The thesis of Aditya Sonpal was reviewed by the following:

Johannes Hachmannnn
Assistant Professor of Chemical and Biological Engineering
Thesis Advisor, Chair of Committee

David A. Kofke
SUNY Distinguished Professor of Chemical and Biological Engineering
Committee Member
Dedication

This thesis is dedicated to my parents Niraj and Pratibha Sonpal for their unwavering support and encouragement. They have always inspired me to pursue my dreams with hard work and dedication.
Acknowledgments

I would like to express my deepest gratitude to my advisor and mentor, Prof. Johannes Bachmann, for his relentless support, patience and encouragement. He inspired me to take up challenging problems without fear and tackle them head on. His guidance continues to help me in developing my career as a professional and as an academic. With the right amount of encouragement and freedom, he has helped me develop my skills and make leaps to find my own place in this wide research community.

I would also like to thank Prof. David Kofke for taking out the time to be on my committee. In addition to this, I am grateful to Dr. Andrew Schultz for helping me through various stages of my project with his insightful observations and sound guidance.

Among all my colleagues, I can safely say that I would not have been able to make it without the inputs and contributions of Dr. Mohammad Atif Faiz Afzal and Mojtaba Haghighatlari. Their constant support and sound advice has not only helped me complete my thesis, but also continues to inspire me to challenge myself by taking up new research endeavours.

I would like to thank my colleagues and friends from the Hachmann Re-
search Group, especially Gaurav Vishwakarma and Yudhajit Pal, for their contributions in my thesis. Without the constant companionship and support of my friends here in Buffalo, Ronel Samuel, Benjamin Hultman, Shabdiki Chaurasia, and others, this journey would not have been as fun. I would like to thank Ankita Agrawal for constantly supporting me throughout these past two years.

I would like to acknowledge the help and support of the Department of Chemical and Biological Engineering, Center for Computational Research and the School of Engineering and Applied Sciences at the University of Buffalo for giving me this opportunity and the support required to complete this thesis.

Lastly, words cannot capture my gratitude towards my parents and family for their support and for always inspiring me to dream big.
# Table of Contents

Acknowledgments iv
Abstract viii

## Chapter 1
Introduction and Background 1
  1.1 Ionic Liquids .................................................. 1
  1.2 Deep Eutectic Solvents ........................................ 5
  1.3 Melting Point .................................................. 10
  1.4 Data Driven Approach ........................................ 15
  1.5 Group’s Software Ecosystem for Materials Discovery ....... 17
  1.6 Summary ....................................................... 17

## Chapter 2
Methods 18
  2.1 Data Collection ............................................... 18
     2.1.1 SMILES .................................................. 19
     2.1.2 Open Babel .............................................. 20
     2.1.3 ORCA .................................................... 20
  2.2 Data Mining ................................................... 21
     2.2.1 Feature Representation .................................. 21
     2.2.2 Feature Selection ....................................... 22
        2.2.2.1 Pearson Correlation ................................. 22
        2.2.2.2 Genetic Algorithms .................................. 23
     2.2.3 Machine Learning Models .............................. 24
        2.2.3.1 Ridge Regression .................................... 24
        2.2.3.2 Bayesian Ridge Regression ......................... 24
# TABLE OF CONTENTS

2.2.3.3 LASSO ............................................. 24
2.2.3.4 Elastic Net ...................................... 25

Chapter 3
Results and Discussion .................................. 26
3.1 Feature Representation and Feature Selection ........... 26
3.2 Machine Learning ........................................ 29
3.3 Conclusion and Discussion ............................... 31

Bibliography ................................................. 33
Abstract

Up until the end of the 20th century, ionic liquids and later, room temperature ionic liquids had been used as solvents for the synthesis of a number of organic and inorganic materials. Owing to their wide liquid range, immiscibility, non volatility, easy synthesis, etc., they were largely favored by scientists everywhere. Until then they were also believed to be green solvents. Their green status was later contested, which paved the way for the discovery of deep eutectic solvents (DES). These solvents were composed of two or more cheap and safe materials which associated with each other through hydrogen bond interactions to create a eutectic mixture. In other words, DES systems had a lower melting point than that of their individual components. In addition to all the advantages of ionic liquids, DES are also green solvents and can be easily and cheaply synthesized.

We therefore identified the potential of DES systems and the importance of melting point in characterizing such systems. We use a data driven approach to predict melting points of DES. We first shortlist a set of 36 DES systems, each
having Choline Chloride as the hydrogen bond donor. We optimize the geometries of these compounds using DFT frameworks and use that structural information to generate descriptors from the Dragon 7 software to generate our feature space. We use pearson correlation coefficients and genetic algorithm for feature selection and use the resulting feature space to train generalized linear regression algorithms from the ‘scikit-learn’ python library. In this thesis, we first show that the PBE0 functional with def2-TZVP basis set generates the most accurate geometrical configurations for feature representation. We further prove the supremacy of genetic algorithms over pearson correlation coefficients for feature selection. In conclusion, we show that the linear ridge regression model yields the most accurate predictions for our dataset with PBE0/def2-TZVP. We then highlight some of the challenges associated with predicting melting points and provide suggestions that could possibly help to mitigate them.
1.1 Ionic Liquids

Chemistry is dominated by the study of species in solution. Although any liquid may be used as a solvent, relatively few are generally used without hesitation [1]. Until the early 20th century, most chemical reactions were carried out in molecular solvents. For two millennia, most of our understanding of chemistry had been based on the behavior of molecules in the solution phase with molecular solvents. The discovery of ionic liquids is attributed to Paul Walden’s observations in 1914. He reported the physical properties of ethylammonium nitrate [EtNH₃][NO₃] (melting point 13°C-14°C), which was formed by the neutralization of ethylamine with concentrated nitric acid. This is where the widely accepted definition of ionic liquids stems from. Ionic liquids are usually defined as materials that are composed of cations and anions which have a melting point of less than or equal to 100°C. This definition was widely recognized until the physical significance of 100°C in this definition was contested [2]. In the current
definition, ionic liquids are said to consist of a salt where one or both of the ions are large, and the cation has a low degree of symmetry. These factors tend to reduce the lattice energy of the crystalline form of the salt, and hence lower the melting point [3]. Over the next few decades, scientists continued to make relentless efforts in exploring the field of ionic liquids. Wilkes' and Hussey's work led to the synthesis of archetypal ionic liquid system, \([\text{C}_2\text{mim}]\text{Cl}-\text{AlCl}_3\) and its subsequent characterization [4]. Although it had a wide liquid range and it was the first genuine example of an ionic liquid system that was liquid at room temperature, it had some major drawbacks. Chloroaluminate(III) ionic liquids were sensitive to moisture and therefore could not find industrial application. Despite this, it was a major breakthrough and paved the way to research that led to the advent of room temperature ionic liquids (RTILs) [2].

At this stage scientists mainly focused on the formation of RTILs by mixing metal salts, mostly zinc, aluminum, tin, and iron chlorides, with quaternary ammonium salts. Although both salts had very high melting points, when they were mixed in specific ratios and at a specific temperature, they formed a liquid and the mixture was known as a eutectic mixture. These eutectic mixtures were characterized by depression of freezing points [5]. In 1992, Wilkes et al. [4], reported the preparation and characterization of a new range of ionic liquids that still contained 1-ethyl-3-methylimidazolium cation but also contained a range of alternative anions like \([\text{CH}_3\text{CO}_2]\), \([\text{NO}_3]\) or \([\text{BF}_4]\). These salts were easier to prepare on the bench as they were air and water stable. In the quest to find RTILs with the widest liquid range, the effects of anion size, cation asymmetry, cation size, etc. were extensively studied. This further cited the 'designer solvent' status of ionic liquids. Properties like melting point, viscosity, density, and hydrophobicity, etc., could be tailored by simple changes to the ionic struc-
ture. For example, Holbrey et al. [6], showed how the size of the alkyl chain on 1-alkyl-3-methylimidazolium tetrafluoroborate salts affects the liquid range of the resulting isotropic ionic liquid. Shorter chain lengths resulted in isotropic ionic liquids at room temperature and exhibit a wide liquid range compared to longer chain analogues which resulted in low melting mesomorphic crystalline solids which displayed a mesophase. Since the advent of green chemistry, in the early 1990s there had been a shift in focus towards metal-free ionic liquids. This entailed the design of numerous ionic liquids by combining organic cations with a variety of anions. This in turn created ionic liquids with varying physical properties like density, viscosity, melting point, etc [5].

Over the years it was discovered that RTILs had various advantages and could be used as potential solvents for synthesis. They work as effective solvents for inorganic and organic materials. Even unusual combinations of reagents can be brought into the same phase. Since they are mainly composed of poorly coordinating ions, they can potentially be highly polar as well, despite being non-coordinating. Ionic liquids are non-volatile, which is why they can be used in high-vacuum systems and eliminate many containment problems. In addition to this, they help reduce the amount of solvent or catalyst in a chemical reaction [1]. Their non-volatile nature also enabled significant engineering advantages for distillation product separation and prevented uncontrolled evaporation. Specifically varying their physical and chemical properties made them ideal candidates for application in biphasic catalysis [7]. Although later Earle et al. [8], discovered that RTILs could indeed be distilled. This proved the volatility of some thermally stable ionic liquids under relatively mild conditions. This created further opportunities to recycle ionic liquids that were already spent. High product yields could be obtained with a reduced amount
of waste in any given reaction. This recyclability of ionic liquids would reduce expenditure in the overall process. Reactions in ionic liquids were usually quick, simple to implement, and required no special apparatus.

Later, the 'green liquids' status of ILs was highly contested [3]. Some Ionic liquids were found to be highly toxic and poorly biodegradable. Deetlefs et al. [9], suggested that an ionic liquid should only be considered green if both the ionic liquid itself, and the process used to synthesize it are green, i.e., they follow all 12 principles of green chemistry as shown in Figure 1.1. After careful assessment of different ionic liquids and their laboratory synthesis, it was concluded that in order to uphold the green status of laboratory grade ionic liquids, the synthesis would require the application of microwaves. Additionally, it was also mentioned that the purification of hydrophobic ionic liquids is intrinsically greener than that of hydrophilic ionic liquids. Emphasis was laid on the fact that the same green status could not be attributed to industry grade ionic liquids. With the current set of standards, RTILs could be rendered green but not green enough. Application of ionic liquids required high purity. Even a slight variation would alter the physical properties and deem it unfit for the particular application. Synthesis of ionic liquids was not environmentally friendly, large amounts of salts and solvents were required for complete change of anions. These disadvantages along with the expensive nature of ionic liquids paved the way for the next generation of solvents, known as deep eutectic solvents.
The Twelve Principles of Green Chemistry

1. It is better to prevent waste than to treat or clean up waste after it has formed.
2. Synthetic methods should be designed to maximise the incorporation of all materials used in the process into the final product.
3. Wherever practicable, synthetic methodologies should be designed to use and generate substances that possess little or no toxicity to human health and the environment.
4. Chemical products should be designed to preserve efficacy of function while reducing toxicity.
5. The use of auxiliary substances (e.g. solvents, separation agents etc.) should be made unnecessary wherever possible and innocuous when used.
6. Energy requirements should be recognised for their environmental and economic impacts and should be minimised. Synthetic methods should be conducted at ambient temperature and pressure.
7. A raw material of feedstock should be renewable rather than depleting wherever technically and economically practicable.
8. Unnecessary derivatisation (blocking group, protection/deprotection, temporary modification of physical/chemical processes) should be avoided whenever possible.
9. Catalytic reagents (as selective as possible) are superior to stoichiometric reagents.
10. Chemical products should be designed so that, at the end of their function, they do not persist in the environment, and break down into innocuous degradation products.
11. Analytical methodologies need to be further developed to allow for real-time, in-process monitoring and control prior to the formation of hazardous substances.
12. Substances and the form of a substance used in a chemical process should be chosen so as to minimise the potential for chemical accidents, including releases, explosions and fires.

Figure 1.1. 12 Principles of Green Chemistry [9].

1.2 Deep Eutectic Solvents

Eutectic mixtures of salts had been utilized for a long time to decrease the temperature for molten salt applications. In extreme, ambient temperature molten salts have been formed by mixing quaternary ammonium salts with metal salts. This type of ionic liquid can be viewed as a deep eutectic resulting from formation of complex anions, $\text{Al}_2\text{Cl}_7^-$ and $\text{Zn}_2\text{Cl}_5^-$ thus decreasing the lattice energy and decreasing the freezing point of the system [11] [12] [10]. Abbott et al. shows that a mixture of substituted quaternary ammonium salts such as hydroxyethyltrimethylammonium (Choline, see figure 1.3) chloride with urea produce eutectics that are liquid at ambient temperature and had unusual sol-
Figure 1.2. Freezing points of choline chloride/Urea mixtures as a function of composition[10].

Figure 1.3. The choline ion.
vent properties. Figure 1.2 shows the freezing point of the choline chloride and urea mixtures that were observed by Abbott et al. As clearly depicted in this figure, choline chloride and urea form a eutectic at a ratio of 1:2. The freezing point of the resulting eutectic is 12°C, which is considerably less than the melting points of choline chloride (301°C) and urea (133°C). Therefore, this mixture can be used as an ambient temperature solvent. This significant depression of freezing point arises from an interaction between urea molecules and choline ion. This is consistent with crystallographic data for the solid adduct [(CH₃)₃N⁺CH₂CH₂OH]₂C₂O₄²⁻.2(NH₂)₂CS which shows extensive hydrogen bonding between the thiourea molecule and the oxalate anion. The existence of hydrogen bonding was observed using NMR spectroscopy [10]. Not only did this work show the first occurrence of deep eutectic solvents (DES), but it also went on to correlate the depression in freezing point of the resulting eutectics to the ability of the components to form hydrogen bonds.

A DES is composed of two or three cheap and safe components which are capable of associating with each other through hydrogen bond interactions, to form a eutectic mixture. This resulting DES mixture has a melting point lower than that of the individual components. Abbot et al. [13], defined different types of DES using the general formula R₁R₂R₃N⁺X⁻.Y⁻.

Type I DES Y= MClₓ , M=Zn, Sn, Fe, Al, Ga
Type II DES Y= MClₓ.yH₂O , M = Cr, Co, Cu, Ni, Fe
Type III DES Y= R₅Z with Z= -CONH₂, - COOH, -OH

Further, a fourth type is also described, which is a mixture of metal chlorides (eg. ZnCl₂) and HBDs like urea, ethylene glycol, acetamide, etc. [13].

Perkins et al. [14], performed MD simulations along with vibrational spectroscopy experiments. These provided the first insights into the molecular struc-
turing within a choline chloride-urea based eutectic mixture. High number of urea molecules were revealed to be surrounding the chloride ion [Cl]− compared to the number of urea molecules surrounding [choline]+. A hydrogen bond is formed between the [Cl]− (hydrogen bond acceptor) and urea (hydrogen bond donor). These results were corroborated by the vibrational spectroscopy experiments [14]. Garcia et al. [15], conducted a DFT study involving topological analysis of electron density to explain low melting points of 45 different DES at a molecular level. In an attempt to establish a relationship between the hydrogen bonding network features according to the AIM (Atoms In Molecules) population analysis approach and melting point, it was found that the electron density for cage critical points corresponding to the whole interacting network was considered as the most suitable feature for inferring molecular level features justifying DES melting points. This was confirmed by a linear relationship between the two properties. The results show that low melting points are a result of low electron density at the cage critical points. High charge delocalization within the hydrogen bonding network leads to this low electron density at the cage critical point. This in turn leads to the author attributing the melting point depression in DES to high charge delocalization within the hydrogen bonding network. Various other computational efforts to study DES systems further cite this theory [16]. DES share many characteristics of ionic liquids, for example non-reactivity with water, non-volatile and biodegradable nature, large liquid range, high thermal stability, more solvation capacity, non-flammability, etc., [17]. In addition to these, it offers some advantages as well. During the preparation of eutectic mixtures in a pure state, no post synthesis purification is required because the purity of the resulting DES will depend on the purity of the individual components. Additionally, synthesis of DES is relatively simpler and cheaper
than synthesis of RTILs. This is owed to the easy availability of DES components. Therefore, DES are more desirable for large scale applications compared to RTILs [18]. DES systems retain the designer characteristic of RTILs, since it can be designed using various quaternary ammonium salts with different hydrogen bond donors. Application specific DES with different physicochemical properties can be prepared. For example, similar to RTILs, most DES systems have a strong ability to dissolve CO₂. With the added advantage of being green solvents, DES systems are a great prospect for various chemical processes like separation and purification of gas, chemical fixation of CO₂, etc. [5] [19]. More specifically, potential application of DES systems for CO₂ capture in natural gas sweetening has been studied by Aissaoui et al. [20]. Further, DES systems are capable of accepting or donating electrons or protons to form hydrogen bonds and thus achieves good dissolution properties [21]. Abbott et al. [10], first demonstrated the possibility of DES being used to dissolve metal oxides. Morrison et al. [12], demonstrated the possibility of DES being used for drug solubilization. Due to its high polarity DES systems have also been used to separate residual glycerol from raw biodiesel. Many studies have been carried out to explore the possibility of substituting RTILs with DES systems as solvents for various catalysis reactions. Singh et al. [22], used choline chloride based DES for C-C bond formation in rapid synthesis of β-hydroxy functionalized derivatives. Other catalysis reactions where DES has been used as a solvent are base catalyzed reactions [23] [24], acid catalyzed reactions [25], transition metal catalyzed reactions and biocatalysis reactions in DES [26]. DES systems are also used in organic synthesis [27] [28], electrochemical devices [29] [30] [31] [32], in preparation of various materials like phosphates and phosphites [33] [34], metal-organic frameworks, and as a stabilizer for preparing nanoparticles [5]
DES are also being explored for potential applications as electrolytes for supercapacitors. All the potential applications of DES listed in the previous section, make it clear that it is imperative to study DES systems in detail and investigate their physicochemical properties. In general, DES are formed by mixing two solids which are capable of generating a new liquid phase by self-association via hydrogen bonds. The new phase is generally characterized by lowered melting point than that of the individual constituents. Therefore, this thesis focuses on predicting the melting point of DES systems.

1.3 Melting Point

The melting point of a compound is one of its most important and fundamental physical properties. It has been widely used for chemical identification, as a check for purity, and for calculating other essential physicochemical properties such as vapor pressure, aqueous solubility, etc. These properties are used in the domain of ionic liquids and drugs. Additionally, melting and crystallization are of particular significance in condensed matter physics, high pressure physics, geophysical and planetary sciences, and materials science and engineering.

Many methods have been developed for computing melting points, including a variety of empirical and semi-empirical methods. These include group contribution methods, molecular simulation methods, and quantitative structure property relationship based methods. Although group contribution methods and correlations have been used to get accurate results for classes of compounds where little experimental data exists, there has always been a dire need for physics based atomistic methods. Such methods can be classified into two
broad categories. The first group of methods, referred to as direct methods, include the hysteresis method [41] [42], the voids method [43] [44], and solid liquid interface-based methods [45] [46] [47]. These methods involve direct simulation of the melting process in a dynamical manner. Although they have limited accuracy, they are relatively easy to apply. For simple molecules like Argon, these methods have yielded accurate results [43], but for more complicated multiatomic systems free energy based methods are preferred [48]. These form the second category of methods, which are in principle more accurate and comprehensive compared to direct methods but are complicated to apply. This group of methods includes Hoover and Rees single-occupancy method [49] [50], Frenkel and Ladds Einstein crystal method [51], the $\lambda$-integration method [52], and the pseudo-supercritical path (PSCP) method [39].

A straightforward way to compute melting point is to carry out molecular dynamics simulations of a perfect lattice at increasing temperatures. The temperature at which the lattice breaks down corresponds to the melting point. However, the existence of superheating, or the hysteresis phenomenon, results in a significant overestimation of melting points even for simple monatomic molecules. Similarly, when a liquid is cooled down, the phase change temperature is underestimated due to the existence of supercooling. The hysteresis method is based on the homogeneous nucleation theory which takes into account, the superheating and supercooling effects during phase transition and relate this behavior to melting point using the equation

$$T_m = T^+ + T^- - \sqrt{T^+.T^-}.$$ 

Here $T^+$ and $T^-$ are the observed phase change temperatures while heating up
a crystal and cooling down a liquid in a simulation respectively. It is usually easier to get $T^+$ from molecular dynamics simulations, but $T^-$ is difficult to obtain, except for simple monatomic molecules. On few occasions, $T^-$ can also be approximated by glass transition temperature. Overall, the hysteresis method is not usually preferred due to low accuracy [41] [42].

As the name suggests, the voids method involves creating voids or defects in a perfect crystal. Constant pressure simulations (NPT) are carried out with temperature gradients in one trajectory or a series of NPT simulations are carried out with each trajectory having a different, but fixed temperature. In each simulation, the melting point is estimated to be the point at which a discontinuity is observed in system properties like density, RMSD (root mean squared displacement), potential energy, etc. The observed melting points decreases with increasing void density. As the void density increases further, it is often observed that the apparent melting point becomes independent of void density. The temperature at which this occurs is considered as the thermodynamic melting point of the system. The existence of voids serves to lower the nucleation free energy barrier, thereby enabling melting to be observable on time and length scales that are accessible to molecular dynamics simulations. Similarly, simulating a solid-liquid interface also lowers the nucleation free energy barrier. This gives rise to two more direct methods to calculate melting point [43] [44].

The formal thermodynamic definition of the melting point is the temperature at which the crystal phase and liquid phase of a compound share the same free energy. Evidently, the free energy-based methods involve direct computation of the free energy of the molecules and generally avoids simulating the nucleation phenomenon. In the PSCP method, as shown in figure 1.4, the solid and liquid states are connected directly by three intermediate states (the weak
liquid (WL), the dense weak fluid (DWF), and the dense weak crystal (DWC)). Therefore, no analytical reference states are required. The intermolecular interactions are scaled down in the intermediate states and the phase change is controlled to occur in a reversible way so that the error caused by superheating is minimized. In other words, the PSCP method is an inter-phase method where two coexisting phases of interest are linked directly [39].

The energy barriers that are associated with first order phase transitions may give rise to observable kinetic and hysteretic effects, i.e., they may lead to deviation from thermodynamic equilibrium [41]. Therefore, avoiding errors due to the hysteresis phenomenon associated with first order phase transition between the liquid and crystal phases is a major challenge associated with all the above
methods. In addition to this, all these molecular simulation methods require the experimental crystal structure as an input. The process of estimating it involves generating all possible crystal structures and determining the most appropriate one based on specific criteria. The most favored criterion is the lattice energy. Based on this, it has been shown that various crystal structures have similar energies and subsequently melt at similar temperatures. Therefore, in theory, if the estimated or predicted crystal structure is close to the experimental one, the calculated melting point will also be a good estimation [38]. That being said, it is highly cumbersome and challenging to generate all possible crystal structures, especially for molecules with internal degrees of freedom. Therefore, these methods have a high computation time and limited accuracy.

The data revolution and advent of machine learning in chemical and material science has encouraged a number of quantitative structure property relationship (QSPR) based models have been for the prediction or correlation of melting points. Relative to the molecular simulation-based approaches, the QSPR methods are simple to implement. These models use molecular structure in the form of graphs to calculate molecular descriptors. The properties that are predicted are calculated as a function of these molecular descriptors. The relationship between the properties and descriptors may be linear or non-linear [40]. Considering the requirements, challenges and benefits of the above methods, in this thesis, we employ a data driven approach to predict melting points of DES systems.
1.4 Data Driven Approach

Recently there has been an emergence of successful machine learning (ML) or statistical learning type approaches being applied to the discovery and design of new materials. The application of ML in computational materials science is driven by the desire to achieve a more rational design of materials. Modern data science- the 4th pillar of science (Figure 1.5), is now unanimously being accepted into chemical and materials research. With remarkable similarities to the traditional field of quantitative structure activity relationship and drug design, the goal of ML based methods is to identify systems, molecules or materials exhibiting certain specific properties. In order to achieve this, ML based methods involve searching through large amounts of data for hidden relationships between the atomic structure and the property of interest [53]. Efforts such as the Materials Genome Initiative (MGI) have led to the creation of public data repositories like the Harvard Clean Energy Project [54], Materials Project [55], etc. These databases have been rapidly populated with the help of high throughput calculations on solid crystalline materials. Public and private data repositories

![Figure 1.5. The rise of data science as the 4th pillar of science.](image-url)
like the ones mentioned above are what traditionally provide the data for the ML models to search through. Along these lines, our group has developed a nearly self sufficient software ecosystem that incorporates all the major elements of materials design and informatics into an efficient intercommunicable framework.

Figure 1.6. Schematic of our group’s software ecosystem consisting of ChemLG, ChemHTPS, ChemBDDB and ChemML codes.
1.5 Group’s Software Ecosystem for Materials Discovery

Our group has identified four components that are critical for the efficient discovery of materials with targeted properties. These four components are being developed in the group as four different software packages [56] [57] [58]. (Figure 1.6)

1. **ChemLG**: A library generator[59].
2. **ChemHTPS**: A virtual high-throughput screening infrastructure[60].
3. **ChemBDDB**: A big data database infrastructure[61].
4. **ChemML**: A data analysis, mining and modeling infrastructure[62].

1.6 Summary

There has been an observable transition from ILs to DES in recent times, which has shed light on their innumerable potential applications. Therefore it has become essential to explore new DES and to study their properties in detail. In this thesis, we describe a data driven computational protocol designed to predict melting points of choline chloride based DES using 36 data points from literature. Our goal is to suggest novel DES systems which can be used as electrolytes in supercapacitors. The challenge is to ensure that these supercapacitors can be used at extremely low temperatures.
Methods

As shown in figure 2.1, our computational protocol is divided into two main parts, i.e., data collection and data mining.

2.1 Data Collection

While public datasets containing melting point data are available, few of them cater specifically to ionic liquids and DES systems. As explained in chapter 1, DES systems consist of a hydrogen bond donor component (HBD) and a hydrogen bond acceptor (HBA) component. The most widely used HBA system in literature was choline chloride. Most of the research concerning melting point of DES systems used a small number of HBD compounds with choline chloride. For the sake of simplicity, we decided to select 36 unique data points, with choline chloride as the HBA, from literature.
2.1.1 SMILES

Simplified molecular-input line-entry system, or SMILES is a method for chemical representation of molecular structures [63]. In this notation, atoms are represented by their atomic symbol and atoms within a ring are represented by the lower-case letters. We obtained the SMILES representation of the 36 DES systems in our data and used it for the remainder of our computational protocol.
2.1.2 Open Babel

Open Babel is a chemical tool box designed to speak the many languages of chemical data [64]. We use the SMILES representation of our 36 HBDs to generate their 3D geometries. These geometries are then optimized in the MMFF94s force field and the coordinates are saved in the xyz file format. XYZ file format represents molecular geometry by specifying the number of molecules and the cartesian coordinates of the compound. The MMFF94s force field was developed by the Merck Research Laboratories [65].

2.1.3 ORCA

The ORCA software written by F. Neese [66] along with many contributors, is an ab-initio, density functional theory (DFT), and semi-empirical self-consistent field-molecular orbital (SCF-MO) package. ORCA has the ability to perform geometry optimizations and predict a variety of spectroscopic parameters at different levels of theory. It includes several DFT functionals including hybrid DFT and double hybrid functions. We fed the optimized geometry information and xyz coordinates to ORCA and performed geometry optimizations using DFT. To demonstrate the effect of different levels of theory on our predictions, we carry out the optimizations using BP86 functional in the double zeta and PBE0 functional in the triple zeta basis sets respectively. The optimized geometries serve as the inputs used to generate our feature space for the ML models.
2.2 Data Mining

As discussed in section 1.4, data driven and machine learning based methods are the new tools used in materials discovery. We use machine learning to map our feature space to the target property. We use generalized linear models like LASSO, elastic nets, ridge, and bayesian ridge regression from 'scikit-learn' library. The linear models are a set of methods intended for regression in which the target value is expected to be a linear combination of the input variables.

2.2.1 Feature Representation

As mentioned in chapter 1, molecular descriptors contain structural information of molecules in a numeric form, that can be fed to the ML model. Selecting molecular descriptors can be highly challenging as the entire ML model rests on their credibility. Our group’s work on high refractive index polymers [56] and several other ongoing research endeavors showed that dragon descriptors perform well for the prediction of a number of physical properties. Therefore, we use the dragon application for the calculation of 2D and 3D molecular descriptors. The latest version of Dragon is 7.0 which calculates 5,270 molecular descriptors, covering most of the existing theoretical approaches. This list is divided into 30 blocks which contain simplest atom types, functional groups and fragment counts, topological and geometrical descriptors, 3D descriptors, and drug like and lead like alerts. Dragon 7.0 has been integrated into ChemML [62], which makes it easier to generate a clean and versatile comma separated values (csv) file which can easily be used as input for our ML models. Dragon 7 generated 2857 3D and 2D descriptors were generated for our model after weeding out similar or undesired descriptors.
2.2.2 Feature Selection

Out of all the descriptors generated by dragon, there is a possibility that some of those are not useful for predicting our target property. In order to rid our machine learning models of unwanted features, noise, and reduce training time, we generate a subset of more relevant features from our original feature space. Embedded into our dragon input script, is our first step of feature selection. This step ensures that any two features that are more than 95% correlated are removed. This removes redundant features from the feature set. We performed feature selection using pearson correlation coefficient and genetic algorithms.

2.2.2.1 Pearson Correlation

For feature selection, we compute the pearson correlation coefficient of the individual features with the melting point values. In this simple and inexpensive method, we find the correlation of individual features with the target values.

\[
r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]

In the above equation, \( r \) is the pearson correlation coefficient, \( n \) is the sample size, \( x_i \) and \( y_i \) are individual points indexed with \( i \), and \( \bar{x} \) and \( \bar{y} \) are the sample means. While this method finds the linear correlation between individual features and the target value, there is a possibility that a combination of features is what effectively stores relevant information for the specific problem. Therefore we move on to use a more reliable method for feature selection using genetic algorithms.
2.2.2 Genetic Algorithms

Genetic Algorithms (GA) is a popular search based method, which is based on a natural selection process that mimics biological evolution. GAs are one of the most well-known evolutionary algorithms and have been applied to a variety of engineering optimization problems. When GAs are used for feature selection, the features that have the largest projections in a lower dimensional space are selected.

Figure 2.2 shows the general design of a GA. In our case, we use a binary
encoded GA from ChemML [62] having 50 chromosomes in the initial population. We use the two point crossover method with 80% probability of crossover and mutation each. We run this GA for 50 generations while evaluating the mean absolute error of the ML model generated using each chromosome as the feature space.

2.2.3 Machine Learning Models

2.2.3.1 Ridge Regression

This model solves a regression model where the loss function is the linear least squares function and regularization is given by the l2-norm. This estimator has built in support for multi-variate regression. The model is highly sensitive to $\alpha$, which is a complexity parameter and controls the amount of shrinkage, i.e., controls the size of the coefficients of the mathematical objective function.

2.2.3.2 Bayesian Ridge Regression

In case of bayesian regression, the regularization parameters are included in the estimation procedure. In other words, it is set rigidly but is tuned to the data at hand. Bayesian Ridge Regression estimates a probabilistic model of the regression problem.

2.2.3.3 LASSO

The LASSO is a linear model that estimates sparse coefficients. It is useful in some contexts due to its tendency to prefer solutions with fewer parameter values, effectively reducing the number of variables upon which the given solution is dependent. Therefore, the LASSO and its variants are fundamental to the field
of compressed sensing. The model is trained with L1 prior as the regularizer, i.e., least absolute deviation. It uses coordinate descent as the algorithm to fit the coefficients of the objective function.

2.2.3.4 Elastic Net

Elastic Net is a linear model trained with L1 and L2 prior as regularizer. This combination allows for learning a sparse model where a few of the weights are non-zero. Therefore, it behaves like LASSO, while still maintaining the regularization properties of Ridge regression and inheriting its stability under rotation. This is useful when multiple features are correlated with one another. While LASSO would pick one of the two correlated features at random, Elastic Net picks both.
Results and Discussion

3.1 Feature Representation and Feature Selection

In chapter 2 we introduced a computational protocol to predict the melting point of DES using a data driven approach. As mentioned earlier we chose 36 unique data points from literature. We used the SMILES code of these compounds to generate optimized 3D geometries within the DFT framework using the PBE0 functional and def2-TZVP basis set. While this method was more accurate, it was computationally expensive. Therefore, in addition to this we also used BP86 functional with def2-SVP basis set and MMFF94s force field as relatively inexpensive methods to generate optimized geometries. This structural information was then used to generate the feature space for our regression models. The final results from bayesian ridge regression, trained using the feature spaces generated from all of the above methods is shown in figure 3.1. The mean absolute error (MAE) of these models were 22.71°C, 35.36°C, and 46.37°C for TZVP, SVP and MMFF94s respectively. This clearly shows that the geometrical configurations generated from the PBE0/def2-TZVP method captured the
Bayesian ridge regression with top features from genetic algorithm where the 3D geometries were generated using different methods

Figure 3.1. Comparison between the feature space generated using 3D geometries from MMFF94s force field, BP86 Def2-SVP DFT and PBE0 Def2-TZVPD DFT frameworks.

structural information of the candidate compounds more accurately than the others. Details about our machine learning model are provided in section 3.2.

For feature selection, we first calculated the Pearson correlation coefficient of individual features with the melting point values. Figure 3.2 (a) shows how all the features are correlated with the melting point values and 3.2 (b) and (c) show the features that have the best correlation.

While Pearson correlation is a relatively inexpensive method of finding correlated features, it is possible that a combination of two relatively uncorrelated features has a major effect on the target value. Therefore, we employed genetic algorithms (GA) to find the combination of features that produced the most accurate melting point predictions for all our individual machine learning models. The particulars of our GAs are discussed in section 2.2.2.2. In the fitness
Figure 3.2. (a) Pearson correlation of dragon descriptors with melting point values. (b) & (c) Descriptors with Pearson correlation of above 0.5 and below -0.5, i.e., features with the best Pearson correlation.

function, we minimized the mean absolute error (MAE) of leave one out cross validation.

Figure 3.3 compares the results of one of our machine learning models, i.e., ridge regression, which uses GA for feature selection, to the same model which uses Pearson correlation for feature selection. We ensured that the number of
features used in both cases were the same. As shown in figure 3.3 the model that uses GA for feature selection yields an MAE of 18.45°C for the test set and the same model generates an MAE of 22.78°C when it uses the top features selected from the pearson correlation method.

### 3.2 Machine Learning

For our machine learning models, we first split the data into training and test sets of 32 and 4 data points each. We then use leave one out type cross validation for the data points in the training set, wherein, the model is trained using 31 data points and tested on the remaining single data point. This process is repeated such that model is tested on each of the 32 data points at least once. In other words, leave one out type cross validation uses the maximum possible data to train the model. In general this method is preferred for small amounts of data like ours, as it can prove to be computationally expensive when used for
Comparing results of linear regression models with top features from genetic algorithm where 3D geometries were generated using PBE0/def2-TZVP

Ridge regression

Elastic Net

Bayesian ridge

LASSO

MAE(test) = 18.45°C

MAE(test) = 26.31°C

MAE(test) = 22.71°C

MAE(test) = 24.81°C

Figure 3.4. Comparison between pearson correlation and genetic algorithm for feature selection for the bayesian ridge regression model.

a large dataset. Since some of our machine learning models are sensitive to the regularization parameter, $\alpha$, we then perform a grid search to select its optimum value. The model is then tested on the remaining 4 unseen data points. We then computed statistical error measures like mean absolute error (MAE). In order to produce more reliable results, we repeated this for 10 different test sets of 4 data points and averaged the MAEs of each model.

Figure 3.4 shows a comparison between four different generalized linear regression models from the `scikit-learn` library. Each model was trained on the same 32 data points and tested on the same 4 unseen data points after leave one
out cross validation. Figure 3.4 clearly shows that the bayesian ridge and ridge regression models both trained and tested better than the elastic net and the LASSO methods. It was observed ridge regression yielded an MAE of 18.45°C, elastic net had an MAE of 26.31°C, bayesian ridge regression had an MAE of 22.71°C and the LASSO model had an MAE of 24.81°C. All the MAE values are for the 4 unseen data points in the test set. This showed that ridge regression yielded the least MAE values compared to others.

3.3 Conclusion and Discussion

Now that we have developed a fairly robust protocol for the prediction of melting points of DES systems, it is fair to identify some of the challenges associated with it. To begin with, the melting point is a relatively complicated property to predict and model. This is due to its dependence on a variety of factors including structural arrangement of molecules in a crystal, inter and intra molecular hydrogen bonding, polymorphism, motion of molecules in crystals, and the strength of crystal lattices. Further, the strength of a crystal lattice depends on molecular symmetry, shape and size of molecules. These dependencies make it this a complicated problem and pave the way for a relatively inexpensive solution involving data driven models.

This brings us to the challenges associated with data driven approaches and protocols to model the melting point of such intricate systems like DES. Due to the complicated nature of modelling melting point in the first place, there is a lack of melting point data existing in literature. While there are a few datasets and libraries for melting point data, none of them cater to the likes of complex organic systems like DES. With a lack of data, it is always challenging to pro-
duce reliable machine learning models to predict a desired target property. The second challenge associated with data driven models is the apparent lack of reliable descriptors to capture complicated physical phenomena like hydrogen bonding.

Given that melting point is a complicated property to model and the fact that there is a lack of melting point data for DES systems, the results of our models are relatively accurate and reliable. These models can also be considered as a potentially good first step towards the design and discovery novel DES systems for numerous industrial applications.
Bibliography


