preparation and analysis, precluding the use of ML methods designed for Big Data.[11,12]

While, in the case of systems described by an exactly known relation, a physical law is better utilized, in complex systems, a single physical relation may not exist, but several relations could underlie the system. These constituent physical relations can be utilized in conjunction with ML to learn the interplay of interactions within these complex systems, and with the increasing use of data-driven approaches, science has utilized traditional ML to predict molecular solubility[13,14,15], discover new thermodynamically stable materials[16], and determine highly accurate interatomic potentials.[17] While a simple, single physical law could be learned through statistical inference techniques with a small dataset of, say, 10 datapoints, systems defined through several complex relations can require the use Big Data in the range of thousands or more datapoints to accurately model. In general, the amount of data needed depends on the ratio of datapoints to the number of features. If a small number of features can effectively model the data, then fewer datapoints are needed. However, as the complexity of the system increases, the higher number of features needed to model the system would require a higher number of datapoints to effectively model.

It is still possible to use the tools of ML on small datasets, but this requires the development of hybrid algorithms that embed domain knowledge in order to develop predictive models that relate system variables to system responses, thus narrowing the search space that data-driven models must explore. In the context of materials science, domain knowledge can take a number of forms, and here four different types are surveyed: (1) physicochemical properties, (2) similarity, (3) system properties, and (4) physical laws and empirical equations.

This introduction will discuss these types of domain knowledge as well as specific examples of how they are implemented in ML algorithms.

1.2. Response Surfaces and Machine Learning

The general task here is to understand how changes in experimental or system variables change the properties of the system. For complex systems, high costs and time demand limit data collection to small datasets,[18] with examples including adhesives, agrochemicals, pigments, paints, coatings, lubricating oils, paper, and pharmaceuticals, all of which have limitations in the amount of data that can be acquired under realistic resource assumptions.[19,20] Design of experiment (DOE) approaches are a common tool for estimating the response surface of such systems based on the incomplete exploration of the variable space. In the DOE approach, system features are systematically varied so that outputs can be mapped as a continuous response surface. These observations allow the discovery of correlations between features and produce a function that can be subsequently optimized.[21] A common method utilized with the DOE is a full-factorial design. In this approach, each of the k-factors (features) are tested at n levels. For example, if n=2, the design will measure the value of the response as maximum and minimum values of each feature against the maximum/minimum values of every other feature resulting in 2^k simulations being performed. The benefits of DOE allow for correlations between features to be discovered and a response surface to be mapped, but disadvantages of the approach include limits to the non-linearity of the surface being mapped, exponential growth of the system with increasing feature size, and large uncertainty of the surface response mapping in areas that are untested, as illustrated schematically in Figure 1.1. Establishing a technique to embed domain knowledge would allow for the response to be better predicted between test points by training the algorithm to understand the relationships between system variables and how they determine system responses.

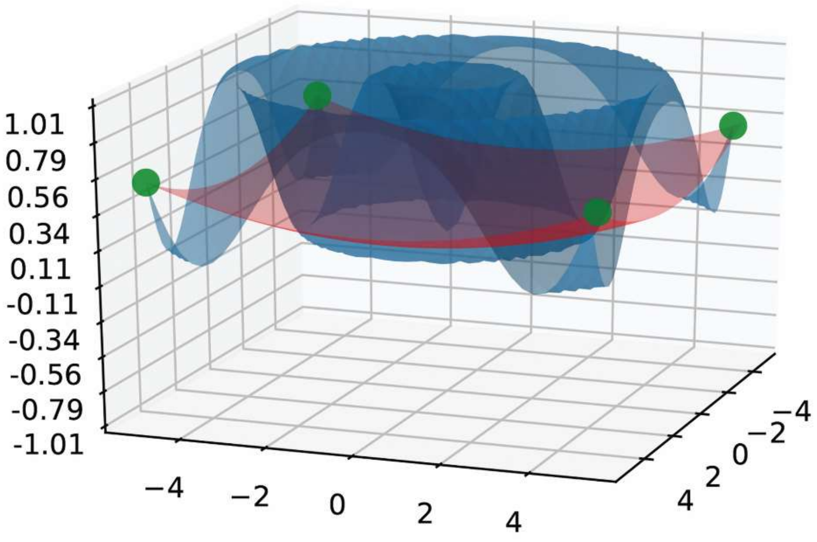


Figure 1.1. Illustration of two possible response surfaces that fit the four points of a training set shown as green circles. The red concave surface represents a simple model from the DOE where the response surface was modeled with a second-order polynomial. This approach fails to capture complex underlying interactions, which could take place throughout the domain space, as shown in blue.

ML techniques are applied across systems as diverse as clinical medicine, facial recognition, self-driving automobiles, and scientific fields such as cheminformatics and bioinformatics. The development of such diverse uses of ML has been predicated on using Big Data (datasets routinely including millions of points) enabled by acquisition over large populations, such as high-throughput measurements.[22] In recent advances, ML techniques utilizing image recognition for detecting melanoma have outperformed medical experts in diagnosing.[23] Where sufficient data are present, ML algorithms have the capability of learning relationships between inputs and outputs; however, unlike human learning, traditional ML techniques relying on raw features perform poorly at determining relationships utilizing small datasets.[24] Foundational research has previously been applied toward developing a general unified theory of learning in conjunction with ML and emphasized the need for the development of multi-strategy techniques for learning on various types of data, but these have not been widely implemented for use in small datasets.[25]

Fields such as cognitive neuroscience, which began incorporating early ML techniques to model human learning as early as the 1960's,[26] have attempted to better predict the relationship between inputs and outputs by explicitly including causal relations, allowing algorithms to learn on small data.[24,27] For example, recent research has demonstrated human-level performance for "one-shot" recognition of handwritten characters on sparse datasets. Causal knowledge was included through parsing characters into the training set by each 'pen stroke,' allowing the model to consider how the characters were drawn to identify each character.[28] The causal relationships create a hierarchical model where domain knowledge can be explicitly included or learned to be included in future models. This approach relates the inputs of a system to a middle layer as opposed to traditional ML where inputs are related through statistical inference in hidden layers (which may or may not be explicitly included in the algorithm) as shown in Figure 1.2.

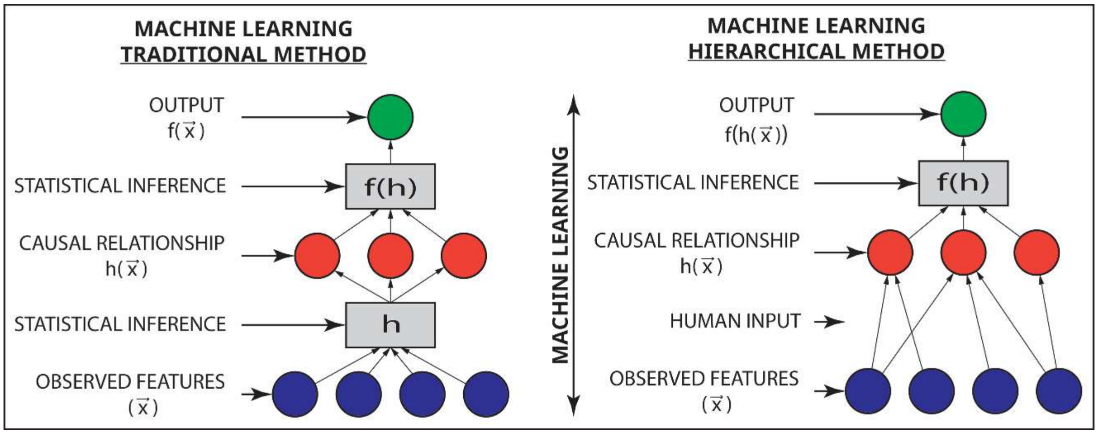


Figure 1.2. The traditional model directly predicts the outputs from the inputs with learned causal relationships. The hierarchical model allows for the incorporation of domain knowledge through human input before predicting outputs.

However, incorporating domain knowledge into ML algorithms needs to be performed in a way which does not block the discovery of unexpected solutions. An heuristic example of this is the statement that “trucks can’t drive over water,” but this domain knowledge ignores the possibility of the water freezing in winter.[29] Translating real-world phenomena, such as temperature variations, into ML problems requires expert knowledge so as to not eliminate possibilities that could be discovered using data-driven approaches.[30]

Research areas in which small datasets are commonly generated face two challenges in adopting ML techniques. The first, as discussed, is the challenge in making accurate predictions using methods based strictly on statistical inference. The second is that experimental design in a laboratory setting tends to be sequential and driven by intuition, and experimental parameters naturally tend toward values that lead to maximizing (or minimizing) an objective function. This artificially limits the dataset to a narrow section of the input variable space and does not adequately train the algorithm on the range of system responses that can be generated, limiting as Jain et al. described, the “completeness” of the dataset.[31] These point to the importance of embedding domain knowledge in ML algorithms to use causal relationships as hypotheses. While incorporating causal relationships to human learning is difficult to formalize[27], extensive and formalized knowledge of them are the basis of scientific model building. Through an expert's appropriate incorporation of domain knowledge, a hypothesis-space of the domain knowledge can be created on which to train.[32] From this perspective, ML can be an effective tool for the unbiased analysis of complex material systems with small datasets.

Here, we discuss methods of embedding physical knowledge into ML algorithms. Instead of experimentally defined descriptors being directly used as the sole inputs for ML, selected inputs also include physical factors modeling the system. This embedded knowledge can take the form of correlative relations, such as identifying similarity metrics, empirical relations in the form of physical equations, or embedding exact relations such as invariance properties. It will be shown that by embedding ML algorithms with these techniques, small data can be utilized to effectively model a complex material system.

1.3. Methods

ML encompasses many varying algorithms. To provide an understanding to some of the basic algorithms being surveyed through this review, an overview of these methods will be provided.

1.3.1. Cross-Validation

ML centers on the development of algorithms that improve in the performance of a given task with experience.[22] Experience, in terms of scientific research, is synonymous with acquiring experimental or computational data. A collection of inputs, sometimes termed features or descriptors, is utilized to improve the prediction accuracy through ML, and to establish a relation between inputs and outputs, a direct relation in the form of a regression can be predicted. There are numerous variations of regression methods used in ML, and each attempt to achieve the same goal: establishing an accurate model of the response surface for a complex system on test data. To establish a best-fit to unseen (test) data, regression models have their parameters learned on training data through a process called cross-validation as shown in Figure 1.3.

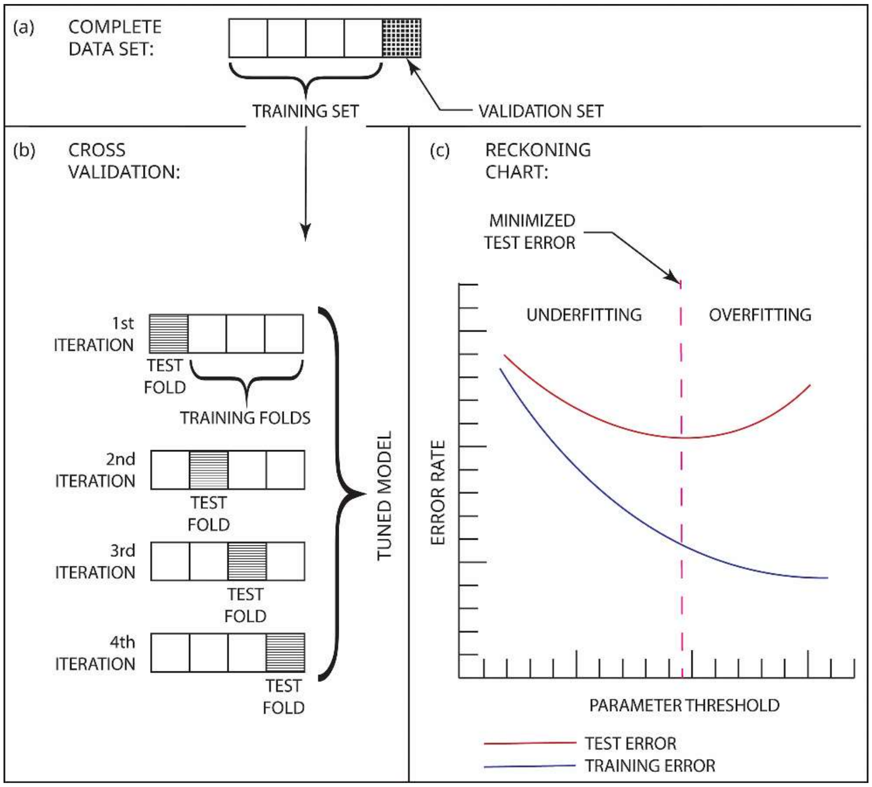


Figure 1.3. (a) The first step of cross-validation is to partition the complete dataset into both a training set in which parameters will be optimized against and a validation set in which error between experimental data and predicted regression is compared. (b) This shows an example of a *k*-fold cross-validation where the training set is split into training and test folds. Each subset of the sample is treated as a test fold through one of the iterations and an optimum parameter is chosen which minimizes the test error. (c) The parameter chosen falls in between a regime of underfitting and overfitting, where underfitting only exhibits small correlation to the dataset and overfitting minimizes the training error but begins to show an increase in test error. This optimized parameter would provide the most accurate fit to the validation set.

However, validity of the model is not assessed until it is applied to new data, and this also requires that it has an appropriate level of complexity. As illustrated in Figure 1.3, while training error decreases monotonically with model complexity, the accuracy of predictions on test data reaches a minimum at intermediate complexity – overly simple models do not predict training or test data, but overly complex models are only valid on training data. Data-driven approaches require both the optimization of model parameters and model complexity. These basic criteria also hold true when domain knowledge is incorporated.

1.3.2. Linear Regression

Ordinary linear regression is a well-known statistical technique that fits a linear model to a dataset. A common approach to optimizing the fit of data to a linear regression is through minimizing the sum of squared residuals, or the sum of squared distance between a datapoint and the best-fit line. These linear least-squares regressions have a trivial solution for the β coefficients where the equation for a line is:

$$Y = X\beta \qquad \text{Eq. 1.1}$$

and where the β coefficients corresponding to a minimized sum of squared residuals are:

$$\beta^{OLS} = \ argmin \ \|Y - X\beta\|_2^2 \qquad \text{Eq. 1.2}$$

The trivial solution to β is:

$$\beta = \ (X'X)^{-1}X'Y \qquad \text{Eq. 1.3}$$

Linear regressions can also be regularized and include additional parameters which are optimized through cross-validation. The least absolute shrinkage and selection operator (Lasso) was originally developed in 1996 by Tibshirani.[33] Similar to ordinary linear regression, Lasso finds the minimized sum of squared residuals with an additional penalty term penalizing the L_1-norm, or sum of the absolute value of β coefficients:

$$\beta^{Lasso} = argmin \, ||Y - X\beta||_2^2 + \lambda||\beta||_1 \qquad \text{Eq. 1.4}$$

The lambda parameter is learned through cross-validation and the value assigned to predicting the lowest test error is selected. At this point, non-important features have their β coefficients reduced to zero and are eliminated from the calculated equation. If the penalty term in Lasso is set to zero, the regression is reduced to an ordinary linear regression where all β coefficients contribute to the final calculated equation.

1.3.3. Neural Networks

Neural networks, a robust form of ML sometimes referred to as deep learning networks, are algorithms used for pattern recognition and regression.[34] These networks are composed of multiple layers of neurons: an input layer, hidden layer(s), and an output layer. The output of each neuron is passed to those in the next layer with an associated weight. The hidden layers and output layer also contain associated activation functions that are tuned while learning relationships between the input and the output using the training data. Activation functions can either be linear or nonlinear, with a common nonlinear choice being a sigmoid function. After the initial input is forward-processed through the network, the output is compared to the target (correct) output value, and the associated error is calculated between the target value and output value from the neural network. A methodology known as backpropagation is then applied to minimize the error, updating the weights between neurons. Through these updates, the error is minimized to an optimal value determining the best-fit curve to the data. By increasing the number of neurons and activation-function complexity in the network, the complexity of the regression increases. Figure 1.4 provides a schematic of a two-layer network (perceptron) and a representation of a higher complexity network.

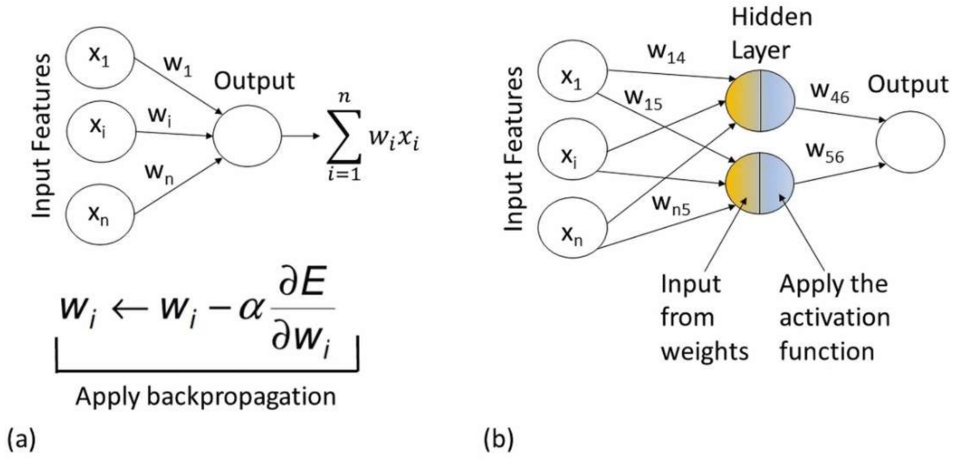


Figure 1.4. (a) A simple neural network is known as a perceptron. This perceptron is made up of an input layer connected to an output layer through a linear activation function. A method of backpropagation, stochastic gradient descent, is shown where α is the learning rate hyperparameter, which controls how fast the weights are updated with the associated error through every iteration. The perceptron shown is the same as performing ordinary linear regression. (b) This shows a more complex network with an included hidden layer. This hidden layer takes inputs and feeds them into an activation function before predicting an output.

1.3.4. Random Forest

Random forests are made up of an ensemble of decision trees, and for regression purposes, the output of all these trees has an average taken to produce a singular best-fit regression for the entire collection of trees. Each tree is split utilizing bootstrapping, a technique of resampling the dataset many times, with replacement, to test on each tree.[35] Each bootstrap tree is split on the collection of all features utilizing a random subset of the features with replacement for every split. The special case where the random subset of features is equal to the total number of features is known as bagging.[36] A recursive binary process of splitting, where the feature to split on is determined through splitting on the feature which maximizes the reduction in some metric such as the sum of squared errors, is continued at each node either until reaching a user-specified end or until complete separation has occurred.[37] To control overfitting, the maximum depth of the trees along with post-processing methods, such as pruning, can be utilized. The number of trees can also be controlled and is typically enough when the prediction made by the forest is approximately equal to the prediction of a subset of the forest.[38] Regression trees work through recursive partitioning and therefore an exact function cannot be fit to the model. The predictions to the regression are determined the ensemble averaging of each tree in a random forest as shown in Figure 1.5.

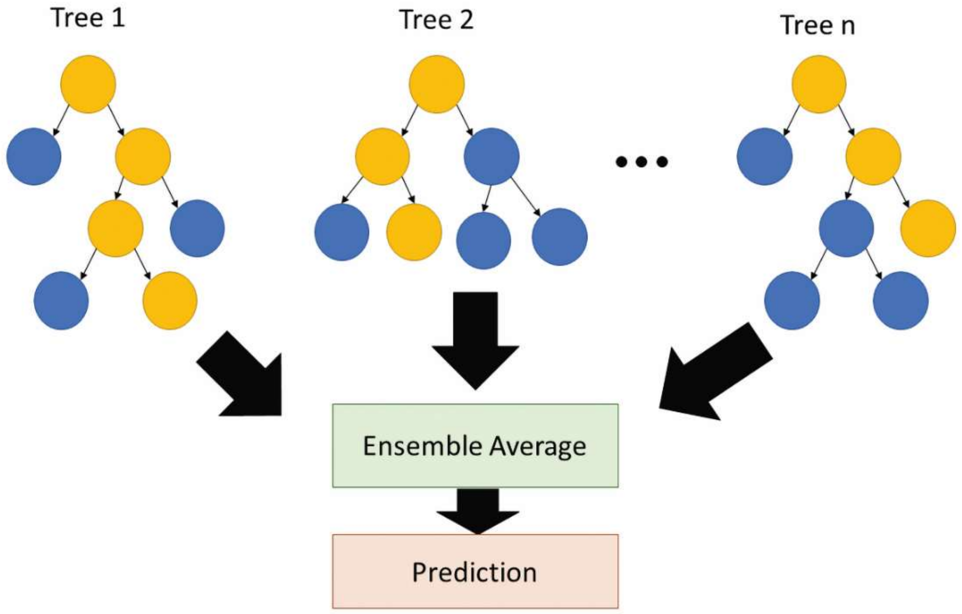


Figure 1.5. A collection of *n* decision trees is created according to bootstrap and bagging techniques. The yellow path in each tree corresponds to the same predicted output. The ensemble average of these collected outputs corresponds to the final prediction as produced from a random forest regression.

1.3.5. Bayesian Probability and Gaussian Processes

A Gaussian process (GP) is based on the concept of including a Bayesian probabilistic approach. Bayesian probability determines the posterior probability of an event based on the probabilities of the factors constituting the event – prior probabilities and likelihood of these occurring as shown in Figure 1.6.

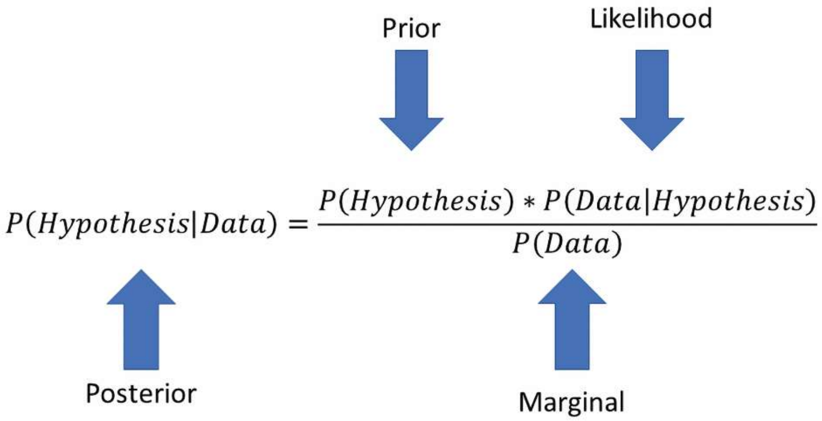

Prior
Likelihood
$P(Hypothesis|Data) = \frac{P(Hypothesis) * P(Data|Hypothesis)}{P(Data)}$
Posterior
Marginal

Figure 1.6. Bayes' theorem states that the posterior probability of a hypothesis occurring given data can be calculated through the prior probability a hypothesis is true before collecting data, the likelihood that the data are collected given that the hypothesis is true, and the probability of collecting that data under all possible hypothesis.

The posterior probability is updated as part of the GP resulting in a regression with error terms represented across the range of the regression. While regression and neural networks are parametric approaches, in that the shape of the curve being fit to the data is defined, GP is a non-parametric approach, meaning that the best shape and the best fit curve are both learned through the regression process.[39] As a GP is a Bayesian approach, the entire process is defined by a mean and covariance function. While a Gaussian distribution is defined over a set of vectors, a GP is defined over a set of functions, that is:

$$f \sim GP(\mu, \Sigma) \qquad \text{Eq. 1.5}$$

where the optimized function, *f*, is distributed as a GP across the mean function μ and the covariance function $\sum$. Multiple covariance functions can be utilized, but the chosen function is a method of relating similarity. Covariance provides insights into how related two variables are through utilizing varying similarity metrics such as linear, squared exponential, periodic, or any combination of covariance functions, the covariance function learns similarity through mathematically representing assumptions about the function being learned.[40] Utilizing the prior mean and covariance, an infinite number of possible functions is introduced to the feature space of the data. As data are introduced to the GP, the mean and covariance are updated at that point with the associated covariance function defining the error in areas without training data introduced, that is:

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma_* \\ \Sigma_*^T & \Sigma_{**} \end{bmatrix}\right) \qquad \text{Eq. 1.6}$$

Where a joint distribution is created between the function fitting the training data, *f*, and the function fitting the test data, f_*, across the normal distribution across these functions. As the training function *f* is known, the conditional distribution, $(f_*|f)$, can be updated to solve for the new mean and error in the updated function.[39] Upon cross-validation, optimized parameters in the covariance function are selected defining the smoothness of the function and error between the training points.

1.4. Applications to Material Systems

1.4.1. Embedding Physicochemical Properties

In predictive models of molecular materials, quantum chemical (QC) parameters have been used to solve chemical Hamiltonians[41], learn force field parameters[42], represent crystal structures[43], and predict heats of formation[44]. Embedding physical knowledge with QC inputs improves predictive capabilities of ML algorithms and accelerates calculations[45], but these calculations add computationally intensive steps in ML analysis. However, more accurate descriptors can enable the use of simpler ML tools and smaller datasets.

The two important ML methodologies utilized in this paper were linear regression and artificial neural networks (ANNs), which both permit embedding of physicochemical properties to improve predictive capabilities. Linear regression is considered to be a simpler formalism, but its performance can rival that of the more complex ANNs when provided with a more accurate feature set. QC calculations[46,47], such as electronic charge distribution, dipoles, vibrational frequencies, and reactivity, have been used to increase the accuracy of predictions on ionic liquids. Mehrkesh and Karunanithi[48] utilized QC-predicted descriptors of the symmetrical value, which describes packing density between anions and cations, as well as the distance between anions and cations, anion volume and surface area, and the dielectric energies of anion and cation liquids along with the temperature as a system condition. Each of these properties relates to important factors that impact the mobility of ions within a complex ionic liquid. Values of viscosity were aggregated from 20 sources for a total of 131 data points, where 48 were used to train and the rest were utilized as a validation set. In this work, multivariate linear regression was implemented to establish the best fit to the data.

The predicted equation for predicting ionic liquid viscosity was:

$$\ln(\eta) = 16.5\sigma + 2.2R_t + 0.01Vol_A - .03Area_A \\ -0.03T - 15.8Di_A - 48.1Di_C - 15 \qquad \text{Eq. 1.7}$$

Where η is the ionic liquid viscosity, σ is the symmetrical value, R_t is the distance between anions and cations, Vol_A is the anion volume, $Area_A$ is the anion surface area, T is the temperature, and Di is the dielectric energy for both the anion and cation. An average relative error comparing to the validation set was found to be 7.40%. These results were found to decrease error from 18.0% on the same group of tested compounds that were considered from purely thermodynamic considerations.[49]

Using more complex ANNs, Fatehi et al.[50] utilized data purely from chemical structures using the features of molecular weight and structural information along with the pressure and temperature as inputs to a neural network to predict the viscosity of ionic liquids without QC features. Experimental data were aggregated from 28 sources encompassing 736 individual datapoints over a range of experimental conditions. An ANN was utilized to relate the model weights and structural features to the viscosity for six families of ionic liquids. For the ionic liquid system, 44 combinations of neural networks with a varying number of hidden neurons and activation functions were tested. The most accurate neural network was selected, having an average error of 1.31% on the validation set, which was 10% of the original data withheld.

In comparing different approaches to modeling the same system, Kalidindi and De Graef[51] have discussed the need for standardization and for data-driven protocols for the transferability of system models, which is the capability of learning on one system and applying the learned model to a separate system. Fatehi et al. and Mehrkesh et al. utilized datasets of different types. While Fatehi et al. utilized a larger dataset and a robust ANN, the QC embedded system from Mehrkesh et al. worked on a smaller set of training data only utilizing a linear regression. Despite this, it was found that the linear regression model, embedding QC features, fit the data well and was able to extrapolate among multiple types of ionic liquid systems. Still, no testing on a standardized dataset for comparison was performed. Creating methods that allow for the comparison of the same systems or allow for transferability between systems is a necessary step forward. A recent approach to resolve this issue has utilized statistical methods to determine the best points to collect data, so that small data can be utilized for valuable analysis.[12] Through embedding physicochemical properties, linear regression on a small dataset was able to predict ionic liquid viscosity with low validation error, establishing that ML can effectively embed physical features as inputs.

1.4.2. Embedding Similarity

Similarity is a measure of how well common features will relate to a common output. For example, comparing similar sequences found in the protein database with the sequences of a protein of unknown structure has allowed for the improved prediction of secondary protein structures.[52] Similarity can be embedded within ML frameworks through the use of a metric, and some common metrics utilized to determine similarity include distance metrics such as Euclidean or Manhattan distance or through cosine similarity. Choosing the appropriate metric to properly model the system being studied also requires expert knowledge to effectively embed the physical properties of the system being studied.

Similarity at a molecular scale operates under the assumption that the more similar molecules are, the closer their structure-property relation is.[53] Hansen solubility is one such property based on underlying assumptions of similarity, and both traditional and QC-embedded ML approaches have been utilized in the prediction of Hansen solubility. Hansen Solubility Parameters (HSPs) are an extension of the Hildebrand solubility parameter, which define the intermolecular attraction between molecules as the square root of the cohesive energy density.[54] The more similar the parameters are the higher the likelihood of compounds being soluble, an extension of the "like dissolves like" definition of solubility. To better predict the solubility of compounds, Hansen split the Hildebrand parameter into three metrics: the dispersion parameter δ_d, the polar parameter δ_p, and the hydrogen-bonding parameter δ_h where the sum of these three parameters is equal to the Hildebrand parameter.[55] Hansen empirically fit a model where the solubility of a system can be determined through a similarity metric, the relative energy difference (RED):

$$RED = \frac{R_a}{R_o} \text{ ; } R_i^2 = 4\delta_d^2 + \delta_p^2 + \delta_h^2 \qquad \text{Eq. 1.8}$$

If *RED* is <1 then the substances are considered to be miscible, and at >1 they are insoluble. Hansen solubility is widely utilized during the design of new drugs and other material formulations along with predictions for the χ parameter in Flory-Huggins polymer solution theory.[56] Various statistical approaches have been utilized in predicting Hansen solubility. Much the same as viscosity, one such approach utilizes the concepts of group contribution methods and chemical structure.[57,58] In a recent study, Sanchez-Lengeling et al.[15], embedded physical knowledge through the algorithm in the prediction of HSP. The model features included direct inputs through chemical structure in terms of chemical fingerprints, and QC determined data including charge density, electrostatic quantities, and molecular shape and size information. Domain knowledge was embedded into the system in knowing that HSP values are based on similarity metrics. To embed this concept into the model, structural, charge density, electrostatics, and molecular shape information were each placed into their own respective vectors. Euclidean distances were then determined as a measure of similarity through the use of the sum of four- squared exponential covariance function for each of the four vectors utilizing a GP regression. GPs are a useful method of ML and have the added advantage of including rigorous uncertainty estimates for the predictions.[59] In most ML algorithms, error analysis is commonly based on calculating the mean squared error of how well the trained regression fits the validation set. Within GP, error analysis is achieved through applying uncertainty to the predicted regression surface itself. Considering that most scientific relations are assumed to follow normal distributions, this creates an automatic connection with standard research practices.

It was found that through embedding similarity through use of a squared exponential covariance function to model similarity between molecular properties, the GP model utilized by Sanchez-Lengeling et al. was able to predict the Hansen parameters. The model found a R^2 of 0.70, an average accuracy of 80%, and an average modeling error of 2.58 $MPa^{0.5}$ between predicted and actual Hansen values. In terms of determining a *RED* ratio, this is a model capable of many correct predictions. The GP was compared to other ML methods in this paper such as Kernel Ridge, Lasso, and a Regularized Greedy Forest. The GP technique that embedded the similarity metric outperformed all of these techniques in the prediction of each of the solubility parameters. This technique illustrates that even inclusion of expected correlative relations, in this case similarity between inputs, can be utilized in a model.

A second example within materials science which utilizes similarity to improve ML results is predictions made on cluster expansions. Cluster expansions are widely used for the prediction of material properties which display substitutional disorder, such as crystals. However, when studying low symmetry systems such as nanoparticles, the computational cost involved in density functional theory (DFT) calculations, which need to consider many-body interactions, have limited capability in quickly predicting properties of the material. This has led to the importance of developing ML algorithms which can accurately measure cluster expansions for various materials on small datasets.[60]

Mueller and Ceder applied a Bayesian method to cluster expansions in order to embed physics.[61] The coefficients for the cluster expansion are known as effective cluster interactions (ECIs). The aim of cluster expansion techniques is to predict the appropriate ECI values that best reproduce the property value. The authors utilized a Bayesian approach to apply *a priori* belief on the nature of the ECIs. Three separate conditions were embedded into algorithm:

1. Property predictions should be close to that predicted by a simple model.
2. The greater number of sites in a cluster and the greater distance between sites should lead to a smaller ECI.
3. ECIs for similar clusters should be similar values.

To satisfy the first condition, the prior means of the ECIs are set to zero. The second condition is satisfied through the use of a Gaussian distribution to model the ECIs with a variance assigned as a decreasing function of the number of sites in a cluster and distance between clusters. Finally, the third condition is satisfied through the use of a second prior distribution where the variance is the function of similarity between clusters, where the more similar the clusters are the closer the predicted ECIs. The above three conditions were applied utilizing Bayes' theorem to derive a maximum likelihood estimate for the ECIs. Various functions were utilized as a representation for updating the variance.

Ten thousand cluster expansions of 201-atom cuboctahedral Ag-Au nanoparticles were created for testing. It was found that the best results on test data were still obtained utilizing the largest set of candidate clusters and training set size. However, in the absence of these large datasets (as could be the case for complex nanomaterials with computational limits to DFT calculations) the best regularization functions embedding similarity performed half to two times better while only utilizing half the training data compared to cluster expansions where similarity is ignored. The use of physically meaningful prior distributions successfully limited the size of data needed for effective modeling.

1.4.3. Embedding System Properties

Physical systems operate under a set of laws that govern their responses, and these laws remain true for any physical system. Data representation, a form of converting raw data into suitable features, becomes an essential component to quickly and accurately utilize ML.[62] As opposed to being learned through data-driven approaches, embedding these system properties into ML has been shown to improve results on smaller material datasets. Stress-strain relationships are an essential part of understanding material deformation. In solid mechanics, one such field studied is crystal elasticity. Utilizing molecular dynamics (MD), stress-strain relations of crystal deformation can be predicted with high precision,[63] and the Cauchy-Born model establishes a relation between atomic pair potentials and continuum models for elastic deformation in crystals.[64] Ling et al.[63] utilized high-throughput MD analysis of 15,000 data points to perform ML analysis in determining crystal deformations under applied loads. All simulations were performed on a nickel crystal at 0 K, and the results predicted agreement with the Cauchy-Born rule for homogenous deformations on a perfect crystal. Under these assumptions, the system can be treated as having invariant properties. For crystal deformations, invariant properties imply that the system does not change upon rotation around the tensile stress axis, only stretching. The strain energy function, *W*, is the function of the deformation gradient, *F,* which is a nine-component tensor composed of the derivatives of atomic positions in a material to their reference positions. The strain energy function can be differentiated with respect to *F* in order to determine the strain, and the goal of this ML regression was an attempt to discover the function *W*, relating material deformation to strain.

Two separate approaches were explored in this work: a traditional one based on statistical inference and a physically embedded approach. Knowing that invariance properties are essential to the behavior of the system, both approaches built assumptions of invariance into the training set, already demonstrating the importance of expert knowledge applied to material systems. The traditional ML approach artificially increased the number of training examples through manually transforming a system's rigid body or cubic rotations; this will not change the output of the model--as rigid body and cubic rotations are invariant properties--but the artificial rotations allow the model to learn on more examples to recognize the invariance through training. For this procedure, the nine components of *F* were utilized as inputs. This deformation technique of artificial rotations has been utilized in prior ML algorithms in order to recognize hand-written images.[65] The artificial deformation is introduced through the nine components of *F* as features multiple times from multiple angular orientations of the same structure. This allows the algorithm to learn invariance properties on its own. The hierarchical approach utilized a symmetry basis set of six invariant relationships, based on the two invariance properties of *W* for a cubic crystal. Kambouchev et al.[66] determined the six basis set equations which fully define the invariance of rigid body rotations and other rotations and inversions based on the cubic symmetry group. Ling et al. utilized these equations to embed the invariance properties of *F* into the model in order to learn *W* from a causal, physics-based relationship as opposed to artificial deformation of the data. These six invariant properties were utilized as inputs into the ML algorithms.

To assess the effectiveness of embedded similarity, two ML algorithms were utilized for regression: neural networks and random forests. Both the neural network and random forest models were trained to find the strain energy function, *W*, by utilizing the nine components of *F* or the six-component invariant vector as inputs and the stress as the output. After testing both types of invariance, the approach embedding rigid body and cubic invariance was shown to have lower validation error than any of the traditional ML approaches for 2D and 3D transformations. Training on embedded domain knowledge arrives at an error <3% compared to MD simulation. Ordinary linear regression of the physics embedded data itself only performed 7% worse than the next most accurate neural network model trained with traditional techniques. Another issue that became apparent was the extremely large data size of random forests trained with the traditional approach, which was large enough that it could not be trained on 3D transformation data. It also had a significant impact on the training times for both random forests and neural networks with a traditional method trial time of 434 hours in the neural network as opposed to 0.6 hours in the invariant neural network. Embedding system properties was thus shown to improve training time and reduce error as compared to purely data-driven learning.

Although not strictly a material property, it is important to mention that the same authors and others also looked at turbulence modeling. Similar procedures were followed as to predicting strain in materials. A basis of invariants was modeled under physical assumptions that certain changes in orientation do not affect Reynolds stress anisotropy. One approximate method of calculating Reynolds stress is through the Reynolds-averaged Navier Stokes (RANS).[67] Recent research has focused on improving RANS measurements through the incorporation of ML techniques by including what is called physics-informed ML (PIML) and has shown success compared to traditional techniques.[68,69]

1.4.4. Embedding Physical Equations

The methods reviewed up to this point have looked at QC properties that can be obtained in a high-throughput approach and in embedded invariance properties, which are well defined to a system. For systems with physicochemical relations that are not well defined, selections of appropriate descriptors allow for causality to be discovered. Ghiringhelli et al. applied this principle towards discovering physical causality through the use of a feature selection for predicting energy differences in semiconductors.[70] This was performed through the utilization of Lasso.

Due to the L_1 penalty term, Lasso has the benefit of providing a natural method of feature selection as non-important features are suppressed to zero. Lasso performed as well as more advanced ML techniques in predicting the same energy differences in semiconductors with fewer descriptors.[70] The benefit of feature selection in materials allows for easy optimization of the discovered equation and a possibility to reduce tests to only those necessary for successful ML to be performed on a system.

A combination of the prior reviewed approaches for embedding domain knowledge into systems has led to another ML approach, hierarchical machine learning (HML). Unlike PIML, HML incorporates physical domain knowledge through utilizing equations to predict the physical interactions that determine the properties or responses of a complex material system.

True for all methods of embedding domain knowledge, the appropriate selection of physical interactions driving a system must at least include the essential descriptors behind any experiment. HML is a methodology that has been successfully implemented to extremely small datasets with the appropriate descriptor selections. The first implementation of HML, as described by Menon et al.,[71] embedded domain knowledge into polymer dispersants to probe their effect in a cement based system. Magnesium oxide is a popular non-setting model of portland cement and dispersant design was the first system modeled using HML.[72] The generic model of an HML system is shown in Figure 1.7.

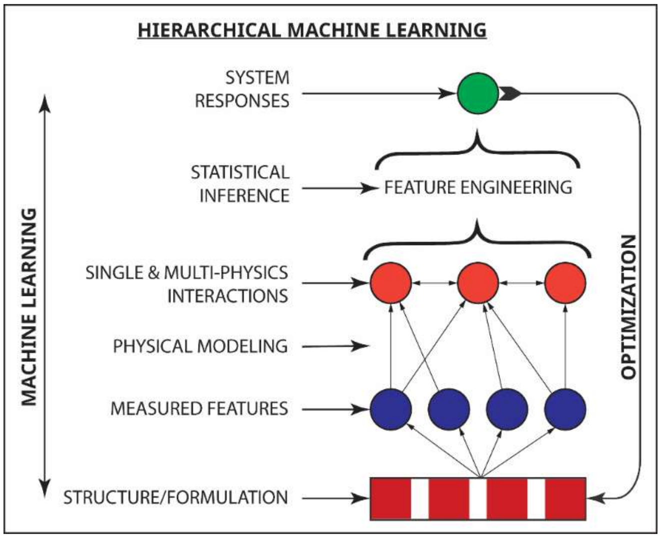


Figure 1.7. Akin to the hierarchical approach shown in Figure 1.2, HML parameterizes a complex system in terms of either system structure or formulation. A bottom layer of observed features is directly measured. This bottom layer is related to the middle layer through embedding domain knowledge into the system through physical equations. This bottom to middle layer allows for the embedding of system physics without it having to be learned in a blackbox approach. The middle layer is utilized as inputs into the statistical learning techniques, such as Lasso, with cross-terms included so that multi-physics interactions are accounted for. After learning, an equation based on physical interactions utilizing ML the upward movement is complete. The predicted equation can be reparametrized in terms of the initial material structure or formulation on the bottom layer and optimized, as shown by the downward arrow.

A polymer dispersant used in cement systems is known as a superplasticizer. Superplasticizers are utilized to reduce yield stress without an increase in water addition, which reduces strength.[73] Individual measurements of adsorption (θ), zeta potential (ζ), sedimentation experiments (s), intrinsic viscosity (η), and the osmotic second virial, A_2, were performed. Each of these individual measurements was related to the associated force through physical equations, which define how superplasticizers reduce yield stress within cementitious systems. For example, an increase in viscous force was assumed to vary linearly with free polymer concentration (c_o) so that:

$$\eta_{pol} = c_o(1 - \theta)[\eta_{pol}(\vec{x})] \qquad \text{Eq. 1.9}$$

Where $[\eta_{pol}(\vec{x})]$ is parameterized in terms of polymer structure.

A library of 10 polymers was utilized for the training set. Each polymer was parameterized in terms of their chemical group composition. Upon representing each polymer in terms of their respective force interactions through connecting the bottom to middle layer, an input of these interactions along with their squared and cross-terms was included in order to increase the system dimensionality and incorporate multi-physics interactions into the hypothesis-space of the material. The selected regression technique utilized was a regularized linear regression, Lasso.

Lasso has the added benefit of a natural form of feature selection in order to reduce the final predicted regression to a line of only the physical interactions most contributing towards dispersibility effects. The regression resulted in Eq. 1.10 for the change in yield stress $\Delta\tau$:

$$\Delta\tau = -0.26\zeta^2 + 1.93\eta + 0.13\pi^2 - 1.00\,\eta\pi + 1.40\eta s + 0.03\pi\zeta \qquad \text{Eq. 1.10}$$

After solving for the regression, optimization was performed in order to parameterize chemical composition in terms of superplasticizer structure. The optimized polymer structure was found to correspond with a novel polymer, a polycarboxylate-grafted lignosulfonate. The synthesized polymer approached reductions in yield stress similar to those of the leading class of superplasticizers, polycarboxylate ethers, showing that embedded physical equations within ML are capable of learning fits and optimizing systems.

1.5. Overviews and Conclusions

Early ML approaches in cement research came through the utilization of artificial neural networks (ANNs). In 1998, Yeh[74] utilized an ANN on a collection of 1030 concrete samples from 17 data sources. Utilizing the compositional proportions of cement, SCMs, aggregate, water, age, and admixture, they were able to reproduce results with a R^2 slightly higher than 0.90 on both training and test sets. These models outperformed traditional regression analysis. Numerous ML algorithms have since been applied to this published dataset comparing how well their algorithm can predict compressive strength.[75]

In 2001, Haj-Ali et. al[76] utilized an ANN to predict sulfate-induced concrete expansion as a function of water to cement ratio (w/c), cement C_3A content, and time. While improving on the predictions produced by analytical equations, despite having over 8000 datapoints from 51 different mixtures, the ANN was as much as 0.3% in the predicted expansion percentage on the test set. Although a solution to improve ML algorithms is to increases the size of data, these samples take over 40 years to complete making iterative testing impractical.

The commonality of prior ML techniques for cement-based materials is that the inputs are all based on the compositional space of the cement-based materials. Training on the compositional space causes lack in both interpretability and generalizability of the model. For example, while training on the Yeh dataset, Dutta et. al[77] found gaussian process regression (GPR) along with two other probabilistic ML models had a R^2 on the test sets of 0.95. Upon performing a sensitivity analysis was also performed indicating that 'cement content' was the most important factor in determining cement compressive strength. This, however, provides no microstructural or chemical insight into the development of cement strength.

Also, with the disparity of source materials, cement composition, and processing conditions there are limitations on the ability to generalize a model trained with specific source materials to a disparate dataset of new materials. Young et. al[78] trained multiple models on the Yeh dataset, finding typical results as other research with a R^2 around 0.85 on the validation set. However, they also trained on a dataset consisting of 9994 datapoints of various compressive strengths from job-site mixtures across the United States. Despite having ten times as many datapoints as the Yeh dataset, the best performing model was only able to achieve an R^2 of 0.60. Similarly, Chou et. al trained on five compressive-strength datasets from various nations around the world. Utilizing various ensemble methodologies to find the best performing ML algorithms, each dataset was found to have a different ML model minimizing the error in prediction. This exemplifies the lack of generalizability in current ML models for cement-based systems. For generating new understanding that improves the design, utilization, and performance of cement-based systems, models which account for the disparity in source materials and processing conditions must be developed.

While ML can successfully model datasets easily available which consist of over 10,000's of unique samples, smaller datasets require embedded domain knowledge to improve ML modeling.[79] Although many ML models for cement-based materials have developed and trialed new algorithms for property prediction, one of the most important factors in developing a successful ML algorithm is domain-specific feature engineering which has been lacking in the study of these materials.[80]

For further examples of domain-specific feature engineering, Bone et. al[81] utilized HML to physically relate ink concentration and print parameters to viscosity, shear rate, and proportionality in a model to predict and optimize print fidelity in 3D printed biopolymers. Similarly, Menon et. al[82] predicted the young's modulus of polyurethanes through relating the molecular composition to a middle layer of physicochemical properties utilizing stochastic simulation and molecular modeling. Finally, cheminformatics approaches can be utilized as a chemistry-specific methodology to relate molecular structure to function. Cheminformatics approaches have been utilized for such tasks as predicting the glassy transition temperature of polymers,[83,84] drug discovery,[85] and improving quantum mechanical calculations for molecules.[86] These methodologies are developed for cement-based materials throughout this thesis.

In Chapter 2 an analysis of the working mechanisms of an optimized polycarboxylate-grafted lignosulfonate polymer is discussed and compared to the predictions of the physicochemical forces discovered as the working mechanisms of cement dispersion through the Lasso equation. An overview of cementitious systems is also introduced in order to provide for an understanding into cement rheology and hydration for the development of machine learning techniques through the rest of the thesis. Chapter 3 utilizes HML to model rheology in a Metakaolin (MK) modified cementitious system. MK is a calcined clay additive for improved cement strength. However, MK decreases cement workability.[87] MK-Portland Cement (MK-PC) systems were studied to design an effective dispersant utilizing HML.[88,89] A similar set of procedures was followed for the ML algorithm as in the MgO study and optimization followed on the resulting force-driven equation as predicted by Lasso. It is interesting to note that in both systems, an increase of viscous force, η, was an important determinant in maximizing the slump. Upon optimization, a novel styrene sulfonate-methacrylic acid-poly(ethylene glycol) methacrylate copolymer was synthesized. In line with the Lasso prediction, this optimized polymer resulted in having a higher intrinsic viscosity than any of the training set, yet still imparted high workability to MK-PC systems.

In Chapter 4, ML is applied to develop a cheminformatics model for the virtual screening of molecules for use as set retarding admixtures in Calcium Sulfoaluminate (CSA) cements. Concepts of molecular similarity are utilized as a form of domain knowledge in order to compare molecular structure towards set retarding capability. The commercial compound glyphosate was identified through virtual screening and predicted to have a set time of 61 +/- 26 min, and experimentally glyphosate was found to impart a set time of 55 minutes at a cost that is competitive with the leading set retarder for CSA cement, citric acid.

In Chapter 5, a HML model was developed for the design and optimization of ultra-high performance concrete utilizing a Bayesian uncertainty ML ensemble. Modeling was performed to allow generalizability to produce high strength cements from locally sourced materials.

In Chapter 6, a HML model was developed to represent cement composition as a latent middle layer which can encode any arbitrary composition from a bottom, compositional, layer. Cement blends were represented using latent variables of particle packing, water film thickness in both the prediction of workability and strength, while various parameters encoding particle-particle and pore solution forces for superplasticizers in the workability model. A random forest model was utilized in the prediction of workability, while gaussian process was found to provide accurate predictions of strength. Analysis of the effect of compositional variables was modeled through the gaussian process regression giving a posterior probability distribution for all predictions. Finally, these were combined with a linear model capturing the CO_2 release for all compositional variables. A genetic algorithm was used to identify a Pareto front corresponding to the points of maximum strength and workability, with minimized CO_2. These blends were reproduced and tested, showing the models ability to predict blends with 50% less CO_2 emissions as compared to ordinary cement, with set workability and strength requirements.

In the utilization of ML for science, physical theories can be utilized to improve models in conjunction with data as opposed to being relearned through only the incorporation of data. ML approaches for physical systems have quickly developed to incorporate scientific fields. Starting with purely statistical analysis of raw data, techniques have progressed over time to include expert knowledge, incorporate physical parameters as features, incorporate metrics of correlation between data, discover physical laws which model simple systems, and now approach a level as to embed multiple physical laws to predict outputs from complex systems. The guiding hypothesis behind this thesis' research is displayed in Figure 1.8, where through the utilization of domain knowledge, small datasets can be effectively modeled and transferred via the utilization of this parameterization.

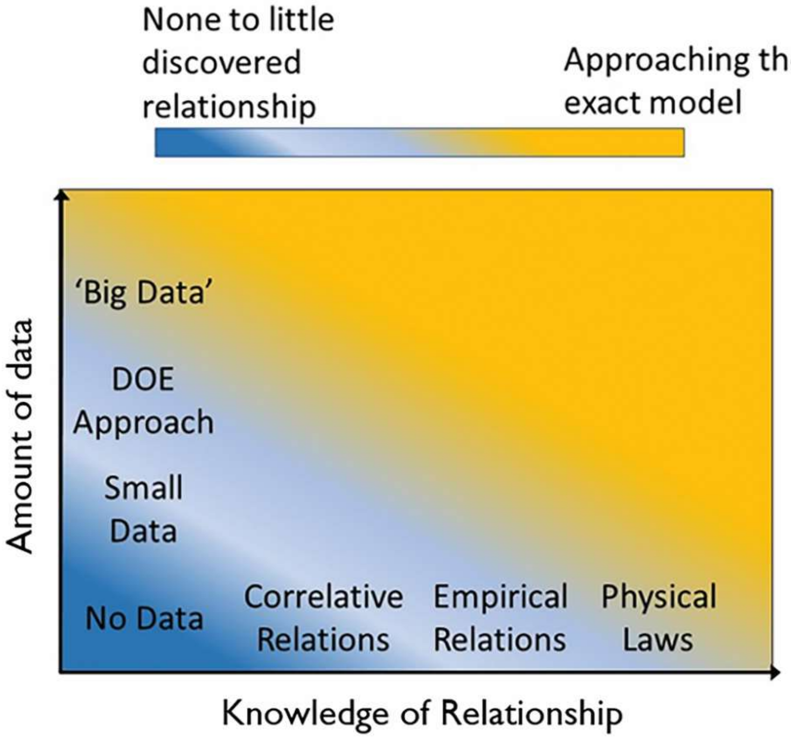


Figure 1.8. The goal of modeling a complex system is to predict the best relation possible. Knowing the physical laws of the system and interactions between the features allows an analytical model to be proposed based on physics alone, as shown on the horizontal axis. Statistics can also allow for the prediction of a good model if sufficient data are collected as is shown on the vertical axis. For many complex systems, neither of these is achievable. By having a combination of both physical and statistical modeling, a good model is able to be predicted and illustrates the importance of embedding physical knowledge into a system.

With complex materials being expensive, time-intensive systems on which to test, methods to improve the cost-effectiveness and reduce the time to find an optimized system are essential. Deploying these hybrid physical-statistical approaches, more accurate modeling and relationships can be extrapolated and understood using domain knowledge embedded machine learning.

1.6. References

[1] Kittel, C. Physical Theory of Ferromagnetic Domains. *Rev. Mod. Phys.* **1949**, *21* (4), 541–583.

[2] Flory, P. J. Molecular Theory of Rubber Elasticity. *Polym. J.* **1985**, *17* (1), 1–12.

[3] Stickel, J. J.; Powell, R. L. Fluid Mechanics and Rheology of Dense Suspensions. *Annu. Rev. Fluid Mech* **2005**, *37*, 129–178.

[4] DeCost, B. L.; Francis, T.; Holm, E. A. Exploring the Microstructure Manifold: Image Texture Representations Applied to Ultrahigh Carbon Steel Microstructures. *Acta Mater.* **2017**, *133*, 30–40.

[5] Saravanan, K.; Kitchin, J. R.; von Lilienfeld, O. A.; Keith, J. A. Alchemical Predictions for Computational Catalysis: Potential and Limitations. *J. Phys. Chem. Lett.* **2017**, *8* (20), 5002–5007.

[6] Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3* (1), 54.

[7] Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (1), 011002.

[8] McDowell, D. L.; Kalidindi, S. R. The Materials Innovation Ecosystem: A Key Enabler for the Materials Genome Initiative. *MRS Bull.* **2016**, *41* (04), 326–337.

[9] Qin, M.; Lin, Z.; Wei, Z.; Zhu, B.; Yuan, J.; Takeuchi, I.; Jin, K. High-Throughput Research on Superconductivity. *Chinese Phys. B* **2018**, *27* (12), 127402.

[10] Gani, T. Z. H.; Kulik, H. J. Understanding and Breaking Scaling Relations in Single-Site Catalysis: Methane to Methanol Conversion by Fe IV □O. *ACS Catal.* **2018**, *8*, 975–986.

[11] Ramakrishna, S.; Zhang, T.-Y.; Lu, W.-C.; Qian, Q.; Sze Choon Low, J.; Heiarii Ronald Yune, J.; Zong Loong Tan, D.; Bressan, S.; Sanvito, S.; Kalidindi, S. R. Materials Informatics. *J. Intell. Manuf.* **2018**.

[12] McBride, M.; Persson, N.; Reichmanis, E.; Grover, M.; McBride, M.; Persson, N.; Reichmanis, E.; Grover, M. A. Solving Materials' Small Data Problem with Dynamic Experimental Databases. *Processes* **2018**, *6* (7), 79.

[13] Kuhne, R.; Ebert, R.-U.; Schuurmann, G. Model Selection Based on Structural Similarity-Method Description and Application to Water Solubility Prediction. *J. Chem. Inf. Model.* **2006**, *46* (2), 636–641.

[14] Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why Are Some Properties More Difficult to Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *J. Chem. Inf. Model.* **2008**, *48* (1), 220–232.

[15] Sanchez-Lengeling, B.; Roch, L. M.; Perea, J. D.; Langner, S.; Brabec, C. J.; Aspuru-Guzik, A. A Bayesian Approach to Predict Solubility Parameters. *Adv. Theory Simulations* **2019**, *2* (1), 1800069.

[16] Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in

Unconstrained Composition Space with Machine Learning. *Phys. Rev. B* **2014**, *89* (9), 094104.

[17] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6* (12), 2326–2331.

[18] Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials Discovery and Design Using Machine Learning. *J. Mater.* **2017**, *3* (3), 159–177.

[19] Rowe, R. C.; Colbourn, E. A. Neural Computing in Product Formulation. *Chem. Educ.* **2003**, *8* (03), 1–8.

[20] Tanco, M.; Viles, E.; Ilzarbe, L.; Alvarez, M. J. Implementation of Design of Experiments Projects in Industry. *Appl. Stoch. Model. Bus. Ind.* **2009**, *25* (4), 478–505.

[21] Montgomery, D. C. *Design and Analysis of Experiments*, 8th ed.; Wiley: New York, 2012.

[22] Jordan, M. I.; Mitchell, T. M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **2015**, *349* (6245), 255–260.

[23] Haenssle, H. A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Ben Hadj Hassen, A.; Thomas, L.; Enk, A.; et al. Man against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842.

[24] Griffiths, T. L.; Baraff, E. R.; Tenenbaum, J. B. Using Physical Theories to Infer Hidden Causal Structure. *Proc. Annu. Meet. Cogn. Sci. Soc.* **2004**, *26* (26), 500–505.

[25] Michalski, R. S. TOWARD A UNIFIED THEORY OF LEARNING: An Outline of Basic Ideas. In *First World Conference on the Fundamentals of Artificial Intelligence*; Paris, 1991.

[26] Carbonell, J. G.; Michalski, R. S.; Mitchell, T. M. An Overview of Machine Learning. In *Machine Learning: An Artificial Intelligence Approach*; Michalski, R. S., Carbonell, J. G., Mitchell, T. M., Eds.; Springer-Verlag: Berlin, 1983.

[27] Tenenbaum, J. B.; Griffiths, T. L.; Kemp, C. Theory-Based Bayesian Models of Inductive Learning and Reasoning. *Trends Cogn. Sci.* **2006**, *10* (7), 309–318.

[28] Lake, B. M.; Salakhutdinov, R.; Tenenbaum, J. B. Human-Level Concept Learning through Probabilistic Program Induction. *Science (80-.).* **2015**, *350* (6266), 1332–1338.

[29] Frawley, W. J.; Piatetsky-Shapior, G. *Knowedge Discovery in Databases*, 1st ed.; The MIT Press: Cambridge, 1991.

[30] Sacha, D.; Sedlmair, M.; Zhang, L.; Lee, J. A.; Peltonen, J.; Weiskopf, D.; North, S. C.; Keim, D. A. What You See Is What You Can Change: Human-Centered Machine Learning by Interactive Visualization. *Neurocomputing* **2017**, *268*, 164–175.

[31] Jain, A.; Hautier, G.; Ping Ong, S.; Persson, K. New Opportunities for Materials Informatics: Resources and Data Mining Techniques for Uncovering Hidden Relationships. *J. Mater. Res* **2016**, *31* (8), 977–994.

[32] Wu, Q.; Suetens, P.; Oosterlinck, A. Integration of Heuristic and Bayesian Approaches in a Pattern-Classification System. In *Knowledge Discovery in Databases*; Piatetsky-Shapior, G., Frawley, W. J., Eds.; The MIT Press: Cambridge, 1991; pp 249–260.

[33] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B J. R. Stat. Soc. Ser. B J. R. Stat. Soc. B* **1996**, *58* (1), 267–288.

[34] Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (5), 468–481.

[35] Mooney, C. Z.; Duval, R. D. *Bootstrapping A Nonparametric Approach to Statistical Inference*; Sage Publications, Inc: Newbury Park, CA, 1993.

[36] Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

[37] Xu, M.; Watanachaturaporn, P.; Varshney, P. K.; Arora, M. K. Decision Tree Regression for Soft Classification of Remote Sensing Data. *Remote Sens. Environ.* **2005**, *97* (3), 322–336.

[38] Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2/3*, 18–22.

[39] Rasmussen, C. E. Gaussian Processes in Machine Learning. In *Advanced Lectures on Machine Learning*; Bousquet, O., von Luxburg, U., Rätsch, G., Eds.; Springer-Verlag: Berlin, 2003; pp 63–71.

[40] Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*, 2nd ed.; MIT Press: Cambridge, 2006.

[41] Li, H.; Collins, C.; Tanha, M.; Gordon, G. J.; Yaron, D. J. A Density Functional Tight Binding Layer for Deep Learning of Chemical Hamiltonians. *J. Chem. Theory Comput.* **2018**, *14* (11), 5764–5776.

[42] Li, Y.; Li, H.; Pickard, F. C.; Narayanan, B.; Sen, F. G.; Chan, M. K. Y.; Subramanian, ⊥; Sankaranarayanan, K. R. S.; Brooks, B. R.; Roux, B. Machine Learning Force Field Parameters from Ab Initio Data. *J. Chem. Theory Comput.* **2017**, *13*, 4492–4503.

[43] Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties. *Phys. Rev. B* **2014**, *89* (20), 205118.

[44] Hu, L.; Wang, X.; Wong, L.; Chen, G. Combined First-Principles Calculation and Neural-Network Correction Approach for Heat of Formation. *J. Chem. Phys.* **2003**, *119* (22), 11501–11507.

[45] von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angew. Chemie Int. Ed.* **2018**, *57* (16), 4164–4169.

[46] Gardas, R. L.; Coutinho, J. A. P. A Group Contribution Method for Viscosity Estimation of Ionic Liquids. *Fluid Phase Equilib.* **2008**, *266* (1–2), 195–201.

[47] Paduszyński, K.; Domańska, U. Viscosity of Ionic Liquids: An Extensive Database and a New Group Contribution Model Based on a Feed-Forward Artificial Neural Network. *J. Chem. Inf. Model.* **2014**, *54* (5), 1311–1324.

[48] Mehrkesh, A.; Karunanithi, A. T. New Quantum Chemistry-Based Descriptors for Better Prediction of Melting Point and Viscosity of Ionic Liquids. *Fluid Phase Equilib.* **2016**, *427*, 498–503.

[49] Preiss, U.; Bulut, S.; Krossing, I. In Silico Prediction of the Melting Points of Ionic Liquids from Thermodynamic Considerations: A Case Study on 67 Salts with a Melting Point Range of 337 °C. *J. Phys. Chem. B* **2010**, *114* (34), 11133–11140.

[50] Fatehi, M.-R.; Raeissi, S.; Mowla, D. Estimation of Viscosities of Pure Ionic Liquids Using an Artificial Neural Network Based on Only Structural Characteristics. *J. Mol. Liq.* **2017**, *227*, 309–317.

[51] Kalidindi, S. R.; De Graef, M. Materials Data Science: Current Status and Future Outlook. *Annu. Rev. Mater. Res.* **2015**, *45* (1), 171–193.

[52] Magnan, C. N.; Baldi, P. SSpro/ACCpro 5: Almost Perfect Prediction of Protein Secondary Structure and Relative Solvent Accessibility Using Profiles, Machine Learning and Structural Similarity. *Bioinformatics* **2014**, *30* (18), 2592–2597.

[53] Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. *Sci. Rep.* **2013**, *3* (1), 2810.

[54] Vandenburg, H. J.; Clifford, A. A.; Bartle, K. D.; Carlson, R. E.; Carroll, J.; Newton, I. D. A Simple Solvent Selection Method for Accelerated Solvent Extraction of Additives from Polymers. *Analyst* **1999**, *124*, 1707–1710.

[55] Hansen, C. *Hansen Solubility Parameters*; CRC Press, 1999.

[56] Lindvig, T.; Michelsen, M. L.; Kontogeorgis, G. M. A Flory – Huggins Model Based on the Hansen Solubility Parameters. *Fluid Phase Equilib.* **2002**, *203* (1–2), 247–260.

[57] Albahri, T. A. Accurate Prediction of the Solubility Parameter of Pure Compounds from Their Molecular Structures. *Fluid Phase Equilib.* **2014**, *379*, 96–103.

[58] Stefanis, E.; Panayiotou, C. Prediction of Hansen Solubility Parameters with a New Group-Contribution Method. *Int. J. Thermophys.* **2008**, *29* (2), 568–585.

[59] Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*; New York, 2016.

[60] Cao, L.; Li, C.; Mueller, T. The Use of Cluster Expansions To Predict the Structures and Properties of Surfaces and Nanostructured Materials. *J. Chem. Inf. Model* **2018**, *58*, 2401–2413.

[61] Mueller, T.; Ceder, G. Bayesian Approach to Cluster Expansions. *Phys. Rev. B* **2009**, *80* (024103).

[62] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547–555.

[63] Ling, J.; Jones, R.; Templeton, J. Machine Learning Strategies for Systems with Invariance Properties. *J. Comput. Phys.* **2016**, *318*, 22–35.

[64] E, W.; Ming, P. Cauchy–Born Rule and the Stability of Crystalline Solids: Static Problems. *Arch. Ration. Mech. Anal.* **2007**, *183* (2), 241–297.

[65] Cireşan, D. C.; Meier, U.; Gambardella, L. M.; Schmidhuber, J. Deep, Big, Simple Neural Nets for Handwritten Digit Recognition. *Neural Comput.* **2010**, *22* (12), 3207–3220.

[66] Kambouchev, N.; Fernandez, J.; Radovitzky, R. A Polyconvex Model for Materials with Cubic Symmetry. *Model. Simul. Mater. Sci. Eng.* **2007**, *15* (5), 451–467.

[67] Karpatne, A.; Atluri, G.; Faghmous, J. H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Trans. Knowl. Data Eng.* **2017**, *29* (10), 2318–2331.

[68] Xiao, H.; Wu, J.-L.; Wang, J.-X.; Sun, R.; Roy, C. J. Quantifying and Reducing Model-Form Uncertainties in Reynolds-Averaged Navier–Stokes Simulations: A Data-Driven, Physics-Informed Bayesian Approach. *J. Comput. Phys.* **2016**, *324*, 115–136.

[69] Wang, J.-X.; Wu, J.-L.; Xiao, H. Physics-Informed Machine Learning Approach for Reconstructing Reynolds Stress Modeling Discrepancies Based on DNS Data. *Phys. Rev. FLUIDS* **2017**, *2*, 34603.

[70] Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114* (105503).

[71] Menon, A.; Gupta, C.; Perkins, K. M.; DeCost, B. L.; Budwal, N.; Rios, R. T.; Zhang, K.; Póczos, B.; Washburn, N. R. 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.

[72] Hirata, T.; Ye, J.; Branicio, P.; Zheng, J.; Lange, A.; Plank, J.; Sullivan, M. Adsorbed Conformations of PCE Superplasticizers in Cement Pore Solution Unraveled by Molecular Dynamics Simulations. *Sci. Rep.* **2017**, *7* (1), 16599.

[73] Marchon, D.; Juilland, P.; Gallucci, E.; Frunz, L.; Flatt, R. J. Molecular and Submolecular Scale Effects of Comb-Copolymers on Tri-Calcium Silicate Reactivity: Toward Molecular Design. *J. Am. Ceram. Soc.* **2016**, *100* (3), 817–841.

[74] I.C. Yeh, Modeling of strength of high-performance concrete using artificial neural networks, Cem. Concr. Res. 28 (1998) 1797–1808. doi:10.1016/S0008-8846(98)00165-3.

[75] J.-S. Chou, C.-K. Chiu, M. Farfoura, I. Al-Taharwa, Optimizing the Prediction Accuracy of Concrete Compressive Strength Based on a Comparison of Data-Mining Techniques, J. Comput. Civ. Eng. 25 (2011) 242–253. doi:10.1061/(ASCE)CP.1943-5487.0000088.

[76] R.M. Haj-Ali, K.E. Kurtis, A.R. Sthapit, Neural network modeling of concrete expansion during long-term sulfate exposure, ACI Mater. J. 98 (2001) 36–43. doi:10.14359/10158.

[77] S. Dutta, P. Samui, D. Kim, Comparison of machine learning techniques to predict compressive strength of concrete, Comput. Concr. 21 (2018) 463–470. doi:10.12989/cac.2018.21.4.463.

[78] B.A. Young, A. Hall, L. Pilon, P. Gupta, G. Sant, Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods, Cem. Concr. Res. 115 (2019) 379–388. doi:10.1016/j.cemconres.2018.09.006.

[79] C.M. Childs, N.R. Washburn, Embedding domain knowledge for machine learning of complex material systems, MRS Commun. 9 (2019) 806–820. doi:10.1557/mrc.2019.90.

[80] P. Domingos, A few useful things to know about machine learning, Commun. ACM. 55 (2012) 78–87. doi:https://doi.org/10.1145/2347736.2347755.

[81] J.M. Bone, C.M. Childs, A. Menon, B. Póczos, A.W. Feinberg, P.R. LeDuc, N.R. Washburn, Hierarchical Machine Learning for High-Fidelity 3D Printed Biopolymers, ACS Biomater. Sci. Eng. (2020). doi:10.1021/acsbiomaterials.0c00755.

[82] A. Menon, J.A. Thompson-Colón, N.R. Washburn, Hierarchical Machine Learning Model for Mechanical Property Predictions of Polyurethane Elastomers From Small Datasets, Front. Mater. 6 (2019) 87. doi:10.3389/fmats.2019.00087.

[83] J. Pugar, C.M. Childs, C. Huang, K.W. Haider, N.R. Washburn, Elucidating the Physicochemical Basis of the Glass Transition Temperature in Linear Polyurethane Elastomers with Machine Learning, J. Phys. Chem. B. 4 (2020) 2020. doi:10.1021/acs.jpcb.0c06439.

[84] C. Kim, A. Chandrasekaran, T.D. Huan, D. Das, R. Ramprasad, Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions, J. Phys. Chem. C. 122 (2018) 17575–17585. doi:10.1021/acs.jpcc.8b02913.

[85] Y.C. Lo, S.E. Rensi, W. Torng, R.B. Altman, Machine learning in chemoinformatics and drug discovery, Drug Discov. Today. 23 (2018) 1538–1546. doi:10.1016/j.drudis.2018.05.010.

[86] F.A. Faber, L. Hutchison, B. Huang, J. Gilmer, S.S. Schoenholz, G.E. Dahl, O. Vinyals, S. Kearnes, P.F. Riley, O.A. Von Lilienfeld, Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error, J. Chem. Theory Comput. 13 (2017) 5255–5264. doi:10.1021/acs.jctc.7b00577.

[87] Ding, J.-T.; Li, Z. Effects of Metakaolin and Silica Fume on Properties of Concrete. *ACI Mater. J.* **2002**, *99* (4), 393–398.

[88] Washburn, N. R.; Menon, A.; Childs, C. M.; Poczos, B.; Kurtis, K. E. Machine Learning Approaches to Admixture Design for Clay-Based Cements. In *Calcined Clays for Sustainable Concrete*; Martirena, F., Favier, A., Scrivener, K., Eds.; Springer, Dordrecht, 2017; Vol. RILEM Book, pp 488–493.

[89] Menon, A.; Childs, C. M.; Poczós, B.; Washburn, N. R.; Kurtis, K. E. Molecular Engineering of Superplasticizers for Metakaolin-Portland Cement Blends with Hierarchical Machine Learning. *Adv. Theory Simul.* **2019**, *2* (1800164).

Chapter 2. Interplay of Anionic Functionality in Polymer-grafted Lignin Superplasticizers for Portland Cement [1]

2.1. Introduction to Cement and Admixtures

Concrete is the most widely used construction material in the world today,[1] with over 25 billion metric tons used each year.[2] As the world's infrastructure continues to expand, the demand for concrete will also increase showing an over 10-fold growth in consumption over the previous 70 years.[3]

Most concrete is composed as a mixture of aggregate, sand, and Portland cement (PC). Production of PC now accounts for approximately 5% of global CO_2 production due to energy-intensive processing and the release of CO_2 during limestone calcination.[4] In order to reduce the carbon footprint of this ubiquitous infrastructure material without compromising performance, it has been proposed that concrete increasingly incorporate alternative binder chemistries (ABC), supplementary cementitious materials (SCMs), and also improve cement and cement admixtures to increase specialization and decrease environmental effects. PC is a complex mixture of inorganic materials- the major components, abbreviated in cement-chemist notation (Table 2.1), being tricalcium silicate (C_3S), dicalcium silicate (C_2S), tricalcium aluminate (C_3A), and tetracalcium alumino ferrite (C_4AF).[5] Following hydration, a complex set of reactions occur between the water and mineral species. These dissolution and remineralization reactions allow nucleation and growth processes resulting in strong hydration products.[1,5,6] A critical parameter in hydraulic cement is the water: cement (w/c) ratio.

[1] This chapter includes work that was published and reformatted:
Childs, C. M., Perkins, K. M., Menon, A. & Washburn, N. R. Interplay of Anionic Functionality in Polymer-Grafted Lignin Superplasticizers for Portland Cement. *Ind. Eng. Chem. Res.* **58**, 19760–19766 (2019).

Table 2.1. Cement Chemist Notation (CCN)

CCN	**Actual Formula**
C	CaO
S	SiO_2
A	Al_2O_3
F	Fe_2O_3
$\overline{S}$	SO_3
H	H_2O

Following the addition of water, the mineral constituents of PC shed a diversity of ionic species, including calcium, sodium, potassium, silicate, aluminate, and hydroxide ions, resulting in a complex mosaic of charged multi-phase mineral particles.[7] Despite the charge density of the particles, the strong ionic environment of the pore solution, with a pH of at least 12.4 and an ionic strength of 250 mM, screens interparticle coulomb forces, and London forces drive aggregation,[8] which results in the development of a high yield stress in cement pastes and concrete that reduces the workability, for a given water-to-cement ratio. Water-reducing admixtures are a diverse class of polymer dispersants, specified under ASTM C494, used in cement–based materials to improve workability (i.e., exhibiting fluidity and cohesion) without additions of water. 'Superplasticizers' are a subset of these, designed to significantly improve workability while maintaining water content or to reduce water content by larger amounts, up to 12-30%, while maintaining workability. Superplasticizers are a diverse class of polymer dispersant that are used to improve the workability of cement paste while reducing water requirements. These are based on anionic polymers, such as lignosulfonate or polynaphthalene sulfonate, but the most effective superplasticizers are polycarboxylate ether (PCE)- copolymers based on methacrylic acid (MAA) and poly (ethylene glycol) methacrylate (PEGMA). The main variables tuned in their composition are the MAA:PEGMA ratio and the length of the pendant PEG chain. Current understanding into their mechanism of action postulates that the carboxylate functional groups mediate adsorption onto cement particle surfaces while pendant PEG groups reduce particle coagulation and network formation via steric interactions.[9]

The global scale of cement production makes these admixtures a widely used class of material. However, PCE superplasticizers are currently synthesized from petroleum-based feedstocks. Developing strategies to augment the performance of plant-derived lignin dispersants would be an important applicant of renewable materials. Currently there are two main classes of purified lignin feedstock: lignosulfonate (LS) and kraft lignin (KL). Prepared through sulfite pulping, lignosulfonates have high concentrations of sulfonate functional groups, making them only soluble in water. This characteristic makes them the most widely used lignin-based dispersant in cement. In contrast, the anionic functionality in kraft lignin is predominantly carboxylate groups and this biopolymer has not been shown to be an effective superplasticizer despite being produced at significantly greater quantities than lignosulfonates.

Recently, a machine-learning algorithm was used to guide the design of a novel superplasticizer based on lignosulfonate grafted with a poly(methacrylic acid) (PMAA) corona, which reduced the yield stress of Portland cement paste to comparable levels as a commercial PCE. The algorithm incorporated polymer effects on solution properties (viscosity and osmolality), particle properties (electrostatic and electrosteric interactions), as well as coupling between solution and particle properties, providing a comprehensive approach to molecular design.[10, 11] The representation of physicochemical interactions contributing to the reduction in yield stress are shown in Figure 2.1.

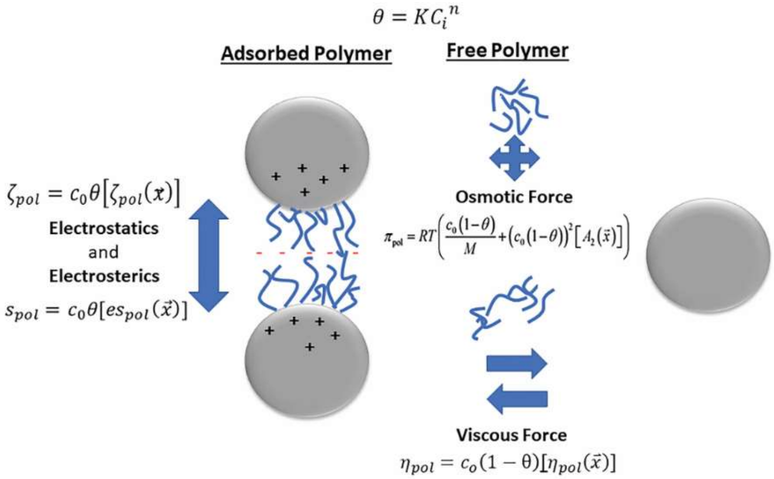

$\theta = KC_i^{\ n}$
Adsorbed Polymer
Free Polymer
$\zeta_{pol} = c_0\theta[\zeta_{pol}(x)]$
Electrostatics
and
Electrosterics
$s_{pol} = c_0\theta[es_{pol}(\vec{x})]$
+
Osmotic Force
$\pi_{pol} = RT\left(\frac{c_0(1-\theta)}{M} + (c_0(1-\theta))^2[A_2(\vec{x})]\right)$
Viscous Force
$\eta_{pol} = c_o(1 - \theta)[\eta_{pol}(\vec{x})]$

Figure 2.1. The degree of polymer adsorption, θ, onto cement particles is calculated through a Freundlich fit to experimental data at various polymer concentrations, C_i. The adsorbed polymer, where negatively charged polymer chains attract to the positively charged portion of cement particles, induces dispersion through both electrostatic effects, ζ_{pol}, as fit to zeta potential measurements and electrosteric effects, s_{pol}, as fit to sedimentation experiments. Free polymer induces solvent-mediated dispersing effects through both osmotic forces, π_{pol}, as fit utilizing the second virial (A_2) as calculated through vapor pressure osmometry measurements and viscous forces, η_{pol}, as fit to intrinsic viscosity measurements.

Here, the effects of anionic polymer grafts with carboxylate or sulfonate functional groups on the dispersant function of kraft lignin and lignosulfonate were compared with PEGylated analogues. Following lignopolymer synthesis, changes in yield stress in cement pastes containing varying superplasticizers was assessed by mini-slump testing and paste viscosity measurements through rheometry. The mechanism of dispersion was explored by measuring adsorption onto cement as well as measurement of zeta potential and intrinsic viscosity. Molecular design provides a framework for understanding trends in superplasticizer activity as a function of lignin and graft chemistries.

2.2. Experimental Methods

2.2.1. Materials

Borresperse sodium lignosulfonate (effective M_n 25,000) was purchased from Borregaard Lignotech and Biochoice Kraft lignin (effective M_n 20,000) was obtained from Domtar Corporation. 3-Sulfopropyl methacrylate potassium salt (SPMA), 2,2′-bipyridine (bpy), copper (I) bromide, ethyl 2-bromoisobutyrate (EBriB), N,N,N′,N″,N″-pentamethyldiethylenetriamine (PMDETA), trifluoroacetic acid (TFA), DOWEX 50WX8 hydrogen form, basic alumina, and neutral alumina were purchased from Sigma Aldrich. Tert-Butyl methacrylate (*t*BMA) from Sigma Aldrich was filtered through basic alumina prior to use. Sodium hydroxide, methylene chloride, and dimethylformamide were obtained from Fisher Scientific. Acetone was purchased from Pharmco-AAPER. ADVA 190, a commercial PCE (GCP Applied Technologies) was dialyzed and lyophilized to collect the pure polymer. Dialysis filtration was achieved using pre-treated dialysis tubing, Spectra/Por®, from Spectrum Labs. All materials were used as received from the manufacturer unless otherwise indicated. Cement pastes were prepared using Type I/II Saylor’s Portland cement in accordance with ASTM C150.[12]

2.2.2. Polymer Synthesis

Preparation of the grafted polysulfonate and polycarboxylate lignin is described below and the schematic for PMAA grafted lignosulfonate is shown in Figure 2.2. The synthesis of the PEGylated lignins was performed according to previous methods.[13] Polymer NMR and GPC data can be found in Figure 2.10-Figure 2.16.

2.2.3. Synthesis of Poly (3-sulfopropyl methacrylate)

To a flask, SPMA (3.69 g, 15 mmol), CuBr (0.043 g, 0.3 mmol), and bpy (0.23 g, 1.5 mmol) were added and placed under an inert nitrogen atmosphere using vacuum-nitrogen cycles. EBriB, H_2O, and dimethylformamide (DMF) were deoxygenated by purging with nitrogen for 60 mins. H_2O (5 mL) and DMF (5 mL) were then added to the reaction flask to solvate the reactants. EBriB (44µL, 0.3 mmol) was then added to the flask and the reaction mixture was allowed to stir for 3 h at room temperature. The reaction was then exposed to air and filtered through Dowex and neutral alumina. The filtrate was concentrated under vacuum, reconstituted into H_2O, and then purified via dialysis for 72 h using 2 kDa tubing. The product was then concentrated under vacuum to yield a white solid with an estimated weight of 2300 g/mol as predicted through kinetic information predicted in Masci et. al.[14]

2.2.3. Synthesis of Poly(methacrylic acid)

For synthesis preparation, *t*BMA, EBriB, and acetone were deoxygenated by purging with nitrogen for 90 min. Cu(I)Br (0.059 g, 0.41 mmol) was added to the flask and placed under an inert nitrogen atmosphere using 3 vacuum-nitrogen cycles. *t*BMA (10 mL, 61.5 mmol), EBriB (120 µL, 0.82 mmol), and acetone (2 mL) were added to the flask and allowed to stir for 10 min. PMDETA (86 µL, 0.41 mmol) was then added to the flask and the solution was allowed to stir for 5 h at 50 °C. The solution was then exposed to air and diluted with acetone prior to filtration through Dowex and neutral alumina. The filtrate was concentrated under vacuum to yield P*t*BMA, a white solid with representative characteristics (M_n = 3880 g mol^{-1}, polydispersity index (PDI) = 1.25). P*t*BMA was dissolved into a solution of TFA (8 mL) in DCM (20 mL) and allowed to stir for 72 h. The resulting solution was then concentrated under vacuum to yield a white solid.

2.2.4. Grafting onto Lignin

A solution of LS (3.8 g, 0.15 mmol) and water (100 mL) was brought to pH 11 using conc. NaOH. The solution was then heated to 70 °C and PSPMA (2.3 g, 1 mmol) was added to the solution while ensuring that the pH was maintained at pH 11 by adding additional NaOH if necessary. The solution was allowed to stir overnight at 70 °C and was subsequently purified via dialysis with 8 kDa dialysis tubing over the course of 72 h. The product was then concentrated under vacuum to yield a brown solid. The same procedure was used to graft PMAA chains and the kraft lignin derivatives, only with variations in the mass of lignin and polymer to yield the same mole ratio of 1 mmol of the graft reacted with 0.15 mmol lignin.

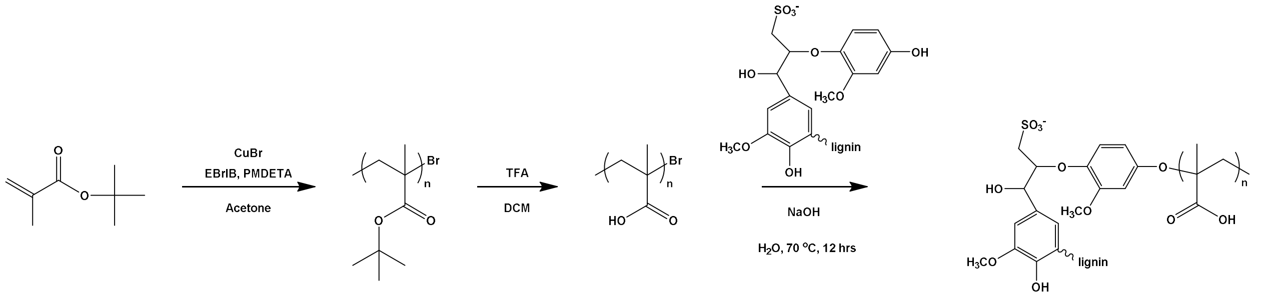


Figure 2.2. Schematic showing preparation of PMAA grafted lignosulfonate.

2.3. Polymer Characterization

2.3.1. Adsorption Measurements

Adsorption of the different polymers onto PC was performed by total organic carbon (TOC) analysis. Samples of 0.5, 1 and 5 mg ml^{-1} of each superplasticizer were prepared. A reference sample for each sample was made to compare the difference in amount of carbon before and after adsorption. For each different sample, the superplasticizer solution was mixed with 0.1 g cement for 1 h. The samples were then centrifuged for 11 min at 4400 rpm. The supernatant was collected, filtered through a 0.45 μm syringe filter, diluted with water, and immediately analyzed using a combustion-based TOC (Shimadzu TOC-L).

2.3.2. Zeta Potential

Zeta potential across a range of pH values from 3-11 were measured in aqueous solutions with a polymer concentration of 10 mg mL^{-1} using a Zeta-sizer Nano-ZS (Malvern Instruments) and a DTS1070 zeta cell (Malvern Instruments).

2.3.3. Osmotic Pressure

Osmolality of aqueous solutions for all lignopolymers was measured at 0.1, 0.2, and 0.4 g mL^{-1} using a vapor pressure osmometer (Wescor 5520). To differentiate the behavior of polymer in the pore solution of cement, the second virial coefficient of each was acquired by plotting osmolality values versus concentration of all the polymer solutions. The slope of this line was determined and multiplied by the molar volume and molality of water to obtain the A2 coefficient.[15]

2.3.4. Polymer Intrinsic Viscosity

Due to the low solubility at higher concentrations of most of the lignopolymers, Kraemer and Huggins curves were not well fit to the data. Instead, a single-point determination of the intrinsic viscosity at a low concentration of 3.3 mg ml^{-1} was used. The lower concentration reduced polymer aggregation so that intrinsic viscosity could be determined utilizing the Solomon–Ciuta equation.[16] An Ubbelohde Viscometer (Canon Instrument Company) was used to determine the relative viscosity by taking the ratio of the time to pass through the capillary of the polymer solution to the pure water elution time. Three trials for each polymer along with the solvent elution time were performed and averaged to yield the reported values.

2.3.5. Assessment of Workability

To assess the workability and changes in yield stress of the cement pastes mini-slump testing was performed. In accordance with ASTM C 305, cement pastes were prepared at room temperature using a planetary mixer (Hobart) with a paddle speed of 62 rpm.[17] To the mixer, 200 g of cement and 70 mL of water were added to produce a 0.35 w/c ratio paste. Superplasticizer was added at a ratio of 0.25% to weight of cement. To ensure optimal effectiveness of the superplasticizers, a fraction of the total water volume was added and mixed for 1 min. The remainder of the water containing the dissolved superplasticizer was added immediately afterwards and mixed for another 1 min for delayed addition of water.

Immediately after mixing, the cement pastes were added into a mini-slump cylinder (BASF, Construction Chemicals Division) with dimensions 3 cm in diameter and 5 cm in height. The cylinder was slowly lifted allowing the cement pastes to slump as shown in Figure 2.3. The change in height was recorded along with the spread.

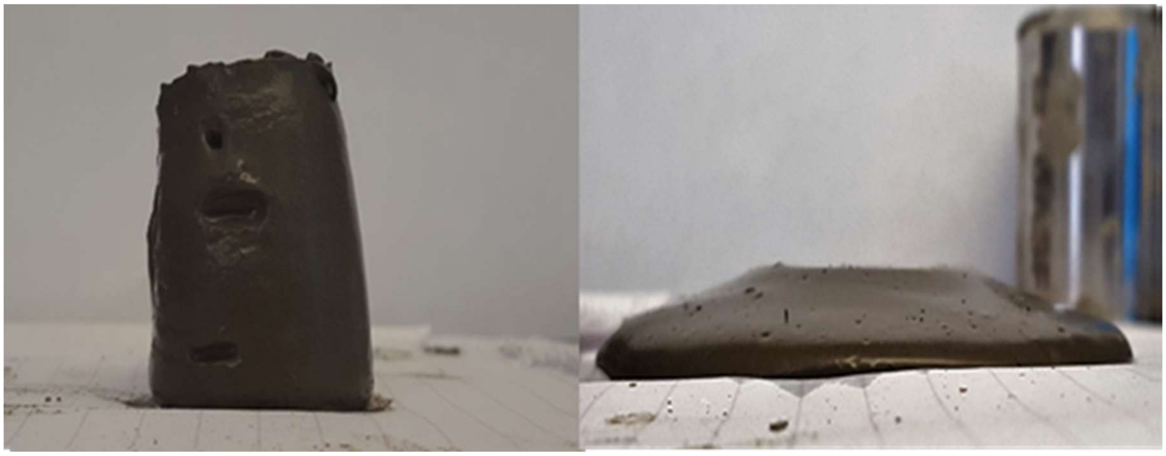


Figure 2.3. The image on the left exhibits a poor slump from cement with no added superplasticizer. The image on the right exhibits a high slump from cement at the same water to cement ratio (0.35 w/c), but with added commercial PCE (Adva 190) superplasticizer at 0.25% by weight of cement.

2.3.6. Rheology

Viscosity measurements of superplasticizer solutions with PC were performed. Cement pastes were made by mixing PC and water at a 0.5 w/c ratio with 1% w/w of cement superplasticizer added. The pastes were sonicated for 1 h and then immediately vibrated for 2 min and placed in a DHR Rheometer (TA Instruments) using a vane fixture to test viscoelastic properties of the paste. All samples were subjected to pre-shear of 100 s^{-1} to ensure they had the same mixing history. Then the shear rate was increased from 0 to 90 s^{-1} over the course of 1 min to obtain oscillatory strain data.

2.4. Results

2.4.1. Workability

Cement paste is a complex fluid based on a mosaic of charged ceramic particles suspended in an aqueous medium. The pore solution has a pH of 11 and an ionic strength of order 250 mM, which effectively screens coulombic interactions between particles. This screening makes van der Waals forces dominant, driving rapid coagulation and resulting in the formation of a percolating network of hydrating cement particles.[18,19,20] Cement paste behaves as a Herschel-Bulkley material,[21] characterized by a yield stress and Newtonian flow post-yield.

The yield stress is the most common metric for workability of cement paste, and slump tests are a widely used method for assessing this.[22,23] To gauge the effects of lignopolymers on cement yield stress, slump was measured by examining the height difference between the top of the slump cylinder and cement paste. The results of the slump for the KL and LS series of dispersants, pure polymer grafts, and the dialyzed commercial PCE used for comparison are shown in Figure 2.4.

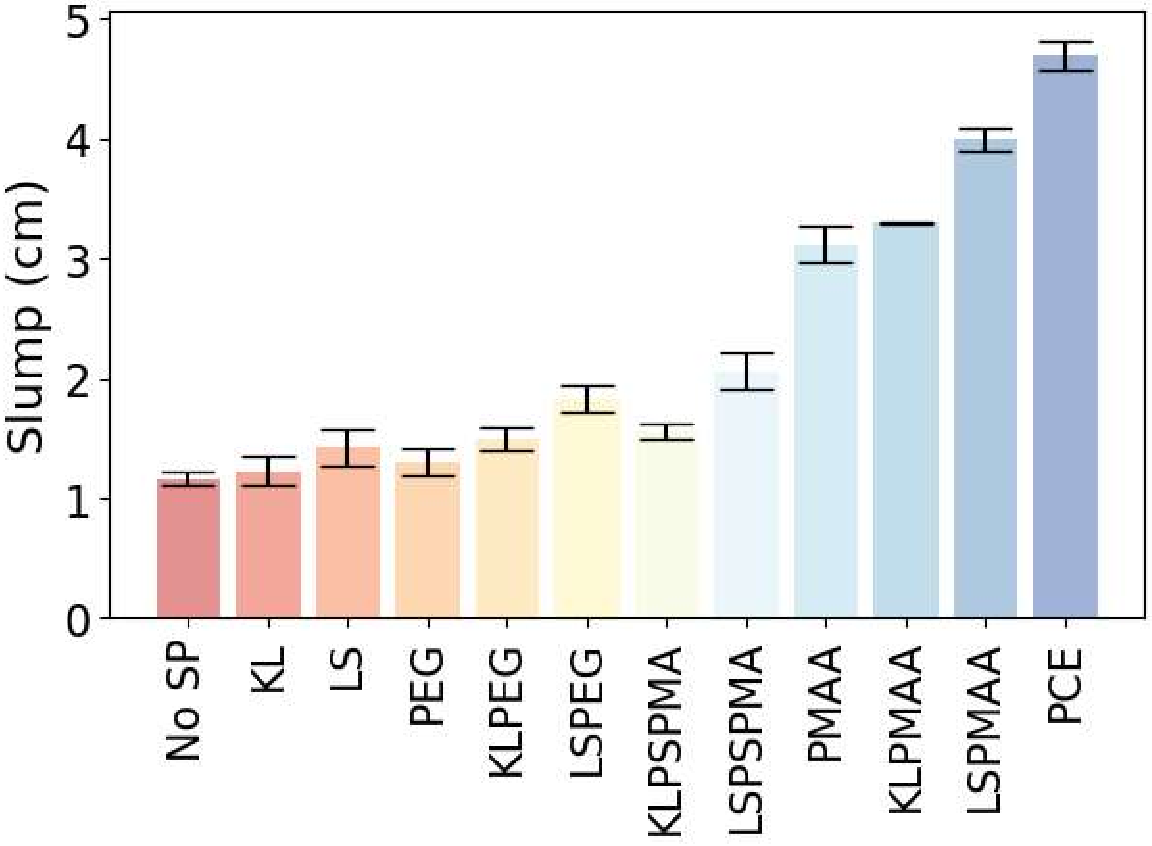


Figure 2.4. Slump height difference for the LS and KL-series, pure polymer grafts, and dialyzed PCE at 0.35 w/c ratio.

As expected, cement paste without superplasticizer exhibited the smallest slump value. Interestingly, the graft chemistry had a greater effect on slump than the lignin core. While homopolymers of PMAA and SPMA have not been found to be effective dispersants for ceramic-particle suspensions, it appears that synergies with lignin cores in a grafted molecular architecture lead to enhanced interfacial activities with PMAA being significantly more active than SPMA in the grafted polymer. In Table 2.4, Figure 2.17, and Figure 2.18 it is shown this same synergy does not occur with unreacted mixtures of the homopolymer and lignin. In comparing across the lignosulfonate series, LSPMAA, as predicted in prior machine learning study, had the largest slump of the lignosulfonates exhibiting twice the slump of LSPEG. The PMAA-grafted kraft lignin also exhibited over twice the slump to any other kraft lignopolymer, but still imparted 0.7 cm less slump than LSPMAA. When comparing lignin cores with similar graft chemistry LS promoted dispersion over KL, indicating that the lignin was still active.

2.4.2. Rheology

Viscosity measurements of the cement pastes were conducted at higher w/c ratios and higher superplasticizer percentages than the slump tests to set the values in the range accessible by the rheometer, and the results are shown in Figure 2.5. While yield stress measurements, such as slump, probe network formation in colloidal materials, responses in steady-shear rheology are thought to be dominated by floc or aggregate formation, which are increasingly disrupted as the shear rate was increased. Interestingly, paste plasticized by LSPSPMA had the lowest viscosity across both the LS and KL series. This suggests that while LSPSPMA may not effectively inhibit network formation under static conditions, it may reduce cohesive forces in aggregates so that even small applied stresses can disrupt their formation.

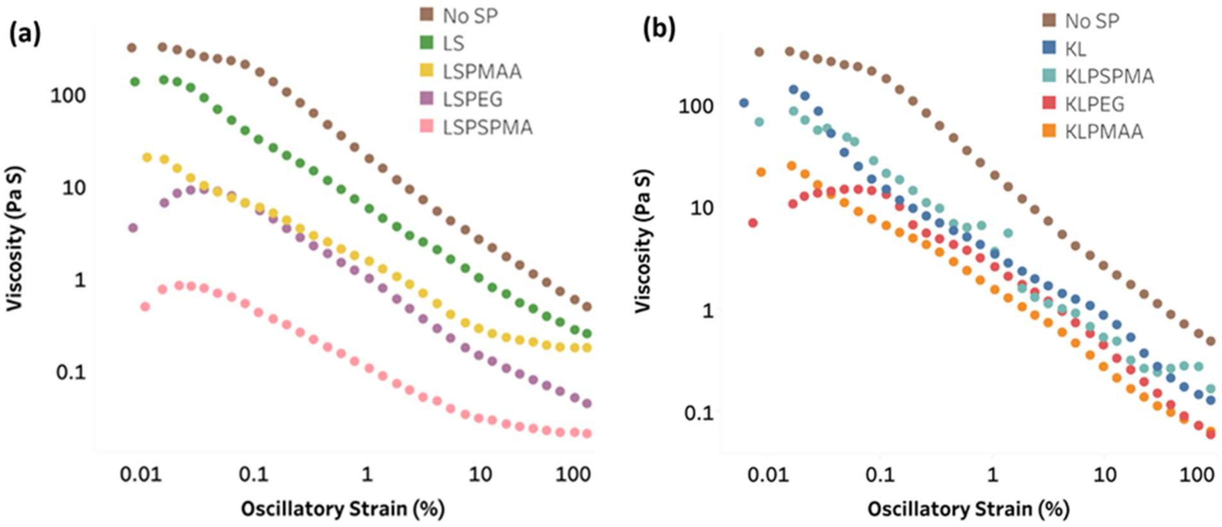


Figure 2.5. Viscosity of admixtures in cement pastes: (a) lignosulfonate series, (b) kraft lignin series.

There was not a direct correspondence seen between slump and viscosity, however, lower viscosity measurements were seen for the polymers with higher amounts of adsorbed polymer.

For the LS series, LSPMAA and LSPEG both had adsorption plateaus at approximately 10 mg superplasticizer per gram of cement, while adsorption of LSPSPMA continued adsorption up to 50 mg superplasticizer per gram of cement. Viscosity testing was performed at 20 mg superplasticizer per gram of cement. At this range LSPMAA and LSPEG would increase the pore solution viscosity, along with causing an increase in depletion flocculation, which would raise cement paste viscosity higher than LSPSPMA which multilayer adsorption of the polymer. A similar trend is seen for the kraft lignin series as well.

2.4.3. Adsorption

Polymer adsorption onto cement particles is known to be an important mechanism in inhibiting cement particle flocculation, generating electrostatic as well as electrosteric repulsions.[20] As can be seen in Figure 2.6, the lignosulfonate series of polymers adsorb at a much lower amount than the kraft lignin series and while LSPEG and LSPSPMA follow Langmuir adsorption behavior, the kraft lignin derivatives continued to adsorb with increasing concentration. This is likely due in part to the aggregation of kraft lignin in high-concentration solutions where solubility becomes a limiting factor.[24, 25] LSPEG and LSPMAA both appeared to reach a plateau consistent with a monolayer adsorption profile, while LSPSPMA appeared to exhibit lower solubility and multilayer adsorption, similar to unmodified lignosulfonate.[13]

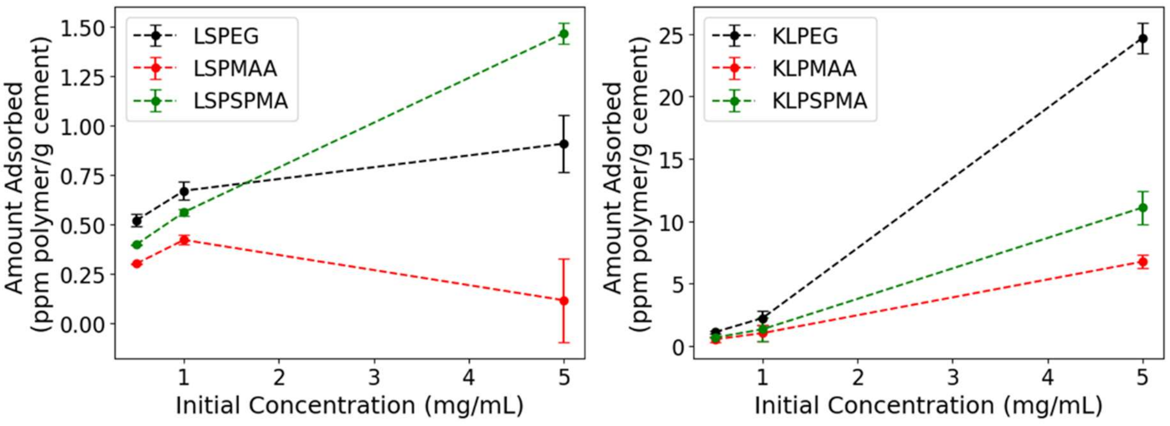


Figure 2.6. Adsorption isotherms for: (a) lignosulfonate series, (b) kraft lignin series.

For most superplasticizers in cement, only low concentrations in the range of 0.25 wt.% were needed for use as a superplasticizer. Table 2.2 shows values for percent adsorption, which reflects the partitioning between solution and adsorbed states.

Table 2.2. The percentage of polymer adsorbed was determined through taking the difference of the amount of carbon in the reference and test samples and then dividing by the amount in the reference sample. Initial concentrations of 1, 5, and 10 mg ml^{-1} were each used and compared to calculate the adsorption at 0.25 wt.% of the lignin to the cement.

Polymer	**LSPEG**	**KLPEG**	**LSPSPMA**	**KLPSPMA**	**LSPMAA**	**KLPMAA**
Percent Adsorbed	25%	85.40%	44.10%	64.10%	32%	52.80%

The kraft lignin analogues adsorbed at higher amounts than the corresponding lignosulfonates, which is attributed to the hydrophobicity of the lignin core. However, it is interesting to note that PEGylated kraft lignin had the highest overall adsorption, but for the lignosulfonate series, the PSPMA-grafted analogue had the greatest adsorption onto cement particles while the PEGylated analogue had the least. This suggests the presence of specific interactions between the polymer graft and the lignin core, such as binding of PEG onto the core, which has been shown to be important in PEGylated proteins.[26] The interactions of the polymer graft with lignin, pore solution, and cement particles influences the adsorption profile, which also contribute to changes in the forces that underlie dispersion. Understanding the impact of these interactions can be assessed through measurement of solution and particle forces, which will provide design principles for polymer-grafted lignin dispersants.

2.4.4. Polymer Characterization

The zeta potential of polymer additives to cement is widely used as an indication of electrostatic repulsion between adsorbed polymers on the cement surface. Characterization of the polymer zeta potential as a function of solution pH provides insight into how dispersants can interact with hydrating cement particles. Zeta potentials greater than an absolute value of 20 mV indicate a net repulsive electrostatic effect according to Derjaguin-Landau-Verway-Overbeek (DLVO) theory.[27] Shown in Figure 2.7, zeta potential was performed over a range of pH values for each of the free lignopolymers in aqueous solution.

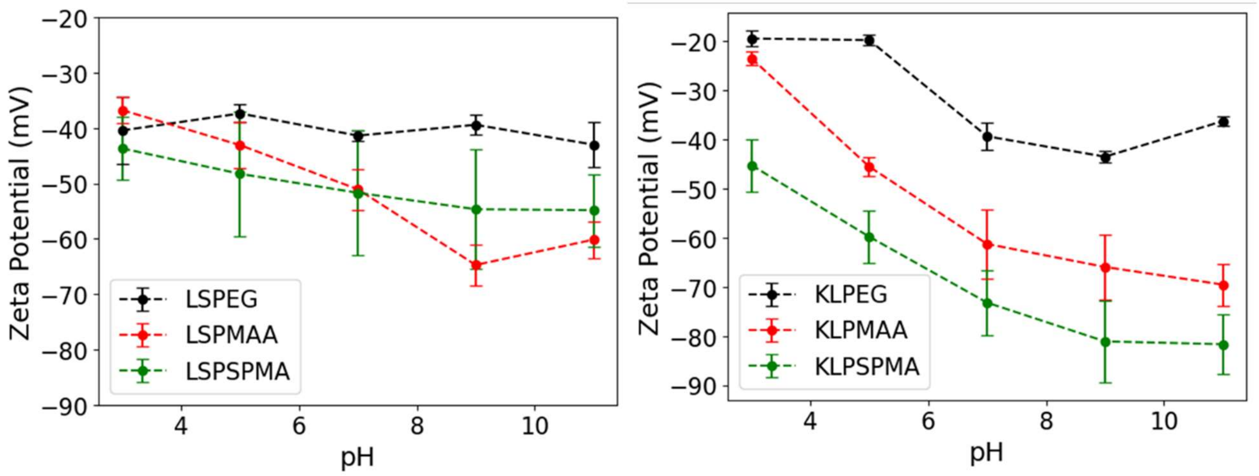


Figure 2.7. Zeta potential measurements for as a function of pH for: (a) lignosulfonate series, (b) kraft lignin series.

The zeta potential of the lignosulfonates displayed weaker pH dependence than that of the kraft lignin analogues, indicating that the chemistry of the anionic groups in the lignin core strongly determined the net charge of the conjugates. In contrast, the kraft lignin analogues had a strong dependence of zeta potential on pH, with KLPEG ranging from ca. -20 mV at pH 3 to -40 mV at a pH 9. Interesting, KLPMAA had the largest range of zeta potential values, ranging from ca. -25 mV at pH 3 to -70 mV at pH 11. This is consistent with the pH-sensitivity of the carboxylate graft, which was nearly as strong for LSPMAA. It may be concluded that the overall zeta potential of the grafted materials was due to distinct contributions from the lignin core and the polymer graft. The lignosulfonate core and the PEG and PSPMA grafts all displayed weak pH dependences, while kraft lignin and the PMAA grafts had stronger pH dependence. This suggests that the grafts do not strongly interact with the core, and the overall architecture is hypothesized to be a lignin core with an extended polymer corona. This structural model suggests further that both the lignin core and the polymer corona are accessible for interactions with the charged surface of cement particles.

In Figure 2.8 is shown a comparison of the zeta potential values at pH 7 with the slump values reported in Figure 2.4 There is a weak correlation between these measurements, suggesting that the effects of adsorbed polymer on particle charge is only partially responsible for changes in slump due to the lignopolymer superplasticizer.

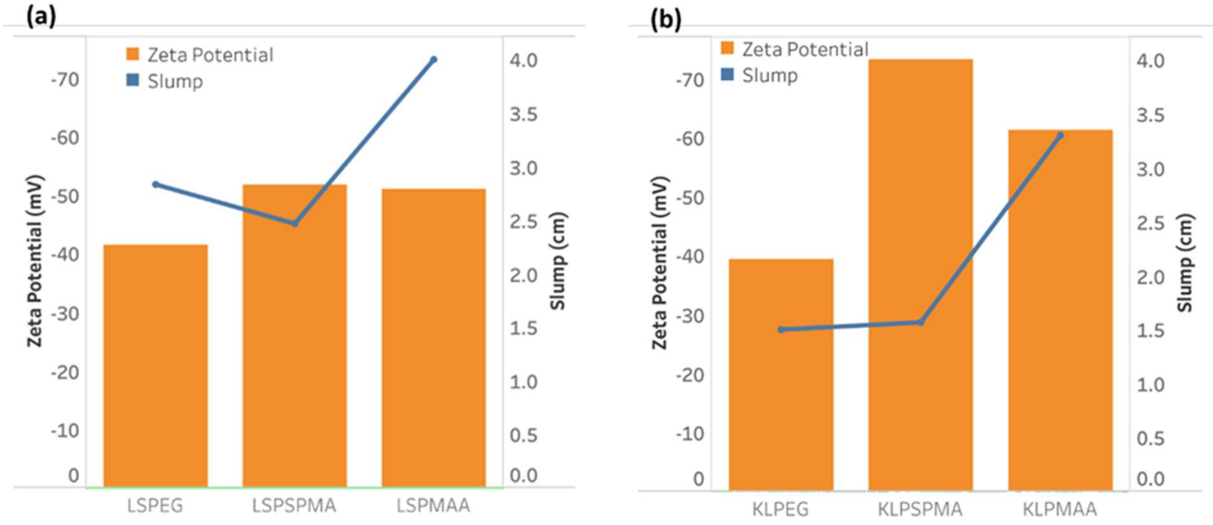


Figure 2.8. Zeta potential of lignopolymers at pH 7 with corresponding slump values shown as a trend line: (a) lignosulfonate analogues, (b) kraft lignin analogues.

While dispersant design has focused on the role of adsorbed polymer in mediating particle-particle interactions, the machine learning model that led to the design of LSPMAA incorporated solution forces in addition to particle forces.[10] Indeed, experimental studies have demonstrated that intrinsic viscosity is important in determining the yield stress in concentrated suspensions,[28, 29] suggesting it needs to be considered as a design principle on par with effects on particle interactions.

Intrinsic viscosity $[\eta]$ reflects the hydrodynamic volume of dissolved species.[30] While the conformation of polymer grafts from the lignin core can tune this, so too can aggregation in solution, which is known to be a factor in aqueous media,[31] although the details depend on the compactness of aggregate formed. Thus, it is difficult to determine if differences in intrinsic viscosity of lignopolymers were due to partial collapse of the polymer grafts onto the lignin core or differences in extent of aggregation, presumably mediated by the lignin core.

In Figure 2.9 are shown intrinsic viscosity values at pH 7 for the lignosulfonate and kraft lignin series with the slump values superimposed as a trend line. The PEGylated analogues of lignosulfonate and kraft lignin had the lowest intrinsic viscosities for each lignin type, while the highest value was recorded for KLPMAA. The extremely low values of $[\eta]$ for KLPEG and LSPEG (less than 5 mL g^{-1}) suggest that these have aggregated in aqueous solution. For comparison, PEG homopolymer having M_n of 750 g mol^{-1} (one graft on the PEGylated lignin) has an $[\eta]$ of 5.5 mL g^{-1}.[32]

It is interesting that the correlation between intrinsic viscosity and slump was much stronger for the kraft lignin series than the lignosulfonate series. Because LSPMAA and KLPMAA imparted the greatest increases in slump, it indicates that the high intrinsic viscosity and strongly negative zeta potential were central factors in the dispersing effects of KLPMAA. However, LSPMAA had moderate values of both these parameters and the largest slump, which suggests that other, undetermined factors may also be important for determining dispersant effects.

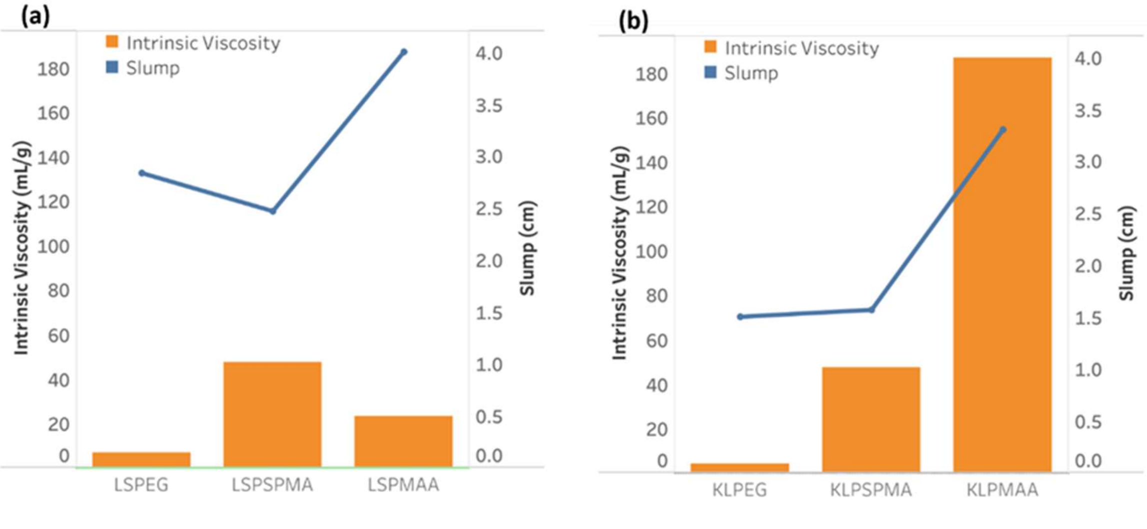


Figure 2.9. Intrinsic viscosity measurements of lignopolymers at pH 7 with corresponding slump values shown as a trend line: (a) lignosulfonate analogues, (b) kraft lignin analogues.

2.5. Discussion

It is well established that lignosulfonate is a much more effective dispersant for cement than kraft lignin,[33] which can be attributed to the presence of sulfonate groups that increase aqueous solubility and interact more weakly with calcium ions than the carboxylate groups in kraft lignin. Based on this, it was expected that the polymer-grafted lignosulfonates would be more active superplasticizers than the polymer-grafted kraft lignins. However, the effects of lignopolymers on cement-paste slump appeared to be most strongly correlated with the chemistry of the polymer graft, with the PMAA-grafted lignins both resulting in the largest slump values, corresponding to the lowest yield stresses.

In the model of superplasticizer action used here, both solution and particle forces mediate colloidal network formation, which is responsible for reductions in cement workability. Lignopolymer adsorption is an underlying variable that determines the balance of the dispersant effects between solution and particle forces. This can be attributed to the greater hydrophobic nature of kraft lignin, which would have a greater tendency to be driven from aqueous solution. Of the six derivatives investigated, only LSPEG had an adsorption profile that resembled the classic isotherm characterized by monolayer formation. LSPMAA actually had lower adsorption at 5 mg mL^{-1} than at 2.5 mg mL^{-1}, which may suggest formation of an aggregated species with lower affinity for the surfaces of the cement particles. In contrast, LSPSPMA and all the kraft lignin analogues continued to show monotonic increases in adsorption even at concentrations of 5 mg mL^{-1}. This suggests the formation of multilayer coatings on the cement particles, which could be due to continued adsorption onto the initial monolayer or could be due to aggregates formed in solution adsorbing onto the particle surfaces. At 2.5 mg mL^{-1}, the superplasticizer loading level for the slump measurements, the PMAA-grafted kraft lignin and lignosulfonates both displayed the lowest adsorption levels for both series.

One clear trend in the adsorption data was the much greater fraction of all the kraft lignin analogues that adsorbed compared with the ones based on lignosulfonate. However, it is interesting to compare this with the zeta potential results. At neutral pH and higher, all lignopolymers had a zeta potential values of -40 mV or more negative, which would predict the formation of a stable species in solution.[34] It is possible that the combination with strongly hydrophobic alkyl and aromatic species makes these unstable against aggregation in solution, but it may be that the mechanism of multilayer formation is based on continued adsorption of essentially monomeric lignopolymer species. Further characterization of solution structure will be necessary to resolve this.

General trends observed in the zeta potential and intrinsic viscosity results was that the polymer-grafted lignosulfonate series had the same trends in both physical properties, with PEG<PMAA<PSPMA. This would be expected based on the harder sulfonate anions providing a more negatively charged corona as well as increase the hydrodynamic volume by swelling in water, resulting in an increase in intrinsic viscosity.

In contrast, the kraft lignin series displayed a complex relationship between the kraft lignin core and the polymer corona, despite having the PMAA-grafted analogue result in the greatest slump value as was observed for LSPMAA. The zeta potential for KLPSPMA had the most negative value, as expected based on swelling of hard anionic grafts, but the intrinsic viscosity was anomalously large for KLPMAA, which was not consistent with the structure-property model developed. It is not clear if the high intrinsic viscosity is due to nascent network formation that, even in dilute solution, results in interactions that increase the effective hydrodynamic volume relative to other members of the lignopolymer series. Regardless, the increase in slump due to KLPMAA may be due predominantly to its high intrinsic viscosity.[35]

2.6. Conclusion

Lignosulfonate and kraft lignin were grafted with water-soluble polymers to explore synergies between graft and lignin chemistries using PEG, PSPMA, and PMAA grafts prepared via controlled polymerization. Grafting anionic polymers resulted in significant increases in superplasticizer activity over PEG grafts. Increases in the slump of cement paste comparable to commercial PCE superplasticizers were observed for PMAA-grafted analogues of both lignin derivatives. The mechanism of dispersion was explored by measuring the effects of lignopolymers on both solution and particle properties. For the lignosulfonate derivatives, the PSPMA graft resulted in the largest values for zeta potential and intrinsic viscosity but the PMAA graft imparted the largest slump value for the lignopolymers investigated in this work. For the kraft lignin derivatives, the PMAA graft also imparted the largest slump, but there were not readily understood trends in intrinsic viscosity, despite this being the largest value across all the lignopolymers. Thus, it appears that PMAA grafting is a promising strategy for enhancing the superplasticizer capability of lignins but the molecular mechanisms depend strongly on the chemistry of the lignin core.

2.7. Research Contributions

C.M. Childs performed polymer property characterization and measurements of cement properties. K.M. Perkins performed lignopolymer synthesis and chemical characterization. A. Menon provided support for rheological measurements.

2.8. Appendix

^{1}H NMR spectroscopy was used to confirm polymer synthesis and successful grafting of the polymers onto the lignin core shown in Figure 2.10, Figure 2.12-Figure 2.16. The molecular weight distribution of the P*t*BMA was determined using gel permeation chromatography with THF as a solvent as shown in Figure 2.11. The molecular weight of PMAA was calculated assuming P*t*BMA was 100% deprotected. Internal standards, maleic acid and dimethyl sulfone, were used to determine grafting density of PSPMA and PMAA grafts, respectively. ^{1}H NMR spectroscopy of polymers with internal standards are shown in Figure 2.12 and Figure 2.13. The internal standards were chosen for solubility and to avoid overlap with polymer peaks. Quantitative proton NMR to determine grafting density was done by utilizing known concentrations of either homopolymer or lignopolymer and known concentrations of internal standard in 1.5 mL of solvent. The peak area was normalized utilizing the known quantity of protons present in the spectra due to the internal standard. This was then utilized to determine the concentration of homopolymer or lignin (depending on the peak chosen) present in the lignopolymer solution which along with the molecular weight and the initial concentration of lignopolymer was used to determine the grafting density. Similar to the PEG grafted lignopolymers, the grafting density was found to be ~2 for PMAA and PSPMA. Figure 2.14-Figure 2.16 show ^{1}H NMR's of varying synthesized polymer-grafted lignins. An expanded table of the lignopolymer physical properties is presented in Table 2.3. Table 2.4 presents the slumps associated with the pure homopolymers, mixtures of homopolymer and lignin, and the grafted lignopolymers. Figure 2.17 and Figure 2.18 show representative pictures of these slumps from Table 2.4.

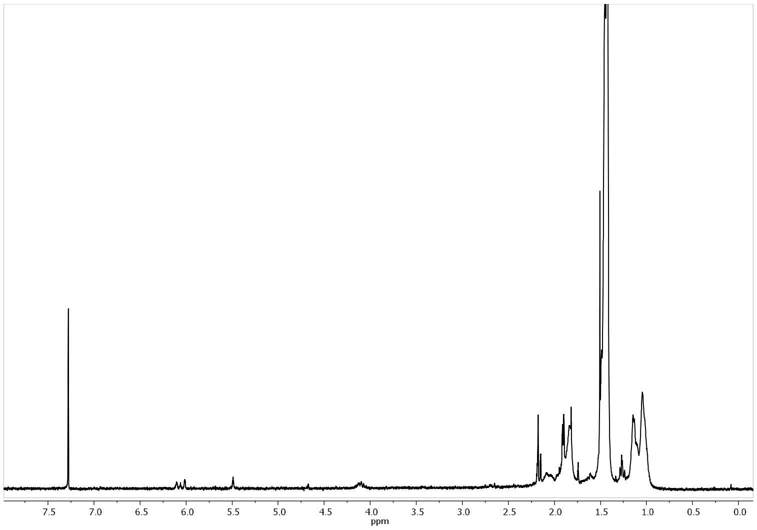


Figure 2.10. ^{1}H NMR of P*t*BMA in CDCl$_3$.

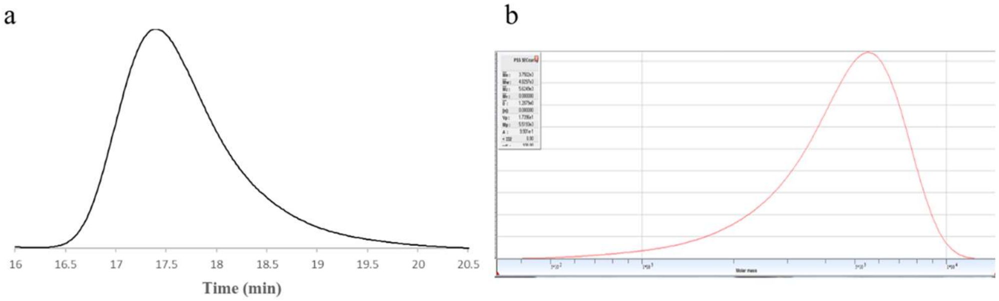


Figure 2.11. GPC elugram (a) and M_w profile (b) of P*t*BMA.

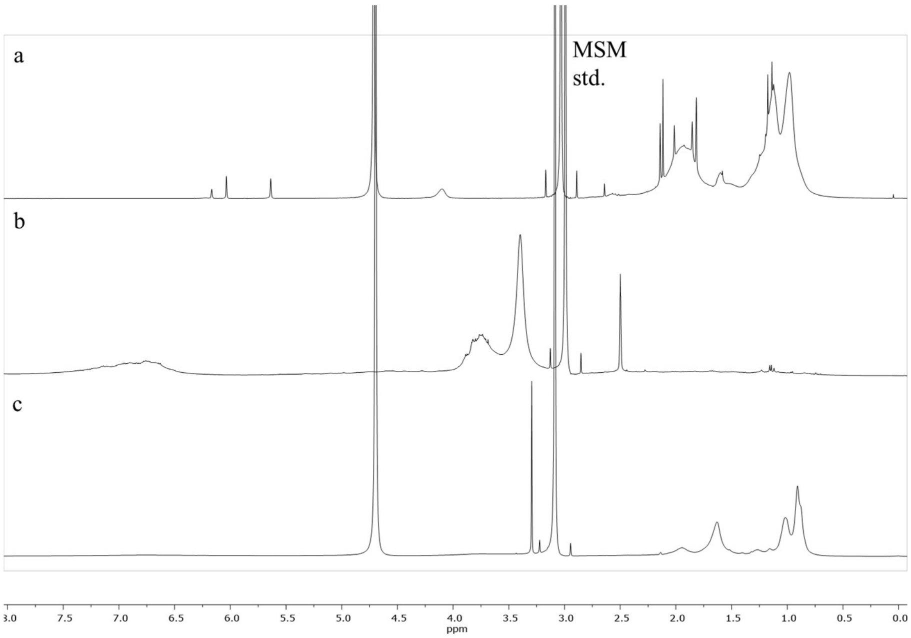


Figure 2.12. ^{1}H NMR of PMAA (a), kraft lignin (b), and PMAA grafted kraft lignin (c) in dimethyl sulfone (MSM) internal standard. PMAA and PMAA grafted kraft lignin were measured in D_2O while kraft lignin was measured in DMSO-d_6. Figure 2.14 shows better shows the presence of the kraft lignin peak for KLPMAA.

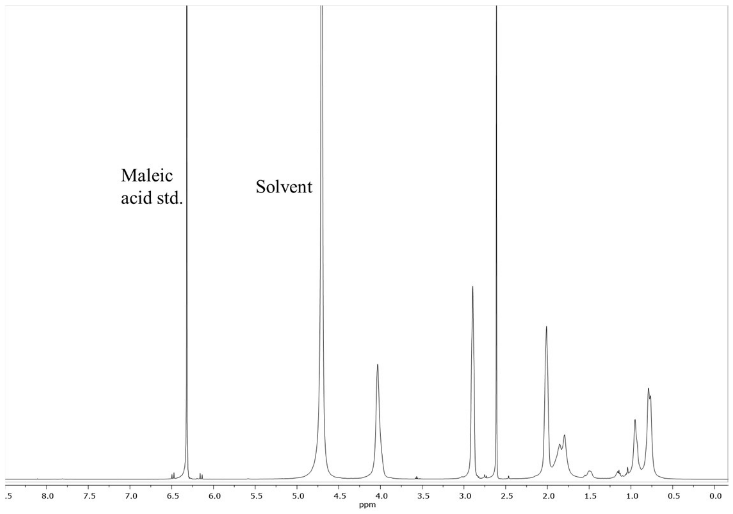


Figure 2.13. ^{1}H NMR of PSPMA in D_2O with maleic acid internal standard.

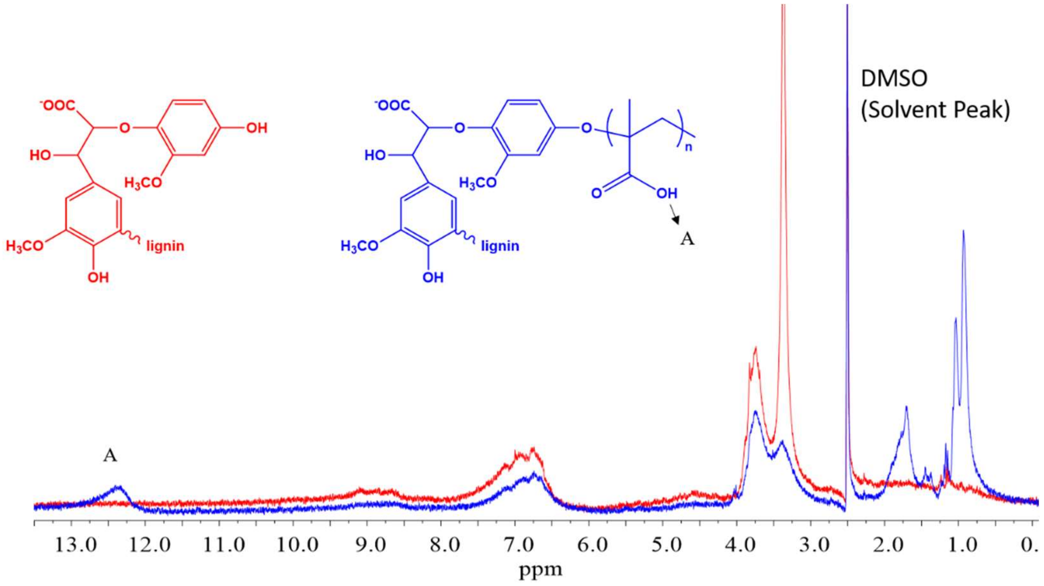


Figure 2.14. ^{1}H NMR comparison between kraft lignin (red) and KLPMAA (blue). Both NMR's were performed in DMSO-d6 for comparison as kraft lignin is not soluble in D_2O.

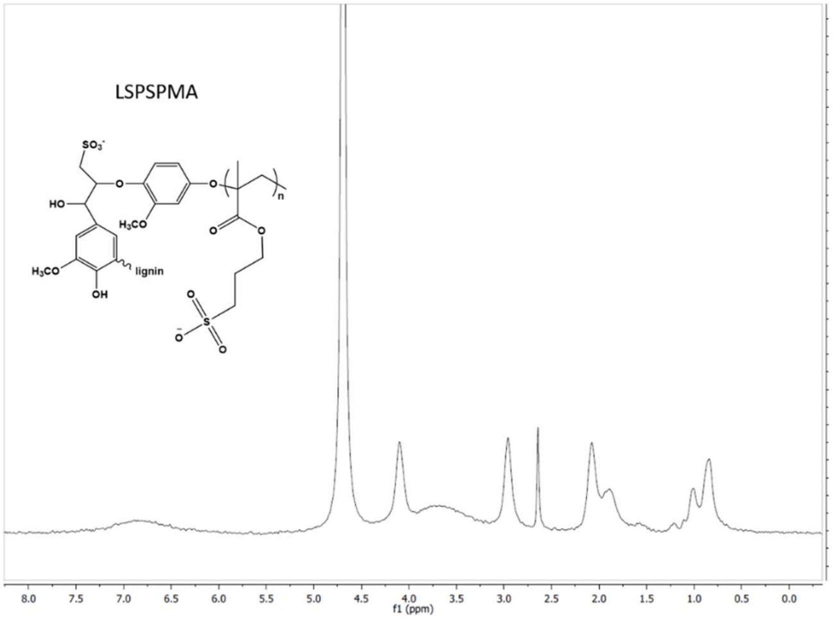


Figure 2.15. ^{1}H NMR of LSPSPMA in D_2O.

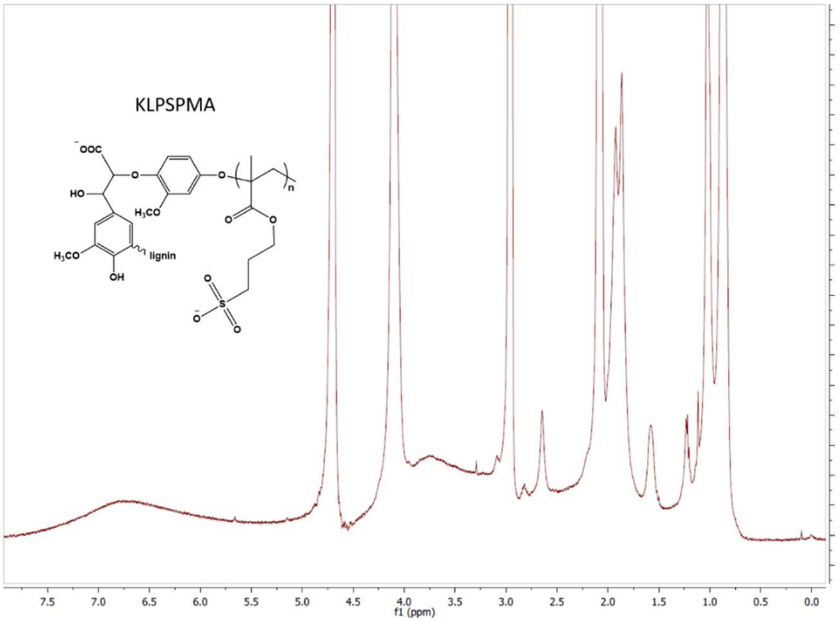


Figure 2.16. ^{1}H NMR of KLPSPMA in D_2O.

Table 2.3. Lignopolymer properties table.

Polymer	Slump (cm)	Intrinsic Viscosity (mL/g)	A2 (mol*mL/g^2) *10^3	Zeta Potential at pH 7 (mV)
LSPEG	2.8	6.459	1.98	-41.36
KLPEG	1.5	3.726	-1.47	-39.35
LSPSPMA	4	47.38	0.317	-51
KLPSPMA	3.3	47.38	-0.892	-61.2
LSPMAA	2.5	23.17	-0.460	-51.7
KLPMAA	1.6	186.5	-1.01	-73.2

Table 2.4. Polymer mixture properties table. The lignin:polymer graft ratio was set at 2:1 by weight to be similar to the synthetic ratio. All tests were performed at 0.35 w/cm ratio and 0.25% superplasticizer by weight of cement.

PEG Polymers	Slump (cm)	PMAA Polymers	Slump (cm)
PEG	1.3	PMAA	3.2
KL + PEG	1.2	KL + PMAA	1.5
KLPEG	1.6	KLPMAA	3.3
LS + PEG	1.5	LS + PMAA	1.8
LSPEG	1.8	LSPMAA	3.8

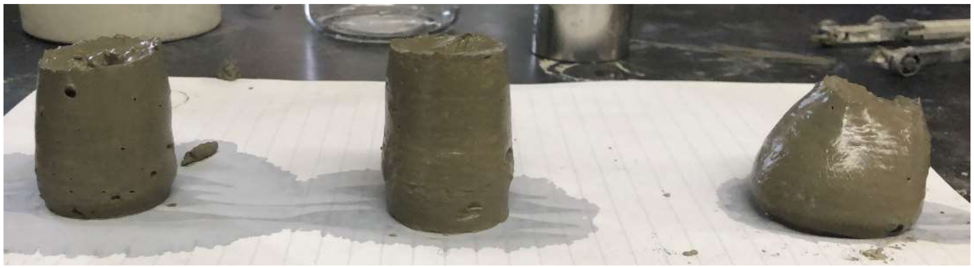


Figure 2.17. On the left is a slump with PEG graft, the middle is a slump with the 2:1 weight mixture of kraft lignin and PEG, respectively, and the slump to the right is KLPEG.

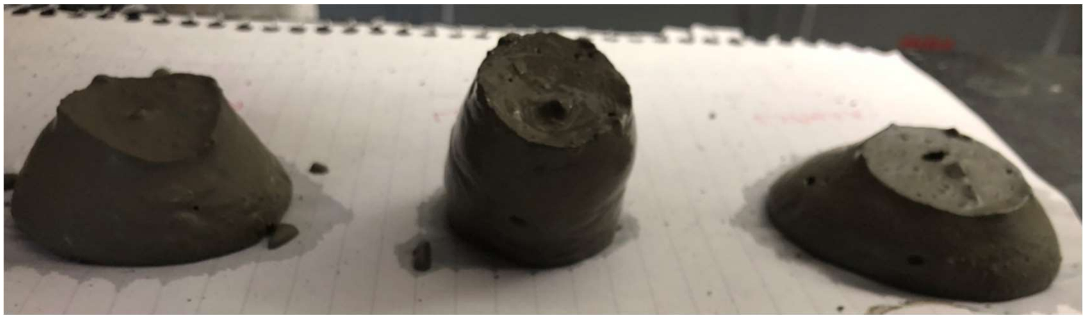


Figure 2.18. On the left is a slump with PMAA graft, the middle is a slump with the 2:1 weight mixture of lignosulfonate and PMAA, respectively, and the slump to the right is LSPMAA.

2.9. References

[1] Mehta, P.; Monteiro, P. *Concrete: Microstructure, Properties and Materials*; Mc-Graw Hill: New York, 2006.

[2] Tiwari, A.; Singh, S.; Nagar, R. Feasibility Assessment for Partial Replacement of Fine Aggregate to Attain Cleaner Production Perspective in Concrete: A Review. *J. Clean. Prod.* **2016**, *135*, 490–507.

[3] P.J.M. Monteiro, S.A. Miller, A. Horvath, Towards sustainable concrete, Nat. Mater. 16 (2017) 698–699. doi:10.1038/nmat4930.

[4] Huntzinger, D. N.; Eatmon, T. D. A Life-Cycle Assessment of Portland Cement Manufacturing: Comparing the Traditional Process with Alternative Technologies. *J. Clean. Prod.* **2009**, *17* (7), 668–675.

[5] Jolicoeur, C.; Simard, M.-A. Chemical Admixture-Cement Interactions: Phenomenology and Physico-Chemical Concepts. *Cem. Concr. Compos.* **1998**, *20* (2–3), 87–101.

[6] Bullard, J. W.; Jennings, H. M.; Livingston, R. A.; Nonat, A.; Scherer, G. W.; Schweitzer, J. S.; Scrivener, K. L.; Thomas, J. J. Mechanisms of Cement Hydration. *Cem. Concr. Res.* **2011**, *41* (12), 1208–1223.

[7] Sakai, E.; Yamada, K.; Ohta, A. Molecular Structure and Dispersion-Adsorption Mechanisms of Comb-Type Superplasticizers Used in Japan. *J. Adv. Concr. Technol.* **2003**, *1* (1), 16–25.

[8] Flatt, R. J. Dispersion Forces in Cement Suspensions. *Cem. Concr. Res.* **2004**, *34* (3), 399–408.

[9] Marchon, D.; Sulser, U.; Eberhardt, A.; Flatt, R. J. Molecular Design of Comb-Shaped Polycarboxylate Dispersants for Environmentally Friendly Concrete. *Soft Matter* **2013**, *9* (45), 10719.

[10] Menon, A.; Gupta, C.; Perkins, K. M.; DeCost, B. L.; Budwal, N.; Rios, R. T.; Zhang, K.; Póczos, B.; Washburn, N. R. 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.

[11] Washburn, N. R.; Menon, A.; Childs, C. M.; Poczos, B.; Kurtis, K. E. Machine Learning Approaches to Admixture Design for Clay-Based Cements. In *Calcined Clays for Sustainable Concrete*; Martirena, F., Favier, A., Scrivener, K., Eds.; Springer, Dordrecht, 2017; Vol. RILEM Book, pp 488–493.

[12] ASTM C150, Standard Specification for Portland Cement. ASTM International: West Conshohocken, PA 2018.

[13] Gupta, C.; Perkins, K. M.; Rios, R. T.; Washburn, N. R. Poly (Ethylene Oxide) -Grafted Lignosulfonate Superplasticisers : Improving Performance by Increasing Steric Interactions. *Adv. Cem. Res.* **2017**, *29* (1), 2–10.

[14] Masci, G.; Bontempo, D.; Tiso, N.; Diociaiuti, M.; Mannina, L.; Capitani, D.; Crescenzi, V. Atom Transfer Radical Polymerization of Potassium 3-Sulfopropyl Methacrylate: Direct Synthesis of Amphiphilic Block Copolymers with Methyl Methacrylate. *Macromolecules* **2004**, *37* (12), 4464–4473.

[15] Schwinefus, J. J.; Checkal, C.; Saksa, B.; Baka, N.; Modi, K.; Rivera, C. Molar Mass and Second Virial Coefficient of Polyethylene Glycol by Vapor Pressure Osmometry. *J. Chem. Educ.* **2015**, *92* (12), 2157–2160.

[16] Pamies, R.; Ginés, J.; Cifre, H.; Del Carmen López Martínez, M.; García De La Torre, J.; Pamies, R.; Hernández Cifre, J. G.; Del Carmen López Martínez, · M; Garcia De La Torre, J. Determination of Intrinsic Viscosities of Macromolecules and Nanoparticles. Comparison of Single-Point and Dilution Procedures. *Colloid Polym Sci* **2008**, *286*, 1223–1231.

[17] ASTM C305, Standard Practice for Mechanical Mixing of Hydraulic Cement Pastes and Mortars of Plastic Consistency. ASTM International: West Conshohocken, PA 2014.

[18] Grierson, L. H.; Knight, J. C.; Maharaj, R. The Role of Calcium Ions and Lignosulphonate Plasticiser in the Hydration of Cement. *Cem. Concr. Res.* **2005**, *35* (4), 631–636.

[19] Yang, M.; Neubauer, C. M.; Jennings, H. M. Interparticle Potential and Sedimentation Behavior of Cement Suspensions: Review and Results from Paste. *Adv. Cem. Based Mater.* **1997**, *5* (1), 1–7.

[20] Flatt, R. J.; Houst, Y.; Bowen, P.; Hofman, H. Electrosteric Repulsion Induced By Superplasticizers between Cement Particles-An Overlooked Mechanism? In *6th CANMET/ACI Int. Conf. Superplasticizers and Other Chemical Admixtures in Concrete*; Malhotra, V. M., Ed.; American Concrete Institute: Farmington Hills, 2000; pp 29–42.

[21] De Larrard, F.; Ferraris, C. F.; Sedran, T. Fresh Concrete: A HerscheI-Bulkley Material. *Mater. Struct.* **1998**, *31*, 494–498.

[22] Wallevik, J. E. Relationship between the Bingham Parameters and Slump. *Cem. Concr. Res.* **2006**, *36* (7), 1214–1221.

[23] Roussel, N.; Stefani, C.; Leroy, R. From Mini-Cone Test to Abrams Cone Test: Measurement of Cement-Based Materials Yield Stress Using Slump Tests. *Cem. Concr. Res.* **2005**, *35* (5), 817–822.

[24] Gaona, R.; Fritz, C.; Salas, C.; Jameel, H.; Rojas, O. J.; Rojas, O. Self-Association and Aggregation of Kraft Lignins via Electrolyte and Nonionic Surfactant Regulation Self-Association and Aggregation of Kraft Lignins via Electrolyte and Nonionic Surfactant Regulation: Stabilization of Lignin Particles and Effects on Fil. *Spec. ISSUE LIGNIN Nord. Pulp Pap. Res. J.* **2017**, *32* (4).

[25] Flatt, R.; Schober, I. Superplasticizers and the Rheology of Concrete. In *Understanding the rheology of concrete*; Roussel, N., Ed.; Woodhead Publishing: Cambridge, 2012; pp 144–208.

[26] Ikeda, Y.; Nagasaki, Y. Impacts of PEGylation on the Gene and Oligonucleotide Delivery System. *J. Appl. Polym. Sci.* **2014**, *131* (9), 40293.

[27] Yoshioka, K.; Sakai, E.; Daimon, M.; Kitahara, A. Role of Steric Hindrance in the Performance of Superplasticizers for Concrete. *J. Am. Ceram. Soc.* **1997**, *80* (10), 2667–2671.

[28] Murray, L. R.; Bice, J. E.; Soltys, E. G.; Perge, C.; Manneville, S.; Erk, K. A. Influence of Adsorbed and Nonadsorbed Polymer Additives on the Viscosity of Magnesium Oxide Suspensions. *J. Appl. Polym. Sci.* **2018**, *135* (3), 45696.

[29] Lange, A.; Plank, J. Contribution of Non-Adsorbing Polymers to Cement Dispersion. *Cem. Concr. Res.* **2016**, *79*, 131–136.

[30] Hiemenz, P. C.; Lodge, T. P. *Polymer Chemistry*, 2nd ed.; CRC Press: Boca Raton, 2007.

[31] Lindströmn, T. The Colloidal Behaviour of Kraft Lignin. *Colloid Polym. Sci.* **1979**, *257* (3), 277–285.

[32] Kawaguchi, S.; Imai, G.; Suzuki, J.; Miyahara, A.; Kitano, T.; Ito, K. Aqueous Solution Properties of Oligo- and Poly(Ethylene Oxide) by Static Light Scattering and Intrinsic Viscosity. *Polymer (Guildf).* **1997**, *38* (12), 2885–2891.

[33] Ouyang, X.; Ke, L.; Qiu, X.; Guo, Y.; Pang, Y. Sulfonation of Alkali Lignin and Its Potential Use in Dispersant for Cement. *J. Dispers. Sci. Technol.* **2009**, *30* (1), 1–6.

[34] Berg, J. C. *An Introduction to Interfaces & Colloids: The Bridge to Nanoscience*; World Scientific Publishing Company: Singapore, 2010.

[35] Menon, A.; Childs, C. M.; Poczós, B.; Washburn, N. R.; Kurtis, K. E. Molecular Engineering of Superplasticizers for Metakaolin-Portland Cement Blends with Hierarchical Machine Learning. *Adv. Theory Simul.* **2019**, *2* (1800164).

Chapter 3. Molecular Engineering of Superplasticizers for Metakaolin-Portland Cement Blends with Hierarchical Machine Learning[1]

3.1. Introduction

Portland cement (PC) is most widely used engineered material in the world and has become a foundation for modern society[1, 2, 3] due to its unique combination of processability in the plastic state and mechanical properties and water resistance in the hardened state. However, production of portland cement now accounts for 5% of global CO_2 production due to energy-intensive processing and the release of CO_2 during limestone calcination. In order to reduce the carbon footprint of this ubiquitous infrastructure material without compromising performance, it has been proposed that concrete increasingly incorporate minimally processed clays as partial replacement for cement, a practice which significantly reduces the CO_2 intensity and actually improves many important material characteristics, such as corrosion resistance and strength. Calcined clays are available worldwide, and thus present a good option for partial cement clinker substitution on a broad scale.[4] However, calcined clays retain much of the structure[5] of clays from which they are produced. Derived as calcined kaolinite clay, metakaolin (MK) is a common source of pozzolanic material utilized in cementitious systems. The particle size of MK is typically 10x finer that of the cement it replaces, which in concrete reduces porosity and contributes to strength and impermeability. Also, MK exhibits a very high surface area to volume ratio and a high concentration of hydroxyl groups, resulting high reactivity including a strong capacity for water binding. Together, these factors significantly reduce fluidity in pastes, mortars, and concrete, practically limiting the amount of MK that can be combined with PC to~6-10% by mass in

[1] This chapter includes work that was published and reformatted:
Menon, A., Childs, C. M., Poczós, B., Washburn, N. R. & Kurtis, K. E. Molecular Engineering of Superplasticizers for Metakaolin-Portland Cement Blends with Hierarchical Machine Learning. *Adv. Theory Simul.* **2**, (2019).

conventional applications. The addition of water to improve fluidity results in significantly increased porosity and reduced strength, which increases permeability and compromises durability.[6, 7, 8] Dispersion of MK, then, is a critically important step toward increasing the rate of MK substitution for cement clinker, leading to improving the sustainability of modern infrastructure. However, PCEs have unpredictable performance in cements that include metakaolin, which can be due to reduced superplasticizer adsorption to particle surfaces.[9, 10] Metakaolin-blended cements are becoming an important strategy for improving the performance and reducing the environmental impact of cement, but there is a need for superplasticizers tailored to these materials, particularly to facilitate cement blending with larger MK fractions.

Data-driven methods have led to radical changes in the process of materials discovery, but these methods are predicated on large datasets.[11, 12] Hierarchical machine learning (HML) was developed to model the response surfaces of complex physical systems using small, primarily experimental datasets.[10] This is accomplished through embedding domain knowledge of the underlying forces that determine system responses to changes in input variables and separately decomposing the system responses into the underlying forces that most strongly determine variance of the responses using methods of statistical learning. Cement-based materials represent a highly complex and broad range of physicochemical systems. As an initial step toward developing dispersants for use with cement-based systems, the HML methodology was validated on the development of a novel dispersant for concentrated MgO suspensions, which can be considered as a non-setting rheological model of portland cement. Machine learning techniques, such as neural networks,[13] fuzzy logic[14] and support vector regression,[15] have been applied to MK-

PC systems but in these studies, only mechanical properties (e.g., compressive and flexural strength) were modeled as system responses, not plastic properties.

In the application of superplasticizers for clay-blended cements, hypothesis-driven research has resulted in advances,[16] but a general design methodology still has not been established. Here, HML was applied to the design of superplasticizers for pastes composed of MK and PC. It is believed that this is the first time machine learning methods have been used for dispersant design in cement-based systems. Polymer dispersants were designed based on their effects in the pore solution and the mineral particle surfaces, and surrogate physical measurements were performed to elucidate both the functional form of these forces and how the workability of cement paste, as assessed through mini-slump (herein, "slump") measurements,[17, 18] was determined by these forces. The model was trained using a library of seven commercially available superplasticizers, and model predictions were validated through polymer synthesis and assessment of paste workability in MK-PC systems produced with the predicted dispersant design.

3.2. Experimental Methods

3.2.1. Materials

Cement pastes were prepared using an ASTM C150 Type I/II ordinary Portland cement (Saylor, Essroc)[12] and metakaolin (MetaMax, BASF). For the seven tested polymers, Borresperse sodium lignosulfonate (LS1), was received from Borregaard Lignotech. PEGylated Borresperse lignosulfonate (LS1PEG) was synthesized according to the procedure from Gupta et al.[13] Other superplasticizers received from BASF include: a commercial lignosulfonate MasterPozzolith 80 (LS2), a polynapthalene sulfonate MasterRheobuild 1000 (PNS), and two different polycarboxylic ethers, MasterGlenium 3030 (PCE1) and 7500 (PCE2) were obtained from BASF, and one PCE was acquired from GCP Applied Technologies, Adva-190 (PCE3).

Sodium styrene sulfonate (NaSS), methacrylic acid (MAA), poly(ethylene glycol) methyl ether methacrylate (M_n=500) (PEGMA), and azobisisobutyronitrile (AIBN) were purchased from Sigma-Aldrich. Dimethylformamide (DMF) was purchased through Fisher Scientific, and acetone was purchased from Pharmco-AAPER. All materials were used as received from the manufacturer. Ultrapure water was obtained through a purification system (Milli-Q Gradient).

3.2.2. Polymer Synthesis

The synthetic procedure was modified from previous methods and is expected to yield a random copolymer.[19, 20] To a flask under N_2 environment were added NaSS (2.500 g, 0.01212 mol), MAA (0.5 mL, 0.0061 mol), and PEGMA (2.8 mL, 0.0061 mol). The monomers were dissolved to 1.0 M in a 50/50 w/w mixture of water and DMF. The reaction mixture was again purged with flowing N_2 then the radical initiator AIBN (0.0035 g, $2.4\text{x}10^{-5}$ mol) was added. The reaction was allowed to stir under a nitrogen atmosphere for 12 h at 80 °C. After exposure to air, the solution was introduced dropwise into cold acetone and precipitated. After washing and decanting with acetone, the polymer was dissolved in water and concentrated under vacuum to yield a white solid. A schematic reaction is shown in Figure 3.1. Characterization via ^{1}H NMR and IR analysis were performed are presented in the Chapter 3 appendix.

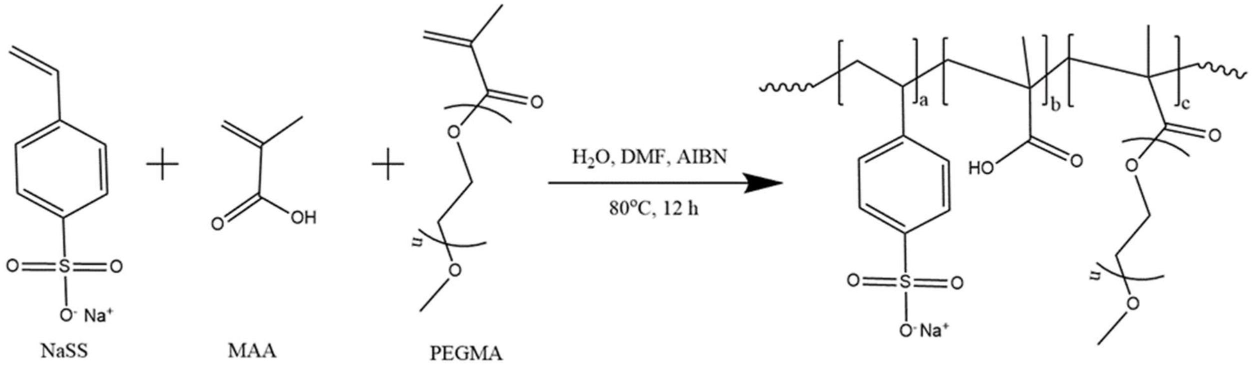


Figure 3.1. Reaction schematic for synthesis of NaSS-PMAA-PEGMA terpolymer.

3.2.3. Adsorption Measurements

Adsorption of the polymers in the training set onto MK and PC was performed by total organic carbon (TOC) analysis. Samples of 0, 0.25, 0.50, and 1.00 wt% to binder of each superplasticizer were prepared. A reference sample for each sample was made to compare the difference in amount of carbon before and after adsorption. For each different sample, the superplasticizer solution was mixed with 25 g cementitious material and vibrated for 3 minutes. The samples were then centrifuged for 10 min at 4400 rpm. The supernatant was collected, filtered through a 0.45 µm syringe filter, diluted to 40 mL with water, and immediately analyzed using a combustion-based TOC (Shimadzu TOC-L).

3.2.4. Zeta Potential

Zeta potential was measured in aqueous solutions with a polymer concentration of 10 mg mL^{-1} using a Zeta-sizer (Malvern Instruments).

3.2.5. Osmotic Pressure

Osmolality of aqueous solutions of each polymer was measured 0.1, 0.2 and 0.4 g mL^{-1} using a vapor pressure osmometer (Wescor 5520). To model changes in osmotic pressure due to superplasticizer dissolved in the pore solution, the second virial coefficient of each was acquired by plotting osmolality values versus concentration of pure polymer solutions. The slope of this line was determined and multiplied by the molar volume and molality of water to obtain the A_2 coefficient.[21]

3.2.6. Polymer Intrinsic Viscosity

Due to the poor solubility at higher concentrations of many of the polymers, Kraemer and Huggins curves[22] did not fit to the data accurately. Instead, a single-point determination of the intrinsic viscosity at a low concentration of 0.0033 g ml^{-1} was used. The lower concentration prevented polymer aggregation so that intrinsic viscosity could be determined utilizing the Solomon–Ciuta equation.[23] An Ubbelohde Viscometer (Canon Instrument Company) was used to determine the relative viscosity by taking the ratio of the time to pass through the capillary of the polymer solution to the pure water elution time. Three trials for each polymer along with the solvent elution time were performed and the results were averaged.

3.2.7. Sedimentation

Direct measurement of electrosteric interactions in a particle suspension is difficult to measure exactly and the results, when parametrized by the details in chemical structure, are not gauged directly on a *per polymer* basis.[24] Here, sedimentation, as a quantitative gauge of electrosteric interactions, was measured for all the polymers at 0.5 wt% for 50 wt% of MgO particles. As components of PC could dissolve over the course of this test, MgO was used as a non-setting model, although hydration also can occur. The same conditions were used to determine the electrosteric effects of polymer and MK by adding 0.5 wt% polymer in a 50 wt% MK suspension. Following a previously published method,[25] the height of the supernatant was measured after 24 h and 120 h. Electrosteric parameters are expressed in percent change in supernatant height relative to suspensions in pure water. Positive numbers imply lower dispersion, resulting in a more-compact sediment and more supernatant, and negative numbers imply good dispersion relative to suspensions in pure water. The negative values measured for MK are attributed to the smaller particle size than MgO, but relative differences in the MK measurements can be attributed to dispersing effects of the superplasticizer.

3.2.8. Assessment of Workability

To assess the workability and changes in yield stress of the cement pastes mini-slump testing was performed.[26] Cement pastes were prepared at room temperature using a planetary mixer (Hobart) with a paddle speed of 62 rpm. To the mixer a 15% MK/85% PC mixture was added. Total water addition produced a 0.50 water-to-cementitious materials (w/cm) ratio paste, where both MK and PC were considered as cementitious materials. Superplasticizer was added at a ratio of 0.25% to total weight of cementitious material. To ensure optimal effectiveness of the superplasticizers, a fraction of the total water volume was added and mixed for 1 min. The remainder of the water containing the dissolved superplasticizer was added immediately afterwards and mixed for another 1 min.[27]

Within 60 s after mixing, the cement pastes were added into a mini-slump cylinder (BASF, Construction Chemicals Division) with dimensions 3 cm in diameter and 5 cm in height. The cement was lightly compressed with the backside of a spoon to even with the top. The cylinder was slowly and evenly lifted in a continuous motion allowing the cement pastes to slump. The change in height due to the slump was recorded along with the spread, and the results were expressed in percent as the change in height divided by the original height of 5 cm. For every mixture of cement tested, three slumps were tested in succession with an average of the trials being used for analysis.

3.3. Computational Methodology

All modeling was performed in MATLAB. The dependence of physicochemical forces on polymer composition were determined using the LSQR method by dividing the seven-dimensional vector of each polymer property by the matrix of polymer composition for all seven polymers in our training set. Variable selection in decomposing the slump into contributions from the physicochemical forces and their cross-products were performed using Lasso with leave-one-out cross-validation (LOOCV). Finally, constrained nonlinear optimization was performed on the master function expressing system response (slump) as a function of compositional variables in the bottom layer of the model using the fmincon solver from optimization toolbox in MATLAB.

3.3.1. Model Development

The HML algorithm developed has three layers. The top layer represents the system response to be predicted or optimized, and the bottom layer represents input variables that are used to control this response, similar to neural network models. However, in HML, the middle layer represents the underlying physicochemical relationships that determine system responses. These can be considered latent variables that can be estimated via surrogate physical measurements, resulting in expressions parametrized for these forces by general compositional variables. The complete structure of the algorithm for superplasticizer design is shown schematically in Figure 3.2.

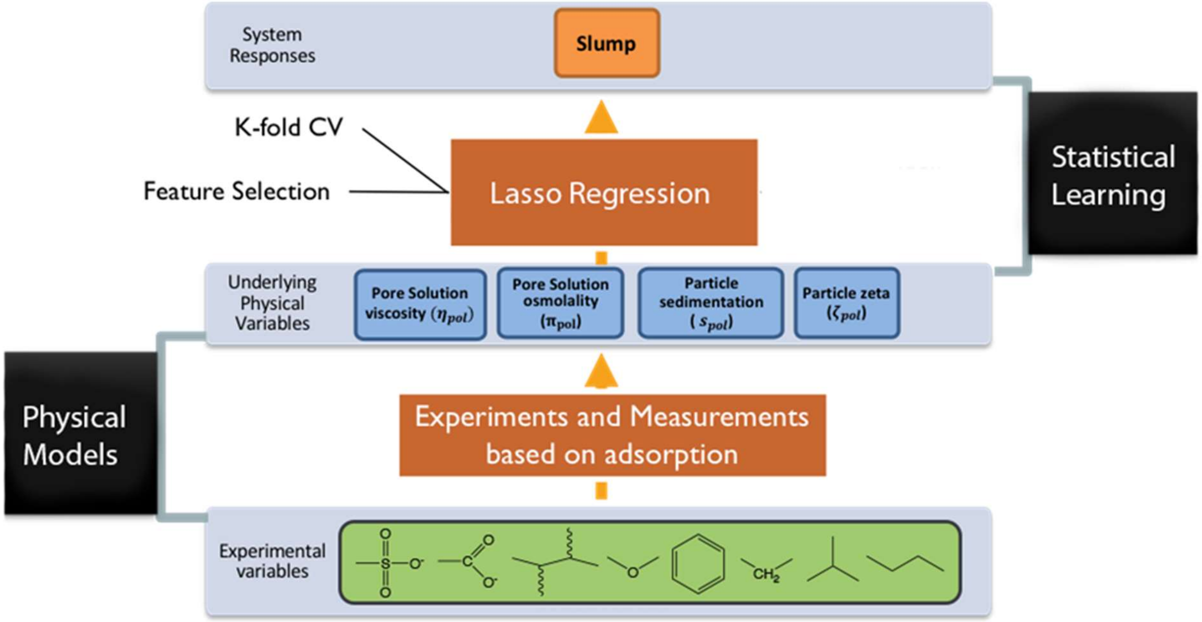

System Responses
Slump
K-fold CV
Feature Selection
Lasso Regression
Statistical Learning
Underlying Physical Variables
Pore Solution viscosity (η_{pol})
Pore Solution osmolality (π_{pol})
Particle sedimentation (s_{pol})
Particle zeta (ζ_{pol})
Physical Models
Experiments and Measurements based on adsorption
Experimental variables
O
S
O
O⁻
O
C
O⁻
O
CH2

Figure 3.2. A schematic for HML approach for the MK-PC/SP system. In this work, the cement and MK composition, w/cm, and superplasticizer dosage were held constant, making superplasticizer composition (represented by combinations of functional groups) and architecture (branched or linear) the only free variables.

The goal was ultimately to predict the optimal superplasticizer composition that maximized slump. However, the algorithm used estimates of the underlying forces in the optimization process, which improves interpretability of the algorithm predictions and allows for molecular engineering in complex material systems.

In suspensions, adsorption determines the partitioning between dissolved and adsorbed polymer, setting the balance between solution and particle forces that superplasticizers exert in tuning the rheology of cement paste. In modeling the underlying forces that determine the slump of cement paste, solution forces are gauged through measurement of intrinsic viscosity ($[\eta]$) and the second virial coefficient (A_2). These parameters give related insight into polymer-solvent interactions, reflecting the hydrodynamic volume of polymer chains under flow and the colligative effects of dissolved polymer, respectively.[28, 29] Both parameters can be measured directly using low-concentration polymer solutions.

The effects of adsorbed polymer on particle forces were represented by changes in electrostatic and electrosteric interactions. While changes in electrostatic forces of particles due to adsorbed polymer can be estimated through direct measurement of the zeta potential (ζ), the high ionic strength of the pore solution in cement paste can make this ineffective in inhibiting particle coagulation. Electrosteric forces (s) are a combination of electrostatic forces, which mediate adsorption onto the charged cement surfaces, and steric forces due to pendant water-soluble oligomers that generate an additional barrier to particle aggregation. The mechanism of PCE dispersion is hypothesized to be due primarily to electrosteric effects, but these are difficult to measure directly.[30] In the previous study on dispersant design using HML, the electrosteric contribution to particle-particle interactions was gauged using sedimentation measurements on dilute solutions of MgO.[10] While this only provides trends, it was found to be sensitive to aggregation effects and thus provided the algorithm with information on electrosteric design principles.

In the first step of developing an HML model, the forces that mediate system response, which are represented in the middle layer, are estimated from separate measurements of these physical properties in the training set. Using least-squares regression, functional representations of these four forces are represented in terms of superplasticizer composition ($\vec{x}$), where $\vec{x}$ is the vector of functional groups represented in mole fraction and polymer architecture (linear or branched). Each of these forces and equations for representing them based on the measurements for the machine learning model are shown in Figure 3.3.

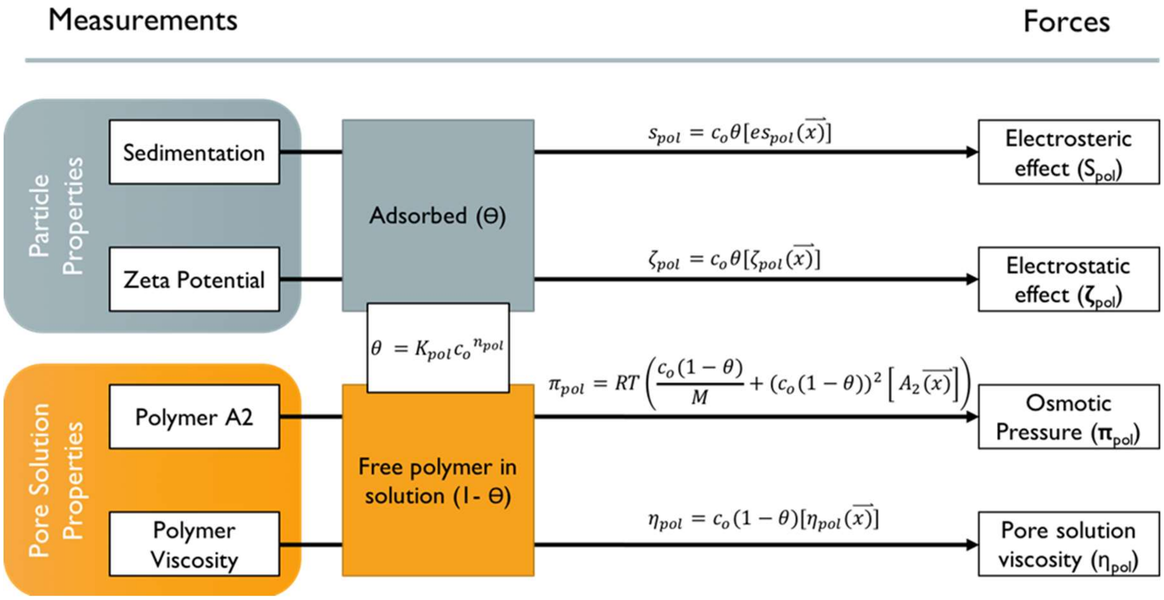

Measurements
Forces
Particle Properties
Sedimentation
Zeta Potential
Adsorbed (Θ)
$s_{pol} = c_o\theta[es_{pol}(\overrightarrow{x})]$
Electrosteric effect (S_{pol})
$\zeta_{pol} = c_o\theta[\zeta_{pol}(\overrightarrow{x})]$
Electrostatic effect (ζ_{pol})
$\theta = K_{pol}{c_o}^{n_{pol}}$
Pore Solution Properties
Polymer A2
Polymer Viscosity
Free polymer in solution (1- Θ)
$\pi_{pol} = RT\left(\frac{c_o(1-\theta)}{M} + (c_o(1-\theta))^2\left[A_2\overrightarrow{(x)}\right]\right)$
Osmotic Pressure (π_{pol})
$\eta_{pol} = c_o(1-\theta)[\eta_{pol}(\overrightarrow{x})]$
Pore solution viscosity (η_{pol})

Figure 3.3. This schematic represents the connection between the first and second layers of HML. Individual polymer measurements are modeled by the physical equations determining the forces.

The next step in the HML algorithm is decomposing the system response in terms of the underlying forces in the middle layer. In doing so, the goal was to develop an expression for the system response in terms of the underlying forces (and combinations of forces) to elucidate which were dominant predictors of the system response. These forces were estimated for the training set based on experimental measurements and were normalized onto the range [-1, 1] or [0, 1], depending on the allowed values, to make the relative values of the coefficients meaningful in the expression for system response. In defining the basis set for representing system response, cross-products of the forces (e.g., product of intrinsic viscosity and zeta potential variables) were included, which can be interpreted as couplings between the different forces. However, these additional combinations of the forces result in a total basis set composed of 10 variables, which can lead to overfitting of the training set and low accuracy in the predictions of the performance of new polymers.

Variable selection is a central component of machine learning and particularly challenging in modeling physical systems for which dimensionality reduction is of critical importance but the functional form of the response surface has physical significance.[31] In this study regularized regression using the least average shrinkage and selection operator (Lasso) was performed to identify the physicochemical forces that most strongly determined the slump. In modeling the dependence of responses y to independent variables x, Lasso includes an additional term in the cost function based on a positive tuning parameter λ and the L_1 norm of the model coefficients β:[32]

$$min_{\beta}(\|y-\beta x\|_2^2+\lambda\|\beta\|_1) \quad (1) \qquad \text{Eq. 3.1}$$

The L_1 penalty drives coefficients of variables that are weaker predictors of responses to zero as the tuning parameter is increased,[33] which, in the context of physical models, allows for identification of a sparse feature set of forces that determine responses, facilitating minimal model development and improving interpretability.

In the final step of the algorithm, the system response (slump) was re-parametrized from the underlying forces to a representation based on system variables (superplasticizer composition) via the relationships established in the first step. This resulted in a functional representation of system response in terms of system variables, thus providing an efficient method for learning the response based on small numbers of experimental measurements. From this response surface, global maxima can be identified, subject to constraints in composition or allowed slump values, using standard optimization techniques. Thus, HML can serve as a design tool for complex physical systems based on knowledge of the underlying forces that drive system responses.

3.4. Results, Discussion, and Validation

3.4.1. Results and Discussion

The system response to be optimized was the workability of a 0.50 w/cm paste consisting of 15% MK and 85% PC, and a constant superplasticizer dose of 0.25% [based on total cementitious mass (cm)]. By holding the cement and MK composition, w/cm ratio, and polymer dose constant, the only free variables were polymer composition. The training set used to develop the algorithm was composed of seven superplasticizers, six of which were commercial products, and the structures of these polymers are shown in Figure 3.4. A simple representation of these polymers was adopted here for the purposes of optimization and guiding subsequent synthesis: the chemical structures were parametrized by the estimated mole fraction of each functional group estimated from ^{1}H NMR spectroscopy, as well as the chain architecture – branched for the three lignin derivatives and linear for the rest of the library. This resulted in a vector representation of polymer chemistry to be optimized for maximizing the workability of MK-PC pastes, which are tabulated in the appendix. Differences in the structures of the three PCE's were due to the MAA:PEGMA ratio and the molecular weight of the pendant PEG oligomer, and the two LS analogues were differentiated by sulfonate:alkyl ratio. The base structures are shown in Figure 3.4.

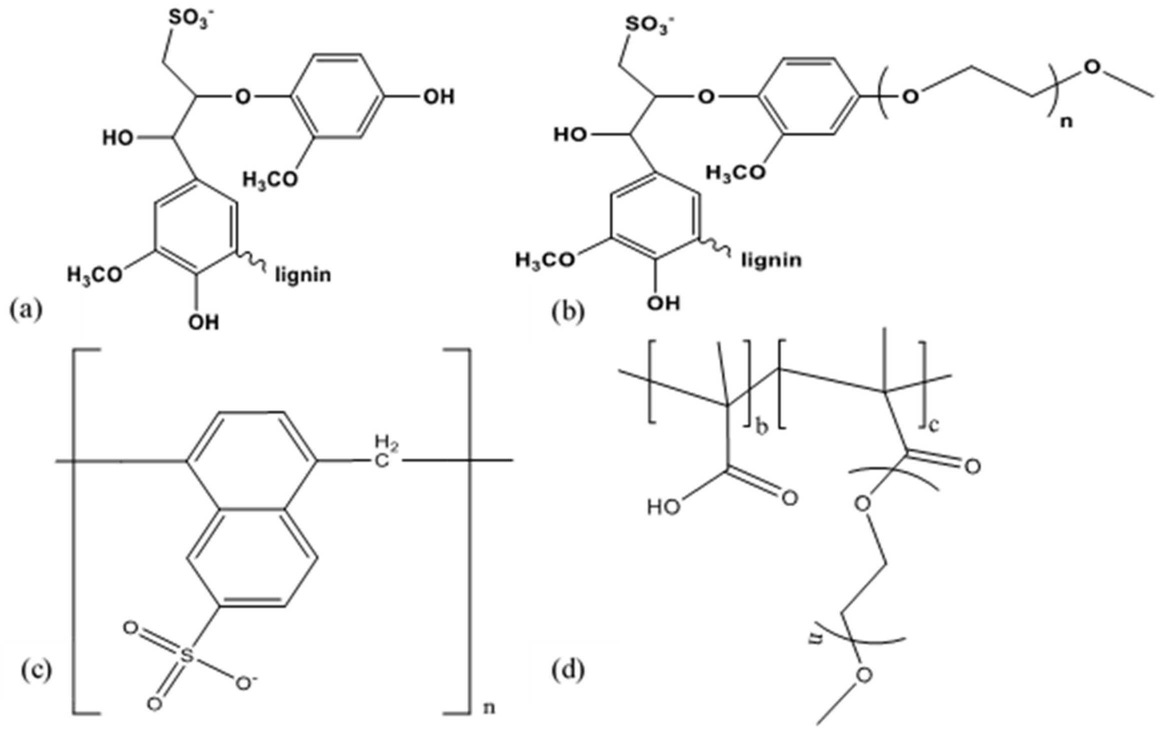


Figure 3.4. Chemical structures of the superplasticizers used to train the algorithm where (a) is lignosulfonates i.e. LS1 and LS2, (b) LS1PEG, (c) PNS and (d) PCE1, PCE2, and PCE3.

To train the algorithm, the first step was determining changes in paste workability across the polymers in the library. Slump measurements (in units of cm) are a quantitative assessment tool of workability, and relate to the yield stress of the paste, which is thought to follow the predictions of the Bingham model:[34]

$$\tau = \eta\dot{\gamma} + \tau_0 \qquad \text{Eq. 3.2}$$

and predicts that mechanical equilibrium will be reached when the shear stress τ is equal to the sum of the product of the viscosity η and shear stress rate $\dot{\gamma}$ with the yield stress τ_0. Slump is a commonly used measure of cement paste workability, providing a gauge of these rheological parameters.[22]

The slump values across the training set are shown in Figure 3.5. For comparison, the values across the training set for paste based on plain portland cement paste at constant w/cm ratio of 0.35 and polymer dose of 0.25%. Large slump values were measured for all PC pastes except LS1 and LS1PEG. However, in MK-PC blends, the slump values were significantly lower for several members of the training set, as expected, with particularly large decreases in efficacy for PCE1, LS2, and PNS with the use of MK. Overall, the lignosulfonates, LS1 and LS1PEG, were less effective, with generally lower slump values (or less workable pastes), as expected since these are generally considered to be mid-range water reducing admixtures. PCE2 and PCE 3 appeared to be relatively effective at achieving workability at this w/cm and at this MK-PC blending rate; in practice, lower w/cm are desirable for both strength and durability.

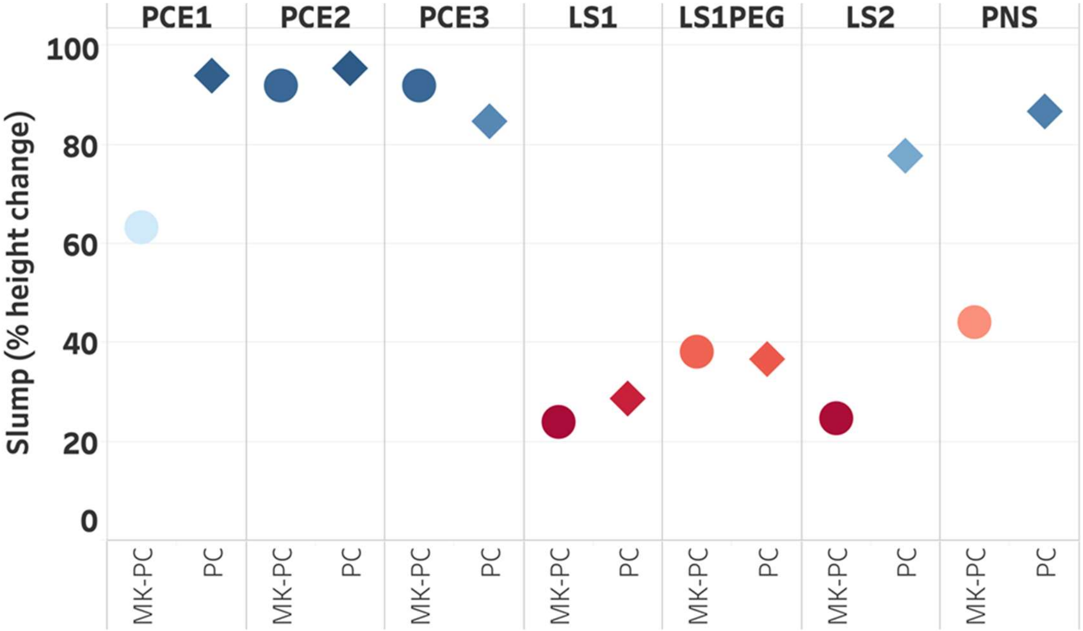


Figure 3.5. Slump values for PC and MK-PC blends containing superplasticizers dose in the training set at 0.25%.

The accepted mechanism of superplasticizer function in PC pastes is based on adsorption to the particle surface.[35, 36] For blends of MK and PC, adsorption was modeled utilizing the

Freundlich equation, which is an empirical model useful for heterogenous materials and not limited to single-layer adsorption:[37]

$$\theta = KC_i{}^n \qquad \text{Eq. 3.3}$$

Here θ is the fraction of adsorbed polymer, C_i is the polymer concentration, and K and n are empirical parameters. The parameters log(K) and n are the intercept and slope, respectively, of the best fit line to the log-log plot of adsorbed amount vs concentration.[38] With adsorption data, a Freundlich model was developed for each polymer-cementitious material pair allowing an integrated parameter for the algorithm in terms of polymer composition. To account for the MK-PC mixture the amount of expected adsorption between the two components was predicted according to Eq. 3.4:

$$\theta_{Total} = \left(\frac{MK}{CM}\right) * \theta_{MK} + \left(\frac{PC}{CM}\right) * \theta_{PC} \qquad \text{Eq. 3.4}$$

where θ_{Total} is the total percentage of expected adsorbed polymer, $\frac{MK}{CM}$ is the ratio of metakaolin to total cementitious material, $\frac{PC}{CM}$ is the ratio of portland cement to total cementitious material (expressed in terms of mass, not surface area), θ_{MK} is the concentration-dependent fraction of polymer adsorbed onto MK, and θ_{PC} is the concentration-dependent fraction of polymer adsorbed onto PC.

Adsorption values across the polymer training set are shown in Figure 3.6. Only PCE1 and PCE3 had a significantly higher affinity for the MK than PC, whereas most superplasticizers exhibited much lower adsorption. However, the relationship between adsorption and slump values was complex. While PCE3 had a high affinity for MK and resulted in large slump values for MK-

PC blends, PCE1 also adsorbed more strongly onto MK than PC but the slump values were lower in the blends.

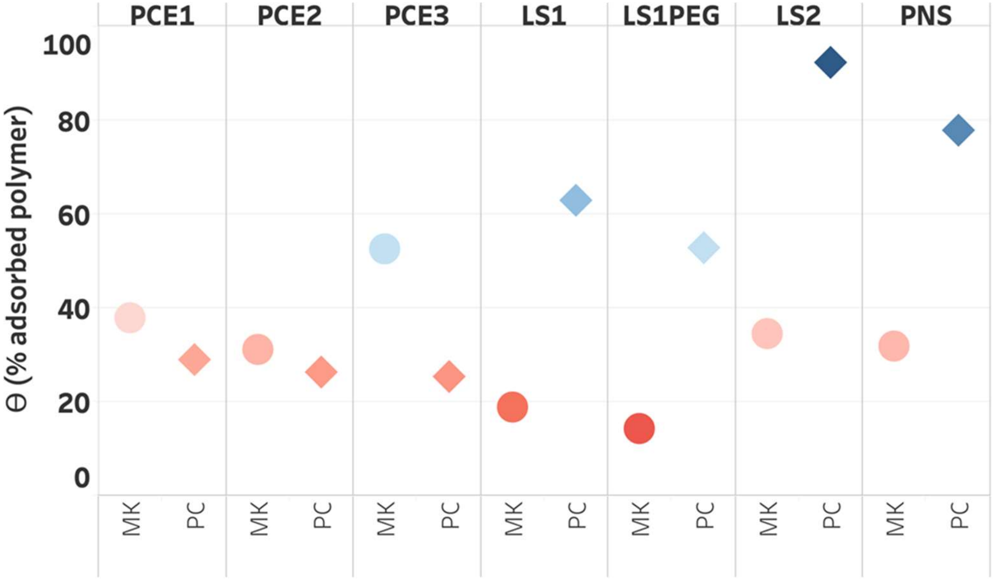


Figure 3.6. Adsorption results plotted in terms of fraction of adsorbed polymer on metakaolin and portland cement at superplasticizer dose of 0.25%.

For the PCE's, the high slump and relatively low adsorbed fractions suggest that higher than necessary dosages of superplasticizer were introduced in these PC pastes. The 0.25% dosage rate was used to restrict the variance to superplasticizer chemistry, but the slump values were near the maximum possible for most of the experiments with PCE's. While this reflects the high performance of these dispersants, it does result in an insensitive measure of their relative dispersing power in PC and MK-PC. Future experiments could identify doses that result in a constant slump value.[9] However, this requires including superplasticizer concentration as an additional variable, and the admixtures may have different mechanisms at different concentrations, which involves more data and potentially a more robust model.

The polymer effects on solution forces of viscosity and osmotic pressure, and on the electrostatic and electrosteric interactions between particles were measured separately, and the results are tabulated in the Chapter 3 appendix along with the cross-correlation matrix that illustrates correlations between the different physicochemical parameters. The intrinsic viscosity and A_2 measurements provide complementary insight into polymer-solvent and polymer-polymer interactions in which the former associated are associated with polymer hydrodynamic volume under applied shear stress and the latter are associated with changes in colligative properties of the solution due to dissolved polymer. Indeed, the difference between these can be appreciated from considering the molecular weight dependence of the two quantities: the intrinsic viscosity depends directly on chain dimensions while the osmotic pressure depends on volume fraction and has at most weak dependence on molecular weight (scaling approximately as $M^{-0.25}$ in good solvents).[39]

While the two measures were positively correlated, there were differences in members of the same group that suggested details of the chemistry were important. For example, the highest intrinsic viscosity, which is associated with the largest hydrodynamic volume, was measured for PCE2 while the largest value of A_2, indicating the most favorable polymer-solvent interaction, was measured for PCE1. Similar trends in the zeta potential of the dissolved polymer and the electrosteric parameters estimated from sedimentation experiments were observed.

Based on these values, a polynomial to predict the slump (S) for MK-PC ($S_{MK\text{-}PC}$) expressed as a function of these polymer properties was developing using Lasso with cross-validation. For a basis set composed of the five physicochemical forces (η, A_2, ζ, s_{MK}, s_{PC}) and the nine cross-products (ηA_2, $\eta\zeta$, ηs_{MK}, ηs_{PC}, $A_2\zeta$, $A_2 s_{MK}$, $A_2 s_{PC}$, ζs_{MK}, ζs_{PC}), the 14-variable basis set exceeded the 7-element training set. To avoid artificially low error estimates in a significantly underdetermined statistical model, Lasso was performed in three stages. In the first, variable selection was performed using only the linear force terms (η, A_2, ζ, s_{MK}, s_{PC}). At a tuning parameter value of 0.002, Lasso selected η and ζ as the two forces that most strongly determined slump in pastes of the MK-PC blend. Then a variable set based on the cross-products of the second virial coefficient were included with the two linear terms from the first round as a second trial basis set (η, ζ, $\eta\zeta$, ηA_2, $A_2\zeta$, $A_2 s_{MK}$, $A_2 s_{PC}$). However, at $\lambda = 0.002$, Lasso discarded the cross-term $\eta\zeta$ and all the terms in A_2, suggesting that these cross-terms were not descriptors of slump in this system. Finally, the remaining cross-terms were included with η and ζ to form a feature set (η, ζ, ηs_{MK}, ηs_{PC}, ζs_{MK}, ζs_{PC}, $s_{MK} s_{PC}$) for describing the changes in slump in terms of the underlying solution- and particle-mediated forces. At $\lambda = 0.002$, the final variable set was identified as (η, ζ, ζs_{MK}, ηs_{PC}). Following cross-validation, a function for slump of the form was identified:

$$S_{MK-PC} = 1.11\eta - 0.55\zeta + 0.36\zeta s_{MK} + 0.12\eta s_{PC} \qquad \text{Eq. 3.5}$$

It is unexpected that the model assigned the greatest weight to the viscosity term (η), which contrasts with accepted models of dispersant design that focus on the role of adsorbed polymers in tuning the electrostatic (ζ) and electrosteric (s) forces between particles.[40] The other terms in (5), based on polymer electrostatic and electrosteric characteristics, also tune the slump in MK-PC paste, but the strongest dependence was founded to be through the polymer intrinsic viscosity.

Reparameterization of Eq. 3.5 by the composition-dependent representations of the forces derived from surrogate physical measurements results in a response surface for the slump of MK-PC paste as a function of superplasticizer composition. The maximum of this surface corresponded to a polymer having a linear architecture and the combination of functional groups shown in Table 3.1.

Table 3.1. Optimal mole fractions of functional groups identified through maximization of the response surface for slump in terms of superplasticizer composition defined by functional groups.

Functional Group	sulfonate	carboxylate	PEG	alkyl	aromatic
	0.20	0.17	0.20	0.22	0.20

3.4.2. Validation

Based on these mole fractions of functional groups, a random terpolymer composed of styrene sulfonate (SS), methacrylic acid (MAA), and poly(ethylene glycol) methacrylate (PEGMA) was synthesized by free radical polymerization. Styrene sulfonate was chosen due to the equal values of sulfonate and aromatic groups and a necessary fraction of alkyl group to form the backbone, and MAA and PEGMA were incorporated at approximately the prescribed mole fractions in synthesizing poly($SS_{0.50}MAA_{0.25}PEGMA_{0.25}$). From a molecular engineering perspective, this terpolymer resembles a PCE combined with poly(styrene sulfonate) (PSS). PSS has been explored as a dispersant for cement,[41] but it has minimal effects on workability. However, the homopolymer had a high intrinsic viscosity that depends strongly on polymer and salt concentration,[42] but no reports of copolymers for cement applications have been reported. While the effects of adsorbed PCE have been studied and modeled in great detail,[24] solution characteristics of PCE's have not been explored in depth.

The physical properties of the HML prediction are shown in Table 3.5, and this terpolymer that was rich in SS monomers displayed a very high intrinsic viscosity of 443.0 mL/g, but otherwise the A_2, zeta potential, and sedimentation values were similar to other polymers in the training set. In analyzing the relationships between structure and composition and the resultant properties, the intrinsic viscosity was generally higher for linear polymers than the crosslinked lignosulfonates, although the lowest value of intrinsic viscosity was found for PNS, but in this case the high aromatic content resulted in a negative A_2 coefficient, which correlates with a reduced hydrodynamic volume due to unfavorable solvent interactions. Incorporation of alkyl carboxylate and PEG functional groups would serve to increase solubility as well as increase the repulsive electrosteric interactions between MK particles due to adsorption of this terpolymer.

To test the effects of the algorithm predictions on the workability of MK-PC paste, adsorption and slump measurements were performed at the same superplasticizer dose in both PC and MK-PC at the same w/cm values as in the training set. As shown in Figure 3.7, while the addition of poly($SS_{0.50}MAA_{0.25}PEGMA_{0.25}$) had no visible effect on the PC paste slump, but it resulted in a significant increase in slump of MK-PC paste. In comparing with the results from the training set shown in Figure 3.5, the slump value of 72% was similar to those measured for the commercial PCE superplasticizers in MK-PC paste. This was unexpected because with currently available superplasticizers, the reverse is common but there have been no reports of selective plasticization of MK-blended cements. This suggests that blended superplasticizer systems could be useful in exploiting multiple plasticization mechanisms.

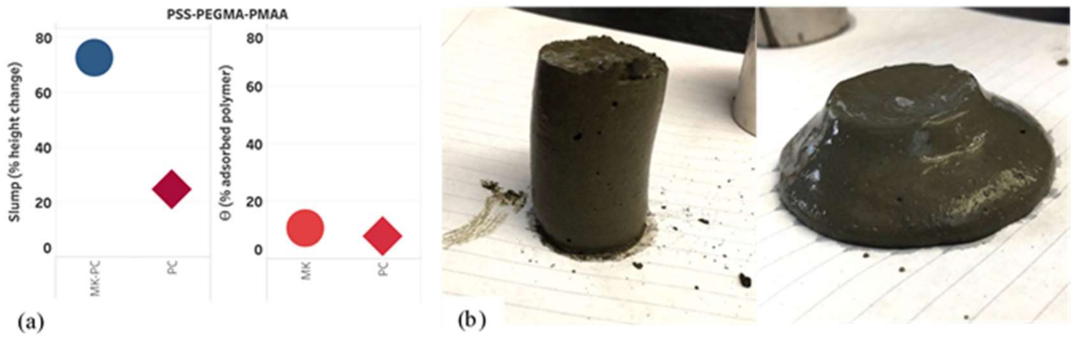


Figure 3.7. (a) Slump and adsorption values for poly($SS_{0.50}MAA_{0.25}PEGMA_{0.25}$) in pastes composed of PC or MK-PC. (b) Representative slump images for pastes PC (left) and PC-MK (right) with poly($SS_{0.50}MAA_{0.25}PEGMA_{0.25}$).

In exploring the mechanism of plasticization, it was also unexpected that poly($SS_{0.50}MAA_{0.25}PEGMA_{0.25}$) adsorbed weakly to both PC and MK, with adsorbed fractions of 7.9% and 10.6%, respectively. This was significantly lower than all members of the training set, and less than half of any of the adsorption values measured for the PCE's. This suggests that the mechanism is not based primarily on adsorbed polymer but instead follows the predictions of Eq. 3.5 in having the dominant force be due to pore solution viscosity. As the other particle-based terms in Eq. 3.5 suggest, adsorbed polymer still plays a role, and the carboxylate and PEG functional groups in poly($SS_{0.50}MAA_{0.25}PEGMA_{0.25}$) are hypothesized to mediate adsorption and steric interactions as with the PCE copolymers, although the details of this interaction with MK will need to be elucidated in the context of the complex chemistry and morphology of calcined clays.[5] However, the HML algorithm led to the design of a superplasticizer specific for MK-PC with a novel mechanism of action involving significant contributions from non-adsorbed chains in contrast with the role of non-adsorbed PCE in PC plasticization where reductions in PCE adsorption due to longer PEG side chains resulted in increases in yield stress and plastic viscosity of the paste.[43]

While the structure of the HML algorithm has provided insight into the mechanism of action via the viscosity of the pore solution, there are still fundamental questions that need to be addressed. The first is understanding how the algorithm determined the connections between polymer composition and the physical properties. The model for superplasticizer effects on the slump as a function of the underlying forces in MK-PC indicated a strong effect for intrinsic viscosity, but the PCE polymers in the training set had the highest intrinsic viscosities while the polymers with aromatic sulfonates had relatively low intrinsic viscosity. This may have been a coincidence, although there are clear correlations between intrinsic viscosity and solubility in the aromatic sulfonates, and increasing the solubility through incorporation of alkyl carboxylate or PEG functionality could significantly increase the intrinsic viscosity. This question can be better addressed through parametrization of superplasticizer chemistry by discrete monomers, not functional groups. Doing this would require a larger training set, but would resolve ambiguity in interpreting the results.

A second fundamental question surrounds the physical basis of how the solution viscosity can have a significant effect on the yield stress of a concentrated suspension. In the original application of HML to designing dispersants for concentrated MgO suspensions, the dominant forces were products of solution forces and particle forces, particularly coupling between the osmolality and the zeta potential mediated by free and adsorbed polymer. While there is theoretical support for this in models of polymer dispersants, there has been less work on the role of purely solution forces on workability, although experimental measurements on model systems indicate that this can be a strong determinant of suspension yield stress.[44] The introduction of metakaolin, which is believed to retain much of its intrinsic clay structure, lends further complexity due to its heterogeneous surface charge density and pH-dependent flocculation and adsorption behaviors. Addressing these issues may require parametrization of the system chemistry to resolve contributions from particle-particle interactions in the percolating network that comprises paste from solution-mediated effects due to changes in viscosity and polymer adsorption,[45] but this would provide a broader understanding of the forces that drive the rheological properties of MK-PC paste.

Finally, the broader applications of HML to molecular engineering, in advanced cementation systems but more broadly to complex material systems, require further study. The results here suggest that the algorithm can develop novel predictions for molecular design based on small experimental datasets and knowledge of underlying forces, but open questions center on determination of learning rates, transfer learning, and other algorithm structures that allow embedding of domain knowledge. Advances in machine learning algorithms for discovery can be leveraged in reducing the uncertainty of predictions, and not simply in optimization.[46, 47] There is great potential in the application of machine learning to molecular engineering of complex systems, and hierarchical models may provide a powerful framework in designing for function, not composition.

3.5. Conclusion

The HML algorithm was used in the molecular engineering of a superplasticizer tailored for MK-PC blends. Based on knowledge of the underlying forces in concentrated particle suspensions and data on seven superplasticizers, the algorithm predicted that the workability of MK-PC blends was strongly determined by the intrinsic viscosity of the pore solution, with contributions due to electrostatic and electrosteric interactions between particles being secondary factors. Following reparametrization of the function for slump by superplasticizer composition and architecture, the global minimum of this response surface was consistent with a random terpolymer poly($SS_{0.50}MAA_{0.25}PEGMA_{0.25}$), which was distinct in having high fraction of styrene sulfonate. Evaluating this terpolymer superplasticizer, the MK-PC paste slump was found to be significantly greater than in pure PC paste. Furthermore, this terpolymer was found to have a high intrinsic viscosity but low adsorption onto PC and MK, indicating that the HML algorithm had predicted a novel mechanism for the plasticization of MK-PC blends and identified a polymer composition that achieved this. This terpolymer could be an important step toward improving the performance of low-energy cements and improving the sustainability of this vital infrastructure material. Furthermore, HML could be broadly useful in the molecular engineering of technologically relevant materials for complex physical systems.

3.6. Research Contributions

C.M. Childs performed polymer and cement property characterization along with polymer synthesis, chemical characterization, and support in ML analysis. A. Menon performed ML analysis and optimization along with providing support in polymer and cement property characterization.

3.7. Appendix

Table 3.2. Composition of polymers in the training set.

Polymer	Sulfonate	Carboxylate	PEG	Alkyl	Aromatic	Crosslinked	Linear
PCE1	0	0.33	0.33	0.34	0	0	1
PCE2	0	0.25	0.5	0.25	0	0	1
PCE3	0	0.3	0.4	0.3	0	0	1
LS1	0.33	0	0	0.34	0.33	1	0
LS1PEG	0.25	0	0.25	0.25	0.25	1	0
LS2	0.4	0	0	0.2	0.4	1	0
PNS	0.5	0	0	0	0.5	0	1

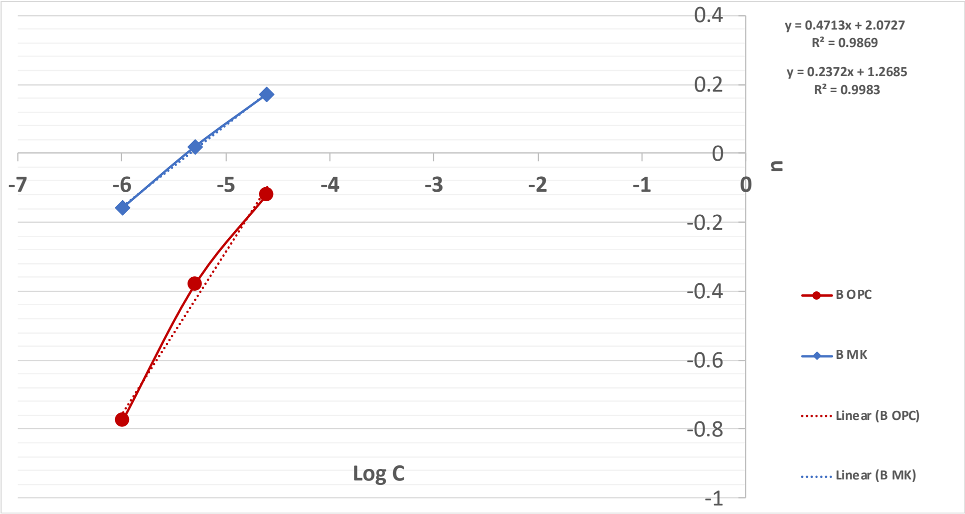


Figure 3.8. Freundlich adsorption curves for LSPEG.

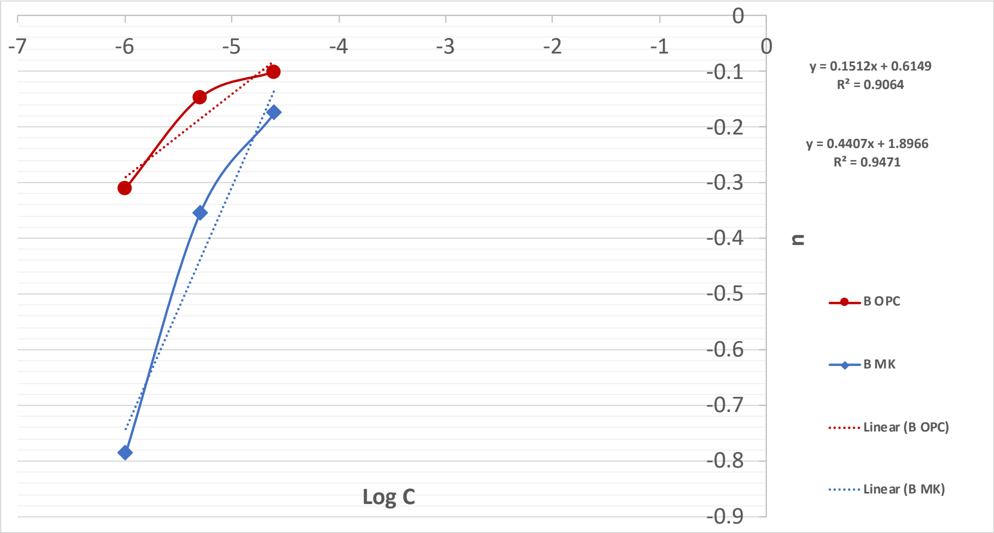


Figure 3.9. Freundlich adsorption curves for MG7500.

Table 3.3. Polymer properties for polymers in the training set.

Polymer	**Intrinsic Viscosity**	**A2**	**Zeta**	**PC sedimentation property**	**MK sedimentation property**
PCE1	26.78	5.84	-12.13	26.47	-82.72
PCE2	54	4.17	-14.10	74.03	-246.78
PCE3	42	1.82	-20.60	32.46	0.00
LS1	4	-1.66	-53.97	53.42	-356.15
LS1PEG	4.5	2.67	-47.67	8.86	-106.35
LS2	2.64	1.12	-19.97	0.00	-79.80
PNS	2	-1.81	-19.77	6.41	-93.96

Table 3.4. Polymer properties for individual compositional variables calculated using LSQR method.

Polymer	Intrinsic Viscosity	A2	Zeta	PC sedimentation property	MK sedimentation property
Sulfonate	-13.96	-2.14	-2.93	-21.47	20.11
Carboxylate	-7.23	30.00	291.20	-455.11	2502.12
PEG	53.05	3.50	-37.90	65.80	-186.16
Alkyl	7.05	-21.56	-236.38	391.47	-2112.40
Aromatic	-13.96	-2.14	-2.93	-21.47	20.11
Crosslinked	6.56	7.49	26.78	-73.78	377.87
Linear	18.39	0.16	-15.73	33.00	-134.10

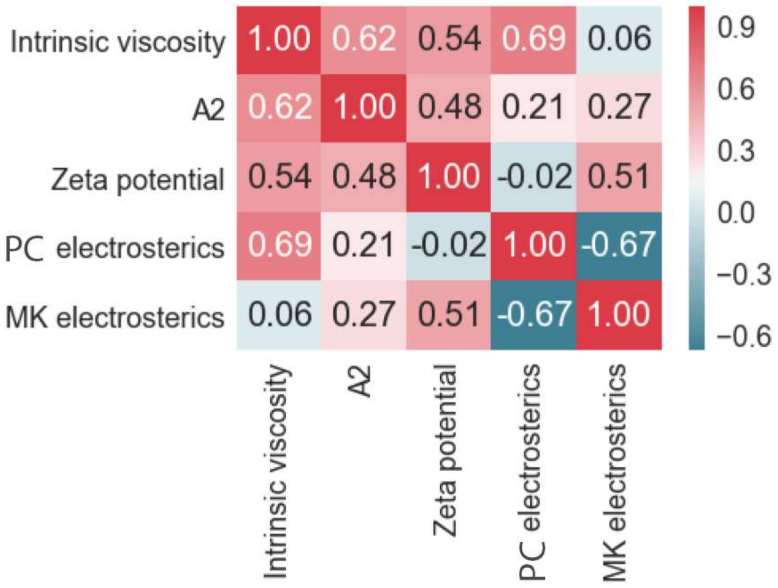


Figure 3.10. Polymer properties for individual compositional variables calculated using LSQR method.

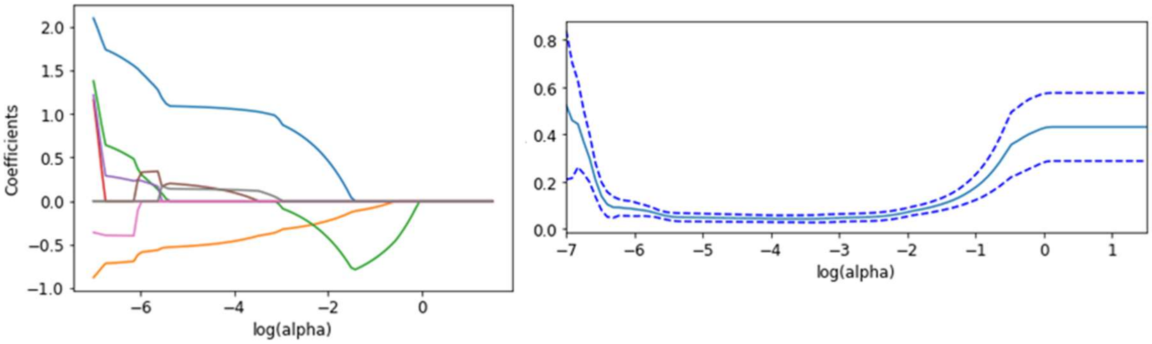


Figure 3.11. Lasso trajectory and mean-squared cross-validation error as a function of tuning parameter.

Table 3.5. Measured properties of the synthesized PSS-PEGMA-PMAA polymer.

PSS-PEGMA-PMAA properties	
Intrinsic Viscosity (mL/g)	443.00
A2	1.40
MK Sedimentation at 0.5 SP (% change)	-100.00%
PC Sedimentation 0.5 SP (% change)	31.60%
Zeta (mV)	-48.47
Slump Height Change MK/PC (cm)	3.63
Slump Height Change PC (cm)	1.23
Freundlich Adsorption Predictions at .25%SP	
PC	0.08
MK	0.11

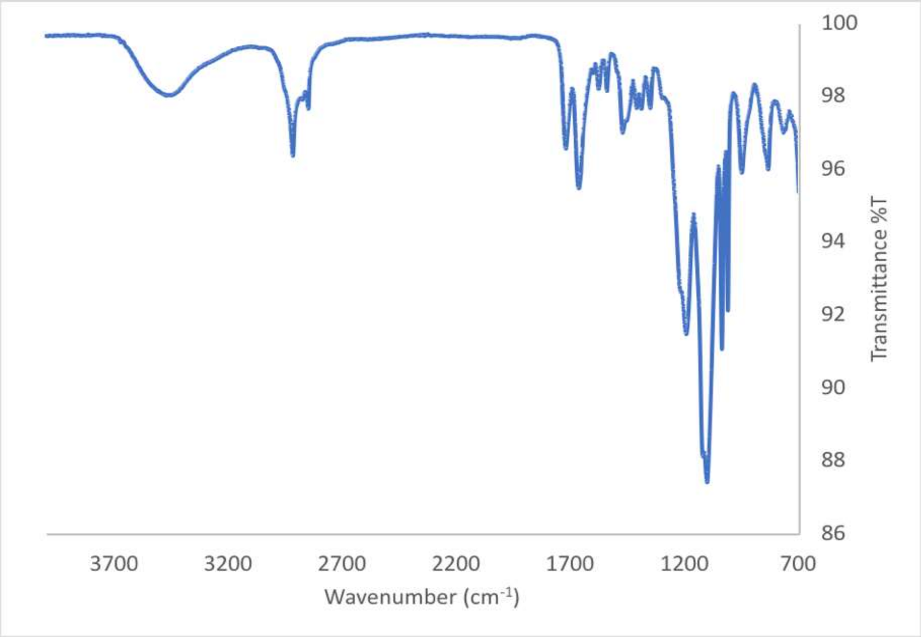


Figure 3.12. IR spectrum of the synthesized PSS-PEGMA-PMAA polymer.

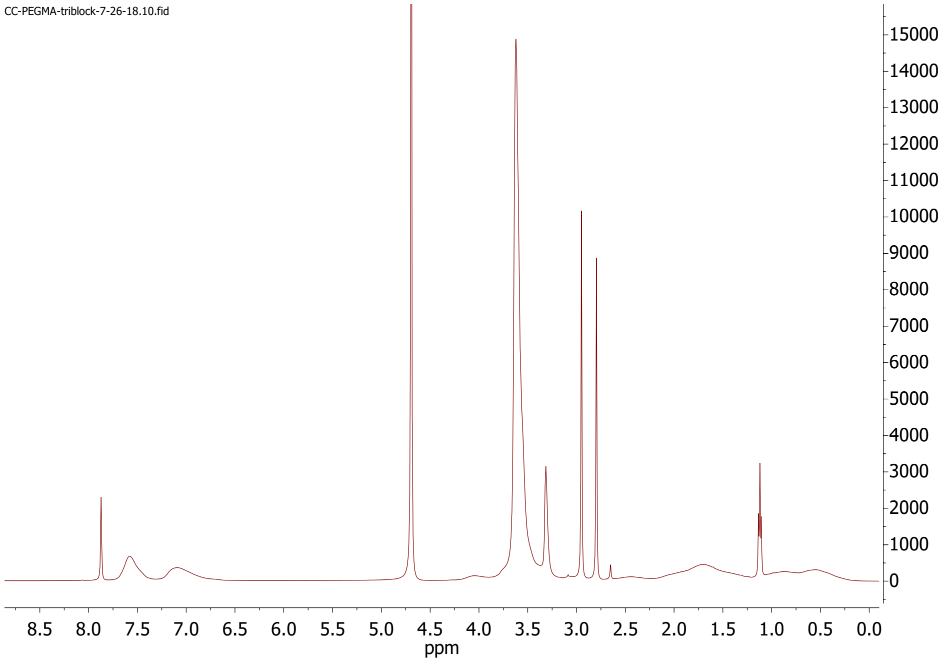


Figure 3.13. The ^{1}H NMR spectrum of the synthesized PSS-PEGMA-PMAA polymer.

3.8. References

[1] Kurtis, K. E. Innovations in Cement-Based Materials: Addressing Sustainability in Structural and Infrastructure Applications. *MRS Bull.* **2015**, *40* (12), 1102–1109.

[2] Flatt, R. J.; Roussel, N.; Cheeseman, C. R. Concrete: An Eco Material That Needs to Be Improved. *J. Eur. Ceram. Soc.* **2012**, *32* (11), 2787–2798.

[3] Kovler, K.; Roussel, N. Properties of Fresh and Hardened Concrete. *Cem. Concr. Res.* **2011**, *41* (7), 775–792.

[4] Scrivener, K.; Martirena, F.; Bishnoi, S.; Maity, S. Calcined Clay Limestone Cements (LC 3). *Cem. Concr. Res.* **2017**.

[5] Gamelas, J. A. F.; Ferraz, E.; Rocha, F. An Insight into the Surface Properties of Calcined Kaolinitic Clays: The Grinding Effect. *Colloids Surfaces A Physicochem. Eng. Asp.* **2014**, *455*, 49–57.

[6] Justice, J. M.; Kurtis, K. E. Influence of Metakaolin Surface Area on Properties of Cement-Based Materials. *J. Mater. Civ. Eng.* **2007**.

[7] Kurtis, K. E. Benefits of Metakaolin in HPC. *HPC Bridg. Views* **2011**.

[8] Antoni, M.; Rossen, J.; Martirena, F.; Scrivener, K. Cement Substitution by a Combination of Metakaolin and Limestone. *Cem. Concr. Res.* **2012**.

[9] Zaribaf, B. H.; Kurtis, K. E. Admixture Compatibility in Metakaolin–Portland-Limestone Cement Blends. *Mater. Struct.* **2018**, *51* (1), 33.

[10] Hosseinzadeh Zaribaf, B. METAKAOLIN-PORTLAND LIMESTONE CEMENTS: EVALUATING THE EFFECTS OF CHEMICAL ADMIXTURES ON EARLY AND LATE AGE BEHAVIOR, Georgia Institute of Technology, 2017.

[11] Agrawal, A.; Choudhary, A. Perspective: Materials Informatics and Big Data: Realization of the “Fourth Paradigm” of Science in Materials Science. *APL Mater.* **2016**, *4* (5), 053208.

[12] Hill, J.; Mulholland, G.; Persson, K.; Seshadri, R.; Wolverton, C.; Meredig, B. Materials Science with Large-Scale Data and Informatics: Unlocking New Opportunities. *MRS Bull.* **2016**, *41* (05), 399–409.

[13] REDDY, N. R. M. Neural Network (NN) Model to Predict the Flexural Strength of Metakaolin Blended Polypropylene Fiber-Reinforced High-Performance Concrete. *Int. Arch. Appl. Sci. Technol.* **2012**, *4* (3), 31–36.

[14] Saridemir, M. Predicting the Compressive Strength of Mortars Containing Metakaolin by Artificial Neural Networks and Fuzzy Logic. *Adv. Eng. Softw.* **2009**, *40* (9), 920–927.

[15] Safarzadegan Gilan, S.; Bahrami Jovein, H.; Ramezanianpour, A. A. Hybrid Support Vector Regression - Particle Swarm Optimization for Prediction of Compressive Strength and RCPT of Concretes Containing Metakaolin. *Constr. Build. Mater.* **2012**, *34*, 321–329.

[16] Lei, L.; Plank, J. A Concept for a Polycarboxylate Superplasticizer Possessing Enhanced Clay Tolerance. *Cem. Concr. Res.* **2012**, *42* (10), 1299–1306.

[17] Bouvet, A.; Ghorbel, E.; Bennacer, R. The Mini-Conical Slump Flow Test: Analysis and Numerical Study. *Cem. Concr. Res.* **2010**, *40* (10), 1517–1523.

[18] Tregger, N.; Ferrara, L.; Shah, S. P. Identifying Viscosity of Cement Paste from Mini-Slump-Flow Test. *ACI Mater. J.* **2008**, *105* (6), 558–566.

[19] Belleney, J.; Hélary, G.; Migonney, V. Terpolymerization of Methyl Methacrylate, Poly(Ethylene Glycol) Methyl Ether Methacrylate or Poly(Ethylene Glycol) Ethyl Ether Methacrylate with Methacrylic Acid and Sodium Styrene Sulfonate: Determination of the Reactivity Ratios. *Eur. Polym. J.* **2002**, *38* (3), 439–444.

[20] Oikonomou, E.; Bokias, G.; Kallitsis, J. K.; Iliopoulos, I. Formation of Hybrid Wormlike Micelles upon Mixing Cetyl Trimethylammonium Bromide with Poly(Methyl Methacrylate-Co-Sodium Styrene Sulfonate) Copolymers in Aqueous Solution. **2011**, *27*, 5054–5061.

[21] Schwinefus, J. J.; Checkal, C.; Saksa, B.; Baka, N.; Modi, K.; Rivera, C. Molar Mass and Second Virial Coefficient of Polyethylene Glycol by Vapor Pressure Osmometry. **2015**.

[22] Sakai, T. Extrapolation Procedures for Intrinsic Viscosity and for Huggins Constant K′. *J. Polym. Sci. Part A-2 Polym. Phys.* **2018**, *6* (9), 1659–1672.

[23] Pamies, R.; Hernández Cifre, J. G.; del Carmen López Martínez, M.; García de la Torre, J. Determination of Intrinsic Viscosities of Macromolecules and Nanoparticles. Comparison of Single-Point and Dilution Procedures. *Colloid Polym. Sci.* **2008**, *286* (11), 1223–1231.

[24] Gelardi, G.; Flatt, R. J. Working Mechanisms of Water Reducers and Superplasticizers. In *Science and Technology of Concrete Admixtures*; Elsevier, 2016; pp 257–278.

[25] Murray, L. R.; Gupta, C.; Washburn, N. R.; Erk, K. A. Lignopolymers as Viscosity-Reducing Additives in Magnesium Oxide Suspensions. *J. Colloid Interface Sci.* **2015**, *459* (January 2016), 107–114.

[26] Kantro, D. L. Influence of Water-Reducing Admixtures on Properties of Cement Paste -- A Miniature Slump Test. *Cem. Concr. Aggregates* **1980**, *2* (2), 95–102.

[27] Flatt, R. J.; Houst, Y. F. A Simplified View on Chemical Effects Perturbing the Action of Superplasticizers. *Cem. Concr. Res.* **2001**, *31* (8), 1169–1176.

[28] Flory, P. J.; Fox, T. G. Treatment of Intrinsic Viscosities. *J. Am. Chem. Soc.* **1951**, *73* (5), 1904–1908.

[29] Rudin, A. Elements of Polymer Science & Engineering. In *Elem. Polym. Sci. Eng. (Second Ed.*; 1999; p 509.

[30] Hirata, T.; Ye, J.; Branicio, P.; Zheng, J.; Lange, A.; Plank, J.; Sullivan, M. Adsorbed Conformations of PCE Superplasticizers in Cement Pore Solution Unraveled by Molecular Dynamics Simulations. *Sci. Rep.* **2017**, *7* (1), 16599.

[31] Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A Compressed-Sensing Method for Identifying the Best Low-Dimensional Descriptor in an Immensity of Offered Candidates. *Phys. Rev. Mater.* **2018**, *2* (8), 083802.

[32] Tibshirani, R. Regression Selection and Shrinkage via the Lasso. *J. R. Stat. Soc. B* **1996**, *58* (1), 267–288.

[33] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer New York: New York, NY, 2009.

[34] Ferraris, C. F. Measurement of the Rheological Properties of High Performance Concrete: State of the Art Report. *J. Res. Natl. Inst. Stand. Technol.* **1999**, *104* (5), 461.

[35] Flatt, R. J.; Schober, I.; Raphael, E.; Plassard, C.; Lesniewska, E. Conformation of Adsorbed Comb Copolymer Dispersants. *Langmuir* **2009**, *25* (2), 845–855.

[36] Dalas, F.; Pourchet, S.; Nonat, A.; Rinaldi, D.; Sabio, S.; Mosquet, M. Fluidizing Efficiency of Comb-like Superplasticizers: The Effect of the Anionic Function, the Side Chain Length and the Grafting Degree. *Cem. Concr. Res.* **2015**, *71*, 115–123.

[37] Nandi, B. K.; Goswami, A.; Purkait, M. K. Adsorption Characteristics of Brilliant Green Dye on Kaolin. *J. Hazard. Mater.* **2009**, *161* (1), 387–395.

[38] Gutierrez, M.; Fuentes, H. R. Modeling Adsorption in Multicomponent Systems Using a Freundlich-Type Isotherm. *J. Contam. Hydrol.* **1993**, *14* (3–4), 247–260.

[39] Casassa, E. F. Thermodynamics Interactions in Dilute Polymer Solutions: The Virial Coefficients. *Pure Appl. Chem.* **1972**, *31* (1–2), 151–178.

[40] Farrokhpay, S. A Review of Polymeric Dispersant Stabilisation of Titania Pigment. *Adv. Colloid Interface Sci.* **2009**, *151* (1–2), 24–32.

[41] El-Hosiny, F. I.; Gad, E. A. M. Effect of Some Superplasticizers on the Mechanical and Physicochemical Properties of Blended Cement Pastes. *J. Appl. Polym. Sci.* **2018**, *56* (2), 153–159.

[42] Davis, R. M.; Russel, W. B. Intrinsic Viscosity and Huggins Coefficients for Potassium Poly(Styrenesulfonate) Solutions. *Macromolecules* **1987**, *20* (3), 518–525.

[43] Winnefeld, F.; Becker, S.; Pakusch, J.; Götz, T. Effects of the Molecular Architecture of Comb-Shaped Superplasticizers on Their Performance in Cementitious Systems. *Cem. Concr. Compos.* **2007**, *29* (4), 251–262.

[44] Murray, L. R.; Bice, J. E.; Soltys, E. G.; Perge, C.; Manneville, S.; Erk, K. A. Influence of Adsorbed and Nonadsorbed Polymer Additives on the Viscosity of Magnesium Oxide Suspensions. *J. Appl. Polym. Sci.* **2018**, *135* (3), 45696.

[45] Zhang, L.; Lu, Q.; Xu, Z.; Liu, Q.; Zeng, H. Effect of Polycarboxylate Ether Comb-Type Polymer on Viscosity and Interfacial Properties of Kaolinite Clay Suspensions. *J. Colloid Interface Sci.* **2012**, *378* (1), 222–231.

[46] Poczos, B.; Xiong, L.; Schneider, J. Nonparametric Divergence Estimation with Applications to Machine Learning on Distributions. *UAI* **2011**.

[47] Kandasamy, K.; Dasarathy, G.; Oliva, J.; Schneider, J.; Póczos, B. Gaussian Process Bandit Optimisation with Multi-Fidelity Evaluations. In *Advances in Neural Information Processing Systems*; 2016.

Chapter 4. Cheminformatics for Accelerated Design of Chemical Admixtures[1]

4.1. Introduction

The rational discovery of new chemicals and materials through data-driven methods is the basis for the fourth paradigm of science. This new paradigm has helped to mitigate both the cost and time involved as compared to a traditional approach which requires multiple iterations (i.e., trial-and-error approaches) for scientific discovery.[1, 2] Developed over the past three decades, cheminformatics is among the early data-driven methods for materials development and has been widely implemented for virtual screening to accelerate drug discovery.[3] One of the well-known techniques in cheminformatics is quantitative structure activity relationship (QSAR), which represents molecules as a vector of descriptors. Quantitative methods are utilized to learn a relationship between these descriptors and molecular function, which in drug discovery can be a complex interaction such as protein binding.[4] These chemical descriptors can be physicochemical in nature, such as logP and dipole moment, or they can directly represent molecular structure. These structural descriptors can be zero-dimensional in nature, such as molecular weight, one-dimensional in nature, such as counting of specific fragments, or be two/three-dimensional in nature as to encode the molecular topology.[5] Common descriptors utilized to model topological information are known as molecular fingerprints. Fingerprints transform molecular structure into a bit string, and depending on the fingerprint method this bit string can be either an integer (count) or binary (true/false) vector.[6]

[1] This chapter includes work that was published and reformatted:
Childs, C.M.; Canbek, O.; Kirby, T.; Zhang, C.; Zheng, J.; Szeto, C.; Pόczos, B.; Kurtis, K.; Washburn, N.R. Cheminformatics for Accelerated Design of Chemical Admixtures. *Cement and Concrete Research.* 2020, 136, 106173.

Informatics approaches are based on the concept of similarity, which imply that the more alike two molecules are, the more similar their function will be.[7] Initially, the most widely utilized quantitative approach towards determining similarity was Tanimoto similarity, which defines similarity between molecules 'A' and 'B' as the number of bits encoded (that is being an integer value) in the fingerprint for 'A' divided by the sum of bits encoded in 'A' and 'B'. This results in a score between 0 and 1, where 1 is a perfect similarity and 0 is no similarity.[8] Early methods for virtual screening in drug discovery utilized Tanimoto or other similarity metrics to discover new molecules with high similarity which could function similar to an existing molecule for testing.[9]

Machine learning (ML) techniques are being explored for predictive modeling of the properties of cement and concrete. An early example was the use of an artificial neural network in determining the susceptibility of concrete to sulfate attack based on compositional factors from over 100 samples with over 8000 datapoints.[10] A number of more recent studies have used significantly larger datasets and a broader range of ML algorithms to predict the compressive strength of concrete as a function of mix design, including both cementitious and aggregate variables.[11,12,13,14] There are few examples of data-driven models used for design of novel chemical admixtures.

A recently developed algorithm, hierarchical machine learning (HML), was used to develop a new superplasticizer for blends of portland cement and metakaolin.[15] The algorithm predicted a dispersant having a novel composition and a mechanism of action based on high viscosity and osmotic pressure of the pore solution rather than electrosteric forces mediated by adsorbed polymer. However, the algorithm represented polymer chemistry by mole fractions of functional groups, requiring additional interpretation in how the monomers were composed of these groups. A more sophisticated representation of molecular structure was taken in the prediction of the surface tension for shrinkage-reducing admixtures, which utilized a cross-validation technique with a multiple linear regression to predict their effectiveness and discover structures of alternative admixtures represented in terms of a molecular signature.[16] However, these tests were based on only the polymer properties, not considering the effects of the predicted structures in a cementitious system. Also, as a forward-stepping selection for features was utilized, only four simple hydrocarbon-based structural fragments were found as contributing to the best model. This can create wide variability in prediction as the molecular structures increase in complexity.

ML regression techniques can be used in tandem with advanced cheminformatics methods for representing chemical structure. Recent cheminformatics approaches in conjunction with ML beyond drug design have been utilized to screen conjugated polymers and predict photovoltaic power conversion efficiency[17,18], predict compound toxicity[19], and predict solubility parameters[20]. Until recently, ML and cheminformatics have been limited to applications with high-throughput and large data availability, which has largely precluded cementitious materials where data collection is limited. However, with recent research into ML techniques on sparse datasets[21,22,23], cheminformatics approaches can be used to develop structure-function relationships and perform rapid virtual screening of chemical admixtures. These ML techniques for sparse datasets allow for higher complexity representations of molecules which can then be reduced through feature selection techniques, allowing for more complex architectures and variability in the dataset.

Calcium sulfoaluminate (CSA) cements are rich in belite, calcium aluminosulfate (i.e., Ye'elimite) and calcium sulfate phases with a general clinker phase composition shown in Table 4.1.

Table 4.1. General clinker phase composition of a CSA cement.

Phase	Amount (%)
Belite (C_2S)	10-60
Ye'elimite ($C_4A_3\overline{S}$)	10-55
Ferrite (C_4AF)	0-40
Anhydrite ($C\overline{S}$)	0-25
Monocalcium aluminate (CA)	0-10
Mayenite ($C_{12}A_7$)	0-10
Free Lime/Calcite (C)	0-25

Because no tricalcium silicates are found in typical CSA clinker, their embodied CO_2 is 15-50% lower compared to portland cement.[24] Just the C_3S calcination alone in PC accounts for 0.578 g of CO_2 per g a raw material. Lifecycle assessments in CSA calcination show CO_2 production at only 0.216 g per g of raw material.[25] CSA cements hydrate rapidly to produce ettringite (AFt or $C_6A\bar{S}H_{32}$) and gibbsite (AH_3), leading to a short set time that is ideal for use in tunnel linings, bridge decks, airport runways, and repair applications. Figure 4.1 shows the initial hydration reactions occurring in CSA with high initial reactivity due to the sulfate and aluminate phases.

$C_4A_3\bar{S} + 2C\bar{S} + 38H \rightarrow C_6A\bar{S}_3H_{32} + 2AH_3$

$C_4A_3\bar{S} + 6CH + 8C\bar{S} + 90H \rightarrow 3C_6A\bar{S}_3H_{32}$

$2C_4A_3\bar{S} + 2C\bar{S} + 52H \rightarrow C_6A\bar{S}_3H_{32} + C_4A\bar{S}H_{12} + 4AH_3$

$C_4A_3\bar{S} + 18H \rightarrow C_4A\bar{S}H_{12} + 2AH_3$

$C_{12}A_7 + 12C\bar{S}H_2 + 113H \rightarrow 4C_6A\bar{S}_3H_{32} + 3AH_3$

$3CA + 3C\bar{S}H_2 + 32H \rightarrow C_6A\bar{S}_3H_{32} + 2AH_3$

$C_4AF + 3C\bar{S}H_2 + 30H \rightarrow C_6A\bar{S}_3H_{32} + FH_3 + CH$

$C_4AF + 3C\bar{S}H_2 + 21H \rightarrow C_6(A,F)\bar{S}_3H_{32} + 2(A,F)H_3$

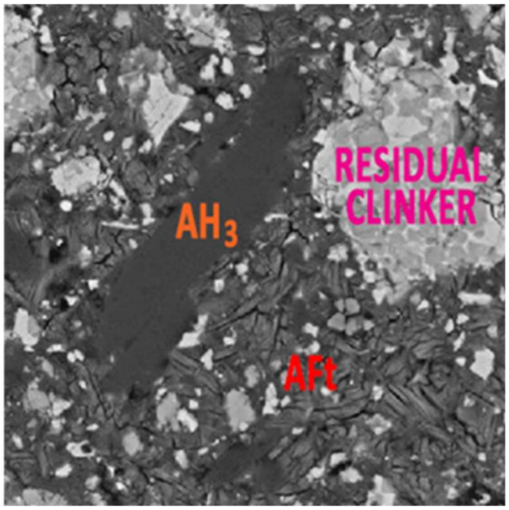


Figure 4.1. CSA early hydration reactions and early product formation.[26,27] AFt and AH_3 phases are shown in the microscopic schematic forming from residual clinker.

Set retarding admixtures are commonly required to prevent premature setting that can occur within 20 min post mixing.[28] A diverse range of chemistries ranging from sugars, lignosulfonates, and polynaphthalene sulfonates (PNS) to both organic and inorganic acids have been used.[28,29,30,31] Retarders are thought to have a diversity of mechanisms that depend on cement phases but are typically thought to operate through selective blocking of reactive surface sites, formation of ion complexes, or hydrate growth inhibition.[32] Small anionic organic compounds such as tartaric, gluconic, and citric acids have been reported for use in CSA cements, with citric acid (CA) commonly used to prolong hydration reactions without adversely affecting mechanical properties when dosed appropriately.[30,33] It has been established that combinations of polar and anionic functionality, such as those in nitrilotris(methylene)triphosphonate (ATMP), have been shown to prolong set times for portland cement, suggesting there may be complex synergies between functional groups and molecular architecture.[34] Due to this complex interplay between chemical features in set retarders, cheminformatics is a promising methodology to predict set times and screen a diversity of retarder chemistries.

The objective of this work was to demonstrate a cheminformatics approach to accelerated design of chemical admixtures, capable of extending set time of CSA cement to greater than 1 hour. Here, a library of 23 small molecules were screened as set retarders for CSA cement at constant dose of 1%. While this dose of citric acid resulted in an excessively long set time, it allowed discrimination of the performance of less active retarders. Three different molecular fingerprint formalisms were explored, and the resulting models were optimized by standard methods of machine learning. To demonstrate the potential of cheminformatics in rapid virtual screening of chemical admixtures, the structures of 500,000 small-molecule compounds from an online library were downloaded and candidates were screened following constraints of molecular weight and functionality. A novel set retarder from this screening was tested, and its effects on CSA cement set time were found to be within the uncertainty estimates of the model.

4.2. Experimental

4.2.1. Materials

All cement pastes were prepared utilizing CSA cement (CTS Cement, RapidSet). The cement composition, as determined by quantitative x-ray diffraction. As shown in Table 4.2, the main phases identified are belite, ye`elimite, bassanite, anhydrite and calcite.

Table 4.2. Phase composition of CSA clinker used, determined by QXRD

Phase	Amount (%)
Ye`elimite	30.3
Belite	42.0
Anhydrite	12.1
Bassanite	5.5
Calcite	2.7
R_{wp}[a] (agreement factor), %	12.4

[a] <15 % R_{wp} is considered as an accurate fit for the quantitative phase analysis.[35]

All chemical retarders were purchased from Sigma-Aldrich Corporation or Fisher Scientific and used without further purification. A representation of the chemistry and abbreviations of the high activity retarders are shown in Figure 4.2. Deionized water with a resistivity of 18.2 MΩ was utilized for all experiments.

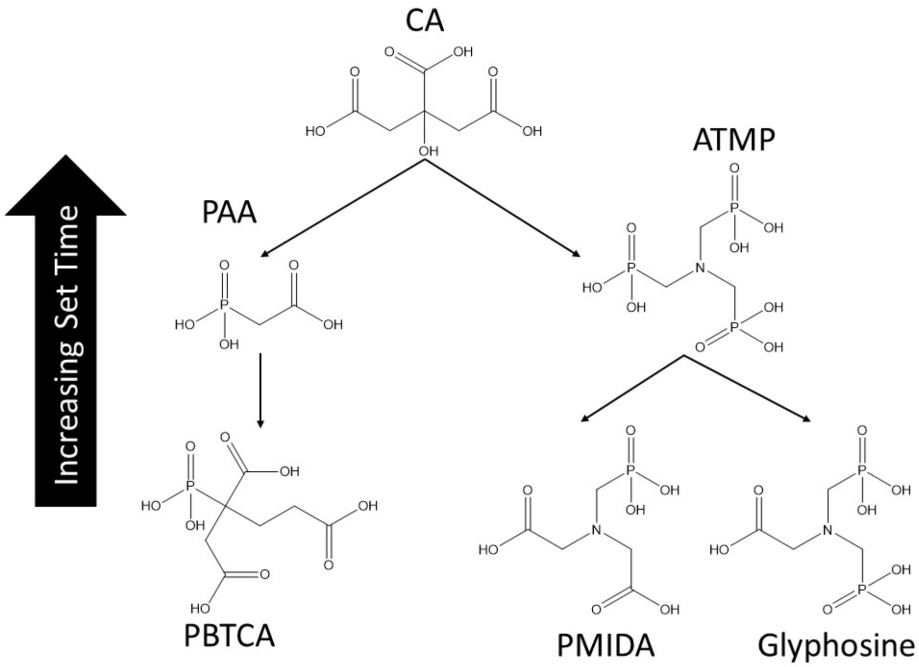


Figure 4.2. Chemical structures of high activity retarders: Citric acid (CA), phosphonoacetic acid (PAA), nitrilotrimethylphosphonic acid (ATMP), 2-phosphonobutane-1,2,4-tricarboxylic acid (PBTCA), phosphonomethyliminodiacetic acid (PMIDA), and N,N-bis(phosphonomethyl) glycine (glyphosine). The arrows in the figure connect compounds with high chemical similarity.

4.2.2. Test Methods

Cement pastes were prepared utilizing a hand-held mixer (HamiltonBeach). The admixture was dissolved in the mix water, and added to the cement, dosed in percentage by weight of cement. The paste was mixed for 30 s at low speed setting and then an additional 60 s at medium. A 30 s rest period was followed with a final 60 s medium-speed mixing period. All mixes were tested immediately followed mixing.

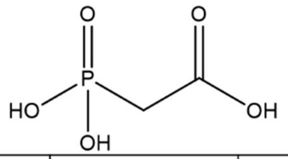
Set time was determined through Vicat needle testing, ASTM C191.[36] Here, a constant w/c of 0.40 was used and tests were conducted every 5 min. In order to maintain consistency in the computational procedure, the time where the needle failed to penetrate 25 mm from the top of the mold was utilized as the initial set time.

Heat evolution of CSA cement pastes mixed with different retarders was determined up to 72 h using a commercial isothermal calorimeter (TAM Air, TA Instruments). Cement pastes were prepared at room temperature (23 °C) outside the calorimeter with a water-to cement (W/C) ratio of 0.40 then placed inside the calorimeter. Retarder dosages were kept constant at 1 % by weight of cement for all mixes. The results were compared with a plain CSA cement paste prepared without adding any retarder.

4.3. Computational Methods

4.3.1. Fingerprint Generation

Set retarders were first represented as a simplified molecular-input line-entry system (SMILES) string, which is a line notation utilized to represent a chemical structure. The SMILES strings were then input into RDKit[37], an open-source cheminformatics software in Python. Each molecule was transformed into a vector of descriptors for use in ML as a molecular fingerprint. Three various fingerprints were utilized and compared with one another. The first was a custom set of 15 descriptors based on the counting of groups, ratios of functional groups, expected charge of each molecule at pH 10.5 similar to that found in CSA cement, and ratio of charge to the number of groups as a similar descriptor to charge density. Figure 4.3 presents PAA represented as this custom fingerprint with each descriptor of the vector explained.



	# Phosphonate Groups	# Sulfonate Groups	# Carboxylate Groups	# Hydroxyl Groups	# Nitrogens
Descriptors 1-5	1	0	1	0	0

	# Carbons	Charge at pH of 10.5	Phosphonate/ Carbon Ratio	Carboxylate/ Carbon Ratio	Hydroxyl/ Carbon Ratio
Descriptors 6-10	2	-3	0.5	0.5	0

	Nitrogen/ Carbon Ratio	Sulfonate/ Carbon Ratio	All Non-Carbon Groups/ Carbon Ratio	Charge/ Carbon Ratio	Charge/ All Groups Ratio
Descriptors 11-15	0	0	1	-1.5	-0.75

Figure 4.3. Listing of each of the 15 descriptors utilized in the custom descriptor set representing PAA.

The second fingerprint, implemented through RDKit, is a binary circular fingerprint based on the Morgan algorithm.[38] In this algorithm, each atom is represented by its molecular environment as defined by the atomic connectivity within circles of radius set by an integral number of bonds. The variant of this fingerprint used was an extended-connectivity fingerprint (ECFP), which represents each possible connectivity within the molecular topology as a binary vector, where the bit is encoded as 1 if the specific connectivity is present, and 0 if absent.[38] A radius of three was chosen so that any path of atoms within three or less bonds of the molecule would be represented. Morgan type fingerprints in RDKit are hashed into a 2048 length bit vector. Through a process known as folding this vector can be decreased based on a modulo operation to lengths of 1024, 512, 256, 128, 64, and 32, however each fold leads to an increasing amount of bit collision where multiple features are encoded on the same bit.[39]

A third fixed-length fingerprint modeled through RDKit is the 79-bit vector Electrotopological State (E-state) fingerprint. This fingerprint sums the electronegativity contribution (as determined through connectivity) for each of 79 various chemical groups that could be present in a molecule[40], thus providing basic physicochemical descriptors.

4.3.2. Machine Learning

Methods of machine learning were used to optimize the cheminformatics model, which was built on a training set of a single CSA cement and a small library of 23 candidate set-retarding molecules. It is important to note that extension to other CSA cements is dependent on the similarity to the phase composition of the CSA utilized in this work, but it is expected that the trends in the model will generalize. In ML it is necessary that the number of features remains less than the number of training points to avoid overfitting. For this reason, a sparse linear model was determined using the least absolute shrinkage and selection operator (Lasso).[41] Lasso is a linear model which utilizes the cost function for least squares regression with an added penalty term in the form of the L_1-Norm as shown in Eq. 4.4.1:

$$min_{\beta}(\|y-\beta x\|_2^2+\alpha\|\beta\|_1) \qquad \text{Eq. 4.4.1}$$

In this equation, y represents the responses, where the set time of unmodified CSA cement is set equal to zero. All admixtures are set relative to this so that the intercept term can be set to zero. Here, x represents the independent variables of the molecular fingerprints for each molecule. The β parameters are the model coefficients minimized through the cost function in order to find the best fit line, and α is the hyperparameter which is optimized through cross-validation. Tuning the hyperparameter to larger values drives non-important β parameters and the corresponding features to zero. This decreases the dimensionality to avoid overfitting in which the model provides an excellent fit to data in the training set, but it can still have low predictive power for data outside the training set. All data analysis was performed in Python using the Scikit-Learn package.[42]

Cross-validation is an approach utilized to find the optimal hyperparameters for a ML model. Here, a leave-one-out cross validation (LOOCV) was applied. In this approach 22 of the 23 independent variables were used as a training set and the 23rd was a test point which has its error, in this paper the mean squared error (MSE), from the predicted value calculated for each hyperparameter value. This process was repeated 23 times, such that each independent variable was treated as a test point and the average of all the MSE's was calculated along with a standard deviation based on testing all CV folds. The goal was to determine the hyperparameters corresponding to a low MSE along with a low standard deviation in the MSE, as these would predict the unseen screened molecules most accurately. For the ECFP, cross-validation was utilized to minimize the error for both fingerprint length and α. Gutlein and Kramer[43], termed a similar technique as a folding and filtering process where folding is a process of setting the fingerprint vector length and filtering, as applied to this research, through utilizing Lasso to reduce the dimensionality. The goal in the minimized CV error is meeting the condition of (number of features) < (number of samples) in order to prevent having an underdetermined system of equations. An overview schematic of this process is presented in Figure 4.4. As the custom and Estate fingerprints are constant-length vectors, no folding was performed leaving α as the only tunable hyperparameter for the filtering process.

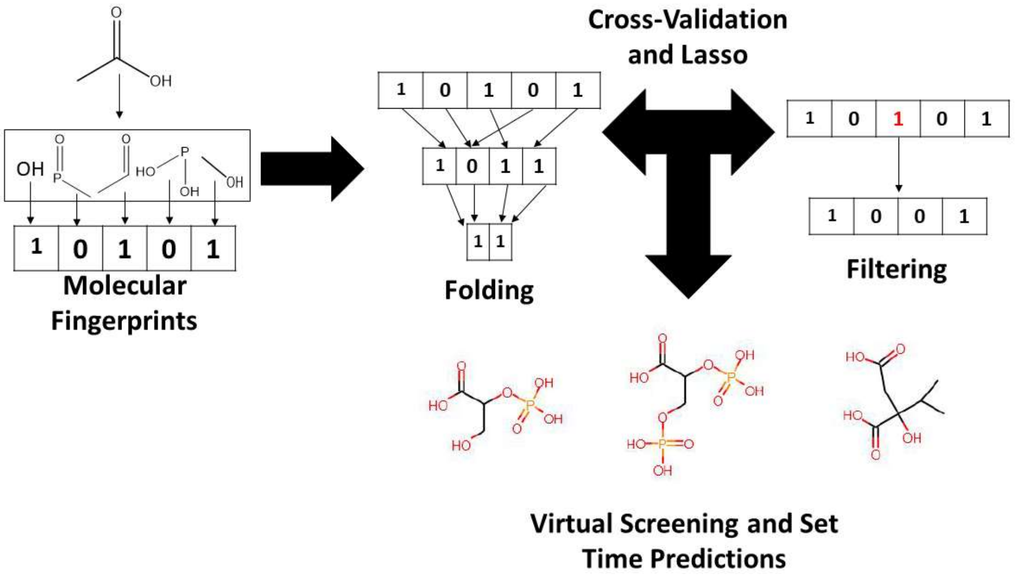


Figure 4.4. Schematic representation of the computational methodology. First the molecules were transformed into binary molecular vectors through fingerprinting techniques. These vectors then pass through both folding where bits from a large vector are mapped onto a vector of smaller size, and filtering, a method of feature selection through utilization of Lasso and cross-validation. The most accurate model was selected and virtual screening of molecules was performed to find molecules with set times above 1 h.

4.3.3. Selection of Compounds of Virtual Screening

Virtual screening was performed by downloading 500,000 compounds in spatial data file ('.sdf') format from the public depository PubChem.[44] Initial constraints applied to the molecules were selected as to only screen molecules similar to those found in the tested library: That they were of the basis set of atoms [H,C,N,O,P,S], no aromatic or aliphatic rings, having >5 but <20 non-hydrogen atoms, having >2 hydrogen donors and >2 hydrogen acceptors, having >1 carbon, and a ratio of oxygen/carbon ≥1, and any mixtures of compounds were not formally screened.

4.4. Results and Discussion

4.4.1. Set Time

The set time for the CSA cement without admixture occurred rapidly, as expected, with initial set at 15 min. Of the 23 candidate admixtures examined, only six produced initial set times in excess of the 1-hour target. Set times for those six admixtures range from 75 to 180 min, as shown in Figure 4.5. The complete results of each candidate admixture in the training set are shown in Figure 4.13.

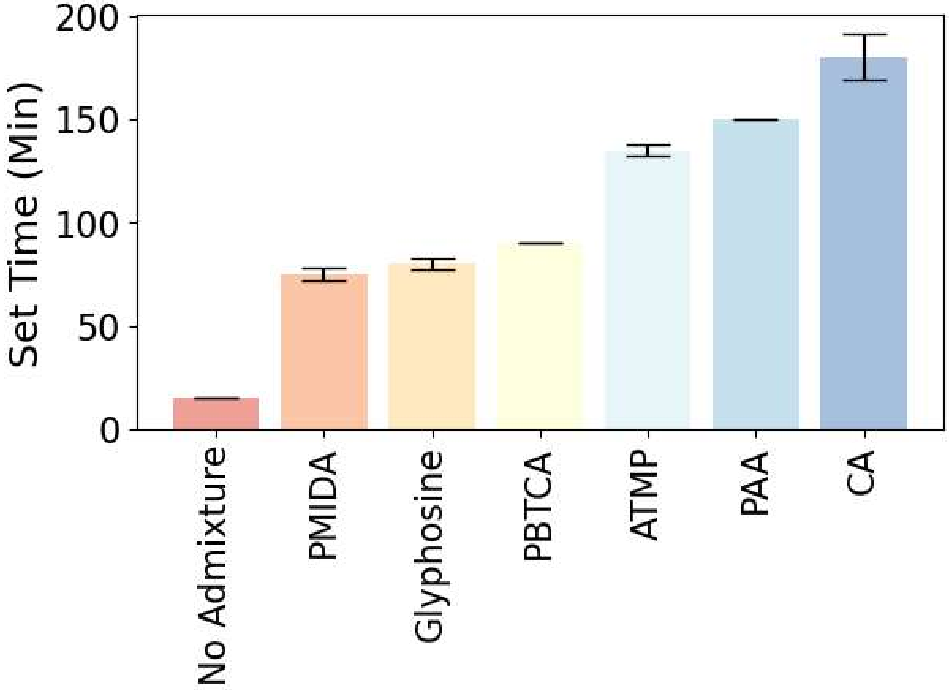


Figure 4.5. Comparison among those retarders producing a set time greater than 1 h. All measurements were performed by using a Vicat apparatus with admixture concentration set to 1% by weight of cement at a 0.40 w/c ratio.

4.4.2. Folding, Filtering, and Fitting

The custom fingerprint, Estate fingerprint, and ECFP's were each optimized utilizing Lasso. The minimum uncertainty in the MSE where the number of selected features < the number of molecules tested, as determined through CV, was reported and shown in Figure 4.6. At larger fingerprint lengths this implies that the RMSE in Figure 4.6 does not correspond directly to the minimized MSE as calculated through Lasso. The alpha hyperparameter had to be selected to prevent an underdetermined system. The complete plots for all the folds of the ECFP's are shown in Figure 4.14 and Table 4.4.

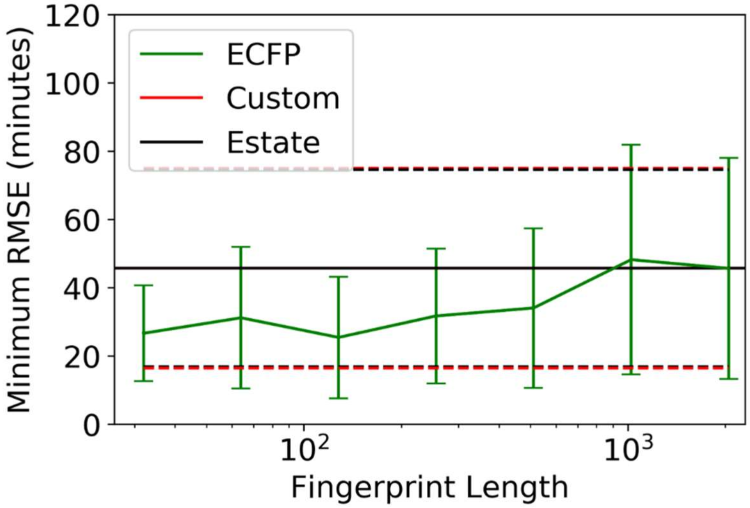


Figure 4.6. The ECFP fingerprint was tested at multiple fingerprint lengths after folding. Both the Estate and custom fingerprints are set length vectors and therefore are shown as a straight line. The custom and Estate fingerprints have a similar constant error in the predicted set time of 45 min and also have similar RMSE uncertainty, but ECFP has a minimum error of 26 minutes in the predicted set time at a fingerprint length of 32 with the smallest RMSE uncertainty.

The minimum MSE was found for the 128 length Morgan Fingerprint with a mean value of 647 while the 32 length vector had the next lowest MSE value of 710. However, the minimum uncertainty, as shown by the standard deviation in the MSE, corresponded to the folded 32-bit ECFP with a value 196 compared to 317 at the 128 length vector. The goal in ML is to minimize both the MSE in mean prediction, along with the uncertainty. As such the 32-bit ECFP was selected for the visual screening of various chemical structures as it corresponds to the lowest amount of predicted uncertainty for the model, but still putting the mean around the minimum of the 128 length vector. The individual Lasso CV plots for the 32-bit ECFP, Estate, and custom fingerprints are shown in Figure 4.7, represented as MSE. Note that the RMSE shown in Figure 4.6 corresponds to the uncertainty in predicted set times, but the MSE is the quantity that is minimized in Eq. 4.4.1.

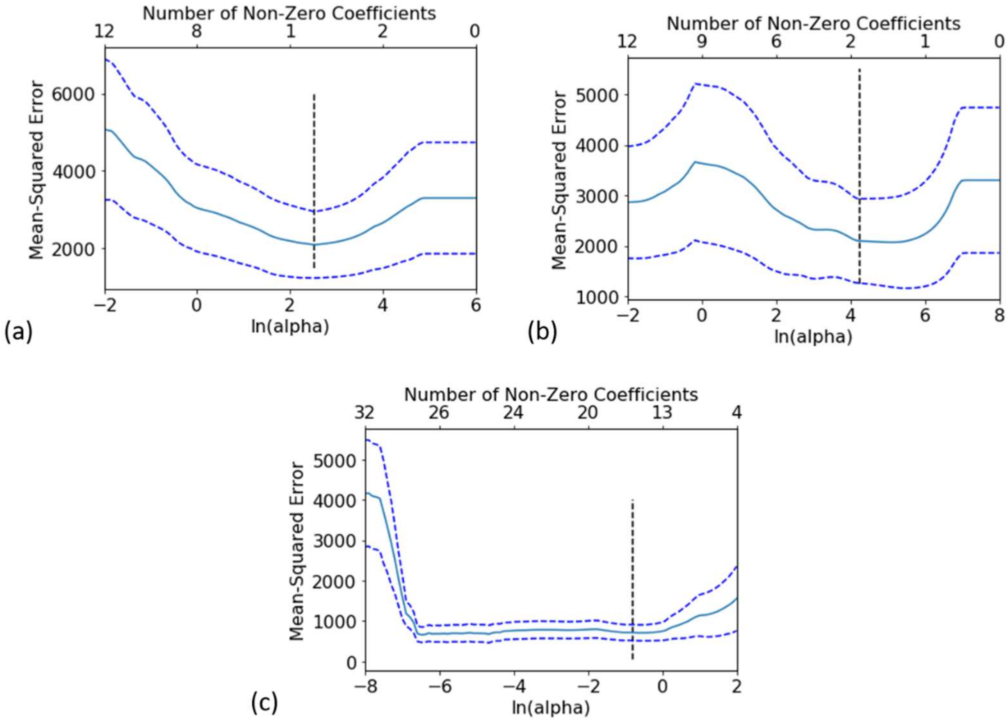


Figure 4.7. Lasso CV plots for a) Custom fingerprint, b) Estate Fingerprint, and c) 32 length ECFP.

Alpha is the Lasso hyperparameter which at small values causes the regression to become equivalent to an ordinary least squares regression. As alpha increases, more parameters are forced to zero as shown on the top axis. The goal was to identify an optimal alpha value to both minimize the MSE, uncertainty in the MSE and have the number of non-zero coefficients (features) < the number of samples tested (23). The dashed vertical line in the plots shows the selected alpha coefficient best meeting these criteria.

4.4.3. Regression Analysis

Due to the folding procedures for circular fingerprints, interpretability suffers due to the high density of information encoded in each bit. However, the custom and Estate fingerprints are interpretable because they map directly from chemical structure, and methods for calculating each bit in Estate have been reported in literature.[40] The selected alpha values, number of selected features, MSE's, standard deviation in the MSE and correlation coefficients (R^2) are shown in Table 4.3. When selecting the alpha value which minimizes the uncertainty in MSE for the Estate fingerprint it resulted in a very sparse model comprised of only two parameters, which represented the hydroxyl and carbonyl groups, respectively. They were both positive, indicating an increase in set time associated with each. As Estate is a summation of the electronegativity contributions due to the surrounding environment these groups are found in, having more hydroxyl and carbonyl groups associated with anionic functionalities such as carboxylate and phosphonate, were associated with longer set times. However, only utilizing these two descriptors led to a significantly higher MSE than the ECFP and a low correlation coefficient on the training set. This is also similar to the custom fingerprint, where upon selecting the appropriate alpha value only one descriptor is selected for set time prediction. This descriptor is the charge of each molecule after dissociation when introduced to a CSA slurry of around pH 10.5. This predictor shows a positive correlation between charge and set time, but it is a poor predictor in terms of MSE and R^2 values as shown in Table 4.3.

Table 4.3. The selected alpha values, number of selected features, MSE's and R^2 values associated with each fingerprint.

Fingerprint Method	**Alpha Value**	**Number of Selected Features**	**MSE**	**RMSE (min)**	**Standard Deviation in MSE**	**R^2**
ECFP	0.45	18	710	26.6	196	0.983
Custom	12.5	1	2088	45.7	857	0.273
Estate	171	2	2094	45.8	833	0.240

4.4.4. Virtual Screening

The model presented in this work is designed to perform rapid virtual screening of candidate chemical admixtures. While the methodology does not allow for optimization, it can provide an accurate estimate for the performance of any chemical species that meets the constraints of the compositional space. Here, after the initial set of 500,000 compounds were reduced according to these constraints, 886 molecules were identified for virtual screening. Each of these molecules were parameterized as a 32-bit ECFP and introduced into the ECFP Lasso regression model to predict the set time. From this group, 365 compounds were predicted to impart set times beyond 1 h. These compounds were then screened for cost and commercial viability in order to find compounds suitable for testing. As a test case, glyphosate was chosen as a commercially available molecule from the screened database for testing and is shown in Figure 4.8 Glyphosate is the most widely utilized herbicide worldwide[45], but its activity as a set retarder in CSA or other cements has not been previously documented. This potential application was only predicted by rapid virtual screening of a chemical library. In addition, while not commercially available, the three molecules with the longest predicted set times are also shown.

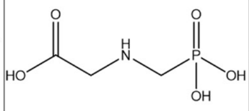
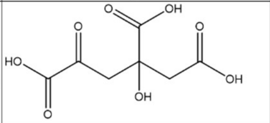
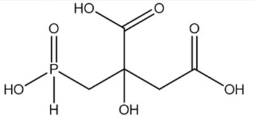
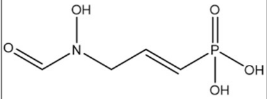
Name	Glyphosate	2-hydroxy-4-oxobutane-1,2,4-tricarboxylic acid	4-Phosphonyl-3-carboxy-3-hydroxybutanoic acid	(3-(Formylhydroxyamino)-1-propenyl)phosphonic acid
Molecular Structure				
Predicted Set Times	**61 min**	**178 min**	**178 min**	**183 min**

Figure 4.8. The structure of glyphosate along with the three structures in the screened molecules leading to the longest predicted set times.

Glyphosate was predicted to have a set time of 61 min. With an MSE of 710 identified through cross-validation within ECFP, molecules would be expected to be predicted within a range of the RMSE, +/- 26 min. Using Vicat testing, the set time of glyphosate was measured to be 55 min, within the predicted error of the cheminformatics analysis. As shown in Figure 4.9, the set time of glyphosate was found to be predicted within the range of molecules with similar measured set times.

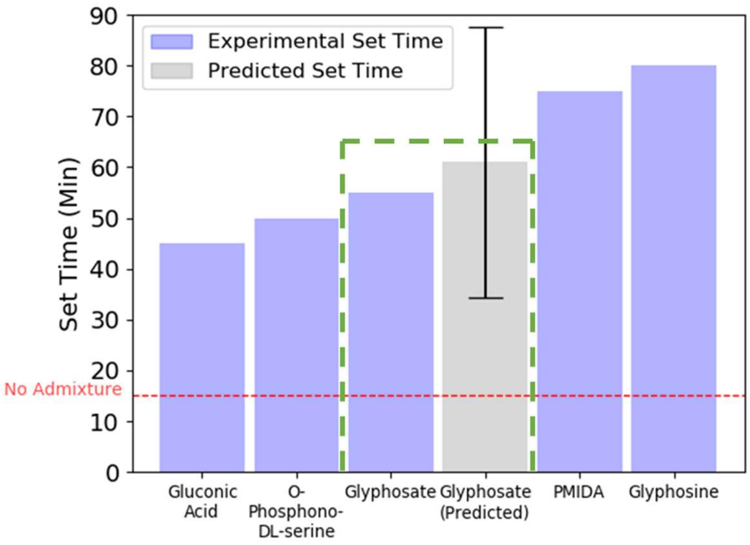


Figure 4.9. Plot of all molecules within the RMSE predicted for glyphosate. The prediction for glyphosate (gray) is shown along with experimentally measured set times of these molecules (blue), including the measured value for glyphosate, and despite the molecular structure variations within this range, glyphosate was shown to have a set time within 6 min of the prediction.

4.4.5. Hydration Kinetics

To assess the validity of the machine learning predictions and to provide insight into the set time results, heat evolution was measured for CSA cement pastes prepared with four high activity retarders selected from the training set and glyphosate selected for validation from the virtual screening. Figure 4.10 compares the rate of heat evolution and cumulative heat evolved for ordinary CSA cement paste with pastes at the same 0.40 w/c ratio, but containing 1% by mass addition of these retarders.

A near-instantaneous first peak was observed in all samples due to the wetting and early dissolution of cementitious phases.[46] In the ordinary CSA, a single peak with a maximum at ~ 54 min was observed. With retarders, dual peaks, extending between 160 and 600 min, were observed, where broader peaks are associated with increasing retardation effect. The presence of bassanite in the CSA cement is associated with the observation of dual peaks, as its rapid dissolution leads to a first peak associated with exothermic gypsum formation that then results in a second peak indicative of ettringite precipitation.[47] Without admixtures, these dual peaks are not typically resolved because the rapid formation of gypsum from bassanite occurs almost simultaneously with other early reactions. However, in the presence of different set retarders, the peak associated with formation of gypsum is shifted towards longer hydration times and can be identified as a shoulder or a distinct peak before the main peak. This phenomenon was also demonstrated by Burris and Kurtis[46] and Velazco et al.[48] for CSA systems containing citric acid. The characteristics of the early CSA cement hydration exotherms in the presence of these 3candidate admixtures are, then, also consistent with effective retardation.

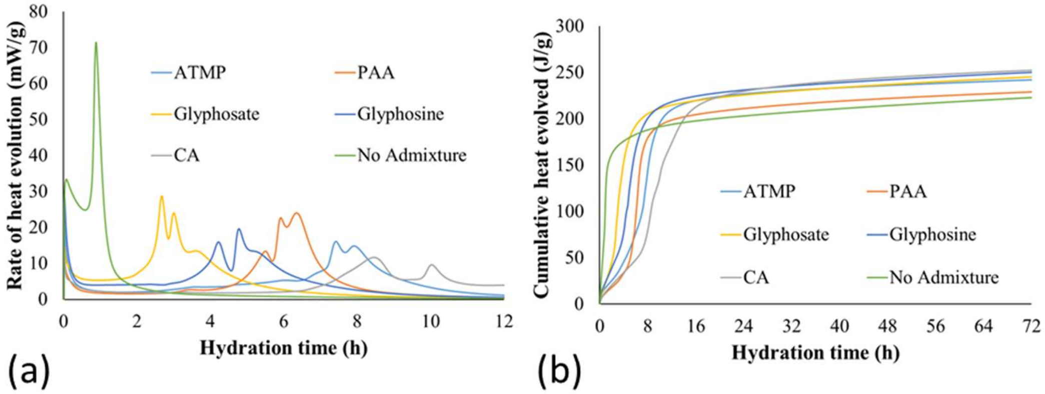


Figure 4.10. a) Rate of heat evolution of CSA cement pastes prepared, and b) cumulative heat evolved comparing plain CSA cement paste (green) with high-activity set retarders in the training library.

After about 14 hours hydration time, the cumulative heat of ordinary CSA is exceeded by each of the CSA cement pastes containing retarders (Figure 4.10 (b)). However, there is not a clear correlation between cumulative heat evolved at 72 hours (which ranges between 2.88% and 13.28% greater than the ordinary CSA), and retardation of early hydration kinetics. For instance, PAA delayed the initial peak more than glyphosine and glyphosate, but the CSA paste including phosphonoacetic acid liberated the least amount of heat among all samples including retarders. However, comparing the time until the first peak after the dormant period may be correlated with set-time results. Based on this metric, the effectiveness of the retarders can be ranked as:

CA (most retarding) > ATMP > PAA > Glyphosine > Glyphosate (least retarding)

This trend generally agrees with the set-time results (Figure 4.11), with the exception of longer setting-time observed for CSA paste mixed with PAA than ATMP.

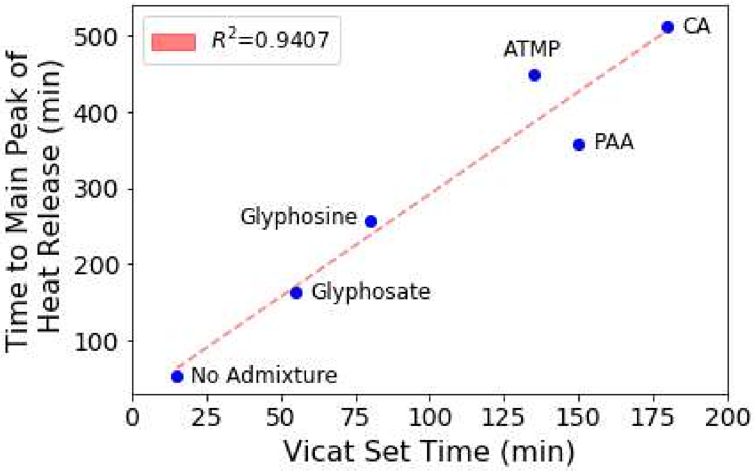


Figure 4.11. Setting time determined by Vicat test versus time to main peak of heat release observed in calorimetry curves, showing a correlation with a plot of the best fit line.

4.4.6. Interpretability and Modeling Sparse Data

Interpretability of data-driven models is often assessed through comparison with established mechanistic models.[49] For set retarders, an important mechanism is through adsorption directly onto cement clinker phases. It has been shown that sulfate ions preferentially adsorb onto aluminate phases, and that the addition of sulfonated water reducers, such as PNS, leads to varying degrees of competitive adsorption with gypsum, which modifies the set time depending on the cement phase composition.[50] The second mechanism occurs through ion complexation of an anionic admixture with cationic cement species, primarily Ca^{2+}. This results in a two-fold effect, the first being complexation of anionic admixtures, such as citric acid forming insoluble precipitates, and the secondary result being changes in cement solubility kinetics due to the change in ionic strength caused through complexation.[46] The third mechanism occurs through adsorption of admixtures onto hydration products inhibiting further crystal growth.[51] However, the exact mechanisms of many of these retarders and their interactions with ions in CSA cements are still poorly understood and may involve more than one physicochemical mechanism and vary based on CSA phase composition.[31,52] The cheminformatics model presented here is not based on mechanism but rather strictly on the effects on set time, and it may be that the retarding effects of different compounds in the library operate through distinct physicochemical interactions in the cement paste. It also may be that the retarding effects of these admixtures have limitations in their generalizability to CSA cements, particularly those with different compositions than utilized in this study. Integration of molecular descriptors derived from cheminformatics and knowledge of physicochemical interactions will be necessary to combine the powerful representations of chemical structures with mechanistic understanding.

Another challenge in data-driven approaches to modeling complex physical systems is developing methods for small datasets. This can also be addressed through embedding domain knowledge in these models, either in the form of chemical and physical knowledge or in terms of similarity.[21] ECFP methodology relies upon the concept of similarity, while the custom and Estate descriptors had mixtures of similarity and chemically embedded knowledge. However, as is shown with the high errors in the custom and Estate fingerprints, chemically embedded knowledge must also be meaningful to the system. Charge and electronegativity had a small amount of predictive capability, but yielded less accurate predictions than embedded similarity. In order to improve the model with embedded physiochemical parameters, features experimentally determined such as ion complexation and other mechanisms leading to CSA set time can be included as a form of causal knowledge to improve quality and interpretability of this sparse dataset.

4.4.7. Future Directions

With cheminformatics methods primarily being developed for quantum chemical and pharmaceutical methods, many of the chemical databases, such as PubChem, contain substantial numbers of pharmacological and theoretical compounds that are not commercially available or yet to be synthesized. In the context of the cement industry, any screened molecules would need to be economically feasible and readily acquired. As opposed to manually examining the remaining compounds leading to long set times, establishing a link to common chemical suppliers to import availability and price would improve this selection process in future research.

While basic physical exploration into the working mechanisms of these retarders was performed, a more in-depth analysis needs to be undertaken. Phosphonated molecules have not garnered much attention for the use as cementitious retarders, and they deserve further study. Measurement of ion complexation and adsorption behavior of these alternative anionic chemistries need to be performed in order to develop a physiochemical understanding of the behavior of these compounds within CSA cement. This understanding could lead to alternative approaches to model these small datasets, such as similar past approaches which utilized physiochemical understanding and ML in development of cement superplasticizers.[15,53]

This methodology could be extended to the screening and prediction of any of the number of small organic admixtures for cementitious systems. A similar approach was performed in the prediction of shrinkage-reducing admixtures[16], but could also be extended to air-entraining admixtures, defoaming agents, and set accelerating admixtures designed for specific cementitious systems. While data-driven research can require extensive experimentation, the tools of machine learning can be used for model building even with small training sets. In this work, folding and feature selection were utilized to learn over a sparse dataset, and an ML process was still able to develop an accurate model relating chemical structure to function. However, as this model still has a RMSE of 26.6 min, increasing the number of datapoints along with higher diversity in the screened compounds would allow for more accurate feature selection and predictions.

The work presented here is an example of the broader field of computer-aided molecular design, which is becoming an important methodology in the development of advanced materials. The broad challenge is efficiently screening the vast space of chemical species to determine which could perform particular functions in materials applications. Bayesian cheminformatics is a powerful formalism for performing this screening[54], and it has been applied to the optimization of thermal conductivity in polymers. Similar approaches can be applied towards the prediction of chemical admixtures, where ML can model the physiochemical factors and molecular structures generated which correspond to maximum efficacy. With the shifting nature of science into a data-intensive fourth paradigm[1], these approaches could be combined with experimental design in order to develop better predictions of different classes of chemical admixtures.

4.5. Conclusion

The efficacy of a cheminformatics approach for chemical admixture design was demonstrated with an example application in the use of virtual screening methods for identification of small organic molecules capable of retarding CSA set time beyond 1 hour. Methodology was developed to utilize molecular fingerprints with the sparse datasets to virtually screen similarity between compounds and predict set times within CSA cement. It was determined that commercially available glyphosate would extend set time to beyond 1 h and experimentally determined to be 55 min, well within the predicted standard deviation of 26 min. Set times determined via Vicat testing were consistent in trends found through isothermal calorimetry. While physical insight into set time trends could not be discovered due to the folding process, multiple anionic chemistries were shown to be effective in prolonging set time where experimentation and insight can be obtained in future experimentation.

While the folding operations performed in Morgan fingerprints make interpretability difficult, they do offer a powerful methodology for performing virtual screening, particularly in exploring specific hypotheses, such as the effects of anionic groups.[38, 43, 55] The results from this study suggest that for small molecules, the phosphono chemistry is a critical component of activity. As the exact retarding mechanisms for each of these molecules is not fully understood, the cheminformatics approach was applied as a mechanism for learning similarity between molecular structures which contribute to the retarding mechanisms.

The development of this machine learning tool that guides the testing of chemical retarders for sustainable, durable cementitious binders allows for efficient, cost-effective virtual screening. This same method also allows for the possibility of building upon this dataset to create large datasets with more bits encoded in the fingerprints, similar to many bioinformatics approaches in drug discovery. Progress in informatics tools are catalyzing new ways of research, which could translate to significant advances in cement research.

4.6. Research Contributions

C.M. Childs performed set time analysis, cheminformatics and ML analysis, and virtual screening. O. Canbek and C. Szeto performed cement characterization and hydration kinetics measurements. T. M. Kirby, C. Zhang, and J. Zheng performed set time analysis.

4.7. Appendix

Vicat testing was utilized for set time experimentation. The Vicat needle apparatus and set up is shown in Figure 4.12.

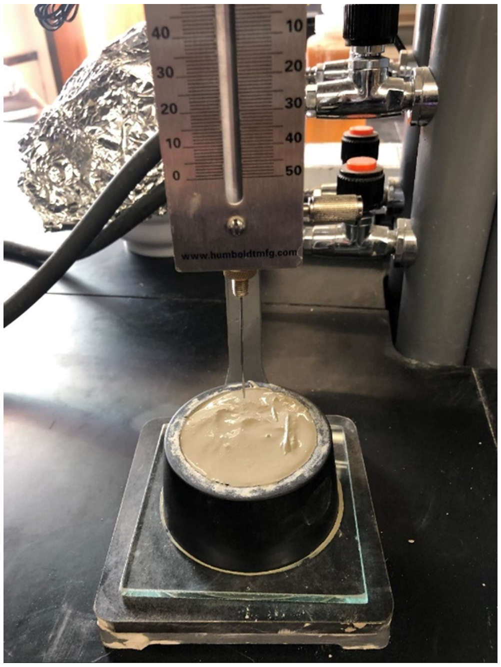


Figure 4.12. Vicat Needle Apparatus utilized for set time experimentation.

A complete graph, showing the set times of all 23 tested molecules is presented in Figure 4.13.

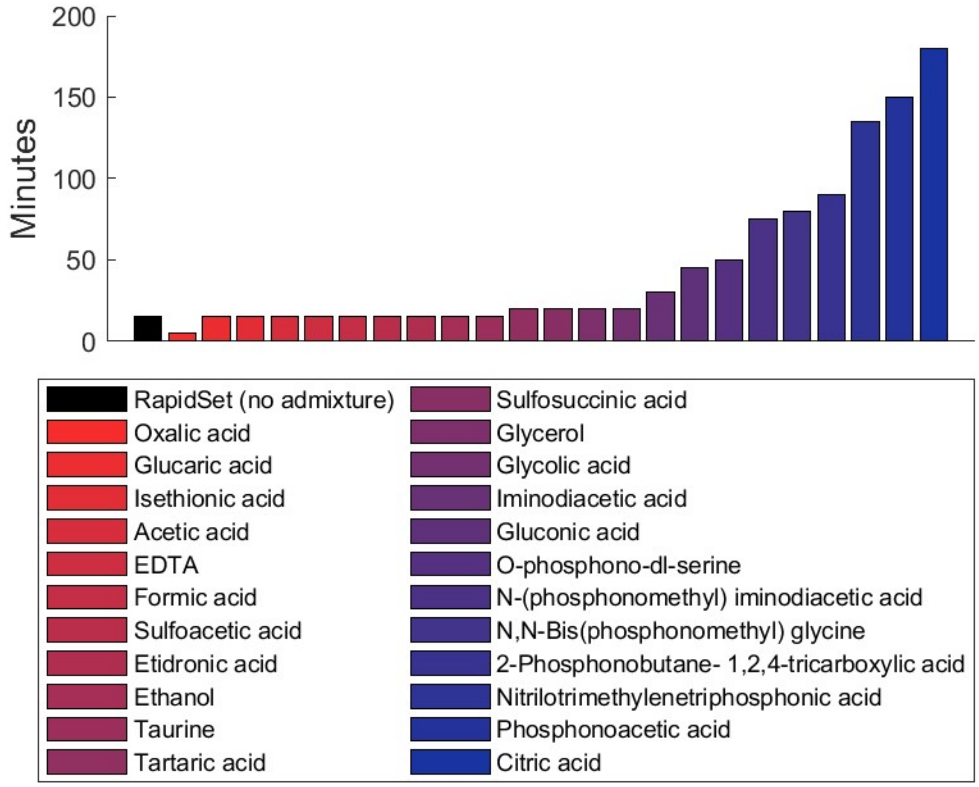


Figure 4.13. Complete summary of Vicat needle measured set times for all 23 tested retarders. All experiments are recorded at 1% loading of retarder by weight of cement.

Complete cross-validation plots for each fold of the ECFP are shown in Figure 4.14. The alpha value corresponding to the lowest uncertainty in MSE is demarcated with a dashed line. The results of this alpha value, MSE, and standard deviation in MSE are presented in Table 4.4.

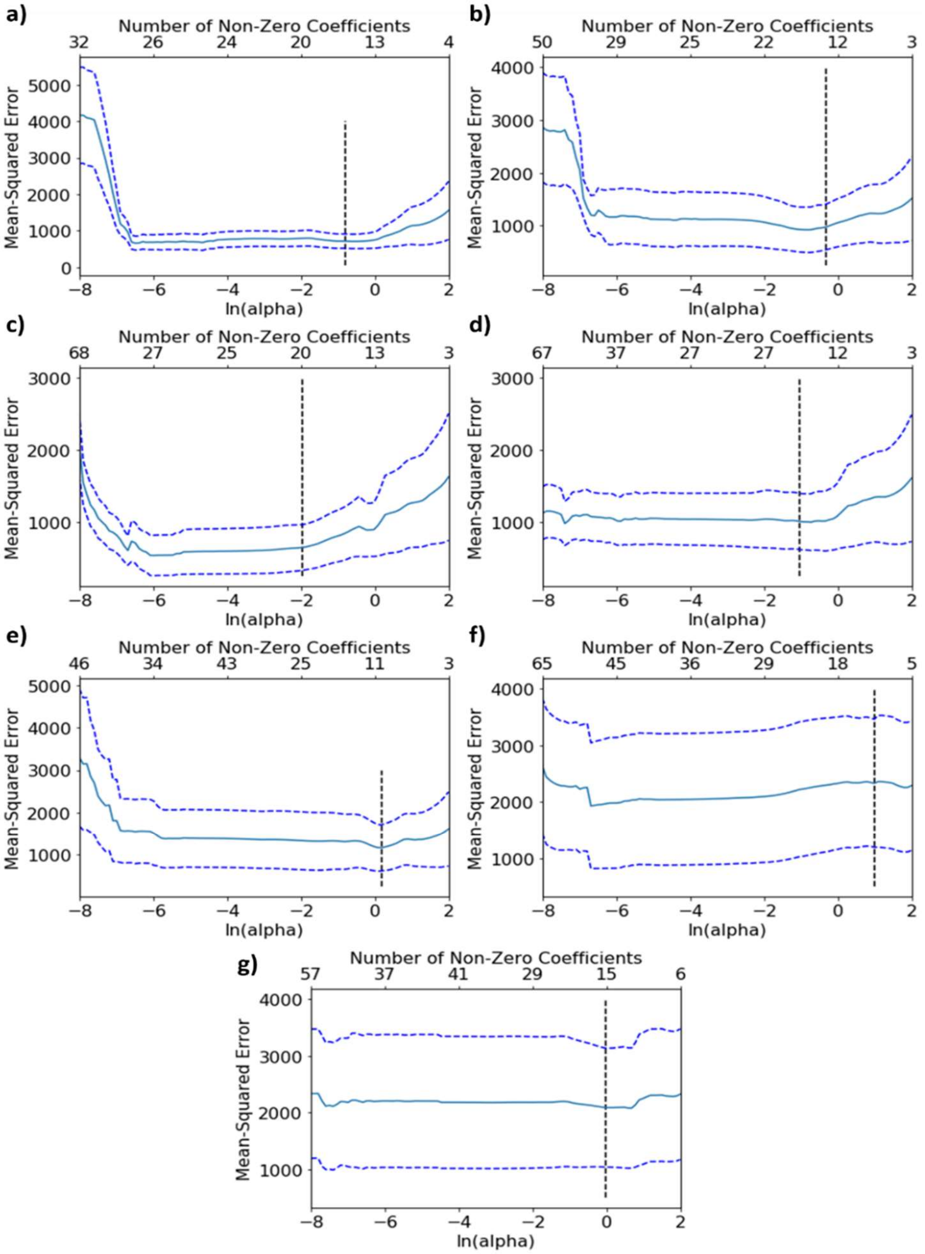


Figure 4.14. ECFP cross validation results for each fold: a) 32 bits, b) 64 bits, c) 128 bits, d) 256 bits, e) 512 bits, f) 1028 bits, and g) 2056 bits. Note that in some cases the minimum MSE and minimum standard deviation of MSE occur in areas with >23 descriptors. As this would create more features than compounds in the test set, alpha must be chosen at a higher value. This selected α value is represented as the dashed vertical black line.

Table 4.4. Summary of the selected α hyperparameter along with the corresponding MSE and standard deviation in MSE at these values for the various length ECFP's. Note the tradeoff between minimizing the MSE or minimizing uncertainty between the 32 bit ECFP and 128 bit ECFP. Finding the area to minimize the standard deviation (uncertainty) was selected as the method utilized to select the best model and thus the 32 bit ECFP was utilized for virtual screening.

# of Bits	Alpha Value	MSE	Standard Deviation in MSE
32	0.45	710	196
64	0.472	973	429
128	0.14	647	317
256	0.35	1007	389
512	1.2	1160	547
1028	2.7	2325	1134
2056	0.98	2085	1049

Table 4.5. Summary of calorimetry data showing plain CSA cement paste and the high-activity set retarders.

Chemical	Abbreviation	Time to first peak after initial wetting (h)	Cumulative heat evolved (J/g)
CSA Cement (No Admixture)	No Admixture	0.88	223.1
N,N-Bis(phosphonomethyl) glycine	Glyphosine	4.26	250.5
Nitrilotrimethylenetriphosphonic acid	ATMP	7.47	242.6
Phosphonoacetic acid	PAA	5.97	229.5
Citric acid	CA	8.53	252.7
Glyphosate	Glyphosate	2.7	245.7

4.8. References

[1] *The Fourth Paradigm: Data-Intensive Scientific Discovery*, 1st ed.; Hey, T., Tansley, S., Tolle, K., Eds.; Microsoft Research: Redmond, Washington, 2009.

[2] Mueller, T.; Kusne, A.; Ramprasad, R. Machine Learning in Materials Science: Recent Progress and Emerging Applications. In *Reviews in Computational Chemistry*; Parrill, A. L., Lipkowtiz, K. B., Eds.; John Wiley and Sons, Inc., 2016; Vol. 29, pp 186–273.

[3] Chen, H.; Kogej, T.; Engkvist, O. Cheminformatics in Drug Discovery, an Industrial Perspective. *Mol. Inform.* **2018**, *37* (9–10), 1800041.

[4] Sippl, W.; Robaa, D. QSAR/QSPR. In *Applied Chemoinformatics*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2018; pp 9–52.

[5] Koch, U.; Hamacher, M.; Nussbaumer, P. Cheminformatics at the Interface of Medicinal Chemistry and Proteomics. *Biochimica et Biophysica Acta - Proteins and Proteomics*. Elsevier B.V. 2014, pp 156–161.

[6] Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71* (C), 58–63.

[7] Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. *Sci. Rep.* **2013**, *3* (1), 2810.

[8] Baldi, P.; Nasr, R. When Is Chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and Its Extreme Values. *J. Chem. Inf. Model.* **2010**, *50* (7), 1205–1222.

[9] Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today*. December 2006, pp 1046–1053.

[10] Haj-Ali, R. M.; Kurtis, K. E.; Sthapit, A. R. Neural Network Modeling of Concrete Expansion during Long-Term Sulfate Exposure. *ACI Mater. J.* **2001**, *98* (1), 36–43.

[11] Young, B. A.; Hall, A.; Pilon, L.; Gupta, P.; Sant, G. Can the Compressive Strength of Concrete Be Estimated from Knowledge of the Mixture Proportions?: New Insights from Statistical Analysis and Machine Learning Methods. *Cem. Concr. Res.* **2019**, *115*, 379–388.

[12] Chou, J. S.; Tsai, C. F.; Pham, A. D.; Lu, Y. H. Machine Learning in Concrete Strength Simulations: Multi-Nation Data Analytics. *Constr. Build. Mater.* **2014**, *73*, 771–780.

[13] Dutta, S.; Samui, P.; Kim, D. Comparison of Machine Learning Techniques to Predict Compressive Strength of Concrete. *Comput. Concr.* **2018**, *21* (4), 463–470.

[14] DeRousseau, M. A.; Laftchiev, E.; Kasprzyk, J. R.; Rajagopalan, B.; Srubar, W. V. A Comparison of Machine Learning Methods for Predicting the Compressive Strength of Field-Placed Concrete. *Constr. Build. Mater.* **2019**, *228*, 116661.

[15] Menon, A.; Childs, C. M.; Poczós, B.; Washburn, N. R.; Kurtis, K. E. Molecular Engineering of Superplasticizers for Metakaolin-Portland Cement Blends with Hierarchical Machine Learning. *Adv. Theory Simul.* **2019**, *2* (1800164).

[16] Kayello, H. M.; Tadisina, N. K. R.; Shlonimskaya, N.; Biernacki, J. J.; Visco, D. P. An Application of Computer-Aided Molecular Design (CAMD) Using the Signature Molecular Descriptor - Part 1. Identification of Surface Tension Reducing Agents and the Search for Shrinkage Reducing Admixtures. *J. Am. Ceram. Soc.* **2014**, *97* (2), 365–377.

[17] Chen, F.-C. Virtual Screening of Conjugated Polymers for Organic Photovoltaic Devices Using Support Vector Machines and Ensemble Learning. *Int. J. Polym. Sci.* **2019**.

[18] Pyzer-Knapp, E. O.; Simm, G. N.; Aspuru Guzik, A. A Bayesian Approach to Calibrating High-Throughput Virtual Screening Results and Application to Organic Photovoltaic Materials. *Mater. Horizons* **2016**, *3* (3), 226–233.

[19] Zaslavskiy, M.; Jégou, S.; Tramel, E. W.; Wainrib, G. ToxicBlend: Virtual Screening of Toxic Compounds with Ensemble Predictors. *Comput. Toxicol.* **2019**, *10*, 81–88.

[20] Sanchez-Lengeling, B.; Roch, L. M.; Perea, J. D.; Langner, S.; Brabec, C. J.; Aspuru-Guzik, A. A Bayesian Approach to Predict Solubility Parameters. *Adv. Theory Simulations* **2019**, *2* (1), 1800069.

[21] Childs, C. M.; Washburn, N. R. Embedding Domain Knowledge for Machine Learning of Complex Material Systems. *MRS Commun.* **2019**, *9* (03), 806–820.

[22] Kensert, A.; Alvarsson, J.; Norinder, U.; Spjuth, O. Evaluating Parameters for Ligand-Based Modeling with Random Forest on Sparse Data Sets. *J. Cheminform.* **2018**, *10*, 49.

[23] Elton, D. C.; Boukouvalas, Z.; Butrico, M. S.; Fuge, M. D.; Chung, P. W. Applying Machine Learning Techniques to Predict the Properties of Energetic Materials. *Sci. Rep.* **2018**, *8* (1), 1–12.

[24] Kurtis, K.; Alapati, P.; Burris, L. Alternative Cementitious Materials: An Evolution or Revolution? . *Public Roads*. 2019, pp 4–9.

[25] Naqi, A.; Jang, J. Recent Progress in Green Cement Technology Utilizing Low-Carbon Emission Fuels and Raw Materials: A Review. *Sustainability* **2019**, *11* (2), 537.

[26] Winnefeld, F.; Kaufmann, J. Concrete Produced with Calcium Sulfoaluminate Cement – a Potential System for Energy and Heat Storage. *1st Middle East Conf. Smart Monit. Assess. Rehabil. Civ. Struct.* **2011**, No. January 2019.

[27] Wang, P.; Li, N.; Xu, L. Hydration Evolution and Compressive Strength of Calcium Sulphoaluminate Cement Constantly Cured over the Temperature Range of 0 to 80 °C. *Cem. Concr. Res.* **2017**, *100*, 203–213.

[28] Gwon, S.; Jang, S.; Shin, M.; Gwon, S.; Jang, S. Y.; Shin, M. Combined Effects of Set Retarders and Polymer Powder on the Properties of Calcium Sulfoaluminate Blended Cement Systems. *Materials (Basel).* **2018**, *11* (5), 825.

[29] Winnefeld, F. Interaction of Superplasticizers with Calcium Sulfoaluminate Cements. In *Proc. 10th International Conference on Superplasticizers and Other Chemical Admixtures in Concrete*; Malhotra, V. M., Ed.; American Concrete Institute: Prague, 2012; pp 28–31.

[30] Zajac, M.; Skocek, J.; Bullerjahn, F.; Ben Haha, M. Effect of Retarders on the Early Hydration of Calcium-Sulpho-Aluminate (CSA) Type Cements. *Cem. Concr. Res.* **2016**, *84*, 62–75.

[31] Ben Haha, M.; Winnefeld, F.; Pisch, A. Advances in Understanding Ye'elimite-Rich Cements. *Cem. Concr. Res.* **2019**, *123*, 105778.

[32] Jolicoeur, C.; Simard, M.-A. Chemical Admixture-Cement Interactions: Phenomenology and Physico-Chemical Concepts. *Cem. Concr. Compos.* **1998**, *20* (2–3), 87–101.

[33] Frank, W.; Stefanie, K. Influence of Citric Acid on the Hydration Kinetics of Calcium Sulfoaluminate Cement. In *1st International Conference on Sulphoaluminate Cement: Materials and Engineering Technology*; Wuhan, China, 2013; pp 288–308.

[34] Bishop, M.; Bott, S. G.; Barron, A. R. A New Mechanism for Cement Hydration Inhibition: Solid-State Chemistry of Calcium Nitrilotris(Methylene)Triphosphonate. *Chem. Mater.* **2003**, *15* (16), 3074–3088.

[35] *A Practical Guide to Microstructural Analysis of Cementitious Materials*, 1st ed.; Scrivener, K., Snellings, R., Lothenbach, B., Eds.; CRC Press: Boca Raton, Fl, 2016.

[36] ASTM International. C191-19 Standard Test Methods for Time of Setting of Hydraulic Cement by Vicat Needle. ASTM International: West Conshohocken, PA 2019.

[37] RDKit: Open-source cheminformatics http://www.rdkit.org.

[38] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

[39] Probst, D.; Reymond, J.-L. A Probabilistic Molecular Fingerprint for Big Data Settings. *J. Cheminform.* **2018**, *10* (1), 66.

[40] Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci* **1995**, *35*, 1039–1045.

[41] Tibshirani, R. Regression Selection and Shrinkage via the Lasso. *J. R. Stat. Soc. B* **1996**, *58* (1), 267–288.

[42] Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

[43] Gütlein, M.; Kramer, S. Filtered Circular Fingerprints Improve Either Prediction or Runtime Performance While Retaining Interpretability. *J. Cheminform.* **2016**, *8* (1), 60.

[44] Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102–D1109.

[45] Heap, I.; Duke, S. O. Overview of Glyphosate-Resistant Weeds Worldwide. *Pest Manag. Sci.* **2018**, *74* (5), 1040–1049.

[46] Burris, L. E.; Kurtis, K. E. Influence of Set Retarding Admixtures on Calcium Sulfoaluminate Cement Hydration and Property Development. *Cem. Concr. Res.* **2018**, *104*, 105–113.

[47] García-Maté, M.; De La Torre, A. G.; León-Reina, L.; Losilla, E. R.; Aranda, M. A. G.; Santacruz, I. Effect of Calcium Sulfate Source on the Hydration of Calcium Sulfoaluminate Eco-Cement. *Cem. Concr. Compos.* **2015**, *55*, 53–61.

[48] Velazco, G.; Almanza, J. M.; Cortés, D. A.; Escobedo, J. C.; Escalante-Garcia, J. I. Effect of Citric Acid and the Hemihydrate Amount on the Properties of a Calcium Sulphoaluminate Cement. *Mater. Constr.* **2014**, *64* (316), e036.

[49] Dimiduk, D. M.; Holm, E. A.; Niezgoda, S. R. Perspectives on the Impact of Machine Learning, Deep Learning, and Artificial Intelligence on Materials, Processes, and Structures Engineering. *Integrating Materials and Manufacturing Innovation*. Springer September 1, 2018, pp 157–172.

[50] Simard, M.-A.; Nkinamubanzi, P.-C.; Jolicoeur, C.; Perraton, D.; Aïtcin, P.-C. Calorimetry, Rheology and Compressive Strength of Superplasticized Cement Pastes. *Cem. Concr. Res.* **1993**, *23* (4), 939–950.

[51] Hou, W.; Bao, J. Evaluation of Cement Retarding Performance of Cellulosic Sugar Acids. *Constr. Build. Mater.* **2019**, *202*, 522–527.

[52] M. Zajac, J. Skocek, F. Bullerjahn, B. Lothenbach, K. Scrivener, M. Ben Haha, Early hydration of ye'elimite: Insights from thermodynamic modelling, Cem. Concr. Res. 120 (2019) 152–163. doi:10.1016/j.cemconres.2019.03.024.

[53] Menon, A.; Gupta, C.; Perkins, K. M.; DeCost, B. L.; Budwal, N.; Rios, R. T.; Zhang, K.; Póczos, B.; Washburn, N. R. 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.

[54] S. Wu, Y. Kondo, M. aki Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa, R. Yoshida, Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm, Npj Comput. Mater. 5 (2019) 1–11. doi:10.1038/s41524-019-0203-2.

[55] Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50* (2), 205–216.

Chapter 5. A Machine Learning Approach for the Prediction of Ultra-High Performance Concrete Compressive Strength Based on Latent Variables

5.1. Introduction

Ultra-high performance concrete (UHPC) is a class of concrete having high binder content, which includes supplementary cementitious materials (SCMs) and fine aggregates, very low water content, high-range water reducers, and fiber reinforcement. These materials are defined by compressive strengths in excess of 150 MPa and high tensile strengths of around 15 MPa.[1,2] Due to the low water ratios, lack of course aggregate, and high binder contents, UHPC cement content is roughly three times greater than normal strength concrete (NSC).[3] One of the earliest utilizations of the compositional factors which are the basis of UHPC was performed by the Pittsburgh office of the US Army Corps of Engineers in 1983. This work was performed on the Kinzua Dam, which was completed in 1966 along the upper Allegheny River, primarily as a flood control measure for Pittsburgh, Pennsylvania.[4] However, by the 1983, the spillway floor had already been replaced once in 1974 and was again in failure due to abrasion and erosion as shown in Figure 5.1.[5]

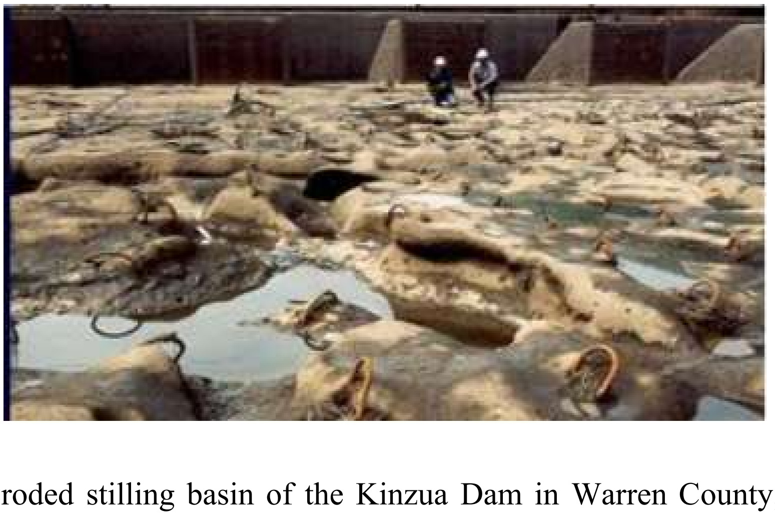


Figure 5.1. Eroded stilling basin of the Kinzua Dam in Warren County, Pennsylvania, in 1983.

The US Army Corps of Engineers performed a series of tests to design high strength cements which they thought would correlate well with abrasion resistance. Their final blend exceeded 85 MPa at 28-day strength testing and was designed with high amounts of silica fume as an SCM, high-range water reducers, water-to-cement ratios (w/c) less than 0.30, and utilization of fiber reinforcement.[6] A diver inspection in 2013 showed little damage to the basin and as of 2018 has lasted five times longer than the original concrete.[5] Although these compositions did not reach compressive strengths consistent with modern-day UHPC, the principles in the design of the Kinzua Dam formulations allowed for the development of UHPC.

Improvement in both materials and composition allowed for the first utilization of UHPC in an engineering structure in Sherbrooke, Canada, in 1997 as shown in Figure 5.2.[7]

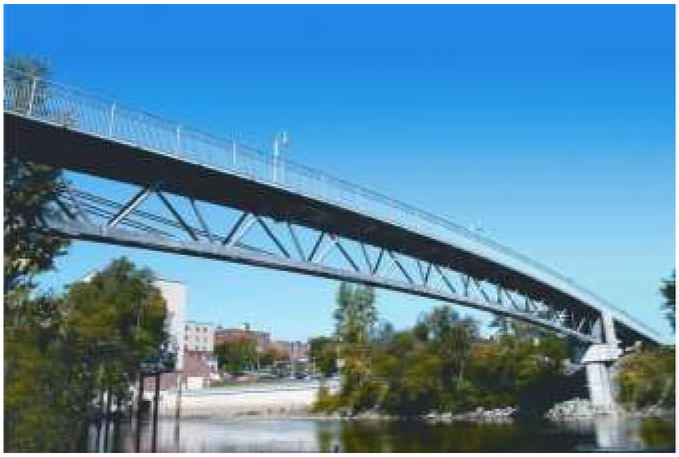


Figure 5.2. First UHPC engineering structure. A pedestrian bridge in Sherbrooke, Canada.

The first UHPC highway bridge in the United States was the Mars Hill Bridge in Wapello County, Iowa in 2006 and shown in Figure 5.3. The strength provided by UHPC allowed for the elimination of stirrups in the shear-reinforcement in the I-girders.[8]

Figure 5.3. Mars Hill Bridge in Wapello County, Iowa.[9]

UHPC strength is developed through both chemical and mechanical improvement to that of NSC. The improvement in chemical strength is due to the pozzolanic activity induced through the utilization of SCMs. SCMs are amorphous materials composed of silica and alumina, which do not react as cementitious materials. Instead, these materials which include silica fume, metakaolin, and fly ash, react with latent CH from hydrated portland cement. This pozzolanic reaction is responsible for formation of a strong C-S-H and C-A-H amorphous phases.[10] In UHPC, both metakaolin and silica fume are widely studied materials due to their high purity and high specific surface areas which promote the pozzolanic reaction.[11,12]

The improvement in UHPC strength from mechanical improvement is due to the minimization of porosity in the microstructure.[13] In UHPC there are several types of void formation. The first is interlayer spacing within the C-S-H and C-A-H phases. The size of these pores is between 5-25 Å and as such are within the realm of van der Waals interactions, therefore are not detrimental to cement strength.[14] A second type of void formation is capillary pores. These pores have been found to be inversely related to cement strength and the size and continuity vary directly with the w/c.[15] Finally, void formation is found in the interfacial transition zone (ITZ) between the hydrated cement paste and aggregate phase in UHPC. Generally, the ITZ is considered as the strength-limiting phase in concrete.[14] However, the void formation within the ITZ can be reduced through the utilization of pozzolanic, high surface area SCMs which lead to increased chemical bonding and physical interactions through increased packing, respectively.[16]

A traditional methodology to minimize the void ratio in the ITZ for UHPC is through particle packing models. Two common models for optimizing UHPC compositions include the modified Andersen and Andreasen model and the compressible packing model (CPM). The modified Andersen and Andreasen model[17] incorporate particle size distributions and an adjustable parameter, q, to generate an ideal gradation curve where actual compositions can be manually fit with the q parameter to find an optimal packing density. The CPM was first developed by de Larrard and calculates a packing index, K, which can be optimized to a specified value, where K=4 is a suggested value for a self-consolidating concrete mix.[18]

A second factor into minimizing UHPC porosity is having a low w/c. This ratio decreases the formation of capillary pores through limiting the amount of unreacted water in the system.[13] While packing density is solely based on the solid content and particle diameters of the mixture, WFT considers the water content and surface areas available for water adsorption.[19] Even with an optimized packing density, excess water could lead to capillary pore formation. While increasing particle packing leads to an increase in compressive strength, an increase in WFT leads to a decrease.[20] Depending on the size and surface area of solids, along with the w/c ratio, a complex interplay occurs between attempts to optimize particle packing, WFT, and SCM reactivity.

Recently, traditional analysis of experimental testing has been supplemented with machine learning (ML) methods. ML is a diverse collection of statistical algorithms which are utilized to predict a system's properties. Ghafari et. al[21] trained an artificial neural network (ANN) on 53 different UHPC compositions, optimizing a composition which experimentally agreed with predicted compressive strengths less than 5%. Although the optimized blend did not extrapolate to predicting an optimized blend outside of the range of tested compositions, the ANN outperformed traditional statistical mixture design approaches based on multiple linear regression.

However, this ANN is parameterized by the specific materials in the training set, which limits its predictions to these components. For each various material with a different size, surface area, or reactivity there would be a lack of generalizability for this model and additional experimentation would have to be performed to retrain an ML algorithm. One method for improving generalizability, particularly for small datasets, is through representing the system in terms of latent variables.[22] With the preexisting literature for optimizing blends based on particle packing and WFT equations, a hierarchical machine learning (HML) model[23,24] was established for the prediction and optimization of UHPC compressive strength. Data from literature, supplemented with experimentally collected data, was encoded with latent variables based on the CPM and WFT, allowing for generalization to unseen, local source materials of various size, reactivity, and surface area. A Bayesian optimization approach was followed to predict UHPC compositions of high compressive strength with low model uncertainty.

5.2. Experimental

5.2.1. Materials

Cement mortars were prepared using a Type I ordinary Portland cement (Lafarge Holcim), metakaolin (MetaMax, BASF), and silica fume (Elkem Materials, Inc.), masonry sand (Vulcan Materials) with a particle diameter of 400 μm and river sand (River Sand Inc.) with a particle diameter of 600 μm. Also included as part of the compositions were a high range water reducer (MasterGlenium 7920, BASF) and steel fibers (Dramix, Bekaert) with a 13 mm length and 0.20 diameter.

5.2.2. Assessment of Strength

All mixes were conducted in 0.03 ft^3 batches in a tabletop mixer. Additional time was allowed for the mix to come together before fibers were added. The specimens cast for compression testing were 2 in by 2 in mortar cubes. There were six specimens cast for each mix, with compression testing being performed at 7 and 28 days. The specimens were loaded in the compression machine at a rate of 18,000 pounds per minute, on average.

5.2.3. Data Collection

A database was compiled of UHPC mixtures from published literature. Four datasets were chosen for training the model and are summarized in Table 5.1.

Table 5.1. Datasets compiled for training.

Data Source	Tafraoui et. al[12]	Ghafari et. al[21]	Berry et. al[13]	Wille et. al[25]
# Datapoints	24	106	54	7
SCMs	Silica Fume Metakaolin	Silica Fume	Fly ash Silica Fume	Metakaolin
Fine Aggregates	Sand- 230 µm Quartz- 11 µm	Sand- 400 µm Quartz- 7 µm	Sand- 500 µm	Sand- 110 µm Sand- 500 µm Glass- 5 µm
Temperature	20C, 90C, 150C	20C and 90C	20C	20C

The seven mixes from Wille et al.[25] were selected for use in validating the model.

5.3. Computational Methods

5.3.1. Data Representation

The amount of cement, supplementary cementitious materials, filler materials, aggregates, water, superplasticizer, and steel fibers in each mix was recorded. Additionally, the curing temperature and 28-day compressive strength results of each mix were included as the output for the dataset. All reported mix design parameters from the training mixes were converted on a per mass basis of the whole mixture (solid and water phases) utilizing an assumed specific gravity for each phase. The reported average particle diameter (D50) and specific surface area (SSA) from each data source for the fine aggregates were utilized. However, for SCM particles where particle sizes are not routinely recorded, the values listed in Table 5.2 were assumed based on existing laboratory data.

Table 5.2. Assumed particle parameters for UHPC components.

Particle	**Specific Gravity**	**D50 (μm)**	**SSA ($m^2\ kg^{-1}$)**
Cement	3.15	15	394
Fly Ash	2.38	25	500
Silica Fume	2.22	0.2	18000
Metakaolin	2.3	12	14000
Sand	2.5	Varies by Source	Varies by Source
Quartz	2.65	Varies by Source	Varies by Source

Six parameters were chosen for inclusion as domain knowledge in the model. These six parameters are the equivalent cement content, the particle packing of the mixture, the water film thickness, the superplasticizer content, the fiber content, and the curing temperature. These parameters were selected for consideration based on established knowledge of established relationships in cement to direct the HML model from compositional to middle layer variables as shown in Figure 5.4.

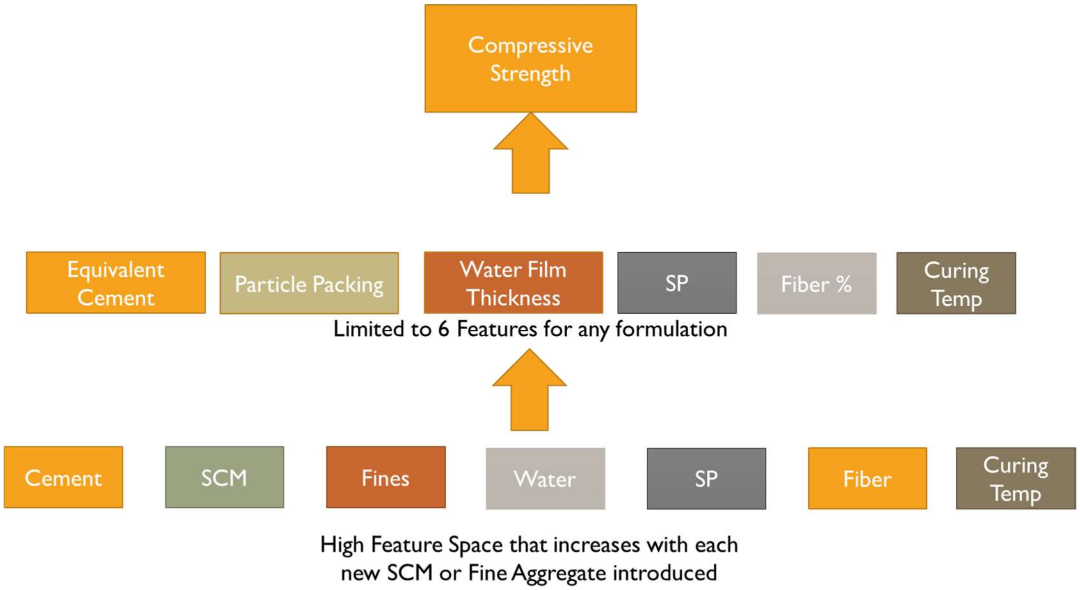


Figure 5.4. The bottom layer represents the compositional space for UHPC. This high feature space increases with each new type and source of aggregate utilized, preventing models from being established without prior trials being conducted. A middle layer represents latent variables and allows for the introduction of new source materials which can be parameterized by six features for any new source or size material introduced.

5.3.1.1. Equivalent Cement Content

The concept of an "equivalent cement" value first appeared in a paper on the thermal control of mass concrete placements.[26] In this application, it serves as an estimate of the approximate amount of heat generated by a concrete that includes SCMs. The equation below normalizes each mix component into an "equivalent" weight of cement based on its assumed heat generation. For example, class F fly ash is assumed to produce half as much heat as regular Portland cement, so the amount of class F fly ash is multiplied by 0.5 as shown in Eq. 5.1. This concept was incorporated as domain knowledge as a way of measuring the reactivity of a mix design in the absence of calorimetry data. The higher the equivalent cement content, the more reactive the mix is assumed to be.

Cement + 0.5(Amount of Class F Fly Ash)*
+ 0.8(Amount of Class C Fly Ash)*
+1.2(Amount of Silica Fume)*
+ 1.2(Amount of Metakaolin)*
+ X(Amount of Slag)
Where X = 1.1 for 0-20% replacement of cement by slag,
1.0 for 20-45% replacement, 0.9 for 45-65% replacement,
and 0.8 for 65-80% replacement. Eq. 5.1

5.3.1.2. Particle Packing

Including particle packing as domain knowledge requires an input parameter that will summarize the packing of the mixture in a single parameter. The packing model will be based upon the CPM, which has been demonstrated to be well-suited for multi-component, polydisperse systems.[27,28] The CPM summarizes the packing of the mixture into a single parameter, *K*. Higher *K* values correspond with denser mixtures and higher compressive strengths. Particle packing is also considered to be a critical in determining the material properties of cement in both the plastic and hardened states, and it can be used as a design variable in increasing the loading of fine aggregate and SCMs and controlling the material properties of these materials. In order to automatically calculate this *K* value, a Python script was developed and included in the model as shown below in Eq. 5.2-Eq. 5.5:

$$a_{ji} = \sqrt{1 - \left(1 - \frac{d_j}{d_i}\right)^{1.02}} \qquad \text{Eq. 5.2}$$

$$b_{ji} = 1 - \left(1 - \frac{d_i}{d_j}\right)^{1.5} \qquad \text{Eq. 5.3}$$

$$\Phi_i^* = \beta_i \left[1 - \sum_{j=1}^{i-1}\left(1 - b_{ji}\left[1 - \frac{1}{\beta_j}\right]\right)\Phi_j - \sum_{j=i+1}^{n} \frac{a_{ij}}{\beta_j}\Phi_j\right] \qquad \text{Eq. 5.4}$$

$$K = \sum_{i=1}^{n} \left(\frac{\frac{\Phi_i}{\Phi_i^*}}{1 - \frac{\Phi_i}{\Phi_i^*}} \right) \qquad \text{Eq. 5.5}$$

Where:

d_i = grain size of rank *i*

d_j= grain size of rank *j*

a_{ij} = coefficient for the loosening effect, exerted by the grains of rank *j* on those of rank *i* ($j > i$)

b_{ji} = coefficient for wall effect, of the grains of class *i* on the grains of rank *j* (j < i), with $d_1 > di > dn$

Φ* = maximum possible volume in the presence of other particles

Φ = volume of particles present

β = virtual packing density

K = Packing index, a unitless number that relates to packing.

The objective here is to maximize packing index (*K*). Based on particle size distributions for each of *n* components, the loosening (a_{ij}) and wall (b_{ij}) coefficients are determined and used to calculate the maximum possible volume for each particle size (Φ_i*), similar to study,[29] suggesting particle packing models can be used to predict flow and strength, particularly at early ages, in these systems. A Python script was created to uniquely represent each blend without the need to explicitly measure the actual packing density. ϕ was represented as 1-water content, while β was held constant for each UHPC blend.

5.3.1.3. Water Film Thickness

The water film thickness is a relationship between the amount of water present in the mixture and the surface area of all particles present in the mixture. A higher water film thickness value corresponds with higher workability and self-consolidating behaviour.[19] For UHPC it is assumed that there is no excess water to fill the voids in the mix. In a cementitious system, the pore solution phase can be split into two distinct types. The first type, filling water, is the water which fills voids between solid particles, and does not contribute to workability. The second portion of water occurs after these voids have been filled and is known as the excess water as shown in Eq. 5.6:[30]

$$u_w' = u_w - u \qquad \text{Eq. 5.6}$$

Where:

u_w'= excess water

u_w= ratio of water in system by volume

u = voids ratio

The amount of excess water is divided by the average specific surface area of all the particles in order to determine the WFT as shown in Eq. 5.7:

$$WFT = \frac{u_w'}{A_m} \qquad \text{Eq. 5.7}$$

Where:

WFT= water film thickness

A_m= average specific surface area

WFT has been shown to correlate well with cement paste rheology and strength as it embeds a standard knowledge, and similar to particle packing models, that as the w/c ratio increases there is an increase in WFT. However, unlike the CPM, the average specific surface area of cementitious particles is also considered and an increase in surface area leads to a corresponding decrease in WFT.[19,31] A python script was written to theoretically represent each cement composition in terms of WFT. As the actual packing density, ϕ, for these blends was not measured, the voids ratio, u, was calculated as shown in Eq. 5.8:

$$u = \frac{(1 - \phi)}{\phi} \qquad \text{Eq. 5.8}$$

Where:

$\phi = (1 - \text{water content})$

This allows for a consistent way to represent ϕ without individually measuring the extent of packing density for each blend.

5.3.1.4. Superplasticizer Content

The model was directed to evaluate the superplasticizer content of all mixes in the database. While superplasticizer lends UHPC its workability, an excess of superplasticizing admixture can cause the strength development to be delayed. Thus, it is important for the HML model to consider the amount of superplasticizer present.

5.3.1.5. Fiber Content

Generally, UHPC contains 2-3% steel fibers by volume. Further addition of steel fibers can cause a loss of workability and an increase in entrapped air, leading to lower compressive strengths. For these reasons, the model was directed to consider fiber content as domain knowledge.

5.3.1.6. Curing Temperature

It is well established that curing temperature has a large effect on the compressive strength of UHPC. Higher curing temperatures lead to accelerated strength development and higher overall compressive strengths.[32,33,34]

5.3.2. Machine Learning Model

Bayes' theorem determines the posterior probability of an event based on the probabilities of the factors constituting the event – prior probabilities and likelihood of these occurring. We use this posterior distribution in Bayesian optimization (BO), for sample-efficient optimization of the concrete formulation. At each iteration in BO, an approximation to the posterior probability density function can be produced by sampling from this posterior distribution. This allows an acquisition function to be defined from which subsequent UHPC formulations can be chosen from to measure in the optimization routine. The approach of utilizing Bayesian analysis is to marginalize over the posterior distribution of parameters so that you get a better prediction result both in terms of accuracy and generalization capability. Error analysis will take place through comparing the Mean Squared Error, a prediction score which ignores Bayesian probability and compares how well the mean values of the data fit to the best model; and Miscalibration area, a quantification of uncertainty in the model based on calibration techniques developed by Kuleshov et. al.[35] Miscalibration area utilizes a predictive uncertainty method that makes a prediction and gives an uncertainty in the form of a "X% credible interval", which aims to capture the true point X% of the time. A hold-out test set is then utilized to measure on how many test points the credible intervals contains the true point. By performing this hold-out test for every X% between 0% and 100%, an average difference between the goal percentage and the measured percentage (averaged over each goal percentage value from 0% to 100%) can be computed giving the miscalibration area. This Bayesian metric tells us how well the uncertainty errors capture where the values should be. Hence, the model learns if it is predicting well or poorly to each test point. Beyond generalization capability, this error metric will give the capability of knowing the optimal points to test to minimize error in the model. Working in conjunction with HML, the capability of determining the best points to minimize error leads to finding the best model with the minimal amount of data collection.

In this work, to perform approximate Bayesian inference, we used a probabilistic ensemble model, consisting of an ensemble of 20 ridge regression models (i.e., linear models with l2 regularization). Each element of the ensemble was instantiated with a randomly drawn regularization strength and initial random state. After training each ensemble element on a given dataset, the mean and variance of the ensemble for a given input were taken as parameters of a Gaussian approximation to a posterior distribution over functions of that input. We then apply a monotonic transformation of the posterior variance parameter, learned with respect to a given validation set, which produces a modified posterior approximation with improved average calibration.

The optimization routine was performed by first defining an acquisition function based on the posterior of our Bayesian model—in our case, we chose the probability of improving the 28-day strength over the best-observed value—and then finding the concrete formulation (i.e., set of input variables) that maximizes this acquisition function. To perform this optimization, we must proceed in two phases, based on the hierarchical structure of our HML model. In the first phase, we begin by determining the formulation of variables in the middle layer that maximize the acquisition function via a mutation-based algorithm procedure. Fixing this set of optimal middle-layer variables, we then performed a second optimization routine, which determined a set of input variables that map to this set of optimal middle-layer variables. This second optimization routine returns a set of input variables that is restricted to a custom set of constraints over the input variable space. Taken together, this procedure yields a set of input variables which maximize the probability that we will measure a value that improves upon the highest-observed 28-day strength.

5.4. Results and Discussion

5.4.1. Regression and Optimization

The results for the uncertainty ensemble utilizing a ridge regression for the bottom layer (Figure 5.5) and middle layer (Figure 5.6) are shown below. The MSE, RMSE, and miscalibration areas are tabulated in Table 5.3.

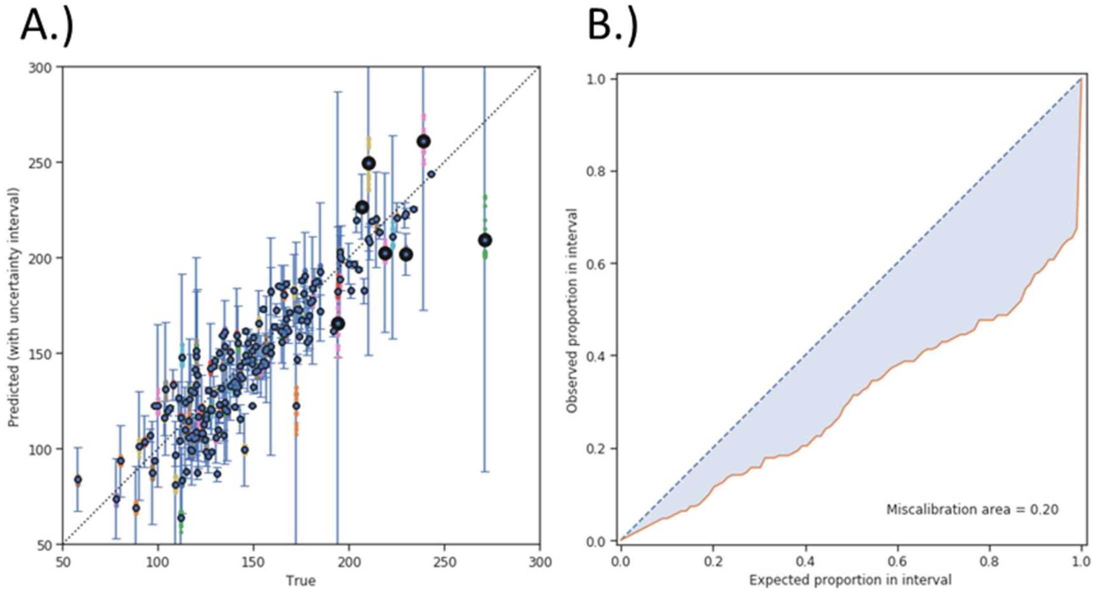


Figure 5.5. Results to the regression showing A.) the predicted and actual values and B.) miscalibration area utilizing the bottom layer compositional variables as inputs. The points with the larger dark circles represent the validation dataset (Wille et. al[25]).

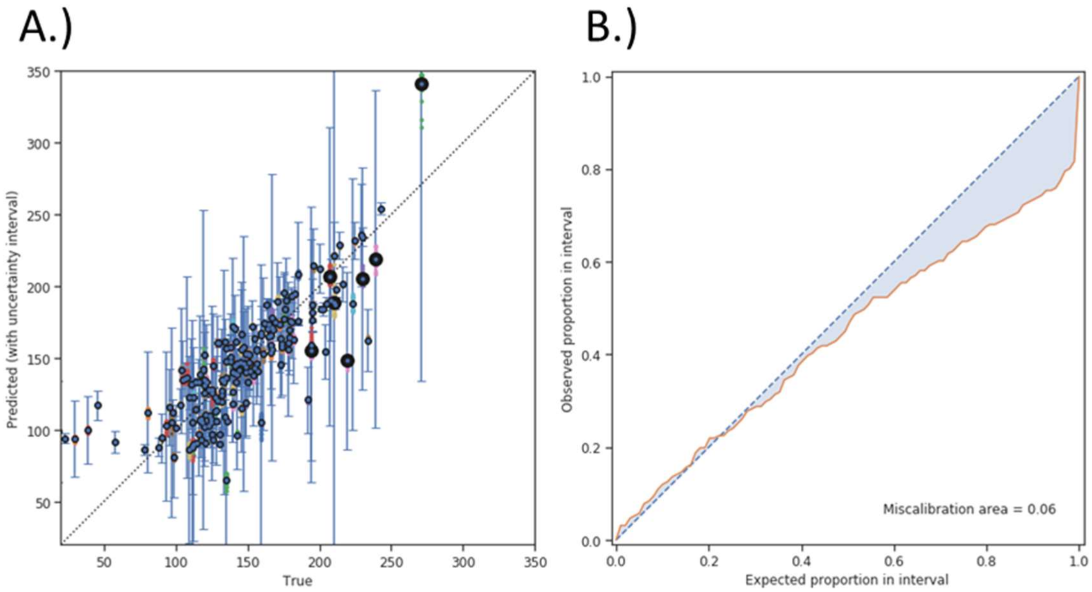


Figure 5.6. Results to the regression showing A.) the predicted and actual values and B.) miscalibration area utilizing the six middle layer variables as inputs. The points with the larger dark circles represent the validation dataset (Wille et. al[25]).

Table 5.3. Statistics representing the bottom and middle layers MSE, RMSE, and miscalibration area.

	MSE	RMSE (MPa)	Miscalibration Area
Bottom Layer	424	20.6	0.20
Middle Layer	660	25.7	0.06

The bottom layer and middle layer resulted in RMSEs of 34.0 MPa and 43.0 MPa on the validation set, respectively. While the bottom layer regression outperforms the middle layer by slightly over 5 MPa in terms of RMSE, utilizing the middle layer parameterization allows for generalizing an optimization routine to unseen material sizes. However, it is also shown that the middle layer has a lower miscalibration area, indicating that this model performs better in measuring uncertainty of the predicted datapoints.

Upon initial optimization, the only constraints applied were that the total sum of the cementitious components added to 100% and bounds were provided utilizing the minimum and maximum for SCMs and fine aggregate as in the dataset trained on. However, these optimized blends were characterized with a high SCM : aggregate ratio with low water ratios, making the initial provided blends unmixable.

For the second round of optimization, additional constraints were developed according to recommendations provided by the Federal Highway Administration for UHPC mix design.[36] These constraints were:

1. Sand to cement ratio limited to between 1.0 and 2.0

2. Maintain a water to cement ratio between 0.2 and 0.3

3. Limit silica fume to 18% of the combined weight of binder materials

4. Limit filler materials to 18% of the combined weight of binder materials

5. Limit superplasticizer to 10% of the amount of water present in the mix

Figure 5.7 shows the optimization results with the prior conditions applied. The top six blends corresponding to the minimized compressive strength: uncertainty ratio were selected for further testing as shown in Table 5.4**.**

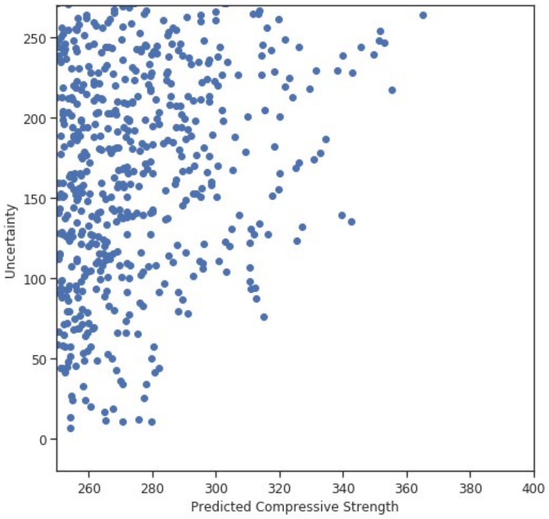


Figure 5.7. Optimization results for with the prior constraints for curing at 20 ºC with silica fume as the utilized SCM.

Table 5.4. Initial proportional mix designs (by weight of cement) along with the predicted and measured 7 and 28 day compressive strengths. For the first round of optimization, metakaolin was constrained to zero, allowing for only predictions of blends containing silica fume.

	Mix A-1	Mix A-2	Mix A-3	Mix A-4	Mix A-5	Mix A-6
Cement	1	1	1	1	1	1
Silica Fume	0.219	0.219	0.255	0.264	0.132	0.305
Metakaolin	0	0	0	0	0	0
Steel Fibers	0.284	0	0	0	0	0
Water	0.226	0.226	0.201	0.191	0.181	0.215
Superplasticizer	0.025	0.025	0.023	0.022	0.017	0.024
Masonry Sand	0	0	0	0	1.169	2.577
River Sand	1.581	1.581	1.500	1.459	0	0
Limestone-45	0.215	0.215	0.071	0	0	0.069
Curing Temperature	20	20	20	20	20	20
Predicted Strength (MPa)	254	254	279	275	270	265
Standard Deviation in Prediction (MPa)	6.5	6.5	11.1	11.9	11.1	11.5
Measured 7-day Strength (MPa)	134.6	75.6	96.8	63.6	77.0	28.1
Measured 28-day Strength (MPa)	189.9	79.2	123.6	82.6	87.1	53.0

The above designs were also not workable and were adjusted with various amounts of superplasticizer for the capability to mix the blends. Table 5.5 provides the initial amount of superplasticizer predicted compared to how much had to be added for the composition to be workable.

Table 5.5. Predicted amount of superplasticizer to add, the actual amount of superplasticizer added, and the corresponding flow test for each of the six optimized mixtures in mL/ft^3.

	Predicted Superplasticizer	**Actual Superplasticizer**	**Flow Test (in)**
Mix A-1	483	600	4″
Mix A-2	483	600	6 ½″
Mix A-3	483	1280	5 ½″
Mix A-4	493	2220	6″
Mix A-5	467	1460	6″
Mix A-6	500	2820	6″

It can be seen that Mix A-1 provided the highest compression strengths and in fact reaches compressive strengths well above the generally-accepted lower bounds for UHPC compressive strength. This strength comes at great cost, however, due to the high percentage of fibers in the mixture. This mix is the only one of the six tested that included fibers, containing around 7% steel fibers by volume of the mix. The usual recommendation for UHPC is 2% steel fibers by volume. This large amount of steel fibers reduced the workability to essentially zero, as can be seen by the four flow test result. These fibers would also serve to make this mix very expensive to produce commercially.

Mix A-2 was identical to Mix A-1, except fibers were excluded from the mixture. These samples exhibited swelling and exhibited much lower strengths due to this. The swelling is believed to be related to additional porosity from the extended 30-minute mix time necessary for the mix to blend. This swelling was not observed in any of the other mixes. The fibers present in Mix A-1 seem to have added enough confinement to prevent this expansion.

Mix A-3 was the second-best performing mix, nearly reaching 18,000 psi compressive strength by 28 days. Mix A-3 also require the second-least addition of superplasticizer. A negative correlation between additional superplasticizer content and compressive strength can be observed. Mix A-6, which required the most superplasticizer performed the worst of all mixes tested. Likewise, Mix A-4 had the second lowest compressive strength at 7 days, the third lowest compressive strength at 28 days, and also required the second highest admixture dosage.

Mix A-5 presents an interesting case because it performed similarly to Mix A-4 but had far less added superplasticizer. The reduction in strength is believed to be due to Mix A-5 having the lowest water to cement ratio. This reduced cement hydration in turn affected how much the silica fume could contribute to the strength.

Figure 5.8 provides the updated regression parity plots utilizing the results from the first round of optimization as additional datapoints for the model.

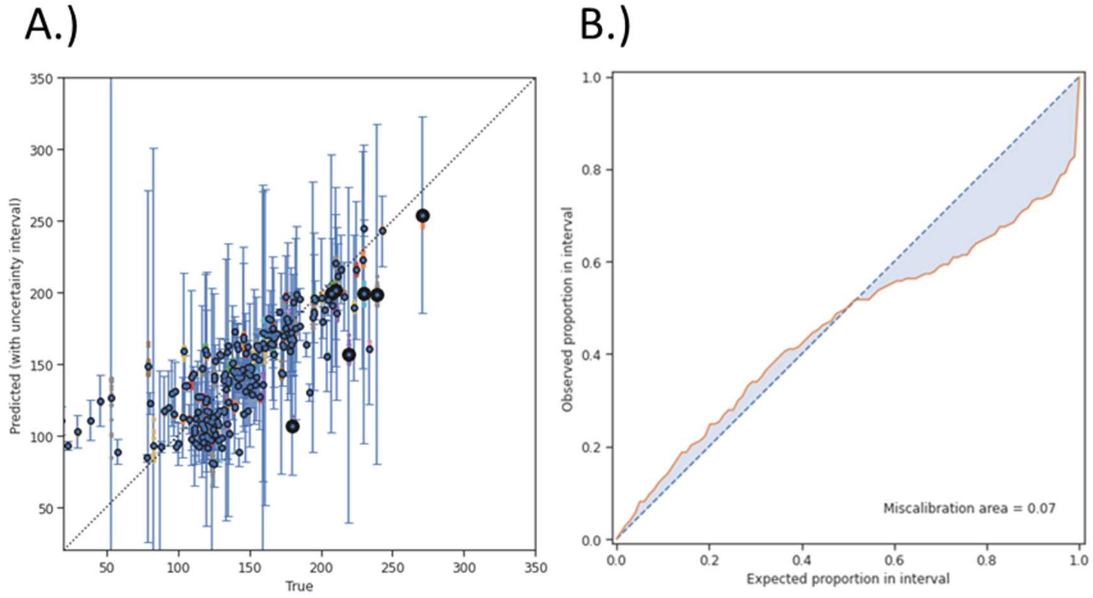


Figure 5.8. Results to the regression showing A.) the predicted and actual values and B.) miscalibration area utilizing the six middle layer variables as inputs. The points with the larger dark circles represent the validation dataset (Wille et. al[25]).

The RMSE for the updated regression is 28.1 MPa, which is a slight 2.4 MPa increase as compared to the first round. However, the RMSE for the validation set in the second iteration with only six additional data points decreased from 43.0 MPa to 41.8 MPa. The slight improvement in the validation indicates that training with the middle layer improves generalization capability of the model.

Figure 5.9 provides the next round of optimization results with an updated regression which includes the six prior tested blends. To reduce issues with workability, the minimum compressive strength was lowered to 200 MPa to allow for the prediction of blends with lower uncertainty and lowered the bounds for fibers to be closer to a maximum of 2% volume. The selected blends are shown in Table 5.6.

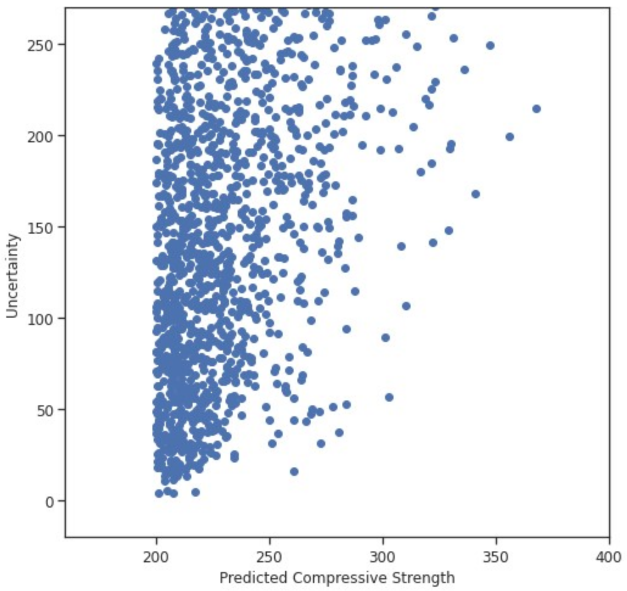


Figure 5.9. Optimization results for with the prior constraints for curing at 20 ºC with both silica fume and metakaolin as the utilized SCMs. The minimum compressive strength was also lowered to 200 MPa to predict blends with lower accompanied uncertainty.

Table 5.6. Initial proportional mix designs (by weight of cement) along with the predicted and measured 7 and 28 day compressive strengths. In this round of optimization two blends containing metakaolin and a single blend containing silica fume were selected for analysis.

	Mix B-1	**Mix B-2**	**Mix B-3**
Cement	1	1	1
Silica Fume	0	0	0.206
Metakaolin	0.206	0.102	0
Steel Fibers	0.245	0	0.245
Water	0.215	0.182	0.215
Superplasticizer	0.021	0.0175	0.021
Masonry Sand	0	0	0
River Sand	1.378	1.198	1.378
Limestone-45	0	0	0
Curing Temperature	20	20	20
Predicted Strength (MPa)	217	261	217
Standard Deviation in Prediction (MPa)	5.1	16.0	5.1
Measured 7-day Strength (MPa)	125.8	94.0	102.2
Measured 28-day Strength (MPa)	148.7	109.2	130.5

Mixtures B-2 and B-3 were also not initially workable; however, superplasticizer was added in much lower amounts than the first round of optimization, while mix B-1 was workable with these ratios. Table 5.7 provides the initial amount of superplasticizer predicted compared to how much had to be added for the composition to be workable.

Table 5.7. Predicted amount of superplasticizer to add, the actual amount of superplasticizer added, and the corresponding flow test for each of the six optimized mixtures in mL/ft^3 for the second round of optimization.

	Predicted Superplasticizer	Actual Superplasticizer	Flow Test (in)
Mix B-1	518	518	8″
Mix B-2	497	696	9″
Mix B-3	517	714	8 ¼″

Unlike the first round of iterations, all blends exceeded 100 MPa in strength as there was less change to the compositional parameters to produce a workable blend. Although this algorithm is capable of predicting UHPC compositions from a disparate source of materials, better feature representation needs to be incorporated for future modeling.

5.4.2 Future Directions

The major disadvantage in optimizing strength for UHPC was placing constraints to allow for the prediction of workable blends. First, more realistic constraints can be designed through several iterations of the optimization to attempt to approach predicted compositions which are workable. Also, future models incorporating UHPC flow test measurements along for a multi-objective design approach. These models can be similarly parameterized by water film thickness and packing density measurements, which have both been shown to have association with workability.[27,30]

The optimized blends selected here were based primarily on exploitation, a concept of finding blends which maximize strength. However, to take full advantage of the Bayesian optimization, additionally considering exploration to minimize model uncertainty can be utilized. Through exploration, the algorithm guides the design of experiment approach for the next best blends to test in order to decrease uncertainty, and can do so in the minimal number of experiments to reach a target uncertainty throughout the model.[37]

The disparate sources of data utilized for predicting compressive strength have inherent error. Although the D50 of many of the fine aggregates were reported and utilized from each individual source, particle sizes of SCMs and specific surface areas were not widely reported. This led to an assumption in the model that these particle sizes would be the same across all datasets. Future data collection for analysis should include more exact measurements of particle size and surface areas. Also, cement and SCM reactivity is currently treated as similar by the equivalent cement equation. Future models could embed thermodynamic modeling into the calculation to model cementitious chemistry based on the cement clinker composition and reactive phase contents of the SCM source materials.

5.5 Conclusion

The ubiquity and the necessity of concrete infrastructure prompts the need for increasing innovation to address the global challenge of meeting societal needs in the most sustainable and economical ways possible. This challenge is to generate new understanding that improves the design, utilization, and performance of UHPC. The feature space for property prediction in UHPC has been largely limited to the initial compositional formulation, leading to models which are valid only when testing the same compositional space. However, UHPC is a dynamic system which is characterized by changes in physicochemical forces, continuous reactivity, and changes in microstructure. The latent variables of particle packing, WFT, and equivalent cement provide better feature representation for UHPC, predict blends in excess of 100 MPa, and allow for greater generalization capability to a diversity of possible compositional materials.

5.6. Research Contributions

C.M. Childs performed ML analysis, latent variable design, and data collection. A. Miller performed compressive strength testing, data collection, and support in UHPC blend design. W. Neiswanger performed ML analysis and Bayesian optimization.

5.7. References

[1] Graybeal, B. A. *Material Property Characterization of Ultra-High Performance Concrete*; United States. Federal Highway Administration. Office of Infrastructure Research and Development, 2006.

[2] Stengel, T.; Schiebl, P. Life Cycle Assessment of UHPC Bridge Constructions: Sherbrooke Footbridge, Kassel Gartnerplatz Footbridge and Wapello Road Bridge. *Archit. Civ. Eng. Environ.* **2009**, *2* (1), 109–118.

[3] Zhong, R.; Wille, K.; Viegas, R. Material Efficiency in the Design of UHPC Paste from a Life Cycle Point of View. *Constr. Build. Mater.* **2018**, *160*, 505–513.

[4] Cowell, C. M.; Stoudt, R. T. DAM-INDUCED MODIFICATIONS TO UPPER ALLEGHENY RIVER STREAMFLOW PATTERNS AND THEIR BIODIVERSITY IMPLICATIONS. *J. Am. Water Resour. Assoc.* **2002**, *38* (1), 187–196.

[5] Bühler, E. R.; Lewis, R. C. MULTIPLE BLEND SUPPLEMENTRY CEMENTITIOUS MATERIALS (RECOVERED MINERAL COMPONENTS), BENEFIT SUSTAINABILITY THROUGH INNOVATIVE CONCRETE DESIGN. In *Fifth International Conference on Sustainable Construction Materials and Technologies*; London, 2019.

[6] Holland, T. C. *AD-A172 804, Abrasion-Erosion Evaluation of Concrete Mixtures for Stilling Basin Repairs, Kinzua Dam, Pennsylvania*; Vicksburg, 1986.

[7] Shi, C.; Wu, Z.; Xiao, J.; Wang, D.; Huang, Z.; Fang, Z. A Review on Ultra High Performance Concrete: Part I. Raw Materials and Mixture Design. *Construction and Building Materials*. Elsevier Ltd December 30, 2015, pp 741–751.

[8] Perry, V. H.; Seibert, P. J. FIFTEEN YEARS OF UHPC CONSTRUCTION EXPERIENCE IN PRECAST BRIDGES IN NORTH AMERICA. In *RILEM-fib-AFGC Int. Symposium on ultra-High Performance Fibre-Reinforced Concrete*; Marseille, 2013.

[9] Chapters 6-7 - Ultra-High Performance Concrete: A State-Of-The-Art Report for The Bridge Community , June 2013 - FHWA-HRT-13-060 https://www.fhwa.dot.gov/publications/research/infrastructure/structures/hpc/13060/006.cfm (accessed Jan 1, 2021).

[10] Liao, W.; Sun, X.; Kumar, A.; Sun, H.; Ma, H. Hydration of Binary Portland Cement Blends Containing Silica Fume: A Decoupling Method to Estimate Degrees of Hydration and Pozzolanic Reaction. *Front. Mater.* **2019**, *6*, 78.

[11] Yu, R.; Spiesz, P.; Brouwers, H. J. H. Effect of Nano-Silica on the Hydration and Microstructure Development of Ultra-High Performance Concrete (UHPC) with a Low Binder Amount. *Constr. Build. Mater.* **2014**, *65*, 140–150.

[12] Tafraoui, A.; Escadeillas, G.; Lebaili, S.; Vidal, T. Metakaolin in the Formulation of UHPC. *Constr. Build. Mater.* **2009**, *23* (2), 669–674.

[13] Berry, M.; Snidarich, R.; Wood, C. *Development of Non-Proprietary Ultra-High Performance Concrete: Final Report*; Montana. Dept. of Transportation. Research Programs: Bozeman, 2017.

[14] Mehta, P.; Monteiro, P. *Concrete: Microstructure, Properties and Materials*; Mc-Graw Hill: New York, 2006.

[15] Jennings, H. M.; Bullard, J. W.; Thomas, J. J.; Andrade, J. E.; Chen, J. J.; Scherer, G. W. Characterization and Modeling of Pores and Surfaces in Cement Paste. *J. Adv. Concr. Technol.* **2008**, *6* (1), 5–29.

[16] Paiva, H.; Silva, A. S.; Velosa, A.; Cachim, P.; Ferreira, V. M. Microstructure and Hardened State Properties on Pozzolan-Containing Concrete. *Constr. Build. Mater.* **2017**, *140*, 374–384.

[17] Funk, J. E.; Dinger, D. R. *Predictive Process Control of Crowded Particulate Suspensions: Applied to Ceramic Manufacturing*; Springer Science & Business Media, 2013.

[18] de Larrard, F. *Concrete Mixture Proportioning A Scientific Approach*, 1st ed.; E & FN Spon: London, 1999.

[19] Ng, P. L.; Kwan, A. K. H.; Li, L. G. Packing and Film Thickness Theories for the Mix Design of High-Performance Concrete. *J. Zhejiang Univ. Sci. A* **2016**, *17* (10), 759–781.

[20] Chen, Y.; Matalkah, F.; Soroushian, P.; Weerasiri, R.; Balachandra, A. Optimization of Ultra-High Performance Concrete, Quantification of Characteristic Features. *Cogent Eng.* **2019**, *6* (1).

[21] Ghafari, E.; Bandarabadi, M.; Costa, H.; Júlio, E. Prediction of Fresh and Hardened State Properties of UHPC: Comparative Study of Statistical Mixture Design and an Artificial Neural Network Model. *J. Mater. Civ. Eng.* **2015**, *27* (11), 1–11.

[22] Childs, C. M.; Washburn, N. R. Embedding Domain Knowledge for Machine Learning of Complex Material Systems. *MRS Commun.* **2019**, *9* (2), 1–15.

[23] Menon, A.; Gupta, C.; Perkins, K. M.; DeCost, B. L.; Budwal, N.; Rios, R. T.; Zhang, K.; Póczos, B.; Washburn, N. R. 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.

[24] Menon, A.; Childs, C. M.; Poczós, B.; Washburn, N. R.; Kurtis, K. E. Molecular Engineering of Superplasticizers for Metakaolin-Portland Cement Blends with Hierarchical Machine Learning. *Adv. Theory Simulations* **2019**, *2* (4).

[25] Wille, K.; Naaman, A. E.; El-Tawil, S.; Parra-Montesinos, G. J. Ultra-High Performance Concrete and Fiber Reinforced Concrete: Achieving Strength and Ductility without Heat Curing. *Mater. Struct. Constr.* **2012**, *45* (3), 309–324.

[26] Gajda, J.; Alsamsam, E. *Engineering Mass Concrete Structures*; 2006.

[27] Jones, M. R.; Zheng, L.; Newlands, M. D. Comparison of Particle Packing Models for Proportioning Concrete Constitutents for Minimum Voids Ratio. *Mater. Struct.* **2002**, *35* (5), 301–309.

[28] Lecomte, A. The Measurement of Real and Virtual Packing Density of Soft Grains. *Mater. Struct. Constr.* **2006**, *39* (1), 63–80.

[29] Favier, A.; Zunino, F.; Katrantzis, I.; Scrivener, K. The Effect of Limestone on the Performance of Ternary Blended Cement LC3: Limestone, Calcined Clays and Cement. In *RILEM Bookseries*; Springer Netherlands, 2018; Vol. 16, pp 170–175.

[30] Kwan, A. K. H.; Li, L. G. Combined Effects of Water Film Thickness and Paste Film Thickness on Rheology of Mortar. *Mater. Struct.* **2012**, *45* (9), 1359–1374.

[31] Li, L. G.; Kwan, A. K. H. Concrete Mix Design Based on Water Film Thickness and Paste Film Thickness. *Cem. Concr. Compos.* **2013**, *39*, 33–42.

[32] Garas, V. Y.; Kurtis, K. E.; Kahn, L. F. Creep of UHPC in Tension and Compression: Effect of Thermal Treatment. *Cem. Concr. Compos.* **2012**, *34* (4), 493–502.

[33] Schachinger, I.; Hilbig, H.; Stengel, T.; Fehling, E. Effect of Curing Temperature at an Early Age on the Long-Term Strength Development of UHPC | Request PDF. In *2nd International Symposium on Ultra High Performance Concrete*; Kassel, 2008; pp 205–213.

[34] Prem, P. R.; Bharatkumar, B. H.; Murthy, A. R. Influence of Curing Regime and Steel Fibres on the Mechanical Properties of UHPC. *Mag. Concr. Res.* **2015**, *67* (18), 988–1002.

[35] Kuleshov, V.; Fenner, N.; Ermon, S. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In *35th International Conference on Machine Learning, ICML 2018*; International Machine Learning Society (IMLS), 2018; Vol. 6, pp 4369–4377.

[36] Graybeal, B. *Development of Non-Proprietary Ultra-High Performance Concrete for Use in The Highway Bridge Sector, Publication No. FHWA-HRT-13-100*; 2013.

[37] Brochu, E.; Cora, V. M.; de Freitas, N. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv:1012.2599* **2010**.

Chapter 6. Hierarchical Machine Learning of LC3 Cements: Multi-Objective Optimization of Rheology, Strength, and Sustainability

6.1. Introduction

Concrete is the most widely used engineering material in the world,[1] however, due to the calcining process involved in the production of clinker, approximately 6% of anthropogenic CO_2 emissions can be attributed to cement.[2] The reactive mineral phase for most concrete is ordinary Portland cement (OPC), and global production is 4.1 Gt per year.[3] To decrease the environmental impact, supplementary cementitious materials (SCMs) can act as partial replacement to OPC and commonly include pozzolanic clays, limestone, and slag.[4] Although the environmental sustainability of these materials is well documented, economic and engineering sustainability of these materials often put them at a disadvantage to OPC.

Recent research has led to the development of limestone-calcined clay cements (LC3). LC3 blends can be designed with around 50% OPC replacement while approaching the mechanical performance of OPC.[5] One advantage of the utilization of calcined clay and limestone as SCMs is the abundance and availability of natural reserves across the world, with billions of tonnes of kaolinite clay stockpiled.[5] Kaolin clays are composed of aluminum and silicon oxides which are easily calcined at relatively low temperatures, as compared to OPC, of 600-800 ºC to produce highly pozzolanic metakaolin.[6] Numerous sources of kaolin clays can be found throughout the world, having various pozzolanic activity and particle sizes which are currently an area of study.[7]

The advantages of LC3 systems, as compared to other SCMs, is the ability to produce cements with superior mechanical properties to OPC. These mechanical properties are improved through additional hydration reactions which can occur from both the introduced calcined clay and limestones. Metakaolin allows for pozzolanic reactivity which consumes portlandite over time to form the strong amorphous calcium aluminum silicate hydrate (CASH) phase. In addition, limestones are able to react with aluminum from the tricalcium aluminate (C_3A) and metakaolin to form carboaluminate hydrates.[8] Gypsum is typically introduced at <5% by weight, which retards the dissolution of C_3A which prevents flash setting and allows for the formation of stronger hydration products.[9] Finally, the small particle size of both the limestone and metakaolin influences known as the 'Filler Effect', where higher surface areas allow for a larger number of suitable surfaces for nucleation and growth.[8] However, clay and limestone behavior in suspension is sharply affected by solution concentration, such as during cement hydration. As a result of their physical and chemical characteristics, use of calcined clays produce sharp reductions in flow in the plastic pastes and concrete.[8]

Numerous studies have made attempts to elucidate the working mechanisms within LC3 systems to reduce environmental impact while improving mechanical properties and durability. However, with the advent of the 4th paradigm of science,[10] the data-guided discovery and optimization of physical and engineering processes necessitates a need for machine learning (ML). With the increasing power in computational resources and refinement of ML methodologies, the ability to transition from a human-centered to human-guided approaches in engineering systems is a core component of a grand challenge to both engineering and scientific research. To resolve the mechanical, economic, and sustainability constraints, along with reducing extensive iterative testing, utilization of ML can lead to improved material systems with lower time and data requirements.[11]

Here, a hierarchical machine learning (HML) framework is presented to establish models for LC3 workability and strength. The models are represented in terms of latent variables which are able to better generalize to the diverse compositional space found in LC3 raw materials. Particle packing and water film thickness are utilized as latent, middle layer variables to embed domain knowledge of particle size and water ratios for both the workability and strength models. The strength model is further represented to embed knowledge of strength evolution, gypsum, and clay kaolin contents. The workability model is represented in terms of admixture behavior consistent with prior research into HML.[12,13,14] The life-cycle assessment of LC3 is studied to determine the effect of composition on the global warming potential (GWP). Finally, a multi-objective optimization is performed to find an LC3 composition which maximizes strength and workability, with constraints that the GWP is less than half of OPC.

6.2. Experimental

6.2.1. Materials

Cement pastes were prepared using ordinary Portland cement (Argos), three types of metakaolin (Imerys and BASF) and four separate size limestones (Imerys). Gypsum was acquired from Sigma Aldrich. In North America, highly pure metakaolin is the most common calcined clay used in LC3, and the model is being built assuming the use of this. While there are numerous calcining options for metakaolin, major producers, such as Imerys, produce it within a narrow range of amorphous content and soluble aluminate, so processing and purity were not considered as variables in the initial model. A list of all materials utilized for LC3 mixtures is shown in Figure 6.1, and all mineral feedstocks were used as received. Additional variables include the particle size distribution of the mineral phases, limestone with D50 values ranging from 3 μm to 40 μm are included and metakaolin with D50 values of 2 μm and 10 μm are being tested. Deionized water with a resistivity of 18.2 MΩ was utilized for all experiments. Table 6.1 provides the specific surface areas (SSA), D50's, and specific gravities for all materials utilized for the optimization.

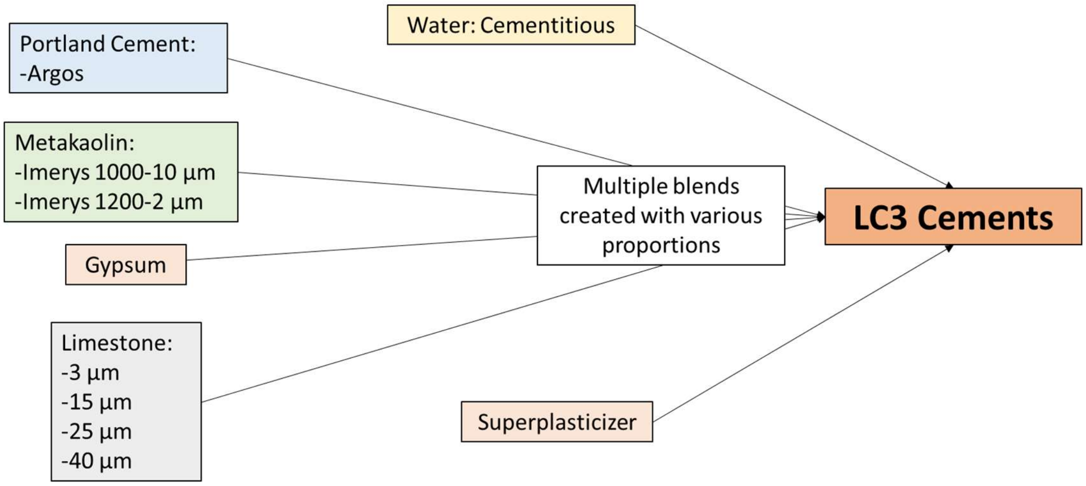


Figure 6.1. Materials used in the LC3 mixes.

Table 6.1. Specific surface area, D50, and specific gravity for all materials utilized in multi-objective analysis and testing.

Material	SSA (cm^2 g^{-1})	D50 (μm)	Specific Gravity
Cement	1140	12.2	3.15
Gypsum	15000	10	2.32
Metakaolin 1000	20000	9.46	2.5
Metakaolin 1200s	25000	4.5	2.5
Metakaolin Meta	26000	3.37	2.5
Limestone 3	4700	3.03	2.7
Limestone 15	1000	13.02	2.7
Limestone 25	800	17.65	2.7
Limestone 40	480	24.99	2.7

6.2.2. Data Collection

6.2.2.1. Workability Model Data Collection

For the first generation of the workability model, the training set included three commercial PCE superplasticizers from BASF (MG7920, MG3030, and MG7500) that were used as received following recommended doses and protocols. Due to MG7920 outperforming at all levels of concentration, the data for modelling was limited to only this superplasticizer. A total of 58 unique blends with superplasticizer concentration ranging from 0.25-1% were tested. Three metakaolin clays and four limestones, all of various D50’s and SSA’s were incorporated into the dataset.

6.2.2.2 Strength Model Data Collection

The complete dataset consisted of 97 unique blends with a total of 442 training points which included multiple curing times of each blend. Data were collected from strength assessment with our Georgia Tech collaborators (GT data), and supplemented with data provided by Dr. Karen Scrivener’s group (Data Benchmark, Top-Down, and Validation). While all datasets had various sizes of materials and w:cm ratios which were tested, the GT data primarily tested a range of OPC: Calcined Clay: Limestone ratios without gypsum addition. The Data Benchmark set primarily focused on changes to the kaolin content of the calcined clays utilized. The Top-Down set primarily focused on changes in the Calcined Clay: Limestone ratio and gypsum. Finally, the Validation set was selected as this set contained different limestone sizes than found in the previous three datasets and at a higher w:cm ratio. The ranges of compositions for each tested blend are shown below in Table 6.2.

Table 6.2. Ranges and sources of the LC3 compositions for all tested blends.

Data Source	**GT**	**Data Benchmark**	**Top-Down**	**Validation**
OPC (%)	40→55	53	53.4→54.65	54→69.7
Clay:LS Ratio	1:2→2:1	2:1	0.15→2:1	1:2→2:1
Kaolin Content (%)	95	0→95	95	46 and 62
# Different Clays	1	44	1	2
# Different Limestones	2	1	3 (Included Limestone Blends)	6 (Included Limestone Blends)
w:cm	0.416	0.50	0.40→0.459	0.53→0.55
Gypsum (%)	0	2	0.35→1.6	0.3→1
Days of Curing Tested	3,7,28	1,3,7,28,90	1,2,3,7,28,90	2,7,28
# Datapoints	54	217	108	63

6.2.3.1 Adsorption Measurements

Adsorption to mineral surfaces is an important mechanism of superplasticizer action.[15] Partitioning of the dialyzed MG 7920 in the training set onto limestone, metakaolin, and portland cement was performed by total organic carbon (TOC) analysis. Samples of 0, 1.25, 2.5, 5.0, and 10 mg admixture/g cementitious material of each superplasticizer were prepared. A reference sample for each sample was made to compare the difference in amount of carbon before and after adsorption. For each different sample, the superplasticizer solution was mixed with 5 g cementitious material at a w/cm ratio of 4 to ensure ability to collect the pore solution for testing. The mixture was immediately vibrated on a mini-vortex mixer for 3 minutes. The samples were then centrifuged for 10 min at 4400 rpm. The supernatant was collected, filtered through a 0.45 μm syringe filter, and 1mL of the pore solution was diluted to 40 mL with water, and analyzed using a combustion-based TOC (Shimadzu TOC-L).

6.2.3.2 Zeta Potential

Zeta potential was measured in aqueous solutions with a dialyzed superplasticizer concentration of 10 mg mL^{-1} using a Zeta-sizer (Malvern Instruments).

6.2.3.3. Osmotic Pressure

Osmolality of aqueous solutions of each dialyzed admixture was measured 0.1, 0.2, 0.3 and 0.4 g mL^{-1} using a vapor pressure osmometer (Wescor 5600). Three trials at each concentration were performed and the results averaged.

6.2.3.4. Polymer Intrinsic Viscosity

An Ubbelohde Viscometer (Canon Instrument Company) was used to determine the relative viscosity by taking the ratio of the time to pass through the capillary of the commercial polymer solutions to the pure water elution time. Three trials for each polymer along with the solvent elution time were performed and the results were averaged.

6.2.4. Assessment of Workability

Cement pastes were prepared utilizing a hand-held mixer (HamiltonBeach). To the mixing bowl various ratios of portland cement: metakaolin: limestone was added. Total water addition produced a 0.40 water-to-cementitious materials (w/cm) ratio paste, where portland cement, metakaolin, and limestone were considered as cementitious materials. The commercial PCE was dissolved in the mix water and added to the cement, dosed in percentage by weight of cement. The paste was mixed for 30 s at the low-speed setting and then an additional 60 s at medium. A 30 s rest period was followed with a final 60 s medium-speed mixing period. All mixes were tested immediately followed mixing.

To assess the workability of the cement pastes, mini-slump testing was performed.[16] Within 60 s after mixing, the cement pastes were added into an acrylic mini-slump cone with a bottom PAT dimension of 3.8 cm in diameter and 5.7 cm in height as shown in Figure 6.2. The cement was leveled with the backside of a spoon to ensure an even finish with the top. The cylinder was slowly and evenly lifted in a continuous motion allowing the cement pastes to slump. The change in height due to the slump was recorded along with the spread. For every mixture of cement tested, three mini-slumps were tested in succession with an average of the trials being used for analysis.

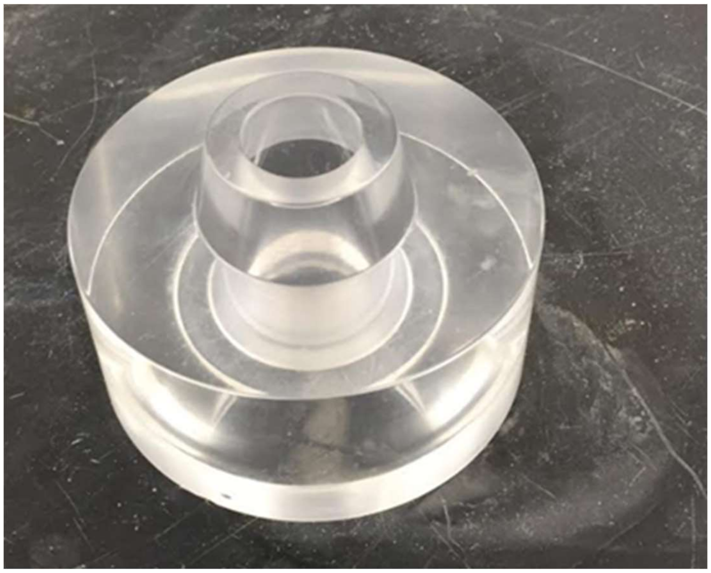


Figure 6.2. Mini-slump cone utilized for measurements of cement workability.

6.2.5. Assessment of Strength

2-inch cube specimens were prepared for compression strength testing at a constant w/cm of 0.4. First, the measured materials were dry blended for 30 seconds in a Hobart mixer to ensure homogeneity. The mix water was then stirred with the admixture prior to adding to the cementitious materials. The paste was mixed on low setting for 30 seconds, then on medium for an additional 60 seconds, and stopped for a 30 second rest period. During this time, the sides and bottom of the mixer bowl were scraped to better incorporate the paste solids and any unmixed material. The mixing regime is ended with a final 60 seconds on medium setting. Paste samples were molded in accordance with ASTM C109 and then kept in a humidity chamber at 23±2 °C and 100% humidity for 24 h. After demolding, the samples were cured in water (23±2 °C) until testing. The compressive strength was measured after 7 and 28 days of hydration for the optimized blend based on the average of six specimens. The cube and compressive strength machine are shown in Figure 6.3.

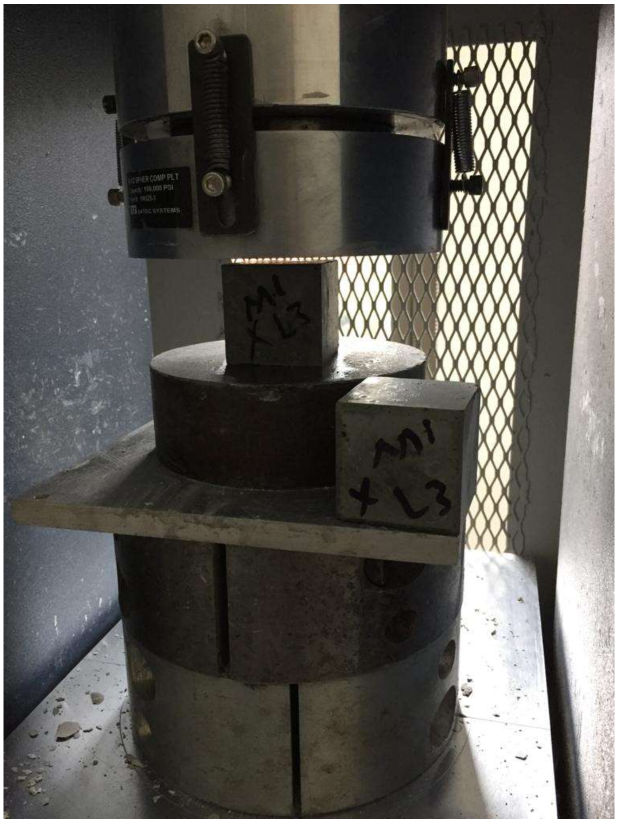


Figure 6.3. Compressive strength samples in the process of being tested.

6.2.6. Lifecycle Assessment

Research on more sustainable alternatives to OPC, are driven by the need to find reliable and durable materials that are also more environmentally friendly. Therefore, a preliminary life cycle assessment has been carried out for LC3.

Life Cycle Assessment (LCA) is an effective method to evaluate the environmental impacts of all products and processes associated to a given system. There are various LCA approaches that can be adopted depending on the analysis and the product of interest. In this case it has been chosen to follow a "cradle to gate" approach, hence considering all the components of the production process but only until the product is released to the market, hence not considering the transportation, placement, maintenance, durability, and disposal of the product outside the cement plant. This is a common approach due to the fact that cement can be part of various end-products. In this preliminary LCA the goal is to compare the GWP, expressed as kg CO_2 eq released during the production of LC3 cements and compared to the production of OPC.

The choice of the functional unit should reflect the similar function and performance that may be obtained using two products. While Portland cement is an established and regulated material, LC3 cements are in a research stage and norms that regulate their applicability are still lacking. As a result, it has been chosen to consider 1 ton of cement as functional unit. Therefore, the results that will be presented in the following sections may be considered to compare the environmental impact of products that are based on these cements, but critically considering the assumptions made.

The LCA software OpenLCA 1.8.0[17] was used to evaluate the environmental impacts of inventory elements of portland cement and LC3 cement. The geographic area considered in this study is the South East of the United States, assuming the location of the cement plant in Atlanta (GA). When possible open source available data from USLCI National Renewable Energy Laboratory[18] are preferred since related to the North American framework. The electrical energy provider is related to the SERC distribution (specific to the South East region of the United States), while the fuel mix for the combustion during the calcination process is a mix obtained from coal, gasoline, natural gas, residual fuel oil, liquefied petroleum gas, petroleum coke, middle distillates, and waste. Data for the average fuel mix for cement kiln in the US have been considered as a reference (obtained from Portland Cement Association).[19]

6.2.7. Computational Methods

6.2.7.1. Particle Packing and Water Film Thickness

The same procedure for calculating the WFT and packing index, *k*, were utilized as in Section 5.3.1.2 and Section 5.3.1.3. The WFT and *K* were utilized as inputs for both the strength and workability models.

6.2.7.2. Workability Model

The schematic representation of the cement workability model is shown in Figure 6.4 for LC3, for which workability remains a key technological hurdle facing adoption. Workability is assessed through mini-slump measurements. In HML, the effects of compositional variables on workability are assumed to be mediated by a diversity of latent variables. In LC3, the input variables are the amounts of OPC, limestone, and calcined clay, the water: cementitious ratio, and the type and dose of superplasticizer. The latent variables, represented in the middle layer, represent underlying forces that drive system responses, as explored in modelled in prior research.[13,14,20] In the absence of chemical admixtures, particle-particle interactions are assumed to drive these responses, but superplasticizers can exert an effect via both particle and solution forces. The approach in this research is to estimate the constituents of the middle layer experimentally or computationally.

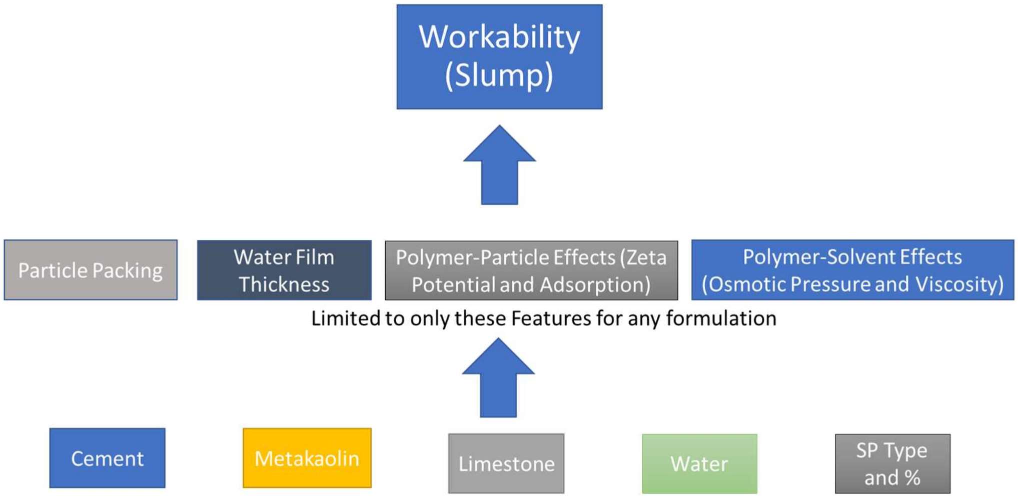

Workability (Slump)
Particle Packing
Water Film Thickness
Polymer-Particle Effects (Zeta Potential and Adsorption)
Polymer-Solvent Effects (Osmotic Pressure and Viscosity)
Limited to only these Features for any formulation
Cement
Metakaolin
Limestone
Water
SP Type and %

Figure 6.4. Schematic structure of the HML model for workability. The input variables on the bottom layer represent the composition of the mineral and admixture components as well as water: cementitious ratio although this is currently held constant at 0.40.

Modeling of the mini-slump (top layer) as a function of compositional variables (bottom layer) and physicochemical variables (middle layer) was performed using the random forest algorithm. Random forests are made up of an ensemble of decision trees, and for regression purposes, the output of all these trees have an average taken to produce a single best-fit regression for the entire collection of trees. In order to establish a best-fit to unseen (test) data, regression models have their parameters learned on training data through cross-validation (CV), a 5-fold CV was utilized for all hyperparameter turning. A random forest model was performed on both the bottom and middle layer variables representing LC3 compositions. The bottom layer is composed of the mass fractions of the LC3 composition and superplasticizer concentration, while the middle layer represents the composition in terms of latent variables that capture mechanistic domain knowledge of workability. Initially, the dataset was composed of three separate PCE architectures, however, due to the large disparity between the top-performing MG 7920 and the remaining 2 architectures, MG 7920 was utilized to build a model for workability and multi-objective optimization with this specific PCE. Due to the small size dataset containing 58 individual slump values the middle layer containing embedded domain knowledge should outperform bottom layer performance among proper selection of physical knowledge. The workability model was trained with the output and each input standardized utilizing the StandardScalar methodology from scikit-learn and Random Forest as performed utilizing RandomForestRegressor in scikit-learn.[21] The data was randomly split into 90% training and 10% test data for validation on unseen values.

6.2.7.3. Strength Model

The schematic representation of the LC3 strength model is shown in Figure 6.5. Strength is quantified through compressive strength measurements over the course of multiple time points. In LC3, and similar to the workability model, the input variables are the amounts of OPC, limestone, and calcined clay, and the water: cementitious ratio. The latent variables, represented in the middle layer, represent particle packing and water film thickness as a way to encode the effects of surface area and particle size. Kaolin content to allow the model to learn the effects of the primary reactive phase and differentiate between the effects of pure and impure clays. The log of the curing time is calculated to help the model in learning the plateauing of cement strength over time and common ratio information for cement, gypsum, and water are elevated from the bottom layer in order to help capture effects which may not be fully represented in the embedded domain knowledge.

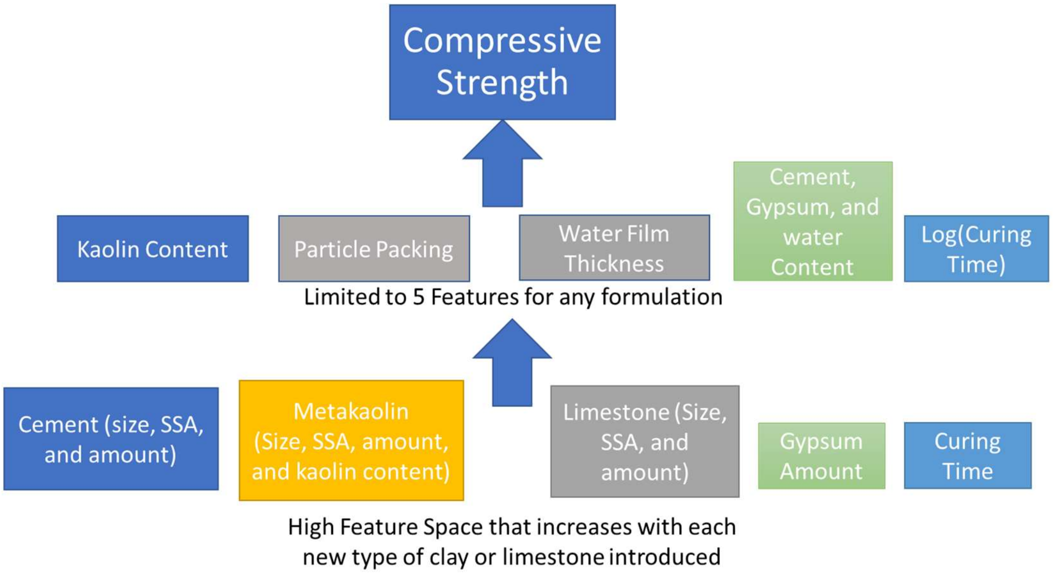

Compressive Strength
Kaolin Content
Particle Packing
Water Film Thickness
Cement, Gypsum, and water Content
Log(Curing Time)
Limited to 5 Features for any formulation
Cement (size, SSA, and amount)
Metakaolin (Size, SSA, amount, and kaolin content)
Limestone (Size, SSA, and amount)
Gypsum Amount
Curing Time
High Feature Space that increases with each new type of clay or limestone introduced

Figure 6.5. Schematic structure of the HML model for strength. The input variables on the bottom layer represent the composition, size, and surface area of the mineral components as well as water: cementitious ratio and curing time.

A gaussian process regression (GPR) was utilized as the ML technique in the prediction of LC3 compressive strength. GPR is a Bayesian methodology which can be utilized to learn both the predicted mean and posterior probability for the expected range of error at each prediction.[22] GPR utilizes a metric of distance known as a covariance function (or kernel) to learn the distribution of functions over the training data.[22] From a prior establishment of mean and covariance function, GPR finds a posterior distribution based on the training data. Instead of utilizing a cross-validation approach as is common in many ML methodologies, GPR updates hyperparameters in the covariance function through an optimization procedure on the log marginal likelihood.[23] The strength model was trained with the output and each input standardized utilizing the StandardScalar methodology from scikit-learn and GPR was performed utilizing GaussianProcessRegressor in scikit-learn.[21] The data were randomly split into 70% train and 30% test sets.

6.2.7.4. Multi-objective Optimization

A multi-objective evolutionary algorithm known as the nondominated sorting genetic algorithm (NSGA-II) was utilized. This genetic algorithm randomly samples a set of points in the compositional space and outputs the predicted strength and slumps for those points. Crossover and mutation occur to create a new generation of points and a sorting algorithm selects the points which move towards a non-dominated (points where no better objectives exist) pareto front solutions.[24] A python package, jMetalPy, was utilized to perform the multi-objective analysis.[25]

6.3. Results and Discussion

6.3.1. Workability Results

6.3.1.1. Adsorption

The most common mechanism for PCE function is through exhibiting steric interaction after adsorption onto the particle surface.[26] For LC3, adsorption was modeled through the Freundlich equation, an empirical model of representing heterogenous surfaces:[27]

$$\theta = K{C_i}^n \qquad \text{Eq. 6.1}$$

Here θ is the amount of adsorbed polymer/g material, C_i is the initial added polymer/g material, and K and n are empirical parameters. A curve-fitting routine from NumPy in Python was utilized in order to fit the optimal parameters to model the adsorption isotherm. With adsorption data, a Freundlich model was developed for each polymer-cementitious material pair and overall adsorption was modeled according to equation below:

$$\theta_{Total} = \left(\frac{MK}{CM}\right) * \theta_{MK} + \left(\frac{PC}{CM}\right) * \theta_{PC} + \left(\frac{LS}{CM}\right) * \theta_{LS} \qquad \text{Eq. 6.2}$$

where θ_{Total} is the total amount of expected adsorbed polymer, $\frac{MK}{CM}$ is the ratio of metakaolin to total cementitious material, $\frac{PC}{CM}$ is the ratio of portland cement to total cementitious material and $\frac{LS}{CM}$ is the ratio of limestone to total cementitious material (all expressed in terms of mass, not surface area- an individual isotherm was measured for each size clay and limestone tested). The Freundlich curves for the PCE's in the workability model are shown below in Figure 6.6:

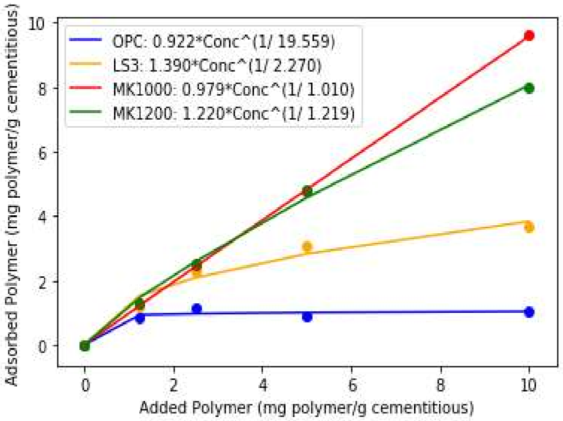


Figure 6.6. Adsorption isotherm for MG7920. The isotherm was measured on the calcined clay, portland cement, and the smallest size limestone. All limestones with average size greater than 3 µm did not exhibit adsorption of the polymer.

The total amount of adsorbed polymer is utilized as the input to the model as a singular value representation for steric interactions.

6.3.1.2. Zeta Potential

The measurement of the MG7920 zeta potential was found to be -17.35 mV. While PCE's primary mechanism is generally associated with steric repulsion, electrosteric interactions can also be induced through changes of the surface charge of cementitious grains via adsorption.[28] Here, the effect of adsorbed admixture on particle surface was assumed to take the form:

$$\zeta_{pol} = c_o\theta\, \zeta_{max} \qquad \text{Eq. 6.3}$$

where, as the polymer concentration would approach complete coverage of the material surface, the zeta potential would approach the maximum value as determined through the pure polymer zeta. The calculated value of zeta is introduced to the model as a representation for electrostatic effects.

6.3.1.3. Viscosity

The plots and 2nd order fits for concentration vs relative viscosity are shown for the MG7920 are shown in Figure 6.7. Viscous forces induced by non-absorbed polymer in the interstitial spaces of adjacent cement particles has been shown to provide a lubricating effect to improve workability.[29]

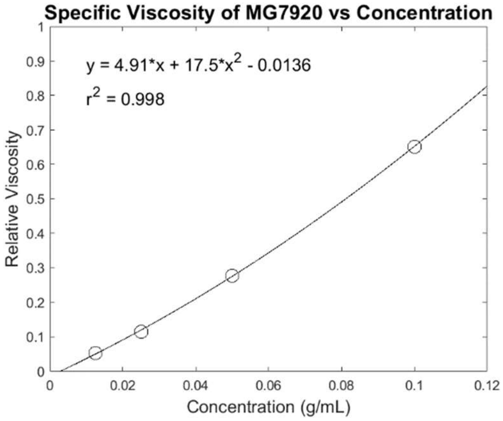


Figure 6.7. Plot of relative viscosity over a concentration range for MG7920.

The relative viscosity, as a function of admixture concentration, was utilized as the input for the ML model.

6.3.1.4. Osmotic Pressure

To model changes in osmotic pressure due to superplasticizer dissolved in the pore solution, the plots of concentration vs osmolality/conc were fit with a linear trend in order to capture the A_1 and A_2 virial parameters and allow for the calculation of osmotic pressures depending on the admixture concentration in the pore solution.[30] The plot for MG7920 is shown below in Figure 6.8 with the associated equations and R^2.

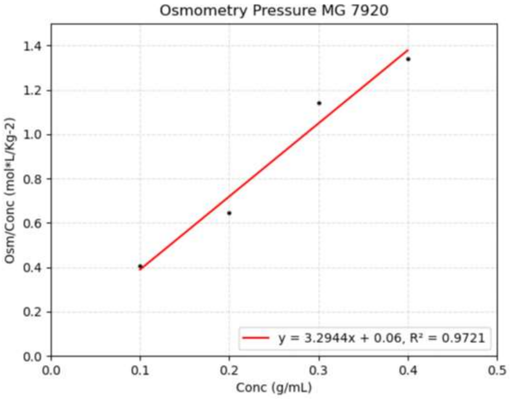


Figure 6.8. Osmotic pressure curves for MG7920. The plot depicts the ratio of osmotic pressure over the polymer concentration against the polymer concentration, and a linear fit suggests the first two virial coefficients provide an adequate model of activity.

The osmotic pressure, as a function of admixture concentration, was utilized as the input for the ML model.

6.3.1.5. Random Forest Model for Workability

Results comparing the training and test set data for both the bottom and middle layers are shown in Figure 6.9 and Figure 6.10 and statistical results shown in Table 6.3. For the bottom layer, the optimal parameters were found to be: [Number of Estimators=600, Bootstrapping=True, Number of features to split= Number of samples, Maximum depth of tree=80, minimum samples in a leaf node=1, minimum samples required for a split=5]. For the middle layer, the optimal parameters were found to be: [Number of Estimators=200, Bootstrapping=True, Number of features to split= sqrt(Number of samples), Maximum depth of tree=50, minimum samples in a leaf node=1, minimum samples required for a split=2].

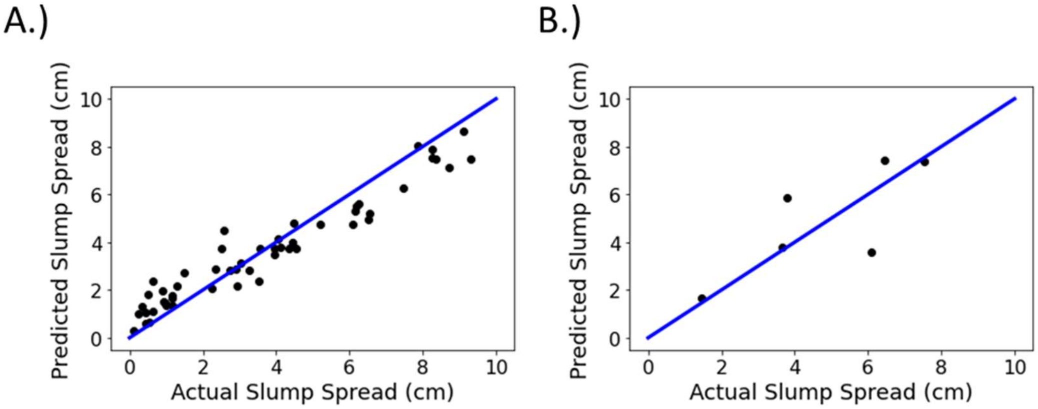


Figure 6.9. Bottom layer plots for random forest model of workability with A.) training set and B.) test set.

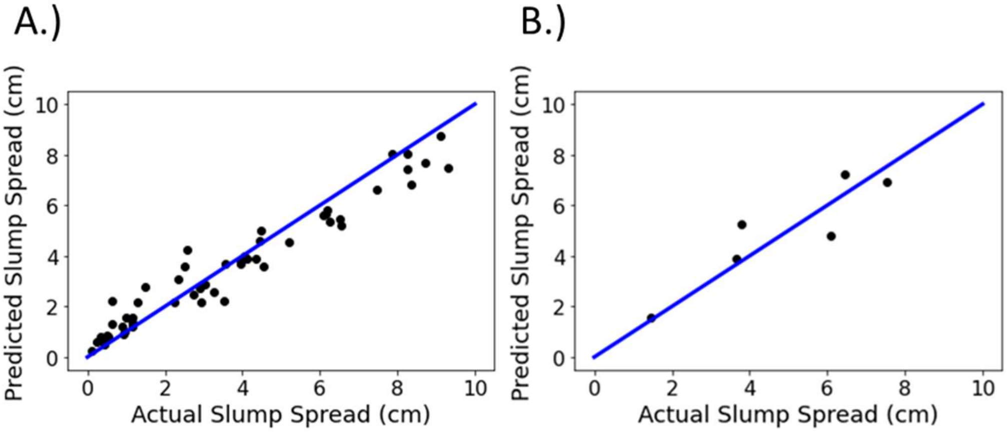


Figure 6.10. Middle layer plots for random forest model of workability with A.) training set and B.) test set.

Table 6.3. Statistics representing the bottom and middle layers for both the test and training sets.

	R^2	MSE	RMSE (cm)
Training Set Bottom	0.90	0.73	0.85 cm
Training Set Middle	0.93	0.54	0.73 cm
Test Set Bottom	0.55	1.92	1.39 cm
Test Set Middle	0.81	0.82	0.91 cm

Measured slump spread differences vary between 0 and 9.3 cm. (Note that 0 cm is considered as 3.80 cm, the width of the slump cone). The middle layer outperformed the bottom layer in terms of R^2 and MSE for both the training and validation sets.

Finally, relative importance of features can be derived from the random forest models based on parameterization by the bottom layer and middle layer as shown in Figure 6.11. When parameterized by the bottom layer, increases in the cement ratio was most strongly correlated with increases in workability, consistent with the expectation that reducing metakaolin and limestone content will increase the mini-slump. When parameterized by the middle layer, the water film thickness was the strongest variable in determining the workability. These variables are in fact related since water film thickness decreases with decreasing average particle size.

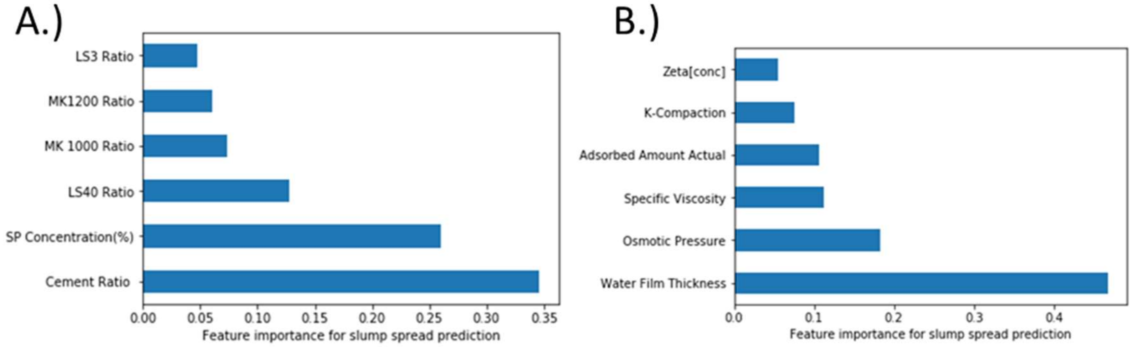


Figure 6.11. Feature importance from random forest models of workability as a function of A.) bottom layer variables and B.) middle layer variables.

6.3.2. Strength Results

In developing the strength model, two rounds of training were performed. The first withheld the 'Validation set' shown in Table 6.2 and was trained on the remaining three datasets. As this set contained different sizes of limestone compared to the first three datasets and a higher w/cm ratio, a robust model would be able to generalize to this data. Second, after generalization was shown to achieve adequate results, all datasets were combined in order to train a full model. For all models, a combined summation kernel of the radial basis function and rational quadratic with an added noise were found to minimize error on the test sets and utilized for all analysis.

The first round of training compared performance between the middle and bottom layers of the HML model and compared their generalization capability to the 'validation set'. First, a linear regression was performed as a baseline model. The bottom layer results are provided in Figure 6.12, while the middle layer results are provided in Figure 6.13. Statistics comparing the bottom and middle layers are shown in Table 6.4.

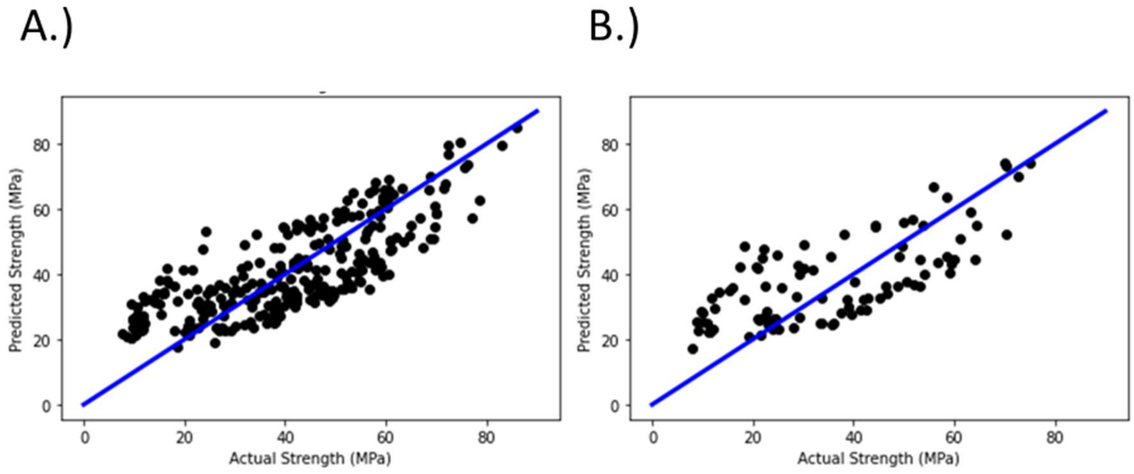


Figure 6.12. Linear regression results performed using the compositional (bottom) layer features for the A.) training and B.) test sets.

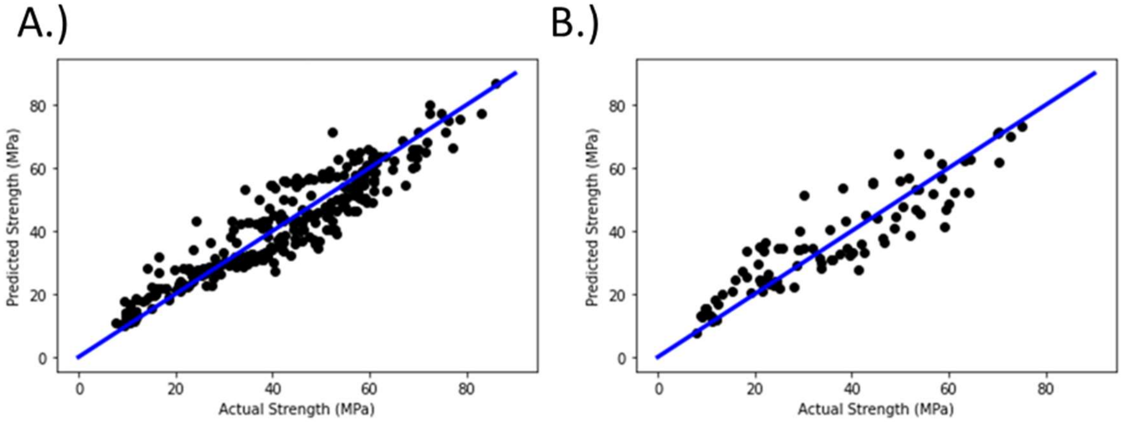


Figure 6.13. Linear regression results performed using the middle layer features for the A.) training and B.) test sets.

Table 6.4. Statistics representing the bottom and middle layers for both the test and training sets through an ordinary linear regression.

	R^2	RMSE (MPa)
Training Set Bottom	0.60	11.29 MPa
Training Set Middle	0.87	6.50 MPa
Test Set Bottom	0.49	12.81 MPa
Test Set Middle	0.83	7.46 MPa

It is clearly shown that an ordinary linear regression on the middle results in RMSE's around half of that as the bottom layer with less features. This is a methodology to show that embedding domain knowledge achieves the goal of simplifying the response surface from a complex compositional space to a more constrained and generalizable middle layer.

Second, a more robust GPR was performed to acquire an accurate model for determining compressive strength. The bottom layer results are provided in Figure 6.14, while the middle layer results are provided in Figure 6.15. Statistics comparing the bottom and middle layers are shown in Table 6.5.

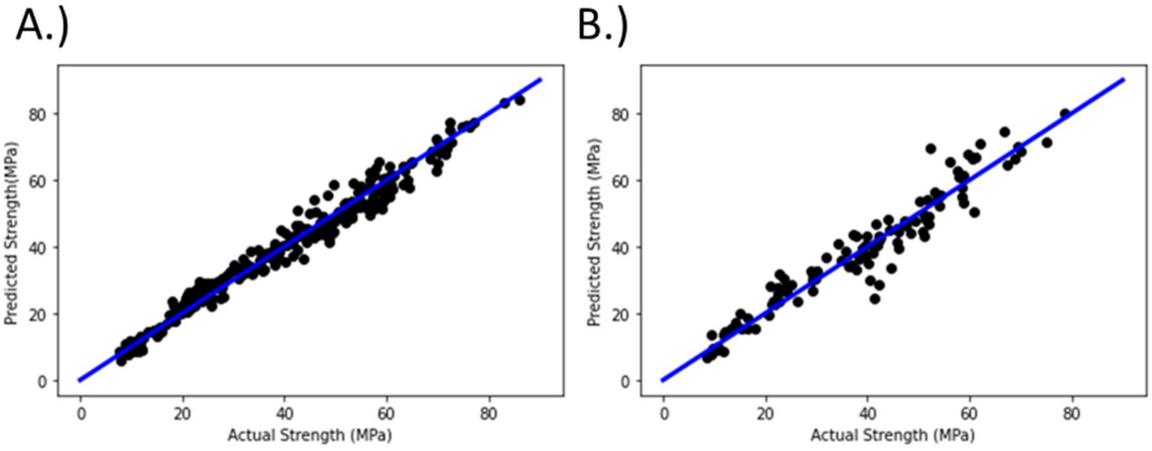


Figure 6.14. GPR results performed using the compositional (bottom) layer features for the A.) training and B.) test sets.

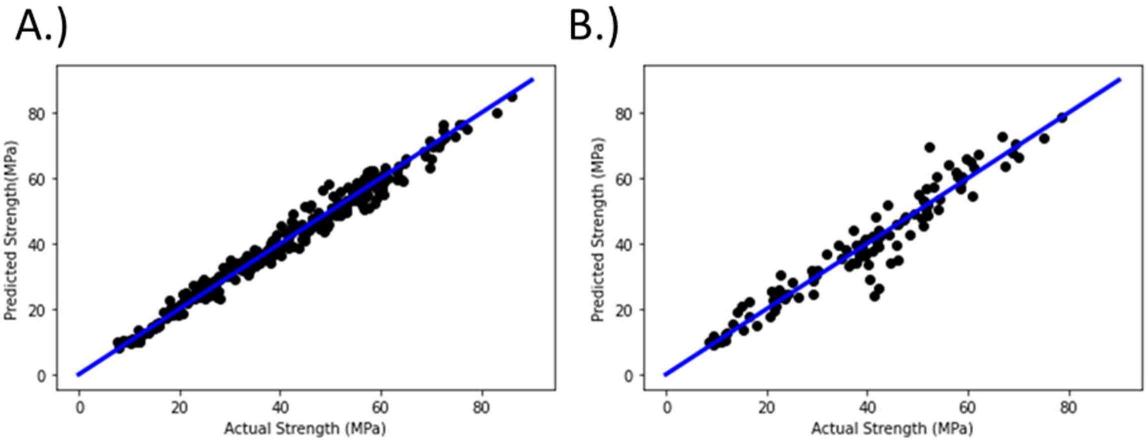


Figure 6.15. GPR results performed using the middle layer features for the A.) training and B.) test sets.

Table 6.5. Statistics representing the bottom and middle layers for both the test and training sets using a GPR.

	R^2	**RMSE (MPa)**
Training Set Bottom	0.97	3.11 MPa
Training Set Middle	0.98	2.59 MPa
Test Set Bottom	0.92	4.80 MPa
Test Set Middle	0.81	4.65 MPa

A GPR was able to outperform the linear regression in terms of RMSE due to the higher robustness of the model. However, it is still clear that middle layer results are able to outperform bottom layer results.

Finally, a comparison of how well the bottom and middle layers were able to generalize to the unseen ‘validation set’ with our GPR are shown in Figure 6.16 and statistics provided in Table 6.6.

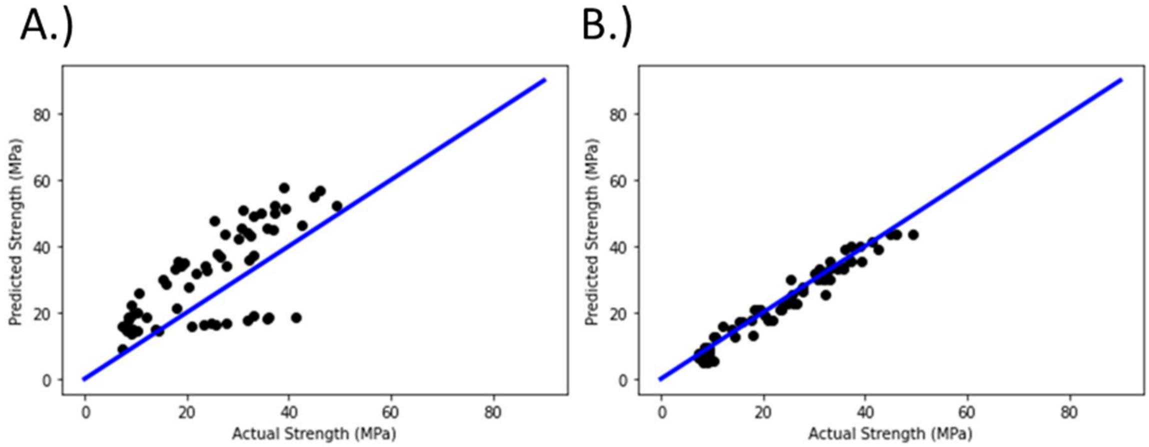


Figure 6.16. GPR generalization predictions performed using the A.) bottom layer features and B.) middle layer features on the validation set.

Table 6.6. Statistics representing the bottom and middle layers on the validation set.

	R^2	RMSE (MPa)
Validation Set Bottom	-0.02	11.74 MPa
Validation Set Middle	0.95	2.54 MPa

After showing that utilization of the middle layer provides better results in terms of R^2 and RMSE, along with much improved generalization capability by showing a drop of over 9 MPa in the RMSE as compared to the bottom layer, all four datasets were combined and retrained utilizing the middle layer with a GPR. The results are shown in Figure 6.17 and Table 6.7.

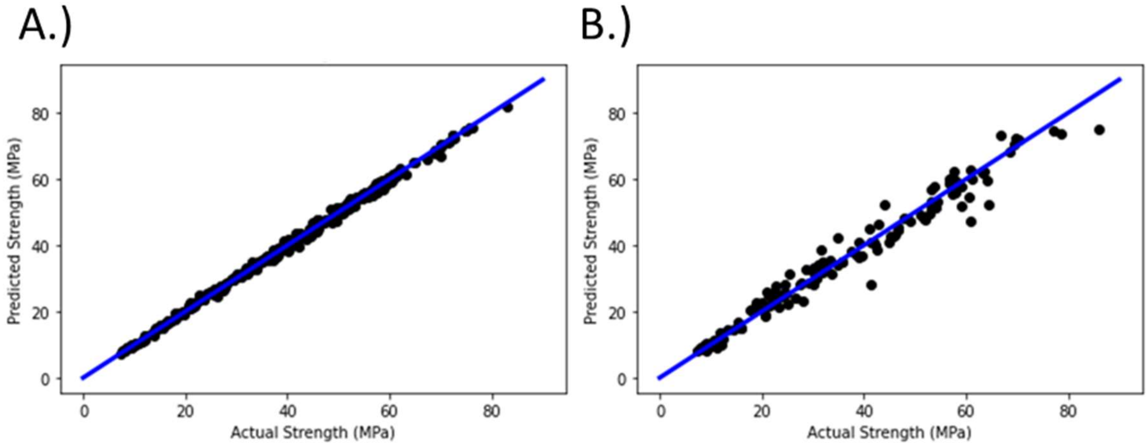


Figure 6.17. GPR results performed using the middle layer features for the A.) training and B.) test sets on all four combined datasets.

Table 6.7. Statistics representing middle layer results for both the test and training sets utilizing a GPR on all four combined datasets.

	R^2	**RMSE (MPa)**
Training Set	1.00	0.88 MPa
Test Set	0.97	3.50 MPa

With the additional data included, there was a continued decrease in RMSE and increase in R^2 values compared to only training on the three datasets. With this improved model, GPR can be utilized to visualize the effects changes in the compositional space will have on compressive strength. Figure 6.18 predicts the effect of changes in strength over varying gypsum contents have in the predicted compressive strength.

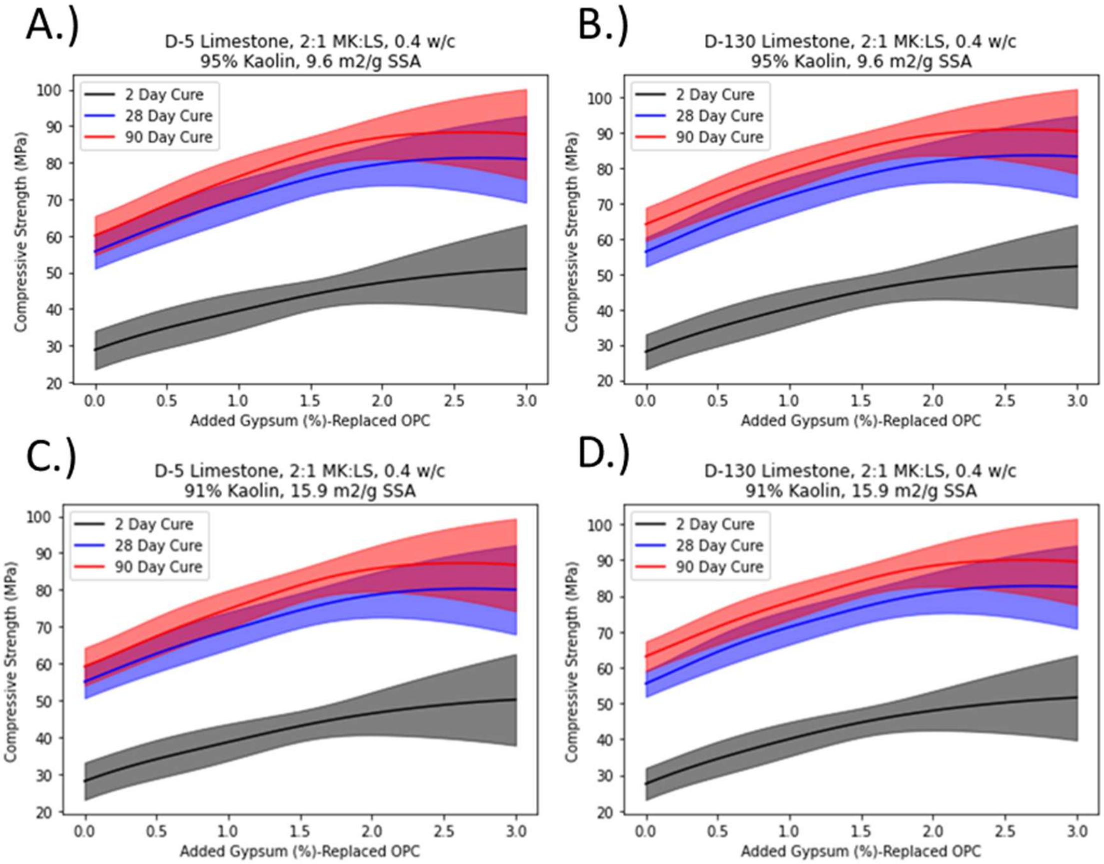


Figure 6.18. Changes in the GPR predicted compressive strength based on limestone size, kaolin content and gypsum addition at three separate curing times. The dark lines represent the predicted mean by GPR while the associated distributions represent the expected standard deviation of error in the prediction.

It is observed that the gypsum ratio steadily increases as the concentration increases until a plateau begins around 2.0% addition, but with growing uncertainty, as no tested blends had gypsum added beyond 2.0%. Surface area and size of limestone appear to have only minor effects at these compositional levels, with only a slight noticeable impact of higher 2-day strengths for the smaller limestone size blends, which is consistent with prior studies into the filler effect.[31] A second tested compositional variation was varying the w/cm ratio and is shown in Figure 6.19.

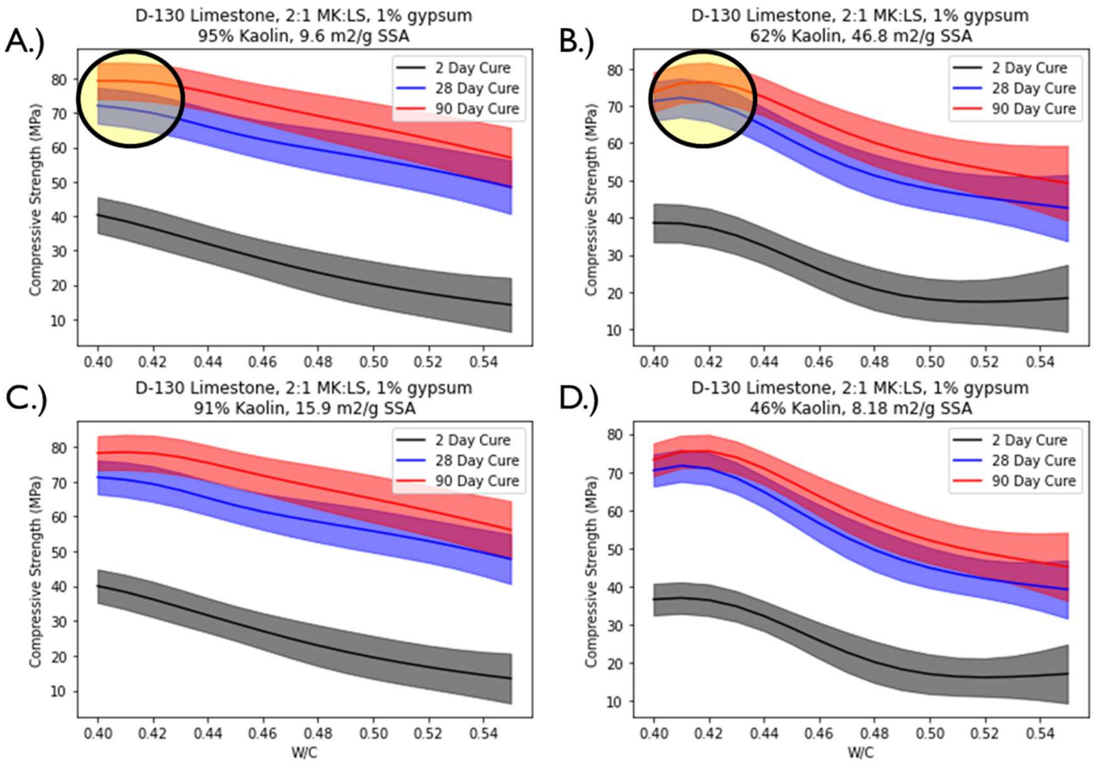


Figure 6.19. Changes in the GPR predicted compressive strength based on limestone size, kaolin content, and w/cm ratio at three separate curing times. The dark lines represent the predicted mean by GPR while the associated distributions represent the expected standard deviation of error in the prediction.

As expected, there is a decrease in strength predicted at higher water ratios. However, there appears to be an initial plateau of strength at low w:cm values before decreasing monotonically, particularly at 90-day strength as indicated by the yellow highlighted circles. These 2:1 blends may indicate a need for higher water amounts for complete hydration of the metakaolin and indicates that increasing w:cm ratio does not result monotonic loss in strength. It can also be noted that at these low w: cm ratios, that impure clays approach strengths on par with pure metakaolin clays. This result can also be viewed through a middle layer perspective as shown in Figure 6.20.

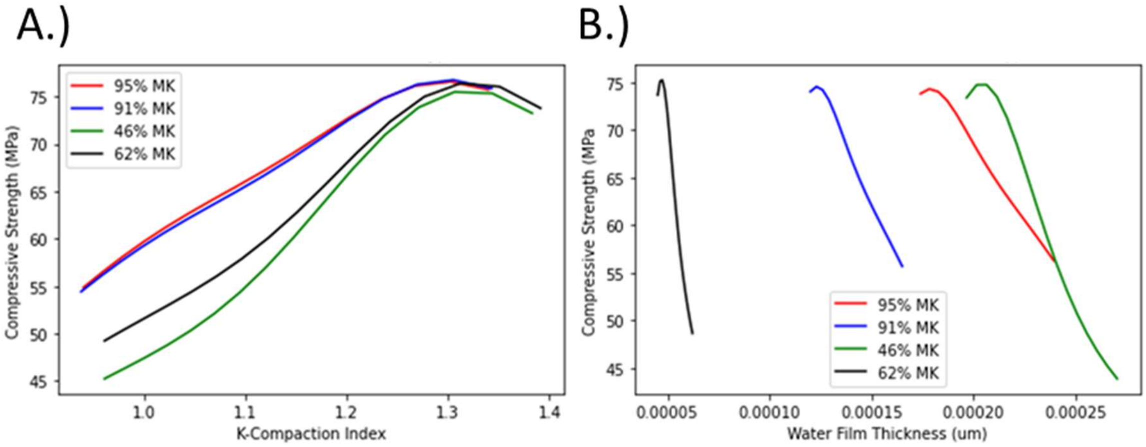
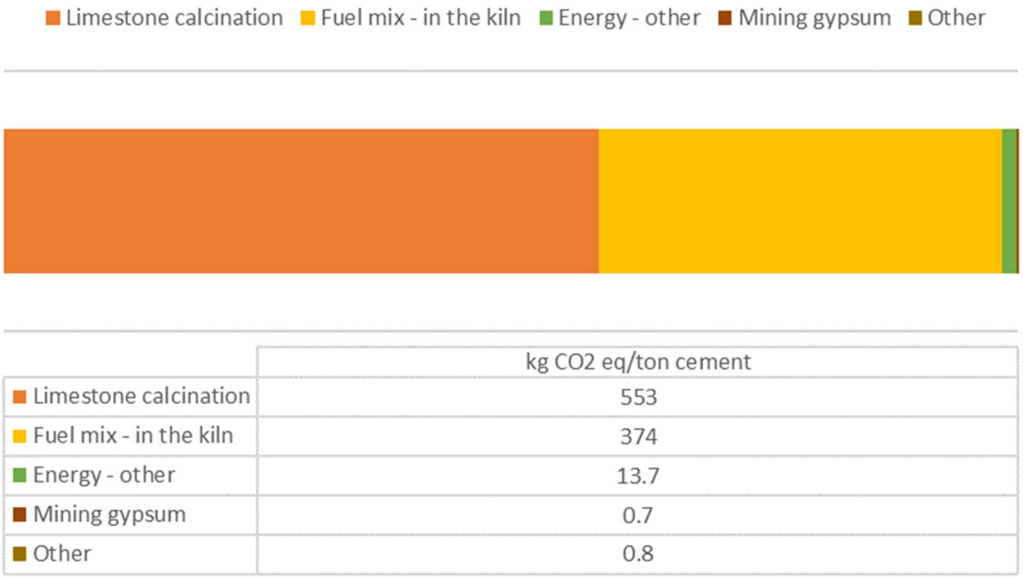


Figure 6.20. Predicted compressive strengths of various kaolin content clays at various A.) particle packing densities and B.) water film thickness levels.

It can be seen above that when the particle packing reaches a certain level high value of compaction, there is little difference in the effect of low and high kaolin content clays.

6.3.3. Lifecycle Assessment Results

As previously mentioned, the GWP of OPC has been used as a benchmark for this study. Figure 6.21 presents the GWP in kg CO_2 eq. related to OPC production. In total 942.21 kg CO_2 eq. per ton of OPC are associated with OPC production, with the calcination of limestone as primary contributor followed by the CO_2 emissions directly related to the fuel mix used.



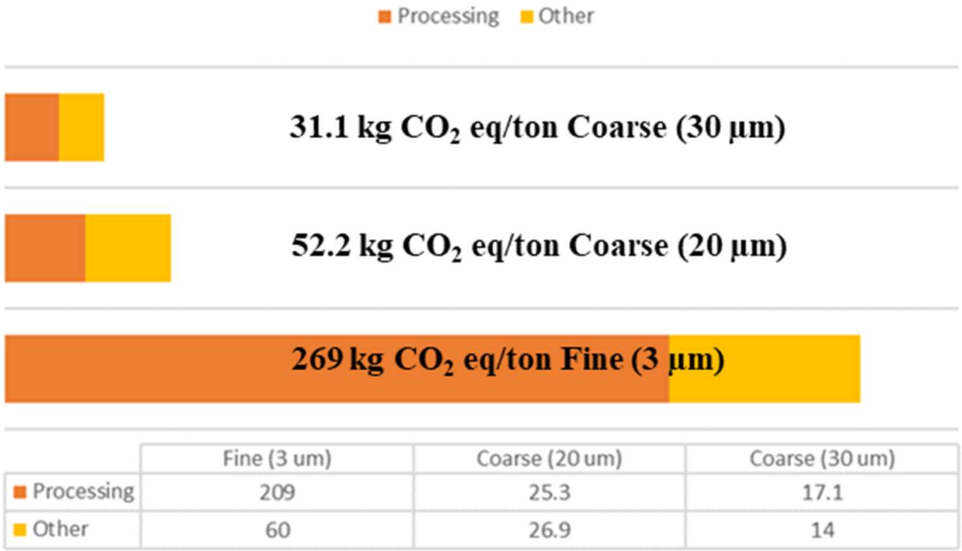
	kg CO2 eq/ton cement
🟧 Limestone calcination	553
🟨 Fuel mix - in the kiln	374
🟩 Energy - other	13.7
🟫 Mining gypsum	0.7
🟫 Other	0.8

Figure 6.21. GWP potential (kg CO2 eq. per ton) of OPC production.

A typical LC3 consists of 50% OPC clinker, 5% calcium sulfate source (anhydrite or gypsum), 30% metakaolin and 15% limestone by mass. LC3 gives comparable mechanical and durability properties to OPC with the additional advantage of a reduction of ~50% in wt. of clinker compared to OPC.

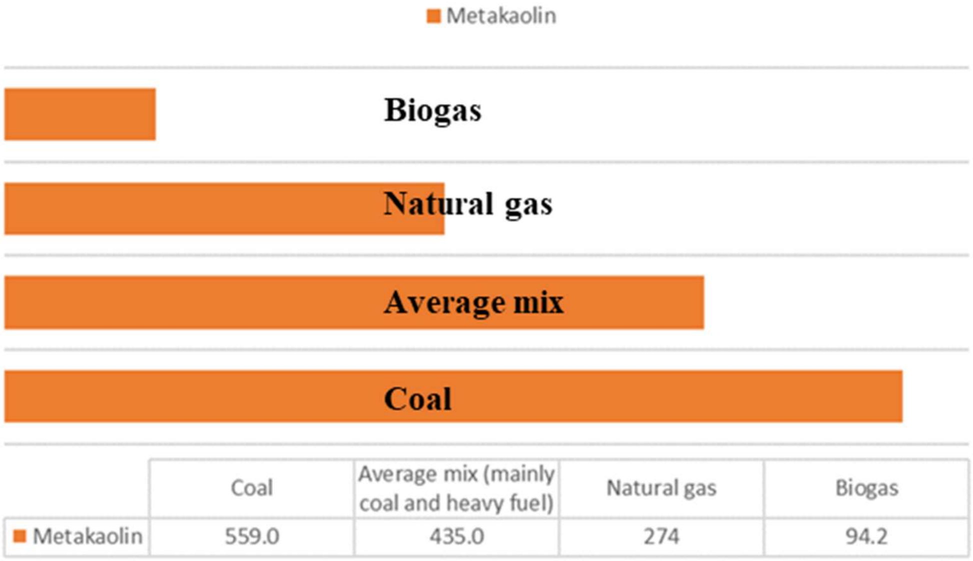
To better assess the GWP of LC3, we analyzed our material feedstock such as limestone and metakaolin. Figure 6.22 and Figure 6.23 summarize the GWP of limestone and calcined clay when different production scenarios are considered. The variable considered are particle size for limestone and the fuel mix for clay calcination. A reduction of ~80% GWP is observed for coarse limestone (30 μm or 20 μm) compared to fine (3 μm) limestone (see Figure 6.22). This difference is due to the additional processing needed to produce a finer limestone.

For the clay calcination (see Figure 6.23), depending on the choice of fuel, a potential reduction of up to ~80% in CO_2 eq. might be achieved when coal is replaced with biogas. In other words, considering the high mass percentage of metakaolin (30%) in a typical LC^3 blend, any significant reduction in CO_2 emissions from metakaolin production might result in a considerable GWP saving for the overall LC3 mix. However, for this study, an average fuel mix (mainly based on heavy fuel and coal) has been considered to assure consistency with the OPC results.



	Fine (3 um)	Coarse (20 um)	Coarse (30 um)
🟧 Processing	209	25.3	17.1
🟨 Other	60	26.9	14

Figure 6.22. GWP potential (kg CO2 eq. per ton) of limestone production.



	Coal	Average mix (mainly coal and heavy fuel)	Natural gas	Biogas
■ Metakaolin	559.0	435.0	274	94.2

Figure 6.23. GWP potential (kg CO2 eq. per ton) of calcined clay production.

The raw material analysis was followed by the investigation of several LC3 mix design with varying mass percentages of OPC, calcined clay, and limestone (see Table 6.8). LC3 blend "55:30:15" refers to 55% OPC, 30% metakaolin, and 15% limestone mass percentages.

Table 6.8. LC3 blends investigated.

Material (wt. % of binder)	55:30:15	55:15:30	50:25:25	45:25:30	45:20:35	40:20:40
OPC	55	55	50	45	45	40
Metakaolin	30	15	25	25	20	20
Limestone	15	30	25	30	35	40

The GWP of the different LC3 blends introduced in Table 6.8, is given in Figure 6.24. The analysis is based on cement paste, hence not including any aggregate. The variability shown by the error bars represents the impact of limestone particle size and the red bar indicates the project target of 450 kg CO_2 eq. per ton of cement as this is under 50% the GWP of OPC. It can be deduced that embodied energy ≤ 450 kg CO_2 eq./ ton cement can be achieved with all LC3 formulations examined. Increasing the limestone content in exchange for OPC or metakaolin contents seems as beneficial in further reducing the GWP.

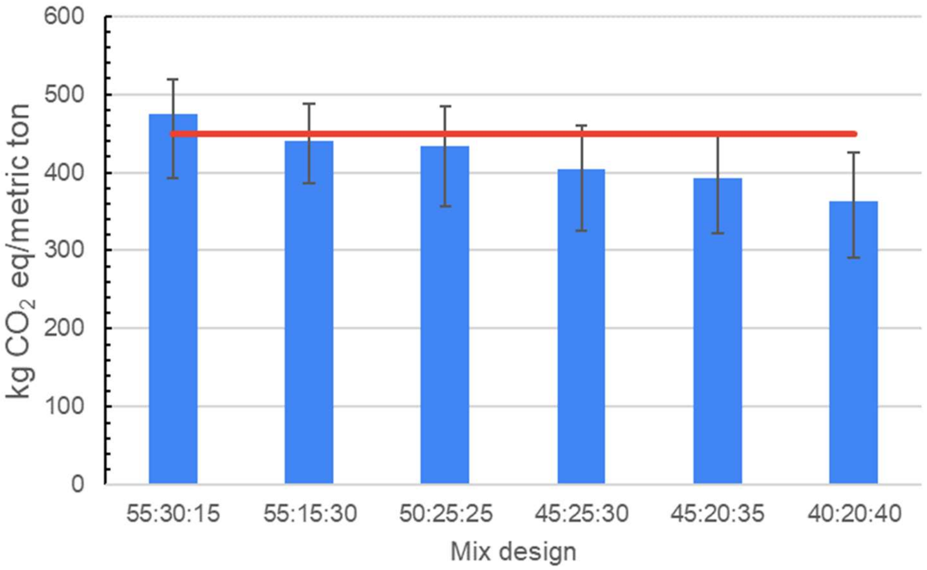


Figure 6.24. GWP potential (kg CO2 eq. per ton) of different LC3 blends investigated in laboratory. The red line indicates 450 kg CO2 eq. per ton, or roughly 50% of OPC GWP.

To aid designing LC3 for GWP considerations, we have ranked the OPC, limestone, and calcined clay in terms of their contributions to GWP, as shown in Figure 6.25. OPC production is the primary contributor to GWP with ≥ 95%.

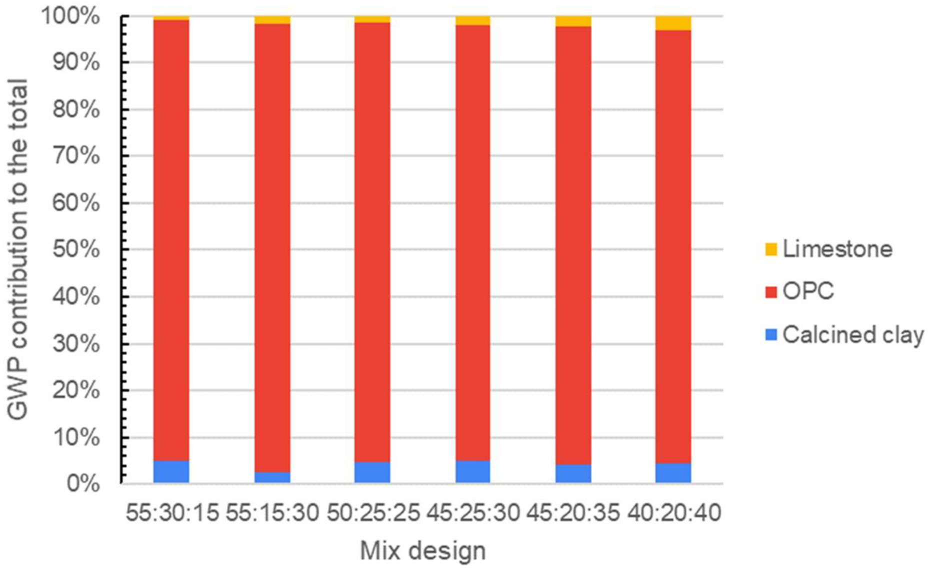


Figure 6.25. Relative GWP contributions of OPC, limestone and calcined clay for different LC3 blends.

Finally, Figure 6.26 shows a comparison of the average GWP of an alternative cement (Calcium Sulfoaluminate, CSA), LC3 and OPC. It can be seen that both CSA and LC^3 display considerable savings in terms of kg CO_2 eq when compared to OPC. From this analysis CSA cements can be produced with a GWP that is between 25-35% lower than OPC, while for LC^3 cement savings are higher than 50% in kg of CO_2 eq.

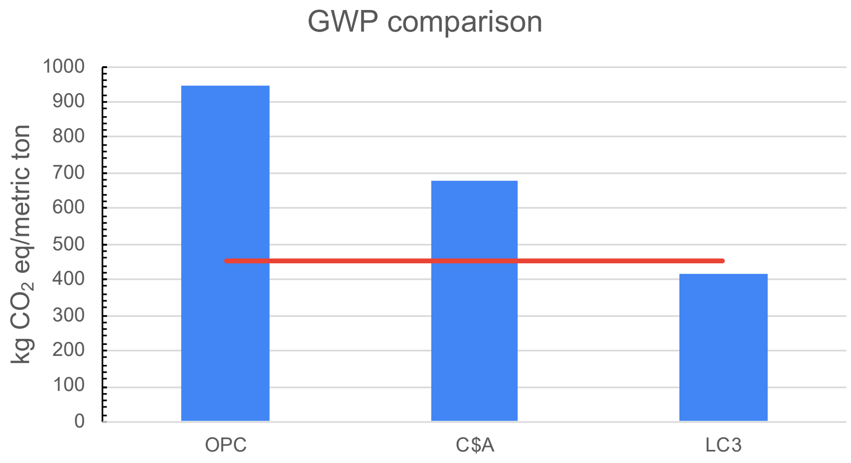


Figure 6.26. GWP comparison of OPC, CSA, and LC3 cements. The red line indicates roughly 50% GWP potential of OPC.

The final equation utilized for modeling GWP is shown below in Eq. 6.4 and accounts for variation among 4 various size limestones which will be included for optimization.

$$
\begin{aligned}
GWP = {} & \frac{942.21\ Kg}{ton}PC + \frac{435\ Kg}{ton}MK1 \\
& + \frac{435\ Kg}{ton}MK2 + \frac{435\ Kg}{ton}MK3 + \frac{269\ Kg}{ton}LS3 \\
& + \frac{117\ Kg}{ton}LS15 + \frac{52\ Kg}{ton}LS25 \\
& + \frac{31.1\ Kg}{ton}LS40 + \frac{11.4\ Kg}{ton}gypsum
\end{aligned}
\qquad \text{Eq. 6.4}
$$

6.3.4. Multi-objective Optimization Predictions

The HML strength model was utilized in conjunction with the workability model and a multi-objective optimization was performed. OPC, three separate metakaolin's and four various limestone sizes along with gypsum were included in the optimization procedure with a constant water: cementitious ratio of 0.40 and MG7920 utilized as the superplasticizer. The optimization conditions are shown below in Eq 6.5-6.10:

$Let\ the\ basis\ set\ of\ input\ parameters\ be$:

$$\theta = \left[Curing\ Time, PC, MK1, MK2, MK3, LS3, LS15, LS25, LS40, gypsum, \frac{W}{Cm}, SP\%\right]$$

Eq. 6.5

$\boldsymbol{Maximize}$:

$$Optimal\ Cement\ Composition = \big(Workability(\theta), Strength(\theta)\big)$$

Eq. 6.6

$$Workability(\theta) = RandomForest(Workability\ model)$$

Eq. 6.7

$$Strength(\theta) = GaussianProcess(Strength\ model)$$

Eq. 6.8

$\boldsymbol{Subject\ to}$:

$$\begin{cases} PC + MK1 + MK2 + MK3 + LS3 + LS15 + LS25 + LS40 \\ \quad + gypsum + \dfrac{W}{Cm} = 1 \\ \dfrac{942.21\ Kg}{ton} PC + \dfrac{435\ Kg}{ton} MK1 + \dfrac{435\ Kg}{ton} MK2 + \dfrac{435\ Kg}{ton} MK3 \\ \quad + \dfrac{269\ Kg}{ton} LS3 + \dfrac{117\ Kg}{ton} LS15 + \dfrac{52\ Kg}{ton} LS25 + \dfrac{31.1\ Kg}{ton} LS40 \\ \quad + \dfrac{11.4\ Kg}{ton} gypsum < \dfrac{450\ Kg}{ton} \end{cases}$$

Eq. 6.9

Where:

$$Curing\ Time = 7\ or\ 28\ days; SP\% = 0.25, or\ 0.50\%\ by\ cement;$$
$$\frac{W}{cm} = 0.40$$

Eq. 6.10

Variations in the MK:LS ratios and percentage of cement were allowed subject to the constraints that the total percentage of the composition summed to 100%, and CO_2 emissions were confined to < 450kg CO_2/ton of material as to optimize blends with under 50% reduction in GWP as compared to OPC. The algorithm was allowed to run for 500,000 iterations taking approximately 10 h. The Pareto front associated with this optimization is shown in Figure 6.27 and the selected composition for testing is shown in Table 6.9.

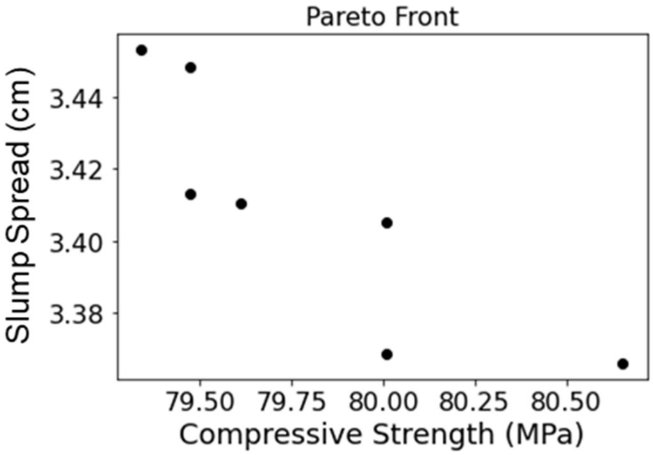
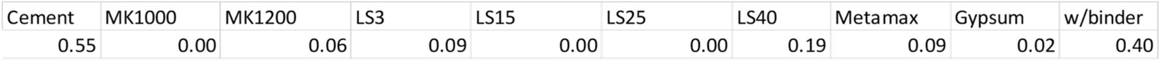


Figure 6.27. Pareto front of the maximized slump and compressive strength values. Slump is measured as the change in slump from the original mini-slump cone diameter.

Table 6.9. Selected composition from pareto front utilized for testing.



Cement	MK1000	MK1200	LS3	LS15	LS25	LS40	Metamax	Gypsum	w/binder
0.55	0.00	0.06	0.09	0.00	0.00	0.19	0.09	0.02	0.40

6.3.5. Optimization Performance

The selected blend was tested at both 0.25% SP and 0.5% SP with the results and comparison shown for workability and strength in Table 6.10-

Table **6.11** with the estimated CO_2 output and discussion following.

Table 6.10. Workability results for the optimized blend.

SP%	Predicted PAT (cm)	Measured PAT (cm)	Standard Deviation (cm)
0.25%	5.8cm	4.56cm	0.10cm
0.50%	7.11cm	11.6cm	0.54cm

Table 6.11. Compressive strength results for 7 and 28 day strength. The SP was held constant at 0.25% for testing.

Cure Time	Predicted Strength (MPa)	Measured Strength (MPa)	Standard Deviation (MPa)
7 days	65.7MPa	46.32MPa	4.4MPa
28 days	78.8MPa	45.88MPa	5.6MPa

The predicted PAT from the random forest model trend with the measured PAT for the optimized blend. The predicted GWP for the optimized blend was found to be 439 kg CO_2/ton which represents over a 50% GWP reduction as compared to OPC. The gaussian process predicted strengths are higher than the measured strengths. The CO_2 constraint created LS:MK ratios of 2:1, however only 22 of the 97 unique compositions had ratios which met or exceeded this ratio, while the rest were predominately a LS:MK ratio of 1:2. This could exhibit a need for further expanding the dataset with a more diverse set of LS:MK ratios. The prediction error in these blends also could indicate an underestimation in the beneficial addition to compressive strength from the pozzolanic activity of the calcined clay. While the model had kaolin content as an input parameter, there was a lack of an embedded latent variable to account for the lower chemical reactivity of the high LS:MK ratio blends.

6.3.6. Future Directions

With the disparate materials and cement utilized from combining these datasets, differences in OPC clinker phase were not accounted for in these models. The complex reactions in cementitious systems have been long researched through experimental techniques in characterizing the microstructure through scanning electron microscopy, nano-computed tomography, and X-ray diffraction.[32] Computational techniques such as thermodynamic modeling[33] and molecular dynamic simulations[34,35] have been utilized in elucidating the microstructure of hydrated cement. Although these approaches have led to insight into the characterization of cement microstructure, little insight has been accomplished into establishing quantitative relationships to the experimentally determined macroscopic properties of cementitious systems. Thermodynamic modeling has been shown to predict the hydrated phase assemblage, including Ca:Si ratios in C-S-H and porosity.[36] Through embedding chemical knowledge of LC3 systems, a more generalizable model will be able to be modeled and tailored with to the initial OPC clinker phase assemblage.

Also, the introduction of more data would be helpful in improving model performance. Due to constraints on GWP, the LC3 composition was limited to 2:1 LS:MK ratios, which were underrepresented in the training set. Through improving the range of data analyzed, more accurate models will be able to be predicted. Beyond a human-centered approach to selecting compositions to train for refinement of the model, a Bayesian optimization routine can be established. Bayesian optimization considers both exploration, where compositions would be selected to minimize uncertainty in the model, and exploitation, where compositions are selected to maximize the target goals of the optimization. This allows for an algorithm guided design of experimentation which can learn the best model in a minimal number of tested compositions.[37]

Finally, the utilization of multi-objective optimization can be extended to design blends that are not only sustainable environmentally, but also economically. Constraints based on material costs can be tailored to keep costs competitive with OPC. While calcined clays in this research were highly pure, impure clays of lower costs can be included for future modeling in strength, workability, GWP, and cost. Studies involving cement are an interdisciplinary field involving material science, civil, and mechanical engineering with a continually progressing connection to ML. With concrete as the leading construction material in the world it is necessary to combine knowledge from natural sciences, engineering, and statistics to improve the manufacturing, placement, strength, longevity, and costs associated with the varying uses of cement

6.4. Conclusion

A multi-objective optimization was able to predict LC3 blends based on commonly available raw materials in the North American market which maximized strength and workability requirements while keeping GWP less than 50% of that for OPC. Physical insight gathered from the ML modeling agreed with prior literature in the factors which improve LC3 strength and workability. While improvements in particle packing due to smaller size particles were found to be beneficial to LC3 strength, tradeoffs in water film thickness negatively affected both GWP and workability predictions. These models can be supplemented in the future with a higher diversity of LS:MK ratios in order to more effectively predict both workability and strength associated with blends corresponding to low GWP.

6.5. Research Contributions

C.M. Childs performed latent variable design, ML analysis, multi-objective optimization, rheological testing, and polymer characterization. O. Canbek performed compressive strength testing and data collection. F. Lolli performed the lifecycle assessment. T.M. Kirby performed rheological testing and provided support in polymer characterization.

6.6. References

[1] Mehta, P.; Monteiro, P. Concrete: Microstructure, Properties and Materials; Mc-Graw Hill: New York, 2006.

[2] Scrivener, K. L.; John, V. M.; Gartner, E. M. Eco-Efficient Cements: Potential Economically Viable Solutions for a Low-CO2 Cement-Based Materials Industry; Paris, 2017.

[3] Li, J.; Zhang, W.; Li, C.; Monteiro, P. J. M. Green Concrete Containing Diatomaceous Earth and Limestone: Workability, Mechanical Properties, and Life-Cycle Assessment. J. Clean. Prod. 2019, 223, 662–679.

[4] Juenger, M. C. G.; Snellings, R.; Bernal, S. A. Supplementary Cementitious Materials: New Sources, Characterization, and Performance Insights. Cement and Concrete Research. Elsevier Ltd August 1, 2019, pp 257–273.

[5] Scrivener, K.; Martirena, F.; Bishnoi, S.; Maity, S. Calcined Clay Limestone Cements (LC3). Cement and Concrete Research. Elsevier Ltd December 1, 2018, pp 49–56.

[6] Naqi, A.; Jang, J. Recent Progress in Green Cement Technology Utilizing Low-Carbon Emission Fuels and Raw Materials: A Review. Sustainability 2019, 11 (2), 537.

[7] Zunino, F.; Scrivener, K. L. Influence of Kaolinite Content, Limestone Particle Size and Mixture Design on Early-Age Properties of Limestone Calcined Clay Cements (LC3). In RILEM Bookseries; Springer, 2020; Vol. 25, pp 331–337.

[8] Scrivener, K.; Avet, F.; Maraghechi, H.; Zunino, F.; Ston, J.; Hanpongpun, W.; Favier, A. Impacting Factors and Properties of Limestone Calcined Clay Cements (LC 3). Green Mater. 2019, 7 (1), 3–14.

[9] Zunino, F.; Scrivener, K. The Influence of the Filler Effect on the Sulfate Requirement of Blended Cements. Cem. Concr. Res. 2019, 126, 105918.

[10] Karpatne, A.; Atluri, G.; Faghmous, J. H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. IEEE Trans. Knowl. Data Eng. 2017, 29 (10), 2318–2331.

[11] Childs, C. M.; Washburn, N. R. Embedding Domain Knowledge for Machine Learning of Complex Material Systems. MRS Commun. 2019, 9 (2), 1–15.

[12] Menon, A.; Childs, C. M.; Poczós, B.; Washburn, N. R.; Kurtis, K. E. Molecular Engineering of Superplasticizers for Metakaolin-Portland Cement Blends with Hierarchical Machine Learning. Adv. Theory Simulations 2019, 2 (4).

[13] Washburn, N. R.; Menon, A.; Childs, C. M.; Poczos, B.; Kurtis, K. E. Machine Learning Approaches to Admixture Design for Clay-Based Cements. In Calcined Clays for Sustainable Concrete; Martirena, F., Favier, A., Scrivener, K., Eds.; Springer, Dordrecht, 2017; Vol. RILEM Book, pp 488–493.

[14] Menon, A.; Gupta, C.; Perkins, K. M.; DeCost, B. L.; Budwal, N.; Rios, R. T.; Zhang, K.; Póczos, B.; Washburn, N. R. 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.

[15] Marchon, D.; Sulser, U.; Eberhardt, A.; Flatt, R. J. Molecular Design of Comb-Shaped Polycarboxylate Dispersants for Environmentally Friendly Concrete. Soft Matter 2013, 9 (45), 10719.

[16] Kantro, D. L. Influence of Water-Reducing Admixtures on Properties of Cement Paste -- A Miniature Slump Test. Cem. Concr. Aggregates 1980, 2 (2), 95–102.

[17] Ciroth, A. ICT for Environment in Life Cycle Applications OpenLCA - A New Open Source Software for Life Cycle Assessment. International Journal of Life Cycle Assessment. Springer Verlag 2007, pp 209–210.

[18] U.S. Life Cycle Inventory Database. Federal LCA Commons.

[19] Marceau, M. L.; Nisbet, M. A.; Vangeem, M. G. Life Cycle Inventory of Portland Cement Manufacture; Skokie, Illinois, 2006.

[20] Menon, A.; Childs, C. M.; Poczós, B.; Washburn, N. R.; Kurtis, K. E. Molecular Engineering of Superplasticizers for Metakaolin-Portland Cement Blends with Hierarchical Machine Learning. Adv. Theory Simul. 2019, 2 (1800164).

[21] Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; et al. Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.

[22] Rasmussen, C. E.; Williams, C. K. I. Gaussian Processes for Machine Learning, 2nd ed.; MIT Press: Cambridge, 2006.

[23] Rasmussen, C. E. Gaussian Processes in Machine Learning. In Advanced Lectures on Machine Learning; Bousquet, O., von Luxburg, U., Rätsch, G., Eds.; Springer-Verlag: Berlin, 2003; pp 63–71.

[24] Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Trans. Evol. Comput. 2002, 6 (2), 182–197.

[25] Benitez-Hidalgo, A.; Nebro, A. J.; Garcia-Nieto, J.; Oregi, I.; Del Ser, J. JMetalPy: A Python Framework for Multi-Objective Optimization with Metaheuristics. Swarm Evol. Comput. 2019, 51.

[26] Gelardi, G.; Flatt, R. J. Working Mechanisms of Water Reducers and Superplasticizers. In Science and Technology of Concrete Admixtures; Elsevier, 2016; pp 257–278.

[27] Nandi, B. K.; Goswami, A.; Purkait, M. K. Adsorption Characteristics of Brilliant Green Dye on Kaolin. J. Hazard. Mater. 2009, 161 (1), 387–395.

[28] Jolicoeur, C.; Simard, M.-A. Chemical Admixture-Cement Interactions: Phenomenology and Physico-Chemical Concepts. Cem. Concr. Compos. 1998, 20 (2–3), 87–101.

[29] Lange, A.; Plank, J. Contribution of Non-Adsorbing Polymers to Cement Dispersion. Cem. Concr. Res. 2016, 79, 131–136.

[30] Schwinefus, J. J.; Checkal, C.; Saksa, B.; Baka, N.; Modi, K.; Rivera, C. Molar Mass and Second Virial Coefficient of Polyethylene Glycol by Vapor Pressure Osmometry. 2015.

[31] Lothenbach, B.; Scrivener, K.; Hooton, R. D. Supplementary Cementitious Materials. Cem. Concr. Res. 2011, 41 (12), 1244–1256.

[32] Monteiro, P. J. M.; Geng, G.; Marchon, D.; Li, J.; Alapati, P.; Kurtis, K. E.; Qomi, M. J. A. Advances in Characterizing and Understanding the Microstructure of Cementitious Materials. Cement and Concrete Research. Elsevier Ltd October 1, 2019, p 105806.

[33] Lothenbach, B.; Winnefeld, F. Thermodynamic Modelling of the Hydration of Portland Cement. Cem. Concr. Res. 2006, 36 (2), 209–226.

[34] Pellenq, R. J. M.; Kushima, A.; Shahsavari, R.; Van Vliet, K. J.; Buehler, M. J.; Yip, S.; Ulm, F. J. A Realistic Molecular Model of Cement Hydrates. Proc. Natl. Acad. Sci. U. S. A. 2009, 106 (38), 16102–16107.

[35] Abdolhosseini Qomi, M. J.; Krakowiak, K. J.; Bauchy, M.; Stewart, K. L.; Shahsavari, R.; Jagannathan, D.; Brommer, D. B.; Baronnet, A.; Buehler, M. J.; Yip, S.; et al. Combinatorial Molecular Optimization of Cement Hydrates. Nat. Commun. 2014, 5 (1), 1–10.

[36] Lothenbach, B.; Matschei, T.; Möschner, G.; Glasser, F. P. Thermodynamic Modelling of the Effect of Temperature on the Hydration and Porosity of Portland Cement. Cem. Concr. Res. 2008, 38 (1), 1–18.

[37] Brochu, E.; Cora, V. M.; de Freitas, N. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. arXiv:1012.2599 2010.

Chapter 7. Conclusions and Future Directions

7.1. Benefits of Domain Knowledge

It has been shown that embedding domain knowledge into cementitious systems allows learning physical interactions on small datasets. This knowledge can be incorporated into machine learning (ML) algorithms through the use of the correct data representation as shown in Figure 7.1.

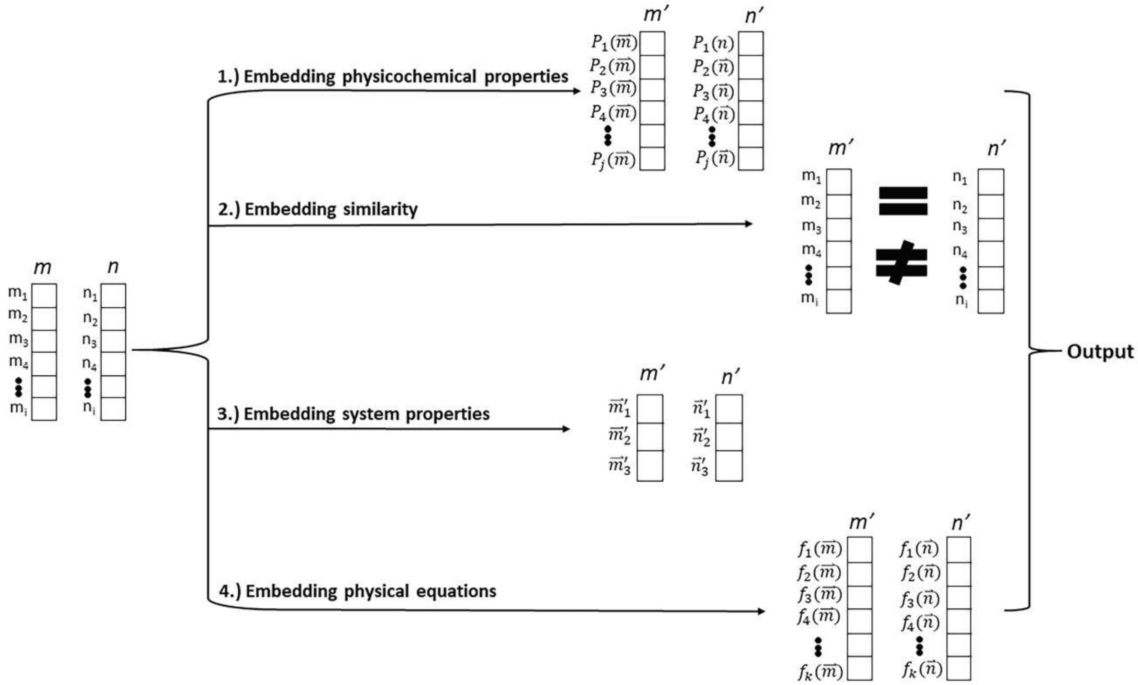


Figure 7.1. A common representation for a material or formulation is as a vector representing the structure or makeup of the system. The original vectors for materials m and n are shown through the four pathways of data representation which were discussed in this thesis to form transformed vectors m’ and n’. Method 1 represents the materials in the form of computational or experimental properties, *P(x)*, for the original material. Method 2 represents the vectors through finding a (di)similarity metric to compare the two. Method 3 represents the vectors through a direct transformation of the *x* components in order to embed properties such as invariance. Method 4 represents the original material in terms of physical interactions utilizing known physical equations. These transformed vectors are a latent variable representation that are utilized as the features for ML techniques to learn an associated output.

Chapter 1 provided a review of various representations for embedding domain knowledge in material systems. These representations were applied towards machine learning tools that guide the design of more sustainable, durable cementitious binders that can accommodate variations in materials through customized cement compositions and chemical admixture formulations.

In Chapter 2, mechanistic understanding of cement dispersion due to polymeric admixtures was discovered through the measurement of physicochemical forces responsible for dispersion. Chapter 3 embeds these physicochemical forces as latent variables to discover the importance of the contributions due to molecular architecture, anionic functionality, and anionic ratios with a hierarchical machine learning (HML) approach. Understanding was gained to predict adsorption mechanisms, and nonadsorbed polymer contributions in the development of a superplasticizer designed specifically for a metakaolin (MK) modified portland cement (PC) system.

In Chapter 4, concepts of physiochemical properties and similarity were embedded into modeling the effects of various small organic compounds as retarders in calcium sulfoaluminate (CSA) cement. Models relating chemical properties and chemical structures to the predicted set time were built. The latent representation of chemical structures through binary fingerprints allowed for the ability of the model to determine retarding capability of unseen test molecules proceeded through the utilization of virtual screening. Insight into the importance of phosphono and carboxylate groups in retarder structures was developed and methodology into the development of a machine learning tool that guides the discovery of chemical admixtures for sustainable, durable cementitious binders allows for efficient, cost-effective virtual screening, was established.

Finally, in Chapter 5 and 6, concepts combining a mixture of embedding system properties and physical equations in an HML framework were utilized in the prediction of the compressive strength of cementitious systems. The latent variable representation for these systems allowed for the design of models which allow for generalizability to locally sourced materials.

Broader impacts for this research include both environmental and economic benefits. Large reductions in CO_2 from utilizing the supplementary cementitious materials (SCMs) and alternative binder chemistries (ABCs) that were studied through this thesis are possible. However, due to drawbacks such as limited production and quick setting in CSA cements along with availability and necessity for specialized admixtures in SCM based cements, economic factors limit their utilization. Through the utilization of ML techniques, costs will be decreased due to improvement in the utilization and knowledge of cementitious systems, leading to wider adoption of these systems. Worldwide adoption of these cements will not only decrease overall economic costs associated with procuring raw material and processing cements, but will lead decrease in CO2 production in the partial replacement of ABCs and SCMs over PC.

7.2. Future Directions

One future prospect in research to allow improvement and usefulness will be on transfer learning. A transferrable HML model which can be further tested and trained on various types of cements to include: PC, ABC cements, and customized cement blends containing supplementary cementitious materials which can be utilized to specifically tailor cement properties. Transfer learning would allow for learning on one system similar to another and have the added benefit of increasing the size of the dataset through learning on similar systems. For systems following similar physical relations, introducing statistics from prior ML studies could easily be remodeled in the new system. Figure 7.2 is an adapted schematic from Hutchinson et al.[1] showing 3 techniques of transfer learning. This methodology has already been studied in quantum chemical

(QC) applications of ML where small molecules were studied to understand relations in larger molecules, an example of multi-task transfer learning.[1,2] The ultimate goal for such studies with proper relations being embedded or learned through ML would be a universal reactive force field.[3] QC has also utilized approaches of difference transfer learning where computational and experimental outputs are compared to predict a system output. Learning on the difference allows for variables to be found to fit a closure model. Such learning techniques have been studied on the prediction of turbulence flow.[4]

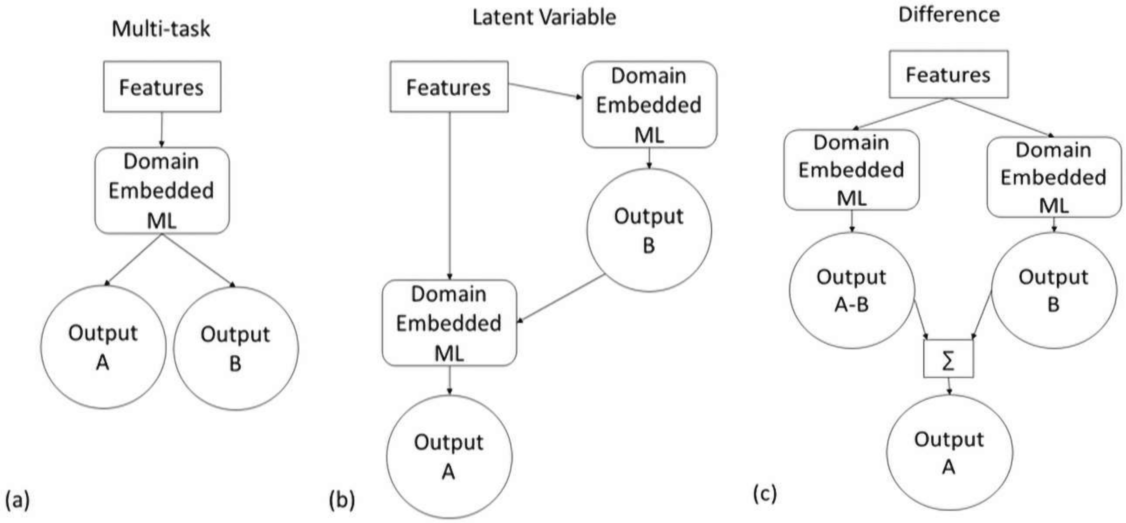


Figure 7.2. Schematics for three types of transfer learning adapted from Hutchinson et al.[1] (a) Multi-task transfer learning is when one model is learned to fit multiple systems. (b) Latent variable transfer learning is a technique where a latent domain variable is learned on one system and included as a feature for predicting the output of another system. (c) Difference transfer learning is a technique where training data are relabeled as the difference between features and a model is learned from this difference.

One issue that needs to be resolved in transfer learning is ensuring proper physics are embedded as the system is changed. Determining the appropriate transfer learning approach for each domain knowledge ML technique may be a system independent approach that again may require prior human knowledge with test and error approaches. If the physics in the interactions within the complex system do not change, multi-task learning may be the best technique to utilize. If the underlying physics do change, then other transfer learning techniques may need to be considered.

The ability to transition from a human-centered to data-guided approaches in engineering systems is a core component of a grand challenge to both engineering and scientific research. To resolve economic and sustainability constraints, along with reducing extensive iterative testing, utilization of data science can lead to novel innovations in material systems in at an accelerated rate of innovation. Data-based engineering can also optimize conditions to address big challenges in engineering such as producing highly sustainable materials and improvement in urban infrastructure.[5]

By nature, studies involving cement are an interdisciplinary field involving chemistry, material science, and civil engineering. While significant improvements in this field – from more efficient production to increased service life – have been realized over the past decades through traditional research paradigms, non-incremental innovations are necessary to meet global goals for sustainable development. Data science is revolutionizing the rate of discovery and accelerating the rate of innovation for material systems. This research connects the continually growing field of machine learning. With cement as the leading construction material in the world, research at the interface of pure science, engineering, and statistics need to work together to improve the manufacturing, placement, strength, longevity, and costs associated with the varying uses of cement. Through the utilization of latent variables, coupled approaches to dimensionality reduction driven both algorithmically as well as through domain knowledge, better feature representation is provided for cement-based materials which allow for more accurate models and greater generalization capability. This enables concrete structures of higher durability and longer service life lowering overall costs, along with reduced environmental impact. Finally, from this research a wide variety of cements can be modeled and tested to discover unique cementitious blends leading to widespread and accepted utilization of novel cementitious materials.

7.3. References

[1] Hutchinson, M. L.; Antono, E.; Gibbons, B. M.; Paradiso, S.; Ling, J.; Meredig, B. Overcoming Data Scarcity with Transfer Learning. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*; Long Beach, 2017; pp 1–10.

[2] Welborn, M.; Cheng, L.; Miller, T. F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *J. Chem. Theory Comput.* **2018**, *14* (9), 4772–4779.

[3] Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3* (12), e1701816.

[4] Parish, E. J.; Duraisamy, K. A Paradigm for Data-Driven Predictive Modeling Using Field Inversion and Machine Learning. *J. Comput. Phys.* **2016**, *305*, 758–774.

[5] W. Perry, A. Broers, F. El-Baz, W. Harris, B. Healy, W.D. Hillis, C. Juma, D. Kamen, R. Kurzweil, R. Langer, J. Lerner, B. Lohani, J. Lubchenco, M. Molina, L. Page, R. Socolow, J.C. Venter, J. Ying, NAE GRAND CHALLENGES FOR ENGINEERING, Washington, D.C., 2017. www.engineeringchallenges.org. (accessed October 9, 2020).