TOOLS FOR COMPUTER-AIDED MOLECULAR AND MIXTURE DESIGN

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Chemical Engineering

Nick D. Austin

Carnegie Mellon University
Pittsburgh, PA

Tools for Computer-Aided Molecular and Mixture Design, Ph.D. Thesis

ABSTRACT

This thesis explores mathematical optimization techniques to address the computer-aided molecular and mixture design problems (CAMD/CAMxD). In particular, we leverage the power of mixed-integer linear programs (MILPs) to quickly and efficiently design over the massive chemical search space. These MILPs, when coupled with state-of-the-art derivative-free optimization (DFO) methods, make for an efficient optimization strategy when designing mixtures of molecules or when considering a single molecule design problem that involves difficult thermodynamics or process models.

In the first chapter, we provide a very general overview of the field of CAMD as addressed from the perspective of mathematical optimization. We discuss many relevant quantitative structure-property relationships (QSPRs) and provide constraints typically used in CAMD/CAMxD optimization problems.

The second chapter introduces our DFO-based molecular/mixture design algorithm and describes how this approach enables a much greater molecular diversity to be considered in the search space as compared to traditional methods. Additionally, this chapter looks at a few case studies relevant to crystallization solvents and provides a detailed comparison of 27 different DFO algorithms for solving these problems.

The third chapter introduces COSMO-RS/-SAC as alternatives to UNIFAC as the method used to capture mixture thermodynamics for a variety of CAMD/CAMxD problems. To fully incorporate COSMO-RS/-SAC into CAMD, we introduce group contribution (GC) methods for estimating a few necessary parameters for COSMO-based methods. We demonstrate the utility of COSMO-RS/-SAC in a few case studies for which UNIFAC-like methods are insufficient.

In the fourth chapter, we investigate reaction solvent design using COSMO-based methods. COSMO-RS is particularly suitable for these problems as they allow for modeling of many relevant species in chemical reactions (transition states, charges, etc.) directly at the quantum level. This information can be immediately passed to the CAMD problem. We investigate a number of solvent design problems for a few difficult reactions.

We summarize the work and provide a few future directions in the final chapter. Overall, this thesis serves to push the field of CAMD forward by introducing new methods to more efficiently explore the massive chemical search space and to enable a few new classes of problems which were previously untenable.

"I think I did pretty well, considering I started out with nothing but a bunch of blank paper..."
— Steve Martin, on writing
Dead Men Don't Wear Plaid (1982)

ACKNOWLEDGMENTS

I am sincerely indebted to my advisor, Prof. Nick Sahinidis, for his constant support and guidance throughout my PhD. Nick, thank you for giving me the freedom to pursue a wide range of research topics, for your knowledge and assistance whenever I needed it, and for your unwavering commitment to my professional career. I am a stronger researcher because you were so open to my exploring almost every idea I had (even the ones you had to know were bad), and I have been beyond fortunate to have had an advisor who took such an active and caring interest in my development as an engineer and a member of the academic community. I've learned a lot from you: about getting to the heart of a new concept, about effective communication as the true currency of one's ideas, about the power of intuition, and about patience as the key to solving difficult problems. I also never mentioned how much I like the Nick-symmetry of our emails.

I would also like to thank the members of my doctoral committee: Prof. Ignacio Grossmann, Prof. Chrysanthos Gounaris, Prof. Dave Yaron, and Dr. Daniel Trahan. All of you have have been valuable influences throughout my stay at CMU and your guidance along the way has contributed significantly to this work. Thanks in particular to Dave, who kept me in touch with my chemistry roots, provided great suggestions/references about a variety of topics, and was always a pleasure to talk to.

I am also very grateful to The Dow Chemical Company for their financial support of my research. It goes without saying that I simply would not have had the same PhD without their support. A particular thanks goes to Dr. Daniel Trahan, my primary collaborator and point of contact with Dow. Daniel, you've been an incredible collaborator and mentor, and I've very much enjoyed working with you and getting to know you over

the last 5 years. You are always full of good ideas and are an impressively quick study of many things. You are also perhaps the only person I've met who can fully commiserate with my woes as a wrongful edit-ee of academic publishers. An additional thanks goes to Dr. Ivan Konstantinov (also of Dow) for his thoroughness and valuable assistance in setting up no small number of quantum calculations.

My time at CMU would not have been nearly as fun without the many, many great friends I've had. Thanks to my friends in the classes above me—Satya, Yash, Francisco, German, Pablo, Bethany, Antonis, and Javier—for showing me the ropes, always inviting me to hang out, playing foosball, watching movies, and skiing. A particular thanks goes to Satya, who has been a great tennis partner and fellow trivia addict. Also, thanks to both Satya and Yash for introducing me to real Indian food.

I'm also very lucky to have had a lot of great classmates in my year. Thanks to Rob for his sense of humor, Jake for always playing any sport with me, Wei for all the fun we've had and for a proper introduction to Chinese food (except for Baijiu–I'm not thankful for that), and Sree for all the visits to Tamarind and for being a great rock climbing partner. Thanks also to Sree for being the only guy I know who would do graph theory homework with me at 3am at Eat'n Park.

Thanks to the many other friends I've had here at CMU. Thanks to John for all the movies, racquetball, and good conversations. I'm glad I found someone in Pittsburgh with such a similar movie/restaurant taste to my own. Thanks to Annia for always having tea and staying up too late. I've also enjoyed all the fun I've had in the office with Sree, Zach, Mustafa, and Juan. A thanks goes to all of the other friends I've had in Pittsburgh: Caro, Steve, Vince, Devin, Justin, Anirudh, Chris, Nikos, Jens, Alex, and so many of the others I'm probably forgetting. You guys made my time here much more enjoyable and were always up for a visit to a local restaurant or bar... you know, nothing too far away.

Finally, I owe a thanks to my parents, Bob and Lynn, for their constant love and support throughout the entire PhD process and everything leading up to it. I am equally thankful to my brothers, Jeff and Greg, for making everything up to this point one very fun and memorable "boys' trip." I am finally very thankful to Clara for her support and compassion and for being one of the most genuinely kind people I've ever met. I'm truly grateful to have all of you in my life.

CONTENTS

LIST OF FIGURES AND ILLUSTRATIONS

LIST OF TABLES

1

INTRODUCTION

1.1 INTRODUCTION

The application of chemistry to manipulate the natural world has its earliest examples in metalworking and pottery (Partington 1970), with some pottery artifacts as old as 20,000 years Wu et al. (2012). Chemical products have since played an important role in history and left an indelible mark on the way we live and work. Early history included basic incendiary fuels, perfumes, and soap as some of the first widespread uses of chemicals. The modern age has witnessed an unprecedented expansion of chemical products, including pesticides, fuels for transportation and electricity, pharmaceuticals, plastics, and a broad array of industrial and consumer products. True to these historical trends, few things are as pervasive as chemical products in 21st century life, and there are an ever-increasing number of new chemical applications which require specialized compounds.

The process of determining new and suitable chemicals for a certain application can be generally termed chemical product design (Cussler & Moggridge 2011). Chemical product design has long been a laborious, trial-and-error procedure, limited often by a fixed amount of chemical, time, and financial resources. Design efforts are often high-throughput and tend to focus on a small class of compounds or structural analogues of known chemicals. Accordingly, the so-called "design space"—the set of unique molecular structures considered—is often quite small for these product design problems, especially considering the massive design space of all possible chemical structures. It is then clear that to keep pace with the growing demand for new chemical products and to adequately explore the full chemical design space, other approaches must be considered. Fortunately, the availability and efficiency of computational resources makes these design problems more tenable than ever before. Noteworthy among computational approaches is the field of computer-aided molecular design (CAMD), which leverages the simplicity of semi-empirical quantitative structure-property relationships (QSPRs) in conjunction with fast and efficient numerical optimization algorithms.

CAMD has its roots in the 1980's, although the general use of computers in chemistry pre-dates this by a few decades. Stated formally, the CAMD problem concerns designing an optimal molecular structure(s) for a certain application. CAMD combines molecular modeling techniques, thermodynamics, and numerical optimization to design good or optimal molecular structures, many of them often completely novel. Advances in chemical modeling in the last few decades have greatly benefited CAMD, and practitioners are now capable of relating chemical structures to properties at several levels of accuracy (molecular mechanics, semi-empirical, *ab initio*). Though CAMD often uses semi-empirical modeling techniques for their simplicity and efficiency, new approaches incorporating more accurate methods are emerging. Modern combinatorial optimization techniques are also essential for CAMD, enabling the optimization over staggeringly large design spaces which would otherwise be inaccessible (using enumeration algorithms, for example).

Overall, this introductory chapter provides a description of popular QSPRs used in CAMD, the CAMD problem itself, and several solution approaches to its various forms. We begin in Section 1.2 by detailing three popular classes of QSPRs which are often used in CAMD: (1) group contribution methods; (2) topological indices; and (3) signature descriptors. Next, in Section 1.3, we present the CAMD problem from a mathematical programming perspective, discussing various classes of the single-molecule design problem as well as CAMD problems considering mixtures of molecules and those involving the simultaneous design of a chemical product and the process it is a part of. In this section, several other important design considerations are presented, including a few important constraints to ensure practical solutions as well as the chemical feasibility of the designed structures. In Section 1.4, various solution techniques for the CAMD problem are discussed, including mathematical optimization strategies, decomposition methods, and heuristic approaches. Finally, in Section 1.5, a diverse though non-exhaustive review of applications of CAMD problems is provided. The interested reader can find a more thorough overview of CAMD in Austin et al. (2016b).

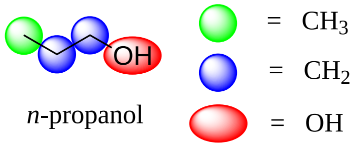
1.2 POPULAR TYPES OF QSPRS IN CAMD

The CAMD problem attempts to choose optimal (or simply good) molecules for some purpose from the space of theoretically possible chemical structures. At first glance, the CAMD problem must consider a very abstract chemical design space of atoms, bonds, aromaticity, structural isomers, electronic effects, etc. Though many of these features are certainly what gives molecules their specific properties and chemical functionality, they are difficult to build into any type of optimization scheme. This is primarily because there is no immediate relationship between an arbitrary chemical structure and its performance or suitability regarding a specific application. In order to "rank" different structures and choose an optimal one, we must have some efficient way to quantify the properties and performance of each structure.

A second issue is the sheer size of the chemical search space. At the time of writing this chapter, the CAS registry (American Chemical Society 2017) reports over 115 million unique organic and inorganic structures. This number only represents compounds which have been synthesized and cataloged, and it is already far too large for every structure to be considered in any type of trial-and-error design scheme. This number is also only a fraction of the theoretically possible chemical space, which some estimates indicate may contain more than 10^{60} unique molecules for small, drug-like structures (Bohacek et al. 1996). Even with very efficient ways to estimate the performance of a certain structure, screening these structures using an enumeration strategy is far beyond current computational capacity. For this reason, we also need to relate the chemical space to a space that can be utilized for combinatorial optimization, allowing us to design over the massive search space far more efficiently.

CAMD practitioners have relied on semi-empirical quantitative structure property relationships (QSPRs) to address both of these issues. First, many semi-empirical methods delineate a clear connection between the abstract chemical space and the more practical space of quantitative properties. These methods are also often simple and can be applied to estimate properties very efficiently. Second, many of these methods break

Figure 1.1: Propanol represented by its groups



molecular structure into sub-molecular collections of atoms and bonds. These molecular sub-structures are assumed to dictate a molecule's properties. Using these types of representations of the molecular space, combinatorial optimization can be directly applied to these design problems.

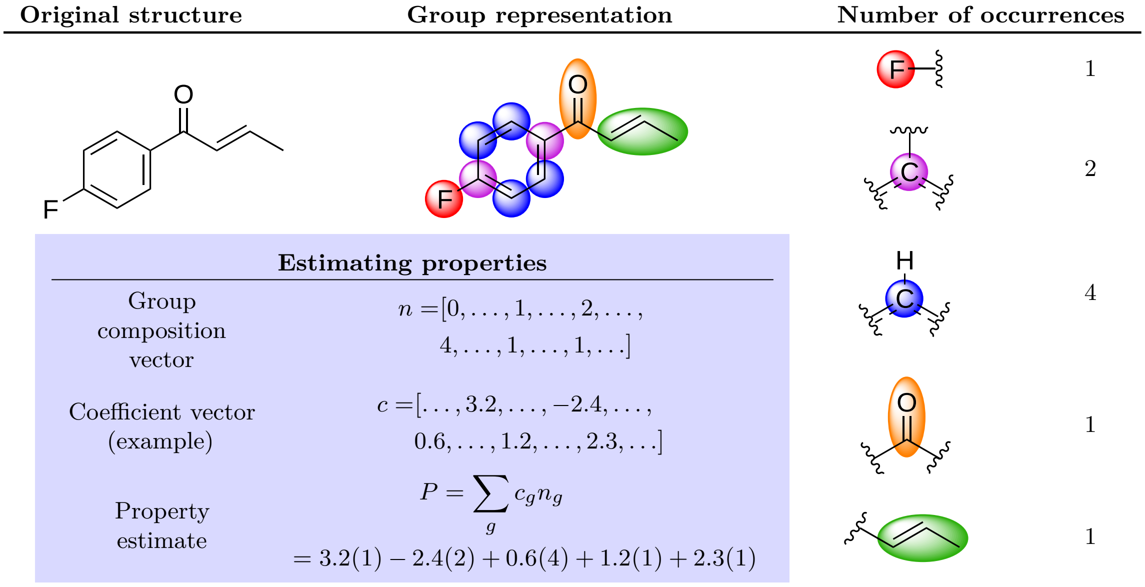
1.2.1 *Group-contribution methods*

The most commonly used QSPRs in CAMD are group contribution (GC) methods. These work under the assumption that a molecule's properties can be predicted by the number of occurrences of various molecular sub-structures called "groups." For example, we may think to represent the simple molecule propanol as a combination of the groups $-CH_3$, $-CH_2-$, and $-OH$. In this case, the dashes $(-)$ represent bonds to other groups. In its group representation, propanol would no longer be thought of as the connected alcohol molecule, but rather as some collection of its constituent groups. The group representation of propanol is shown in Fig. 1.1.

Being QSPRs, group contribution methods translate the group representation of a molecular structure into an estimate for some property P. To do this, group contribution methods define a vector n that represents the number of occurrences of each of the groups. Assuming we only have the three groups shown above, propanol's n vector would be $n = [1, 2, 1]$, where the entries in this vector represent the number of occurrences of the groups $-CH_3$, $-CH_2-$, and $-OH$, respectively. Each of these groups g would also be associated with a coefficient c_g which quantifies its affect or "contribution" to a particular property P. Properties are calculated as follows:

$$P = \sum_g c_g n_g \tag{1.1}$$

Figure 1.2: Example usage of group contribution methods



Estimating properties

Group composition vector	$n = [0, \ldots, 1, \ldots, 2, \ldots, 4, \ldots, 1, \ldots, 1, \ldots]$
Coefficient vector (example)	$c = [\ldots, 3.2, \ldots, -2.4, \ldots, 0.6, \ldots, 1.2, \ldots, 2.3, \ldots]$
Property estimate	$P = \sum_g c_g n_g = 3.2(1) - 2.4(2) + 0.6(4) + 1.2(1) + 2.3(1)$

The vector of coefficients c comes from regression over a large dataset of the property P of different molecules. To regress these parameters, the identity of all of the groups must be specified *a priori*. Returning to our example, one can easily imagine different sets of groups being used to describe propanol. For example, the groups $-CH_3$, $-CH_2-$, and $-CH_2OH$ also completely account for the atoms in propanol, and these may provide a better fit for the regression problem. For this reason, different group contribution methods to estimate different properties usually do not have completely consistent sets of groups, although there is typically a large amount of similarity. Finally, we note that the vector n is generally much bigger as many group contribution methods contain 50-100 groups. Many group contribution methods make the additional assumption that groups cannot overlap, which means that n is typically a sparse vector. A pictorial example of the usage of group contribution methods is given in Fig. 1.2. In this example, we apply a hypothetical GC method to a hypothetical molecule. We show how a molecular structure is decomposed into its constituent groups and provide a count of each of these groups. These counts constitute the elements of the vector n. The n vector is paired with a hypothetical c vector, and an example property is calculated.

One of the earliest examples of GC methods is from Benson & Buss (1958), who are considered to be the originators of so-called “group increment theory.” Group increment theory, or Benson group increment theory (BGIT), is analogous to GC methods, but these terms may be more common in the physical chemistry literature. In the original 1958 paper (Benson & Buss 1958), Benson and Buss proposed a simple group additivity scheme for the prediction of bond dissociation energies. Benson et al. (1969) extended this work to account for a greater diversity of groups and to estimate heat capacities. Additional work from Cohen and Benson includes estimating heats of formation with group increment theory (Cohen & Benson 1993). A large number of additional efforts have used Benson-like increments to estimate the same thermophysical properties, a very small sample of which are provided here: Domalski & Hearing (1988); Jalowka & Daubert (1986); Roganov et al. (2005).

Another very popular GC method was devised by Joback & Reid (1987). This method extended the group increment idea to model many different properties with the same set of groups. The Joback and Reid model also included functional transformations for the original group increment summations. These altered the group contribution definition to the following:

$$P = f\left(\sum_{g} c_g n_g\right) \tag{1.2}$$

where f represents some function of the inner product of the vectors c and n. These functions f appear in many group contribution methods and are important when predicted properties are not simple linear functions of the number of groups in a structure.

Perhaps the most widely-used GC method in CAMD is that of Marrero and Gani (Marrero & Gani 2001; 2002). This method, like the method of Constantinou & Gani (1994) that predated it, provides another extension to the general form of GC methods in that it introduces multiple levels of groups to better capture proximity effects, meaning the effect of two or more groups which are close to one another in a molecular structure. As many GC methods are limited to groups of just a few atoms and bonds, many cannot differentiate between structures with different connectivity. The Marrero-Gani method, called the GC+ method, uses as a first-order approximation a normal group contribution method, where the groups belong to a set of primary groups F. An additional set of groups S contains slightly larger sub-structures. Finally, a set of groups T accounts for large groups and overarching molecular structural features. Unlike groups in the set F, groups in the sets S and T are allowed to overlap with each other. Using the hierarchical depiction of molecules with the GC+ method, a much clearer picture of a molecule is provided. The general form of the GC+ estimates is shown below:

$$P = f\left(\sum_{g\in F} c_g n_g + \sum_{g\in S} c_g n_g + \sum_{g\in T} c_g n_g\right) \tag{1.3}$$

Group contribution methods have also often incorporated interaction terms (Nannoolal et al. 2004; 2007; 2008; 2009; Klincewicz & Reid 1984; Platts et al. 2000). These are ways to include additional terms to account for the simultaneous presence of two (same or different) groups in a particular structure. For example, in predicting toxicity, one group in a molecule may lead to a simple metabolic pathway and therefore make many structures containing that group non-toxic. If that structure were also to have another group which is normally quite toxic, a GC method without interaction terms may not predict toxicity well. This would be because both groups would have an additive effect, meaning that there would be one highly non-toxic contribution and one highly toxic contribution. As a result, the molecule may be predicted to have an average toxicity when, in reality, the presence of the non-toxic group should outweigh the toxic group. Introducing an interaction term for these two groups accounts for the situation of their co-occurrence. In this example, this interaction term would likely remove whatever toxicity value was predicted by the toxic group. For more discussion of interaction terms in predicting toxicity, readers are referred to Martin & Young (2001). Interaction terms generally take the form

$$I_{g,g'} = f_I\left(n_g, n_{g'}\right) \tag{1.4}$$

where f_I usually represents multiplication, but can also represent other functions. Group contribution methods can also include some idea of structural features (Nannoolal et al. 2004; 2007; 2008; 2009; Marrero & Gani 2001; 2002). These account for larger effects, typically at the molecule scale, such as aromatic ring substitution, cis/trans isomerism, aliphatic chain lengths, etc. These can typically be implemented as large groups but sometimes require special considerations. Table 1.1 provides several properties typically used in CAMD problems and a few GC methods to estimate them.

Strengths of GC methods. GC methods are useful in that they are very intuitive to use. They represent a chemical structure in terms of its functional components, very analogously to how chemists compare and analyze structures. GC methods are also able to easily represent a large and diverse chemical space as the groups can be combined in many different ways to produce a large variety of different structures. This is especially useful from a CAMD perspective. Finally, GC methods are easily translated into the mathematical formulations of CAMD problems as the inclusion and count of the groups (the vector n) are easily represented in the context of mathematical optimization.

Weaknesses of GC methods. There are a few shortcomings of modern GC methods. One is that many GC methods are unable to distinguish isomers from one another. As isomers can have very different properties, this represents a gap in the predictive power of GC methods. We note that some GC methods such as the GC+ methods are able to distinguish many isomers due to the inclusion of large groups. A second issue with GC methods is the lack of consistency in groups used to predict various properties. Though this has no major effect estimating these properties for a given structure, it becomes problematic for mathematical formulations of the CAMD problem. Finally, GC methods require specifying the set of groups prior to regressing the GC coefficients. Though many GC methods are quite accurate, there is no guarantee that the set of groups used best captures the property they model. Using different groups can sometimes drastically alter the predictive power of a GC model.

1.2.2 *Topological indices*

Chemical graph theory (Bonchev 1991) is a field which became very influential in the 1970's and has since been used to produce a large number of QSPRs. The basic idea of chemical graph theory is that the atoms and bonds which constitute a molecule can be thought of as nodes and edges in a graph. In general, we use $G = (V, E)$ to define a graph G, its vertices $v \in V$ and its edges $e \in E$. Using this depiction of molecular structures, various properties of that graph, referred to here as topological indices, can be used as descriptors in QSPR models. More specifically, this means that

Table 1.1: Sample of available GC methods for predicting various properties of pure compounds

Property	**GC methods**
Aqueous solubility	Marrero & Gani (2002), Klopman & Zhu (2001)
Boiling point	Joback & Reid (1987), Stein & Brown (1994), Nannoolal et al. (2004), Marrero & Gani (2001)
Bond dissociation energy	Benson & Buss (1958)
Critical pressure	Jalowka & Daubert (1986), Joback & Reid (1987), Klincewicz & Reid (1984), Nannoolal et al. (2007), Marrero & Gani (2001)
Critical temperature	Jalowka & Daubert (1986), Joback & Reid (1987), Klincewicz & Reid (1984), Nannoolal et al. (2007), Marrero & Gani (2001)
Critical volume	Klincewicz & Reid (1984), Nannoolal et al. (2007), Marrero & Gani (2001)
Enthalpy of formation	Cohen & Benson (1993), Benson (1999), Domalski & Hearing (1988), Roganov et al. (2005), Joback & Reid (1987), Marrero & Gani (2001)
Enthalpy of fusion	Joback & Reid (1987), Marrero & Gani (2001)
Enthalpy of vaporization	Roganov et al. (2005), Joback & Reid (1987), Marrero & Gani (2001), Ceriani et al. (2009)
LC_{50}	Martin & Young (2001) (fathead minnow)
Melting point	Joback & Reid (1987), Marrero & Gani (2001)
Gibbs energy of formation	Cohen & Benson (1993), Benson (1999), Domalski & Hearing (1988), Roganov et al. (2005), Joback & Reid (1987), Marrero & Gani (2001)
Heat capacity	Benson et al. (1969), Benson (1999), Domalski & Hearing (1988), Joback & Reid (1987), Kolská et al. (2008), Ceriani et al. (2009)
Octanol/water partition coefficient	Marrero & Gani (2002), Platts et al. (2000), Klopman et al. (1994)
Vapor pressure	Nannoolal et al. (2008)
Viscosity	Joback & Reid (1987), Sastri & Rao (1992), Ceriani et al. (2007), Cao et al. (1993), Nannoolal et al. (2009)

various topological indices are paired with regression coefficients and used to estimate properties in a similar way to GC methods.

Topological indices (TIs) can take many forms. They are defined as some function of the nodes and edges in a chemical graph, and one can easily see that there are a large number of possible functions even just considering standard molecular graph properties like degree counts for nodes, connectivity, atomic types, etc. One of the first topological indices used in chemical graph theory is the Wiener index (Wiener 1947). The Wiener index attempts to describe the total distance between all atoms in the graph, as given by $d(v, v')$, the graph theoretic distance between vertices v and v'. The Wiener index $W(G)$ is defined as:

$$W(G) = 1/2 \sum_{v,v'} d(v, v') \tag{1.5}$$

While the Wiener index describes a graph in terms of its distances, another important consideration is how a graph is connected. To address this, an important class of topological indices called connectivity indices (CIs) was developed. Connectivity indices are widely used in CAMD and have been shown to be useful in QSPR applications (Estrada & Rodríguez 1999). The first connectivity indices were developed by Randić (Randic 1975), who used these indices to account for the degree of branching in alkanes and to model enthalpy of fusion and vapor pressure. Randić defined an edge index to be:

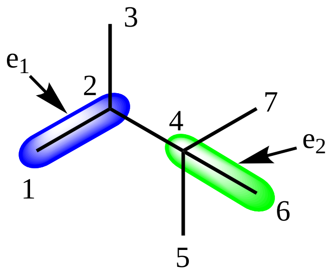
$$CI_E(v, v') = \frac{1}{\sqrt{\delta_v \delta_{v'}}} \tag{1.6}$$

where v and v' are two connected vertices in the chemical graph. This means that the atoms which correspond to v and v' are connected by a chemical bond. Furthermore, δ_v and $\delta_{v'}$ are the degrees of nodes v and v'. In the study of Randić, these degrees signified the number of σ bonds a particular atom had to non-Hydrogen atoms (i.e., the number of atomic neighbors in the hydrogen-suppressed graph), but they are sometimes defined differently for other connectivity indices. The connectivity index of the entire molecule (graph) was then given by

$$^1\chi = \sum_{\{v,v'\}\in E} CI_E(v, v') = \sum_{\{v,v'\}\in E} \frac{1}{\sqrt{\delta_v \delta_{v'}}} \tag{1.7}$$

Figure 1.3: Randić edge indices for a simple alkane

Structure



Calculating $CI_E(v, v')$

Atom Number (v)	1	2	3	4	5	6	7
δ_v	1	3	1	4	1	1	1

CI_E for two edges

$$CI_E(e1) = CI_E(1,2) = \frac{1}{\sqrt{1 \times 3}} = 0.577$$

$$CI_E(e2) = CI_E(4,6) = \frac{1}{\sqrt{4 \times 1}} = 0.500$$

where E is the edge set of the graph. A calculation for the edge connectivity indices for a simple alkane is shown in Fig. 1.3.

${}^1\chi$ is the so-called first-order Randić connectivity index. An even simpler "connectivity index" exists which does not account for bonding at all. This is called the zeroth-order connectivity index, and is given below:

$${}^0\chi = \sum_{v \in V} \frac{1}{\sqrt{\delta_v}} \tag{1.8}$$

Kier et al. (1975), Hall et al. (1975), and Murray et al. (1975) were the first to apply these connectivity indices as descriptors in QSPR models. These models are typically linear in the descriptors, but many variations exist. An important step forward came from Kier et al. (1976) who introduced higher-order connectivity indices. In general, these are defined for an index of order i as:

$${}^i\chi = \sum_{\{v_1, v_2, \ldots\} \in V_C^i} \frac{1}{\sqrt{\prod_v \delta_v}} \tag{1.9}$$

where V_C^i is the set of all sets of i connected vertices. Kier & Hall (1976) developed additional connectivity indices to account for heteroatoms. These modified connectivity indices distinguished atoms by their valence electrons, leading to a new value for vertex degree, δ^{V}, where the superscript indicates that valence is considered. For second period elements, Kier & Hall (1976) define δ_v^{V} for a vertex v as

$$\delta_v^{\mathrm{V}} = Z_v^{\mathrm{V}} - h_v \tag{1.10}$$

where Z_v^{V} indicates the number of valence electrons for atom/vertex v and h_v is the number of hydrogens attached to v. For atoms in the third period and beyond, the main difference from this perspective is the number of core electrons. These are accounted for in the following:

$$\delta_v^{\mathrm{V}} = \frac{Z_v^{\mathrm{V}} - h_v}{Z_v - Z_v^{\mathrm{V}} - 1} \tag{1.11}$$

where Z_v is the atomic number of atom v. These modified δ^{V} values define analogous connectivity indices ${}^i\chi^{\mathrm{V}}$.

The connectivity index χ can be thought to primarily capture vertex adjacency and the local neighborhoods of every atom in a molecule. As such, it represents how atoms are connected and perhaps not the ensemble molecular structure and shape. Another topological index often used in CAMD that aims to address this is the so-called shape index, κ. The shape index accounts for features of the entire molecular structure as functions of the underlying graph architecture and counts of graph substructures. The main parameter in the calculation of κ is the number of paths in the chemical graph of a certain length. We define ${}^iP$ to be the number of paths of length i in a particular chemical graph.

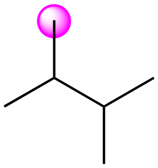
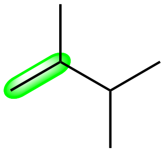
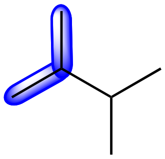
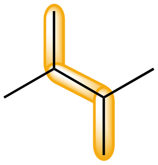
Like other topological indices, the shape index maps this information onto a single value. There are three often used ${}^i\kappa$ values which are defined as:

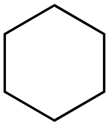
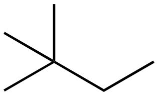
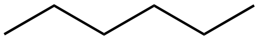
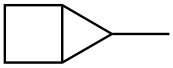
$$ {}^1\kappa = \frac{2({}^1P_{\max})({}^1P_{\min})}{({}^1P)^2} \tag{1.12}$$

$$ {}^2\kappa = \frac{2({}^2P_{\max})({}^2P_{\min})}{({}^2P)^2} \tag{1.13}$$

$$ {}^3\kappa = \frac{4({}^3P_{\max})({}^3P_{\min})}{({}^3P)^2} \tag{1.14}$$

Figure 1.4: Getting ^iP and calculating $^i\kappa$ for an example molecule

Number of atoms	Paths of length 1	Paths of length 2	Paths of length 3
			
^0P: 6	^1P: 5	^2P: 6	^3P: 4
Calculating κ values	$^1\kappa = \frac{2(^1P_{\mathrm{max}})(^1P_{\mathrm{min}})}{(^1P)^2} = \frac{2(30)(5)}{5^2} = 12$	$^2\kappa = \frac{2(^2P_{\mathrm{max}})(^2P_{\mathrm{min}})}{(^2P)^2} = \frac{2(10)(4)}{6^2} = 2.22$	$^3\kappa = \frac{4(^3P_{\mathrm{max}})(^3P_{\mathrm{min}})}{(^3P)^2} = \frac{4(4)(3)}{4^2} = 3$

Other structures	^0P	^1P	^2P	^3P	Other structures	^0P	^1P	^2P	^3P
	6	6	6	6		6	5	7	3
	6	5	4	3		6	7	11	13

with $^iP_{\mathrm{max}}$ and $^iP_{\mathrm{min}}$ representing the maximum and minimum possible number of paths of length i for a hypothetical molecule with an equivalent number of atoms. $^iP_{\mathrm{max}}$ and $^iP_{\mathrm{min}}$ can be easily derived from graph theoretic arguments. The formal expressions for these can be found in Hall & Kier (2007).

In Table 1.2, we list a few references for QSPR models using topological indices. There are many, many examples of such models, and we note that this table is not meant to be exhaustive. For a more complete list of available topological-indices-based QSPRs, the reader is directed to the book of Devillers & Balaban (2000).

Table 1.2 provides some idea of the diversity of applications of QSPR development with topological indices. We note that this table is inclined towards properties common in CAMD problems, so the properties listed may be more relevant to chemical engineering applications. However, we provide a few examples of the many applications of topological indices to modeling biological, environmental, and pharamacological properties. There have been many efforts to build models related to pharmaceutical properties (Basak 1987; Estrada & Uriarte 2001), so only a small subset is listed here. One additional interesting application of connectivity indices (Gani et al. 2005) is in the pre-

Table 1.2: Sample of available TI-based methods for predicting various properties of pure compounds

Property	**TI method**
Anti-inflammatory activity	Gupta et al. (2002), Bajaj et al. (2005)
Aqueous diffusion coefficient	Schramke et al. (1999)
Aqueous solubility	Hall et al. (1975), Katritzky et al. (1998)
Biodegradability	Boethling (1986)
Blood-brain barrier partition coefficients	Rose et al. (2002)
Boiling point	Hall et al. (1975), Hosoya (1971), Hall & Story (1996), Galvez et al. (1994)
Critical temperature	Hall & Story (1996)
Density	Kier et al. (1976), Estrada (1995), Katritzky & Gordeeva (1993)
Enthalpy of formation	Mercader et al. (2000)
Enthalpy of fusion	Gharagheizi et al. (2012)
Enthalpy of vaporization	Galvez et al. (1994)
Flash point	Patel et al. (2009)
Heat capacity	Yao et al. (2003)
LC_{50}(fathead minnow)	Hall et al. (1989a;b), Basak et al. (1984)
Melting point	Katritzky & Gordeeva (1993)
Nonspecific local anesthetic activity	Kier et al. (1975), Katritzky & Gordeeva (1993)
Octanol/water partition coefficient	Murray et al. (1975)
π-electron energy (C-C bonds)	Hosoya et al. (1975)
Refractive index	Katritzky & Gordeeva (1993)
Vapor pressure	Katritzky et al. (1998)
Viscosity	Kauffman & Jurs (2001)
Water-air partition coefficient	Katritzky et al. (1998)

diction of coefficients for group contribution methods where the group is missing (the group is not in the descriptor space).

Strengths of TIs. One of the main advantages of TIs is that they can discriminate between very similar structures, often in cases where GC methods cannot (e.g., isomers). This provides a more holistic picture of the molecule and can be very useful for certain design problems. For example, this may have potential applications for CAMD in areas where a structural feature of the to-be-designed compound is fixed *a priori*. Furthermore, since many TIs are a function of the entire graph, TIs reflect the entire nature of the molecular structure. This can have advantages over GC methods, which assume that each group provides a contribution independently of other groups in the structure (this is offset somewhat by GC interaction functions). Finally, TIs have been extensively applied to modeling pharmacological properties. The quality and volume of this literature means that TIs are very suitable to many pharmaceutically relevant CAMD problems.

Weaknesses of TIs. Though topological indices have been widely applied to QSPR, there are only limited examples in CAMD. Topological indices are usually not as generally applicable as GC methods, meaning that TI-based QSPRs are often restricted to a certain class of chemicals. For design purposes, this is problematic as it means that TI-based design problems can only consider that particular subset of the chemical search space. Furthermore, TIs represent graph-theoretic properties of the chemical graph, and many of these properties are not always readily understandable from a chemical perspective (although Randic & Zupan (2001) have offered some interpretations of several TIs). Finally, TIs are more difficult to incorporate into CAMD than GC methods. Such CAMD problems can sometimes face combinatorial difficulties and have only been demonstrated thus far on small design problems. We further discuss TI-based CAMD in Austin et al. (2016b).

1.2.3 *Signature descriptors*

In a broad sense, GC methods capture the important subsets of atoms in a molecule while TIs rely on some function of the chemical graph. One QSPR method which has been shown to capture aspects of both GC- and TI-based methods is signature descriptors (SD). Signature descriptors are far younger than the other methods discussed, originating in 2002 from Visco et al. (2002) and 2003 Faulon et al. (2003). Like TIs, they conceive of chemical structures as the chemical graph. Rather than ascribe various values to a complete molecular graph, SDs retain all of the structural and connectivity information for every atom in a molecule.

Analogously to TIs, SDs define the chemical graph to be $G = (V, E)$. Standard SD methods also incorporate node coloring for every node v via a coloring function $c_G(v)$ and the colors of each node C_v. This change is reflected in the slightly altered definition of the chemical graph as $G = (V, E, C, c_G)$. These node colorings are intended to distinguish between different atoms as well as different types of the same atom. For example, it may be beneficial for the model to differentiate between an oxygen with two single bonds and an oxygen with a double bond. Additionally, one may want to distinguish hydrogens by what atom they are attached to, aromatic carbons from non-aromatic ones, atoms attached to aromatic rings, and many other chemical features. We note that the colorings of the nodes are one of the only subjective parts of SD models, and different coloring schemes can have significant effects on the performance of the models.

One important class of signature descriptors is known as atomic signatures. Given a certain atom in a chemical graph, its atomic signature represents all of the atoms that are within a certain distance, or height, from it. Varying the value of this distance gives rise to different atomic signatures. In the simplest case, with this distance set to 0, the atomic signature of an atom is simply that atom, colored in keeping with the coloring definition. More formally, for an atom (vertex) v, its atomic signature of height 0, ${}^0\alpha$ is given by:

$${}^0\alpha(v) = c_G(v) \tag{1.15}$$

Of course, these atomic signatures of height 0 are not a robust set of descriptors. Considering higher values for height provides a more detailed picture of an atom's environment. In general, for a height of i, the atomic signature of height $i \geq 1$ for a vertex v is defined as:

$${}^i\alpha(v) = G_S(V^i, E^i, C^i, c_G) \tag{1.16}$$

where G_S defines a subgraph of G that contains all vertices and bonds such that the distance between v and any vertex in G_S is at most i. Thus, an atomic signature of height one for a hypothetical atom v defines v and every atom bonded to v as well as all connecting edges. A height two atomic signature defines v, every atom bonded to v, and every atom bonded to those atoms bonded to v along with all necessary edges. Increasing the height of an atomic signature can thus be thought to add a layer of connected atoms. In graph-theoretic terms, increasing the i value is equivalent to adding a layer to a breadth-first-search. A pictorial explanation is provided in Fig. 1.5. In this example, we assume that all atoms are colored by aromaticity and *sp* hybridization and that hydrogens are colored by the atom they are attached to.

The atomic signatures can be thought of as a descriptor space for a molecular structure. A property, P, of a molecule can be estimated by all of its atomic signatures of up to a particular height. This QSPR has a familiar form:

$$P = \sum_{i} \sum_{d \in D^i} c_d \, {}^{i}\alpha_G(d) \tag{1.17}$$

where d is the index of the set of all atomic signatures and D^i is the set of atomic SDs of height i. c_d is a regression coefficient accounting for the "contribution" of each atomic signature to a certain property. ${}^{i}\alpha_G(d)$ represents the number of occurrences of atomic signature d. Using the atomic signatures, signatures of the entire molecule can also be generated.

A major advantage of signature descriptors is that they can be manipulated via simple functions to represent groups from GC methods as well as various TIs. This means that the large amount of QSPRs derived from GC methods and TI-based methods are accessible using SDs. For this reason, we do not provide a table of QSPRs using SDs because they can be—and often are—used to calculate properties via GC- and TI-based methods. A few example of converting SDs to groups and TIs are given in Faulon et al. (2003).

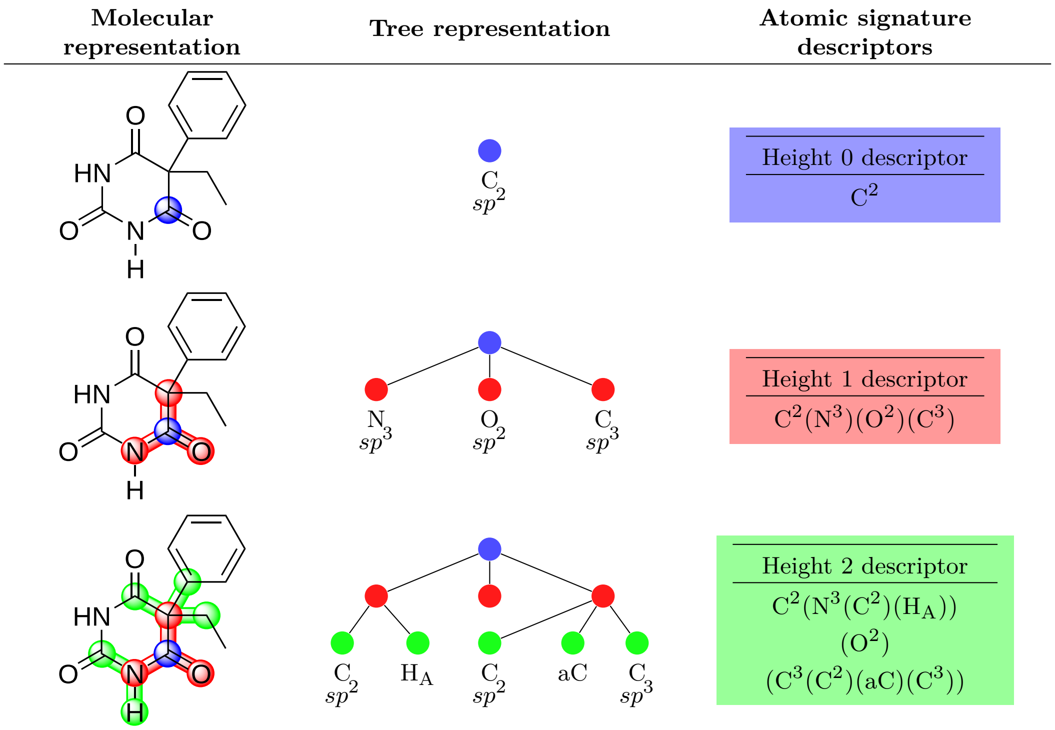
Strengths of SDs. One of the main advantages of signature descriptors is that they have a small inherent bias as compared to GC methods and TIs. The only bias introduced in these descriptors is the choice of an atomic coloring scheme, or the choice of what defines a different type of the same atom. Furthermore, the inclusion of every atomic signature in SD-based QSPRs means there are no theoretical restrictions on the descriptor space defined for a molecular structure. There are also a variety of more modern signature descriptors capable of distinguishing stereoisomers. In the case of stereoisomers, certain SDs have far greater discriminative power than GC methods or TIs, which typically cannot differentiate stereoisomers. Furthermore, the equivalence of SDs to many TIs and groups for GC methods means that SDs can be used directly with these TI and GC QSPR models. This makes a large library of QSPR models accessible to SDs.

Weaknesses of SDs. Many QSPRs using SDs use all of the available atomic signatures of up to a certain height. This can quickly become an issue with the predictive power of SDs as models without sufficient training data may be overfit. A second concern with SDs is that the coloring scheme of the descriptors must be specified before the descriptors are used for QSPRs and CAMD problems. It is likely that some coloring schemes provide better results than others, and the best coloring scheme may not always be easy to determine. Finally, atomic signature descriptors always discriminate between identical atoms in different structural environments. This may again lead to issues of overfitting or not capturing the "true" chemical behavior of the system as sometimes it is better to model identical atoms with a general descriptor which is independent of the atom's environment (although this can be captured to some degree with SDs of lower heights).

1.3 CAMD AS AN OPTIMIZATION PROBLEM

The various types of structure-property relationships discussed above can quickly and often accurately estimate properties from a structure. The application of QSPR techniques in this direction—*predicting properties from structures*—defines what is known as the "forward problem," and is what QSPR techniques are generally intended for. CAMD can broadly be thought to consider the "reverse problem," or the problem of *predicting structures from properties.* At first glance, there is no immediately obvious way of relating properties to a specific molecular structure. One issue is that there are so many structures to consider. A reasonable approach to the CAMD problem should be able to consider a large diversity of structures without running into significant computa-

Figure 1.5: Atomic signature descriptors for a carbon atom in an example molecule



tional difficulties. Another issue involves structural feasiblity. Solutions to the CAMD problem must also be sensible molecular structures, meaning CAMD should produce structures that do not violate any inherent laws of chemical bonding. A final issue involves consistency. All of the QSPRs discussed (though non-overlapping GC methods are an exception) require that if certain features are present, so must be other features. For example, it would be unreasonable to design a structure that has four paths of length three but no paths of length two. In the case of GC methods or SDs, an example erroneous solution would contain one occurrence of the carboxylic acid (-C(=O)OH) group/signature and no occurrences of the carbonyl (-C(=O)-) group/signature. Note that this example assumes that overlap is allowed between these groups/signatures (although this is generally the case with signatures).

Mathematical optimization is the key to addressing all of these issues. Before introducing the problem in a general optimization formulation, we define a few important sets and variables. First, we assume that we have a vector of properties p and a property value p_k for each property k. The vector n encapsulates relevant structural information of the designed molecules and is dependent on the type of QSPR chosen. In the case of GC methods, the value n_d would represent the number of occurrences of each group d. For TIs, the n vector may represent the number of topological features d (edges, paths of certain lengths, etc.) from which TIs would be calculated. For SDs, this n vector usually represents counts of various atomic signature descriptors d. The function f then transforms this structural information into a property estimate using the appropriate QSPR relationship. In general, the CAMD can then be expressed:

$$\min_{n} \quad C(n,p) \tag{1.18}$$

$$s.t. \quad p = f(n) \tag{1.19}$$

$$h_1(p,n) \leq 0 \tag{1.20}$$

$$h_2(p,n) = 0 \tag{1.21}$$

$$s_1(n) \leq 0 \tag{1.22}$$

$$s_2(n) = 0 \tag{1.23}$$

$$p_k^L \leq p_k \leq p_k^U \quad \forall k \tag{1.24}$$

$$n_d^L \leq n_d \leq n_d^U \quad \forall d \tag{1.25}$$

In the above, Eq. (1.19) involves QSPR functions f which estimate a vector of properties p from attributes such as group counts, graph topological features, or atomic signatures. Eqs. (1.20) and (1.21) define general functions, h_1 and h_2, representing inequality and equality constraints on property values, desired structural features, process conditions, and a variety of other possibilities. For example, the presence of certain groups or structural features may necessitate changing a chemical process to accommodate these structures. These constraints can also eliminate certain groups/topological features/signatures from the solution space or require that they appear a certain number of times. The functions h_1 and h_2 can also account for design considerations such as system thermodynamics, cost, and a variety of system-specific interactions between n and p. Eqs. (1.22) and (1.23) define inequality and equality constraints which ensure structural feasibility. More specifically, the functions s_1 and s_2 determine if the vector n is consistent with a molecular structure which can actually exist. These constraints prevent erroneous structures from being considered, eliminating compounds that violate atomic valences, are disjoint, have an incorrect number of aromatic atoms, etc. Eqs. 1.24 and (1.25) set bounds on property values and n, respectively. Each property k is bounded below by p_k^L and above by p_k^U. Similarly, n_k^L and n_k^U define lower and upper bounds for n_k. Finally, Eq. (1.18) is a general objective function for the CAMD problem. The function $C(n, p)$ can define a number of possible functions. These functions somehow quantify the performance of a specific molecule based on its properties p and perhaps its descriptors n.

1.3.1 *Classes of the CAMD problem*

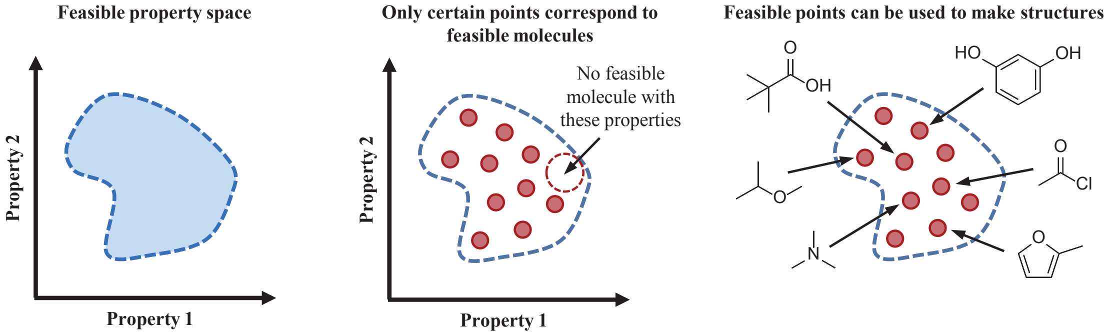
1.3.1.1 *Single molecule design*

The single molecule design problem is the problem of determining a single, optimal structure for a particular purpose. For these problems, the structure of the compound is the only design consideration, meaning that the variables represented by n above are the only degrees of freedom in the problem. Furthermore, it is assumed that there exists some ranking criteria with which to determine which structures are better than others. To make this section as general as possible, we assume that the ranking criteria can either be applied during the optimization procedure as the objective function or afterwards, evaluating the performance of each of a pool of candidate molecules. Though many classifications of these single molecule design problems are possible, we suggest three basic forms: (1) determining all feasible structures; (2) using an objective function which directly quantifies a molecule's performance; and (3) designing a structure with properties as close as possible to certain property targets.

The feasibility problem.

We begin by describing the case where there is no objective function used in the optimization problem. More formally, this is equivalent to solving the above CAMD formulation with the objective function equal to a constant. The solutions to this problem represent all molecules which satisfy the functions h_1 and h_2, are chemically sound structures, and do not violate the property or descriptor constraints. This is a useful type of problem to solve in CAMD when the performance function C is not accessible

Figure 1.6: Pictorial representation of the problem of finding all feasible molecules


Feasible property space
Property 2
Property 1
Only certain points correspond to feasible molecules
No feasible molecule with these properties
Property 2
Property 1
Feasible points can be used to make structures
O
OH
HO
OH
O
Cl
O
N
O

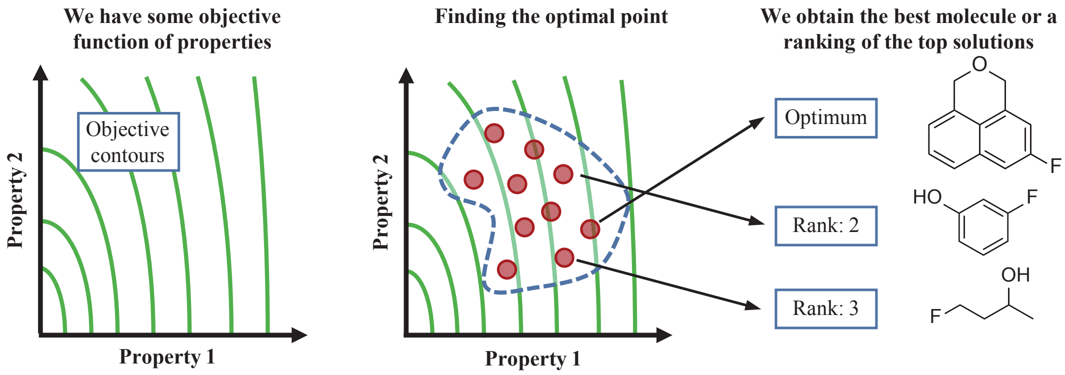
or reliable. It may be the case that C requires a complex simulation or experimental work. The performance function may also be inaccurate or the designer may not know exactly which properties to optimize. In these cases, these CAMD problems leverage the power of optimization to reduce the large number of possible compounds to a more manageable number. This smaller pool of compounds can then be investigated using high-order models or experiments.

Feasibility problems are sometimes difficult to distinguish from other types of problems. In many cases, a feasibility problem is solved at one of the beginning stages of the problem, and all of the feasible structures are then evaluated based on some ranking criteria. For these types of problems, there is no clear line between a feasibility problem and a problem with an objective function. Usually, the molecules designed by these problems are ultimately ranked, so many "feasibility problems" actually have an implicit objective function. A second complication is that many CAMD methodologies are designed to solve both feasibility problems and problems with objectives. We provide a few examples of feasibility problems here for GC methods (Joback 1989), TIs (Kier & Hall 1993; Kier et al. 1993), and SDs (Churchwell et al. 2004). We leave a more detailed discussion for the applications section of this chapter. A graphical example of the feasibility problem is given in Fig. 1.6.

Exact relationship between structures and performance

In many cases, it is possible to model a molecule's performance directly as a function of its structure and properties. This is the most natural and common form of the CAMD objective because the performance function used to rank molecular structures is what is minimized or maximized. Using the exact performance function as the objective

Figure 1.7: Pictorial representation of the problem of finding an optimal molecule


We have some objective function of properties
Objective contours
Property 2
Property 1
Finding the optimal point
Property 2
Property 1
We obtain the best molecule or a ranking of the top solutions
Optimum
Rank: 2
Rank: 3
O
F
HO
F
OH
F

$C(n, p)$ guarantees that the solution to the CAMD problem is the optimal molecule for the application, at least as judged by the provided performance function. We also note that if the function $C(n, p)$ or any of the constraints is non-convex, the optimal molecule may only represent a local optimum. Using a global optimization algorithm will provide the best molecule for these non-convex problems.

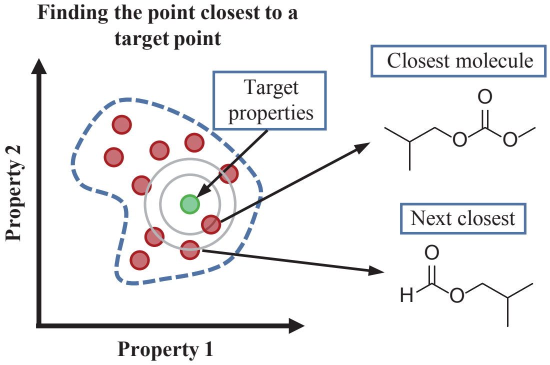
An exact representation of the objective has inspired many numerical optimization strategies to solving the CAMD problem. Numerical optimization strategies are particularly advantageous in these cases because there is an exact algebraic relationship between the descriptor space n and the performance function C. This enables the use of modern, state-of-the-art optimization techniques. Several of these will be discussed in an upcoming section.

There are many examples in the CAMD literature which fall into this category. Again, we will postpone the discussion of these topics until later sections of this document. We provide a small (and non-exhaustive) selection of relevant references in the meanwhile to offer some insight to the interested reader: Odele & Macchietto (1993); Sinha et al. (1999); Sahinidis et al. (2003)(GC), Camarda & Maranas (1999); Siddhaye et al. (2000) (TIs), and Chemmangattuvalappil et al. (2010) (SDs). A graphical example of this type of CAMD problem is provided in Fig. 1.7.

Minimizing distance to property targets

Another class of single molecule design problems concerns finding a molecule with properties as close as possible to target values. These types of formulations have been

Figure 1.8: Pictorial representation of the problem of finding a feasible molecule with predicted properties as close as possible to target values



used extensively in CAMD, often to design alternatives to molecules being used in practice. These molecules may need to be substituted for reasons of environmental friendliness, cost, or availability. Many other types of problems can be addressed when the ideal properties of a molecule are known. However, if these ideal properties do not represent a structure that is physically realistic, these approaches face some difficulty.

In general, these types of problems define the objective to be:

$$C(p) = \sum_k w_k \|p_k^T - p_k\|_2 \tag{1.26}$$

where p_k^T represents a vector of property targets for each property k and w_k is a weight for the distance from the estimated property value to its corresponding property target. The second norm shown above is the most common distance function in CAMD, but many other distance functions are possible.

We present a few demonstrative examples of CAMD with property target objective functions. Matsuda et al. (2007) used groups to design ionic liquids based on conductivity and viscosity targets. Siddhaye et al. (2004) addressed this problem with topological indices to design pharmaceutical products. Brown et al. (2006) used signature descriptors to design polymers with specific properties. A graphical example of this type of CAMD problem is given in Fig. 1.8.

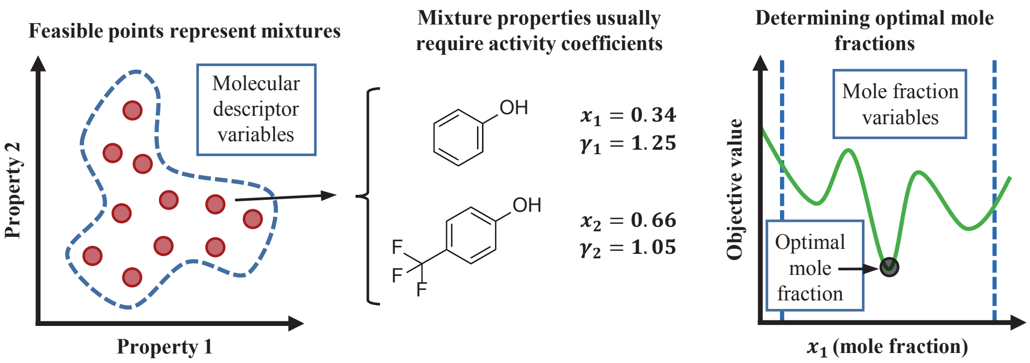
1.3.1.2 *Mixture design*

Real-world applications often demand a product with specifically tailored properties. This sometimes necessitates utilizing a mixture of compounds as no single compound possesses all of the necessary properties. Using CAMD techniques to simultaneously design two or more compounds for use in a blend/mixture is referred to as the mixture design problem. We note here that we alter a definition given in Austin et al. (2016) for consistency with the prevailing definition of mixture design in the literature. All applications classified under mixture design in this section can be assumed to design two or more structures simultaneously. Though we designed two or more structures simultaneously in Austin et al. (2016), we denoted the mixture design problem to be the problem of designing one or more structures to be used in a multi-component system.

While the single-molecule design is difficult in many cases, the mixture design problem is much harder. The difficulties come from several sources: (1) the descriptor variables must now represent the descriptors of every unknown component in the mixture; (2) mixture properties must be calculated and included in the problem; (3) non-ideal mixture behavior must be considered in the form of thermodynamic relationships, activity coefficient models, or equations of state; (4) the design of mixtures also requires a determination of the amount of each component, so mole fractions must be considered. We provide a pictorial representation of the mixture design problem in Fig. 1.9. In this problem, we define the variable x_i to represent the mole fraction of unknown component i. Furthermore, q_j will represent the mixture property j. The formulation of the mixture design problem is similar to the single-molecule design problem. The objective function is altered to now include mixture properties:

$$\min_{n,x} \quad C(n, p, q) \tag{1.27}$$

Figure 1.9: Pictorial representation of the mixture design problem



A few additional constraints are also necessary:

$$q = g(x, n, p) \tag{1.28}$$

$$\sum_i x_i = 1 \tag{1.29}$$

$$h_1(p, q, n) \leq 0 \tag{1.30}$$

$$h_2(p, q, n) = 0 \tag{1.31}$$

$$q_j^L \leq q_j \leq q_j^U \tag{1.32}$$

In the above, Eqs. (1.30) and (1.31) represent modified constraints from the previous formulation to include the mixture property variables q. Eq. (1.28) represents simple mixing functions or complex thermodynamic relationships which relate individual component property variables p, descriptor variables n, and mole fractions x to the mixture property variables q. In many cases, some of these q values represent activity coefficients. Eq. (1.29) ensures that mole fractions sum to 1. Finally, Eq. (1.32) places upper (q_j^U) and lower (q_j^L) bounds on mixture properties. We provide a few mixture design citations (Gani & Fredenslund 1993; Conte et al. 2011a; Buxton et al. 1999) and leave the rest for the applications section.

1.3.1.3 *Integrated process and product design*

Though many CAMD endeavors design products with the ultimate goal of being incorporated into an industrial process, few have explicitly considered the relationship between a particular structure and a process. This is especially important as process performance is typically very sensitive to the molecule(s) chosen. These problems are especially challenging from an optimization point of view because there is no easily discernible algebraic relationship between the descriptor variables n and process variables. This problem also requires the introduction of process variables μ_w, with w defining the index over the process variables. To present the formulation for integrated process and product design, we modify the single-molecule design problem for simplicity.

First, the objective function now reflects process variables:

$$\min_{n,\mu} \quad C(n, p, \mu) \tag{1.33}$$

A few additional constraints then account for the inclusion of process variables:

$$h_1(p, \mu, n) \leq 0 \tag{1.34}$$

$$h_2(p, \mu, n) = 0 \tag{1.35}$$

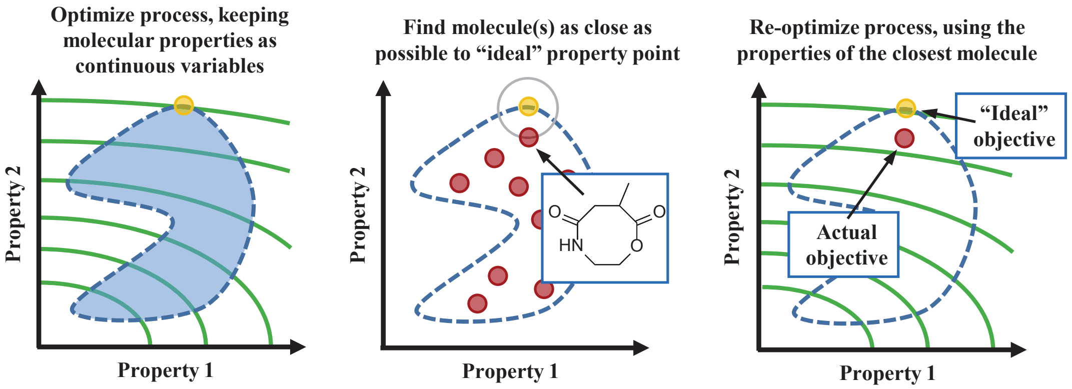
$$\mu_w^L \leq \mu_w \leq \mu_w^U \tag{1.36}$$

In the above, Eqs. (1.34) and (1.35) represent modified constraints from the previous formulation to include the process variables μ. Eq. (1.36) is introduced to place upper (μ_w^U) and lower (μ_w^L) bounds on process variables. There is a similar formulation defined for process/product design problems which also consider mixtures. Again, we provide a small selection of references (Eden et al. 2004; Papadopoulos & Linke 2005; Karunanithi et al. 2005) and leave the remaining discussion for the applications section. One approach to the process design problem is illustrated in Fig. 1.10.

1.3.2 *Common design features in the form of constraints*

These constraints define any process criteria, design necessities, thermodynamic conditions, and any other features which may be important for a particular design problem. As these constraints reflect the diversity of applications of CAMD, we cannot go into detail about all of them here. Rather, we present a summary of design features and conditions which occur most commonly in CAMD.

Figure 1.10: A decomposition approach to the integrated product/process design problem


Optimize process, keeping molecular properties as continuous variables
Find molecule(s) as close as possible to "ideal" property point
Re-optimize process, using the properties of the closest molecule
Property 2
Property 1
HN
O
"Ideal" objective
Actual objective

A very common design feature in CAMD limits the number of groups which can occur in the solution. For example, constraints of this variety can place a limit on the number of molecular descriptors d:

$$n_d \leq n_d^U \tag{1.37}$$

or a limit on the total number of descriptors:

$$\sum_d n_d \leq N^U \tag{1.38}$$

where N^U represents the maximum number of structural features (groups, atoms, descriptors, topological features, etc.) in the designed molecule. Many design problems also set a lower bound on this summation, N^L, which is typically equal to 2. This ensures that more than one descriptor must appear in the solution. Finally, some design problems seek solutions which are analogues within a given family of chemical structures. In this case, it may be necessary to include the descriptor variables corresponding to this structural family:

$$n_d^F \leq n_d \quad \forall d \in D^{\text{FIX}} \tag{1.39}$$

where D^{FIX} defines the set of descriptors which must occur in the solution and n_d^F represents the number of descriptors d which must occur for the structural family to be produced. This constraint is typically part of Eq. (1.25), where other descriptors may have lower bounds for other reasons.

Properties are also bounded similarly in equation (1.24). These constraints ensure that no process conditions, environmental regulations, toxicity thresholds, etc. are violated with the designed structures.

Mixture design problems or single component design problems involving mixtures typically have a number of common constraints. Most notably, these problems often require some idea of the activities of the chemical species in the mixture and thus necessitate incorporating an activity coefficient model. There are three such models often used in CAMD:

1. **UNIFAC.** UNIFAC (Fredenslund et al. 1975) is a group contribution variant of the UNIQUAC equation (Abrams & Prausnitz 1975) and has been used extensively in CAMD. UNIFAC is simple, accurate, and easy to incorporate into CAMD problems due to its use of groups. The original UNIFAC may lack accuracy outside of standard temperature and pressure ranges, but many extensions and re-parameterizations exist to address these cases.

2. **SAFT.** SAFT (Chapman et al. 1989) is an accurate equation of state that is applicable in many temperature and pressure domains. It is gaining popularity in CAMD and several group contribution methods (Tihic et al. 2007; Lymperiadis et al. 2007; 2008; Peng et al. 2009; Papaioannou et al. 2014) have already been developed to estimate the SAFT parameters necessary for its use.

3. **COSMO-RS and -SAC** COSMO-RS (Klamt 1995; Klamt et al. 1998) and COSMO-SAC (Lin & Sandler 2002) are two post-processing methods for the COSMO solvation model (Klamt & Schüürmann 1993). Unlike other approaches, these COSMO-based models use electronic surface charge distributions that are calculated at the quantum chemistry level. These methods are now more amenable to CAMD due to the development of various group contribution methods (Mu et al. 2007; 2009; Austin et al. 2016a).

1.3.3 *Forms of the structural feasibility constraints for GC methods*

The constraints defined above in Eqs. (1.22) and (1.23) ensure that the structural descriptors chosen with the variables n_d are consistent with all other structural features such that they can be assembled to create a chemically feasible molecule. As this document primarily concerns GC methods, we will only describe those herein. For a discussion of structural feasibility for other classes of QSPRs, we direct the reader to Austin et al. (2016b).

If using GC methods to solve CAMD problems, the variable n_d typically represents the number of occurrences of the group d in the solution. Many structural constraints in CAMD also consider the valence of each group, Φ_d, which is simply the number of bonds a group requires to satisfy its valence electron requirements. For example, the group $-\mathrm{CH_1(Cl)}-$ requires 2 external bonds ($\Phi_{\mathrm{CH_1(Cl)}} = 2$) as two of carbon's four required bonds are accounted for by hydrogen and chlorine atoms. One of the most widely-used rules for structural feasibility using valences comes from Odele & Macchietto (1993):

$$\sum_d (\Phi_d - 2) n_d = 2m \tag{1.40}$$

where m is defined by:

$$m = \begin{cases} -1, & \text{if compound is acyclic} \\ 0, & \text{if compound is monocyclic} \\ 1, & \text{if compound is bicyclic} \end{cases} \tag{1.41}$$

These constraints alone, while appropriate for most molecules, still allow for certain sets of groups which cannot be joined to form feasible molecules. For example, a solution of 1 $> \mathrm{CH_0} <$ ($\Phi_{\mathrm{CH_0}} = 4$) group and 2 -Br ($\Phi_{\mathrm{Br}} = 1$) groups is feasible for the above constraints with $m = 0$. To address these situations, Odele and Macchietto also define a constraint to ensure that there are enough groups to meet the valence requirements of every group. This means that to include a group in the solution with a valence requirement of 4, there must also be at least 4 additional groups. This is captured in

$$\sum_d n_d \geq n_{d'}(\Phi_{d'} - 1) + 2 \quad \forall d' \tag{1.42}$$

where d' is also an index over descriptors. These constraints can be generalized with integer variables N_R^{arom} and N_R^{ali}, which represent the number of aromatic rings and the number of aliphatic rings, respectively. Furthermore, we introduce G_{ali} and G_{arom} to be the sets of groups which have any open valences which are aliphatic and aromatic, respectively. We define these valence counts to be Φ^{ali} for aliphatics and Φ^{arom} for aromatics. Note that the same group can appear in both sets. We also define the parameter A_d^{arom} to represent the number of aromatic atoms in a group d and the parameter ρ_d to represent the number of available aliphatic attachment points for every aromatic atom in group d. Below we provide a slightly more general formulation considering aromatic rings:

$$\begin{aligned} \sum_{d \in G_{ali}} (2 - \Phi_d^{ali}) n_d = 2 - 2N_R^{\text{ali}} \\ +2 \sum_{d \in G_{arom}} \rho_d n_d - 2N_R^{arom} \end{aligned} \tag{1.43}$$

$$\sum_{d \in G_{arom}} (2 - \Phi_d^{arom}) n_d = 0 \tag{1.44}$$

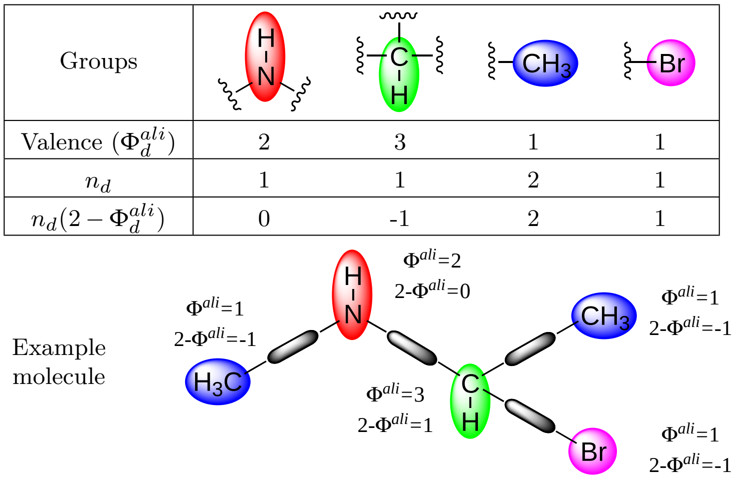
$$\sum_{d \in G_{arom}} A_d^{\text{arom}} = 6N_R^{\text{arom}} \tag{1.45}$$

$$\sum_{d \in G_{ali}} n_d \geq n_{d'}(\Phi_{d'}^{ali} - 1) + 2 \quad \forall d' \tag{1.46}$$

$$\sum_{d \in G_{arom}} n_d \geq n_{d'}(\Phi_{d'}^{arom} - 1) + 2 \quad \forall d' \tag{1.47}$$

We note that in this example, the aromatic rings are assumed to be benzylic and not attached to another aromatic ring (biphenyls and fused aromatics). An example of these basic structural constraints for GC methods are given in Fig. 1.11. A few simple extensions allow these constraints to account for all aromatic rings. These modified Odele-Macchietto constraints work in cases where all groups are bonded to each other with only single or aromatic bonds. More complex connectivity constraints, account-

Figure 1.11: An example of structural constraints for GC methods, considering an acyclic, aliphatic molecule



Groups	H N	C H	CH$_3$	Br
Valence (Φ_d^{ali})	2	3	1	1
n_d	1	1	2	1
$n_d(2-\Phi_d^{ali})$	0	-1	2	1

Feasibility check

Valence balance constraints

$$\sum_d (\Phi_d^{ali} - 2) n_d = 2$$

$$\overset{?}{=} 2 - 2(N_R^{\text{ali}}) = 2 \checkmark$$

Sufficient number of groups for satisfying valence

$$\sum_d n_d = 5$$

$$\overset{?}{\geq} n_{\text{NH}_1}(\Phi_{\text{NH}_1} - 1) + 2 = 3 \checkmark$$

$$\overset{?}{\geq} n_{\text{CH}_1}(\Phi_{\text{CH}_1} - 1) + 2 = 4 \checkmark$$

$$\overset{?}{\geq} n_{\text{CH}_3}(\Phi_{\text{CH}_3} - 1) + 2 = 2 \checkmark$$

$$\overset{?}{\geq} n_{\text{Br}}(\Phi_{\text{Br}} - 1) + 2 = 2 \checkmark$$

ing for cases where groups may be double or triple bonded to one another are given in Sahinidis et al. (2003).

1.4 TECHNIQUES FOR SOLVING THE MOLECULAR DESIGN PROBLEM

1.4.1 *Generate-and-test methods*

Many QSPRs, especially those often used in CAMD, are simple functions which require little computational effort to evaluate. Methods such as those discussed above are able to provide property estimates for millions of structures in a matter of minutes. For this reason, many CAMD approaches have applied QSPR models in the "forward" direction, generating a large number of candidate structures and then evaluating every property of interest for every molecule. This approach — known as "generate-and-test" — has its merits in that it can often optimize over a pool of molecules without solving a potentially difficult optimization problem. This is especially advantageous if the pool of potential molecules has been reduced to a practical number, either by considering a small problem or by one of many knowledge-based reduction procedures (Conte et al. 2011b; Harper

& Gani 2000; Harper et al. 1999). Of course, problems with a large design space are not solved efficiently with this approach. In these cases, optimization methods have a distinct advantage over generate-and-test procedures.

Generate-and-test algorithms are fairly intuitive. The primary requirements are consideration of every possible structure and low redundancy. To that effect, several generate-and-test algorithms exist for groups (Harper et al. 1999; Joback 1989), topological indices (Kier et al. 1993; Hall et al. 1993), and signature descriptors (Faulon et al. 2003).

1.4.2 *Decomposition methods*

Many CAMD problems are characterized by a large combinatorial space of potential descriptors, challenging non-linearities in the thermodynamics of process models, and/or high sensitivity of the objective to process and descriptor variables. As a result, many CAMD problems are too difficult to be addressed directly as optimization problems and must instead be solved as a series of optimization subproblems. Typically, these subproblems successively apply increasingly difficult constraints from the original problem, reducing the feasible set of molecules upon the solution of each subproblem. This step-wise reduction in problem space makes many CAMD problems significantly easier. Alternatively, many decomposition approaches approximate a set of constraints with a lower-bounding surrogate (in the case of minimization), devise a subproblem to generate feasible points, and then iterate between these subproblems until the upper and lower bounds on the objective function are sufficiently close. Decomposition techniques can be broadly divided into the three categories of CAMD problems to which they are most often applied. These are discussed below.

1.4.2.1 *Decomposition in single molecule design*

The most common technique in single molecule design involves systematically reducing the space of feasible molecules by applying increasingly stringent constraints. For example, referring to the molecular design formulation shown in the previous section, some techniques (e.g. Harper et al. (1999)) first apply the structural feasibility constraints, Eqs. (1.23) and (1.22), resulting in a set of all structurally feasible n vectors. This set of n vectors represents a significantly smaller feasible region than that of the entire space defined by the descriptors. These n vectors can be assembled into feasible structures, and this set of structures can be evaluated one-by-one based on the remaining constraints and objective function or used in another optimization problem.

This approach works best if the number of feasible structures is reduced to a reasonably small number before difficult constraints are evaluated or the remaining optimization subproblem is solved. For this reason, decomposition methods have seen the most success in cases where there are few possible descriptors, the constraints are very tight, or the problem involves minimizing a distance to property targets. In cases where many n vectors are possible, decomposition methods should be paired with optimization methods rather than generate-and-test methods.

1.4.2.2 *Decomposition in mixture design*

One of the more challenging problems in CAMD, the mixture design problem benefits from decomposition. Early work from Gani & Fredenslund (1993) proposed a decomposition algorithm based on the prior work of Klein et al. (1992) to decouple the mixture design problem into several single-component molecular design problems. Each of these solutions could then be investigated as a potential mixture component. Many other approaches have followed suit (Karunanithi et al. 2005; Conte et al. 2011b; Austin et al. 2016), relying on the efficiency of single-molecule CAMD techniques to quickly solve single-molecule subproblems and optimization techniques to optimize over the space of the mole fractions.

1.4.2.3 *Decomposition in integrated product/process design*

Integrated product/process design problems are also often decomposed. A popular method relies on optimizing the process variables while allowing for any possible properties of the products. This is a relaxation of the original problem and represents a lower bound (in the case of minimization) on the objective. Solving this problem is far easier and generates a set of ideal properties for an optimal molecule to have. Using these ideal properties as targets, a molecular design problem can be solved to determine the molecule that is closest in properties to these ideal values. Finally, the process design problem can be updated with actual property values of the closest structure and then re-optimized. This technique always produces a feasible value of the objective, though it may not always find the globally optimal structure(s) and process variable values. This two-stage approach has been used by Eden et al. (2004) and Bardow et al. (2010). The latter reference has used this approach throughout the literature, referring to it as the continuous molecular targeting approach (CoMT-CAMD).

Other approaches optimize first in the space of the molecular structures before investigating process performance. Papadopoulos & Linke (2005) designed molecules based on a multi-objective optimization problem, determining the best potential structures based on a number of criteria. This resulted in a Pareto-optimal set of structures, each of which could be tested in the context of the process design problem. Approaches similar in nature (Karunanithi et al. 2005; Austin et al. 2016) have also optimized in the space of structures and evaluated each structure/mixture as an input to process design problems.

Other approaches solve the problem iteratively. For example, Buxton et al. (1999) proposed iteratively solving two subproblems: one to identify process conditions and the other to determine molecular structures. A new approach from Gopinath et al. (2016) uses the outer approximation algorithm (Duran & Grossmann 1986) to treat the product/process design problems as two subproblems. In this approach, one subproblem solved the process problem for a fixed molecule and the other determined another candidate molecule. Approaches such as these two discussed work by generating upper and lower bounds on the objective by solving the subproblems. When the lower and upper bounds are within a certain tolerance, the algorithms terminate.

1.4.3 *Mathematical optimization methods*

In some cases, CAMD problems can be addressed straightforwardly with optimization techniques. In others, optimization approaches become tenable with a slight alteration to the formulation or by exploiting the problem structure. To reiterate an earlier statement, optimization approaches are best suited to problems with many possible descriptors or challenging non-linearities or non-convexities in molecular, mixture, or process models. In cases where there are few possible descriptors, the design space can often be enumerated efficiently using generate-and-test methods. Before discussing a few relevant techniques, we note that there is a large degree of overlap between the categories of decomposition methods and mathematical optimization methods. Both techniques, however, merit an independent discussion as both play a critical role in the solution of CAMD problems.

Many techniques in single-molecule design are able to first solve feasibility optimization problems based on property and structural feasibility constraints. This also qualifies as a decomposition method as the objective function is evaluated afterwards or structure ranking is presumed to be done by expert analysis. Odele & Macchietto (1993) introduced a general optimization formulation for solving molecular design problems. The problem discussed in their paper, the single molecule design problem, can be addressed very efficiently with optimization techniques, even if the descriptor space is large.

Duvedi & Achenie (1996) applied outer-approximation (Duran & Grossmann 1986) to the formulation given by Odele and Macchieto with a few slight alterations. A variant of this approach was also used by Churi & Achenie (1996), who considered connectivity of the structures. In both cases, the problem was formulated as a mixed-integer nonlinear program (MINLP), which was solved with an outer approximation algorithm. Though efficient for many types of problems, the outer approximation algorithms used in these cases are not guaranteed to find globally optimal solutions if the problem is non-convex.

Sinha et al. (1999) and Sahinidis et al. (2003) applied the branch-and-bound algorithm to these problems. This required deriving underestimators (linear in many cases) of the constraints and objective function. Solving the problem with underestimators resulted in a lower bound on the objective (in the case of minimization). This approach allowed optimization strategies to quickly disregard areas of the search space which would not lead to optimal solutions. Furthermore, upon convergence, the branch-and-bound algorithm guarantees a globally optimal solution, at least to within a user-specified tolerance.

The problem of designing molecules to match property targets is far easier. This problem can be solved in the mixed-integer quadratic program (MIQP) shown above or as a mixed-integer linear program (MILP), where some transformations of the QSPR functions must be done before optimization Sahinidis et al. (2003). This MILP has an altered objective function

$$\min_{n,x} \sum_k w_k(d_k^+ + d_k^-) \tag{1.48}$$

where d_k^+ and d_k^- are the positive and negative deviations between the target property values and the estimated values. Both d_k^+ and d_k^- are modeled as positive continuous variables. They require one additional constraint to calculate:

$$d_k^+ - d_k^- = p_k^T - p_k \quad \forall k \tag{1.49}$$

Samudra & Sahinidis (2013b) used this observation to decompose the molecular design problem into three basic steps: (1) identification of optimal sets of groups, n; (2) structure generation; and (3) application of higher-order models to feasible structures. Zhang et al. (2015) extended step (1) of this methodology to account for higher-order groups by introducing variables to account for the connectivity. Maranas (1996) proposed linearization of difficult property models to allow for the efficient use of MILP techniques. Camarda & Maranas (1999) convexified non-linear terms in their CAMD formulation, simplifying the location of globally optimal structures.

1.4.4 *Heuristics*

Though mathematical programming approaches are very useful, there are many CAMD problems which are too high-dimensional or non-linear to be practically considered by mathematical optimization approaches. Problems of this type may attempt to design very large, structurally diverse molecules, necessitating a large search space with many possible combinations of descriptors. Alternatively, these problems may involve a complex process design component, which may require a difficult simulation for each feasible structure. Perhaps the thermodynamic functions and mole fractions optimization must be decoupled from the molecular design problem, meaning that each feasible structure must be investigated in a difficult mole fractions optimization subproblem.

Heuristic approaches to complex optimization problems typically apply high-level selection strategies to generate a series of trial points, evaluating the objective for each trial point and determining a new trial point based on the value of the objective as well as the history of function evaluations. These algorithms terminate either by converging to a point of optimality or by exceeding some user-defined limit (time, iterations, etc.). In the context of CAMD problems, most heuristic optimization approaches have optimized in the space of molecular structures. Practically speaking, this involves generating a molecular structure based on certain design specifications (descriptors, property targets, etc.). Each structure is fixed as an input to the entire CAMD problem, and the objective value is determined. A new structure (or series of structures) is chosen for evaluation based on the relationship between structures/descriptors/properties and the objective value.

We note that one of the most important considerations for using heuristics for CAMD problems is the translation of molecular structures into a particular encoding. In simple cases, the encoding can represent the number of descriptors used in the molecules. In other cases, the descriptors may be translated into a binary string or assigned a value that is a function of the descriptors.

1.4.4.1 *Genetic algorithms*

Genetic algorithms (GAs) are a class of heuristic algorithms roughly based around the idea of natural selection. More specifically, GAs evaluate the performance of each member of a certain pool of solutions called a "generation." The members of the generation are then combined, taking some features from certain members—or "parents"—and some from another. The likelihood of being chosen for this "reproduction" process is based on each solution's rank, with preference given to better solutions. In their application to CAMD problems, GAs typically work in the space of molecular structures. Each structure in a generation is evaluated, features from the best-performing molecules are passed on to the next generation, and the process continues until convergence is achieved.

Venkatasubramanian et al. (1995) introduced GAs to CAMD problems. They encoded molecular structures as a string of their substituent groups and applied genetic operations to the parent strings. Dyk & Nieuwoudt (2000) proposed an encoding based on the UNIFAC (Fredenslund et al. 1975) groups. Xu & Diwekar (2005) developed a

GC-based GA which encoded structures based on various group identities. Herring & Eden (2015) applied GAs in conjunction with signature descriptors to design structures based on property targets. Zhou et al. (2016) applied GAs to solvent design problems for a two-phase reactions and to process design problems for gas absorption (Zhou et al. 2016). Scheffczyk et al. (2016) also applied GAs to liquid-liquid extraction problems based on COSMO-RS thermodynamics.

1.4.4.2 *Tabu search*

Tabu search (TS) algorithms also work by first proposing a pool of initial solutions. These solutions are altered by one of many operations to produce slightly altered solutions. This process is repeated so long as the altered molecules do not appear in a tabu list, i.e., a list of solutions forbidden from consideration based on various factors. These factors can include: frequency of occurrence (to ensure the same solutions are not always visited), infeasibility or low-objective values (to allow the algorithm to determine good solutions), and many others. Using this tabu list, the TS algorithm maintains some "memory" of previous solutions, which can offer some advantages in CAMD problems. Naturally, the analogue of TS in CAMD problems maintains a solution set of molecular structures and alters these provided they are not on the tabu list of forbidden structures. One important feature of TS algorithms is that the tabu list is often dynamic, meaning once tabu-forbidden solutions may be acceptable during a later generation.

Tabu search has only recently been applied to CAMD problems. For example, Chavali et al. (2004) and Lin et al. (2005) applied TS to the design of transition metal catalysts. Another example comes from McLeese et al. (2010a), who considered ionic liquid design.

1.4.4.3 *Other methods*

Several other heuristics have been applied to CAMD problems. For example, Gebreslassie & Diwekar (2015) solved liquid-liquid extraction problems with a modified ant colony optimization (ACO) algorithm. Like other heuristics, ACO works by proposing a pool of solutions (molecules). These solutions are assigned a certain weight, and good solutions attract other solutions in their direction, meaning that bad solutions become more like good solutions in properties, descriptors, or otherwise. Similarly, bad solutions discourage other solutions from becoming similar to them.

Simulated annealing is another heuristic to produce solutions for the CAMD problem. It works by altering a given solution by randomizing the descriptors that define its encoding. If the new solution is better than the previous, it is accepted and becomes the current solution. If the new solution is worse but still within the bounds of an error function, it is also accepted. As the algorithm proceeds, the error function becomes more and more stringent, meaning that worse solutions are less likely to be accepted. Examples of this algorithm in CAMD come from Ourique & Telles (1998) and Marcoulaki & Kokossis (1998).

1.5 LITERATURE REVIEW OF CAMD APPLICATIONS

1.5.1 *Single molecule design*

To begin, we mention a few approaches for single molecule design which qualify as solving the "feasibility problem." As discussed earlier, it is sometimes difficult to make the distinction between this category of problems—feasibility problems—and those that have some ranking criteria. The simple reason for this is that many methodologies are designed to do both. Nonetheless, this type of application is presented here to underscore its importance. A few noteworthy examples come from Joback (1989), who describes techniques to solve the feasibility problem using groups, and Kier and co-workers (Kier & Hall 1993; Kier et al. 1993;?), who solved this problem using topological indices, and Churchwell et al. (2004), who described techniques to solve these types of problems with signature descriptors. All of these studies, though not directly applying optimization, use many techniques to reduce the massive size of the chemical search space. Many techniques described in subsequent sections can solve these feasibility problems with optimization approaches. These methods typically have an explicit objective function and thus fit better into a different category.

The next category of applications concerns problems with explicit objective functions. The many objective functions of this category reflect the diversity of the CAMD field as a whole. Some of the first contributions in CAMD came from Gani & Brignole (1983) and Brignole et al. (1986) who designed extraction solvents based on solvent power and selectivity. Though Gani and Brignole did not use optimization in this approach, they significantly reduced the number of feasible molecular structures based on arguments such as limiting groups, placing restrictions on various properties, and investigating solubility curves. In this way, they narrowed the massive chemical design space to a few possible molecules which could all be considered directly by solving the "forward problem." A similar approach was used by Macchietto et al. (1990) and Joback (1989).

Early efforts at optimization in the CAMD problem come from Gani et al. (1991) and Knight & McRae (1991). Odele & Macchietto (1993) introduced a general numerical optimization to the CAMD problem, formulating several solvent design problems as optimization problems over the space of groups. Since then, many CAMD efforts have focused on the design of solvents. For example, Harper et al. (1999) proposed a multilevel approach to design separations solvents, reducing the number of feasible molecules in each of several sub-problems. Sinha et al. (1999) applied global optimization techniques to the design of a blanket wash solvent. Karunanithi et al. (2005) developed a decomposition algorithm to address difficult optimization problems, applying it to the design of liquid-liquid extraction solvents and crystallization solvents Karunanithi et al. (2006). Subsequent investigations of this problem came from Samudra & Sahinidis (2013b) and Austin et al. (2016).

Many efforts in CAMD have been applied towards industrial processes. One of the most popular application areas is in designing solvents for liquid-liquid extraction. Gani et al. (1991) proposed the design of a solvent for the separation of water and acetic acid. Harper et al. (1999) and Harper & Gani (2000) sought to find a replacement for toluene in the separation of phenol and water. A few other examples include: Marcoulaki & Kokossis (1998), Xu & Diwekar (2005), Scheffczyk et al. (2016), and Gebreslassie & Diwekar (2015).

Extractive distillation is also another important industrial process which is often studied using CAMD. Many approaches investigated focused on separations in general and have also been applied to liquid-liquid extraction. Two of those such studies are Harper et al. (1999) and Gani et al. (1991). Dyk & Nieuwoudt (2000) applied simulated annealing to the design of a solvent separation of five binary pairs of common industrial compounds.

Another emerging area of study in CAMD is the design of solvents to optimize reaction properties. Wang & Achenie (2002) made one of the first efforts to solve this problem, designing solvents to promote ethanol fermentation and subsequent extraction. Gani et al. (2005) proposed a rules-based strategy for the selection of solvents for several common reactions and a pharmaceutical example. Folić et al. (Folić et al. 2007; Folic et al. 2008) designed novel solvents to maximize the reaction rate constant of an S_N2 reaction. They applied GC methods to estimate the parameters of the solvatochromic equation of Abraham et al. (Abraham et al. 1987) and then used this equation to predict rate constants. Struebing et al. (2013a) investigated the same type of reactions, now using quantum chemical calculations for their solvents and the solvatochromic equation as a surrogate model. Zhou et al. (2015) used COSMO-RS (Klamt 1995) thermodynamics in conjunction with CAMD techniques to design solvents to maximize reaction selectivity (Zhou et al. 2015). Austin et al. (2016a) designed solvents to maximize a reaction rate using COSMO-RS thermodynamics and projecting the original problem onto a lower-dimensional space.

Ionic liquid design is another application area for this subset of CAMD problems. For example, Karunanithi & Mehrkesh (2013) used decomposition techniques to design ionic liquids for electrical conductivity, heat transfer, liquid-liquid separations, and solubility. McLeese et al. (2010b) designed ionic liquids using a tabu search algorithm, which they also showed produced a globally optimal solution for their test problem. Matsuda et al. (2007) designed ionic liquids based on conductivity and viscosity targets. These case studies used a small number of possible groups and could thus be alternatively addressed via exhaustive enumeration.

A number of approaches have also investigated pharmaceutical applications of CAMD. We note here that a good review exists on the topic of pharmaceutical solvents already (Harini et al. 2013). Chemmangattuvalappil et al. (2010) investigated a drug modification problem using molecular signatures. Their approach, discussed above, required a fairly large number of linear constraints to ensure a consistent set of descriptors. They applied this approach to the alteration of alkyl substituents on a fungicidal compound. Siddhaye et al. (2004) designed pharmaceutical products, focusing on molecules likely to be the active pharmaceutical ingredient (API). Leveraging the power of connectivity indices, they were able to design a few pharmaceutically relevant case studies, including producing a penicillin derivative with specified properties. Churchwell et al.

(2004) used signature descriptors to design peptide inhibitors to leukocyte functional antigen-1 (LFA-1) and its ligand intercellular adhesion molecule-1 (ICAM-1).

Many CAMD efforts have also been focused on the design of alternative refrigerants. For example, Gani et al. (1991) designed refrigerants based on a few important properties. Joback (1989) first considered the problem of finding a replacement refrigerant for Freon-12. Duvedi & Achenie (1996; 1997) and Churi & Achenie (1996) looked at this same problem, focusing on heat capacity and heat of vaporization as the properties to optimize. Marcoulaki & Kokossis (1998) also designed a replacement for Freon-12 using a simulated annealing approach. Sahinidis et al. (2003) investigated the same replacement problem using a global optimization approach with modified structural constraints and an improved CAMD formulation. Samudra & Sahinidis (2013a) designed heat transfer fluids for refrigeration systems.

Polymer design has been investigated extensively using CAMD techniques. Many of these approaches focused on designing a polymer with certain physical properties. Venkatasubramanian et al. (Venkatasubramanian et al. 1994; 1995) applied genetic algorithms to the problem, designing polymers to approximate target values in various property categories like glass transition temperature, bulk modulus, heat capacity, density, and others. Maranas (1996) approaches the problem in a similar way, minimizing distances to target values. Unlike Venkatasubramanian et al., Maranas used a mixed-integer linear program (MILP) formulation and mathematical optimization techniques to solve the problem. Camarda & Maranas (1999) addressed the design problem using topological indices and a mathematical optimization formulation. Eslick et al. (2009) also used topological indices but designed molecules with a tabu search algorithm rather than a mathematical optimization approach. Brown et al. (2006) considered the problem with signature descriptors. Pavurala & Achenie (2013) used an outer-approximation approach to design polymers to aid in oral drug delivery.

This section presents a selection of main applications of single-molecule design problems. For a more comprehensive list, see Table 1.3.

1.5.2 *Mixture design*

The mixture design problem is a difficult variant of the single-molecule design problem. As a result, there are far fewer examples of applications considering a mixed product. Though we divide by application in this section, we emphasize that many of these techniques are generalizable. Many of the references given here can likely be altered to design a mixed product for an arbitrary application.

Klein et al. (1992) and Gani & Fredenslund (1993) first considered the mixture design problem, solving a few example problems involving solubilities and compounds which form azeotropes. Vaidyanathan & El-Halwagi (1996) designed blends of polymers, relying on simple mixing rules for algebraic simplicity. Vaidyanathan and El-Halwagi also designed single polymers. Duvedi & Achenie (1997) developed an MINLP formulation for mixture design, using an equation of state to estimate some mixture properties and relative simple mixing rules to estimate others. They applied the methodology to the design of refrigerant blends.

Buxton et al. (1999) proposed a decomposition technique to solve mixture design problems. They produced solvent blends which would reduce the environmental impact of an industrial process. Sinha et al. (2003) solved mixture selection problems as an MINLP, choosing the best combination of solvents from a given list. This approach was able to select an optimal single solvent and mixture of solvents for use as cleaning agents in the lithographic printing industry.

Karunanithi et al. (2006) used a decomposition technique to first reduce the mixture design problem to the set of all feasible individual components. Each possible mixture was then used to evaluate the objective function. This approach was used for the design of a crystallization solvent and anti-solvent. Conte et al. (2011a) proposed a task-based decomposition algorithm, which was applied to the design of paint solvents blends and insect repellent solvent blends. Austin et al. (2016) addressed the mixture design problem in the reduced-order space of each individual component's properties, employing derivative-free optimization (DFO) methods to optimize over the lower-dimensional space. This was applied to reproduce the crystallization solvent design problems of Karunanithi et al. (2006). Another approach comes from Jonuzaj et al. (Jonuzaj et al. 2016; Jonuzaj & Adjiman 2017) who used Generalized Disjunctive Programming (GDP) (Balas 1979) to select mixtures of solvents from a fixed list.

1.5.3 *Integrated process and product design*

Though many CAMD endeavors design products with the ultimate goal of being incorporated into an industrial process, few have explicitly considered the relationship between a particular structure and a process. This can be problematic as many recent efforts have observed some sensitivity between product descriptor variables and process variables. To overcome this issue, various approaches have considered the product and process design problems simultaneously.

Eden et al. (2004) solved an integrated process and product design problem to best recover volatile organic compounds (VOCs) from an industrial process, identifying optimal property targets for solvents in a reduced-order space given by clustering methods. Hostrup et al. (1999) proposed a general framework for integrated process and product design that was focused on separations. This method relied on reduction of the feasible solution space via thermodynamic arguments and case-specific considerations. Then, molecules were designed based on proximity to property targets for a certain process architecture.

Kim & Diwekar (2002b) solved liquid-liquid extraction process problems, considering the process performance and designing suitable structures using a heuristic optimization strategy for the generation of solvent structures. Papadopoulos & Linke (2005) developed a methodology for considering integrated problems which relied on decomposing the problems into product and process subproblems. Unlike other approaches, Papadopoulos and Linke solved multi-objective optimization problems, determining the Pareto optimal front for solvent properties likely to be related to process performance. Using these Pareto-optimal structures, they could solve the process problems for a much smaller set of possible molecules. They applied this methodology to extractive fermentation (Papadopoulos & Linke 2005) and liquid-liquid extraction and gas absorption (Papadopoulos & Linke 2006b;a).

Karunanithi et al. (2005) proposed a decomposition methodology to solve difficult process design problems. This methodology first filtered out a large number of possible molecular structures based on property bounds. It then applied a few stages of more complicated constraints to the remaining molecules, further reducing the pool of feasible structures. Finally, the process model was applied to each of the molecules which were feasible for all of the constraints. This methodology was applied to the design of a liquid-liquid extraction process (Karunanithi et al. 2005) and to the design of crystallization solvents (Karunanithi et al. 2006). Bommareddy et al. (2010) addressed the product/process design problem first in the space of the process, finding ranges of properties for the molecules to be designed. These ranges then represented what was most suitable for a particular process and therefore defined a much smaller search space for the molecular design subproblem.

Bardow et al. (2010) proposed the CoMT-CAMD approach to first identify target solvent properties and then select an optimal solvent based on proximity to these ideal solvent properties. This was applied in conjunction to a variant of the SAFT equation of state to design a carbon capture and storage process. Stavrou et al. (2014) used the same approach to consider carbon capture problems. Pereira et al. (2011a) also used SAFT to optimize a process to separate carbon dioxide and methane at high pressures. Papadopoulos et al. (2010) designed fluids for an organic Rankine cycle, considering fluids which fell on the Pareto front of optimal properties for the process. Lampe et al. considered the same problem for fluid selection (Lampe et al. 2014) and fluid design (Lampe et al. 2015). A summary of the references, categorized by application, is provided in Table 1.3.

Table 1.3: Summary of CAMD applications and the methodologies used in each case

Application	**References**
Antigen inhibition activity	Churchwell et al. (2004) [sd,d,o]
Biodiesel additives	Hada et al. (2014) [gc,d,gt]
CO_2 capture	Gani et al. (1991) [gc,d,gt,m], Bardow et al. (2010) [sel,d,p], Pereira et al. (2011a) [gc,o,p], Stavrou et al. (2014) [sel,d,o,p], Burger et al. (2015) [gc,o,p], Lampe et al. (2015) [gc,d,o,p], Gopinath et al. (2016) [gc,d,o,p]
Crystallization solvents	Karunanithi et al. (2006) [gc,d,gt,p,m], Samudra & Sahinidis (2013b) [gc,d,o], Austin et al. (2016)[gc,d,o,p,m]
Extractive distillation	Harper et al. (1999) [gc,d,gt], Gani et al. (1991) [gc,d,gt], Papadopoulos & Linke (2006a) [gc,d,h,p], Dyk & Nieuwoudt (2000) [gc,d,h]
Extractive fermentation	Wang & Achenie (2002) [gc,d,o], Papadopoulos & Linke (2005) [gc,d,h,p]
Gas absorption	Odele & Macchietto (1993) [gc,d,o], Buxton et al. (1999) [gc,d,o,m,p], Papadopoulos & Linke (2006b;a) [gc,d,h,p], Bommareddy et al. (2010) [gc,d,gt,p], Zhou et al. (2016) [gc,d,h,p]
HIV-1 protease inhibition activity	Visco et al. (2002) [sd,d,o]
Ionic liquids design	Matsuda et al. (2007) [gc,gt], McLeese et al. (2010b) [ti,h], Karunanithi & Mehrkesh (2013) [gc,d,h]
Liquid-liquid extraction	Gani & Brignole (1983) [gc,d,gt], Brignole et al. (1986) [gc,d,gt], Odele & Macchietto (1993) [gc,d,o], Marcoulaki & Kokossis (1998) [gc,d,h], Harper et al. (1999) [gc,d,gt], Harper & Gani (2000) [gc,d,gt], Gani et al. (1991) [gc,d,gt], Karunanithi et al. (2005) [gc,d,gt,p], Austin et al. (2016a) [gc,d,o,p,m], Ourique & Telles (1998) [gc,h], Kim & Diwekar (2002a) [gc,d,h,p], Papadopoulos & Linke (2006b) [gc,d,h,p], Xu & Diwekar (2005) [gc,d,h], Scheffczyk et al. (2016) [d,o], Gebreslassie & Diwekar (2015) [gc,h]
Organic Rankine cycle fluids	Papadopoulos et al. (2010) [gc,d,h,p], Lampe et al. (2014) [sel,d,p], Lampe et al. (2015) [gc,d,o,p]
Pharmaceutical products	Siddhaye et al. (2004) [ti,o]
Polymer design	Venkatasubramanian et al. (1994; 1995) [gc,h], Maranas (1996) [gc,o], Vaidyanathan & El-Halwagi (1996) [gc,o,m], Camarda & Maranas (1999) [ti,o], Brown et al. (2006) [sd,o], Eslick et al. (2009) [ti,h], Pavurala & Achenie (2013) [gc,d,o], Zhang et al. (2015) [gc,o]
Reactions solvents	Wang & Achenie (2002) [gc,d,o], Gani et al. (2005) [sel,d], Folić et al. (2007); Folic et al. (2008) [gc,o], Struebing et al. (2013a) [gc,d,o], Zhou et al. (2015) [gc,o], Austin et al. (2016a)[gc,d,o,m], Zhou et al. (2016) [gc,d,h]
Refrigerant design	Joback (1989) [gc,d,gt], Gani et al. (1991) [gc,d,gt], Churi & Achenie (1996) [gc,d,o], Duvedi & Achenie (1996; 1997) [gc,d,o], Marcoulaki & Kokossis (1998) [gc,d,h], Sahinidis et al. (2003) [gc,o], Ourique & Telles (1998) [gc,h], Samudra & Sahinidis (2013a) [gc,d,o]
Separations (general)	Hostrup et al. (1999) [gc,sel,d,o,gt,p]
Solvents for consumer products and industry	Pistikopoulos & Stefanis (1998) [gc,o], Buxton et al. (1999) [gc,d,o,m,p], Conte et al. (2011a) [gc,d,gt,m], Sinha et al. (1999) [gc,o], Sinha et al. (2003) [sel,o], Weis & Visco (2010) [sd,o]
Soybean oil products	Camarda & Sunderesan (2005) [ti,o]
Structural modifications to a fungicide	Raman & Maranas (1998) [ti,o], Chemmangattuvalappil et al. (2010) [sd,o]
Transition metal catalyst design	Chavali et al. (2004) [ti,h], Lin et al. (2005) [ti,h]
VOC recovery	Eden et al. (2004) [gc,d,gt,p]

[**gc**] Group contribution methods are the main QSPR method used

[**ti**] Topological indices are the main QSPR method used [**sd**] Signature descriptors are the main QSPR method used

[**sel**] Compounds are selected from a fixed list rather than designed [**d**] Decomposition methods used

[**gt**] Generate-and-test procedure used [**o**] Numerical optimization used [**h**] Heuristical optimization used

[**m**] Mixture design considered [**p**] Process design considered

1.6 CONCLUSIONS

The advent of the computational age has drastically impacted the design of chemical products and novel molecules, altering a once intuition-based, trial-and-error practice into a rapid and efficient search through millions of possible structures. The availability and accuracy of QSPRs combined with efficient mathematical programming techniques has extended this capability even further, enabling chemical product designers to investigate a previously unimaginable diversity of chemical structures.

This section has provided background on the QSPRs which often serve as the underpinning of CAMD problems. Each of three methods (group contribution, topological indices, and signature descriptors) was discussed in detail, and relevant constraints for optimization problems were provided for the case of GC methods. The CAMD problem was also addressed from the vantage point of mathematical optimization. Various formulations were discussed for a few broad classes of the CAMD problem (single-molecule design, mixture design, integrated product/process design). Solution techniques were discussed to aid in the solution of the often difficult CAMD problem. Finally, we provided a summary of the many design endeavors and applications of the CAMD problem.

The increasing availability of computational resources, efficient optimization algorithms, and accurate QSPRs bodes well for the future of CAMD. CAMD has a proven history of determining improved solutions for many well-known industrial processes as well as designing new products for consumers and optimizing high-impact chemical processes. More recently, there have been a growing number of more advanced modeling and design efforts, concerning ideas such as integrating quantum chemistry techniques, designing transition metal catalysts, and determining optimal structures of pharmaceutical compounds. The potential applications of CAMD are numerous, and the field is poised to play an integral role in the development of the chemical and biochemical technologies of the not-so-distant future.

2

MIXTURE DESIGN USING DERIVATIVE-FREE OPTIMIZATION IN THE SPACE OF INDIVIDUAL COMPONENT PROPERTIES

2.1 INTRODUCTION

In this chapter, we address the mixture design problem via a decomposition approach. The mixture design problem poses a number of challenges for mathematical optimization, chief among which is perhaps the even larger combinatorial design space of the multiple chemical compounds to be designed. A secondary complication is the difficulty of incorporating mixture property models (equations of state, activity coefficient models, etc.) into the design problems. These complications can be effectively overcome using our approach.

As discussed in the introduction, traditional methods of product design are largely experimental, meaning that most new candidate products must first be synthesized and then tested. This traditional, Edisonian search space for new products is often very small as it represents a costly design and validation process that is limited by a fixed amount of time and resources. Furthermore, many industrial product design endeavors focus on a particular type of molecule or a structural analogue of a known compound, which significantly attenuates the chemical search space. Clearly, new chemical product design approaches are necessary to probe a growing library of known structures as well as design novel compounds to fit a particular task. With the increasing need for new chemical products, product design has found applications in wide-ranging fields including new fuel design (Sundaram et al. 2001), solvents for optimal separations or optimal solvation (Odele & Macchietto 1993; Brignole et al. 1986; Mitrofanov et al. 1995; Karunanithi et al. 2006; Struebing et al. 2013b), cosmetics (Mitrofanov et al. 1995), food and industrial additives (Conte et al. 2012; 2011b), pharmaceuticals (Harini et al. 2013; III et al. 2013), microelectronics Warrier et al. (2012), refrigerants and heat transfer fluids (Duvedi & Achenie 1997; Sahinidis et al. 2003; Samudra & Sahinidis 2009; 2013a), solvents for carbon capture (Nuchitprasittichai & Cremaschi 2013; Pereira et al. 2011b; Buxton et al. 1999), and a wide variety of materials (Ashbya et al. 2004).

Computer-aided mixture design (CAMxD) can be formally defined as the problem of determining an optimal mixture of compounds whose mixture properties fall within specified ranges. We note that in this chapter, we consider a mixture design problem to be a problem of designing ≥ 1 molecule to function in a mixture of ≥ 2 components. In these problems, the molecular compounds are assembled from scratch, meaning that solutions of CAMxD can consist of one or many structures that have never before been synthesized. Due to the complexity of these problems, CAMxD leverages a large body of prior work in the area of computer-aided molecular design (CAMD), which addresses the same problem but only for the design of a *single* structure. Early work from Gani & Fredenslund (1993) proposed a decomposition algorithm based on the prior work of Klein et al. (1992) to decouple the mixture design problem into several single-component molecular design problems. This decomposition allowed for many existing methodologies and prediction methods to be applied directly to the mixture design problem, although applying optimization techniques to these problems would sometimes necessitate empirical correlations and simplified thermodynamic models. Other approaches (Vaidyanathan & El-Halwagi 1996; Buxton et al. 1999; Duvedi & Achenie 1997; Sinha et al. 1999) applied mathematical optimization techniques directly to mixture design problems, but challenging non-linearities and non-convexities limited these problems to a reasonably small molecular search space or dictated the use of simplified models. Venkatasubramanian et al. (1994) sought to deal with the large search space in these types of problems using genetic algorithms. These stochastic search algorithms do provide a way to explore a large search space, but they assume no algebraic form of the function and, as a result, are often outperformed by many competing techniques (Rios & Sahinidis 2013). Conte et al. (2012; 2011b) successfully exploited a similar decomposition methodology to handle formulations and solvents. This method relied on generating a large number of candidate molecules or blends and then systematically reducing the pool of feasible molecules. While a very useful methodology for many types of problems, this approach can also be limited to small design spaces in that insufficient reduction of the solution pool leads to a large number of complicated subproblems. Karunanithi et al. (2005) also developed decomposition algorithms to translate the mixture design problem into a series of molecular design problems. Their approach required enumerating all feasible molecular structures, regardless of how these might perform in the mixture design problem. Each combination of these structures was tested for compatibility and the objective was evaluated for all feasible combinations. This approach can lead to the over-design of molecular components and the resultant limitation to optimization over a small subset of the feasible mixture design space. Furthermore, many of these approaches calculated mixture thermodynamic properties using the UNIFAC method (Fredenslund et al. 1975).

Mathematical optimization approaches to the solution of the CAMxD problem have limitations in the diversity of structures they can consider as well as in the complexity of mixture thermodynamics involved. To consider more general problems, the existing literature for solving CAMxD problems has focused primarily on enumeration-based algorithms, meaning the CAMxD problem is solved by exhaustively tabulating every possible structure for every possible mixture component. However, this exhaustive enumeration often leads to an unnecessarily costly search through the molecular design space and also limits many mixture design algorithms to a few specific classes of molecules. To address a truly open-ended mixture design problem, a methodology would have to efficiently explore a large part of the chemical search space and determine one or many points of optimality. The primary contribution of this work is the development of a novel general-purpose methodology for mixture design that achieves this goal. The main idea is to address mixture design in the space of pure component properties. Viewed this way, mixture design is naturally decomposed to pure product design and mixture fraction subproblems. Pure component design can be addressed via highly efficient software recently developed for this class of problems (Samudra & Sahinidis 2013b), and the mixture fraction subproblem can be solved with traditional algebraic optimization techniques. What makes our approach particularly suitable to this decomposition is the incorporation of optimization algorithms that are capable of optimizing in the absence of an algebraic expression of the objective function. In addition to demonstrating the usefulness of the proposed approach for mixture design, another contribution of the paper is in terms of presenting results from testing 27 different derivative-free optimization (DFO) algorithms on the proposed mixture design formulation. These results provide insights into which DFO algorithms are best suited for typical mixture design applications.

In the next section, we introduce various concepts and tools which will be used as building blocks in our mixture design methodology. In the section that follows, we propose a new decomposition algorithm which can be generally applied to mixture design problems. Then, we compare 27 different DFO algorithms as applied to two illustrative examples and two problems drawn from the CAMD literature. Finally, we provide comparisons among various classes of DFO solvers and draw conclusions from this work.

2.2 BUILDING BLOCKS

2.2.1 *AMODEO Methodology for Molecular Design*

As previously mentioned, our mixture design methodology decomposes mixture design in a way that requires the solution of several pure component design problems. This decomposition is advantageous as it allows us to capitalize on recently developed methodologies for CAMD that are highly efficient. Addressing the then shortcoming of CAMD to systematically design structures over a large molecular search space, Samudra & Sahinidis (2013b) proposed an optimization and decomposition approach, implemented in the software AMODEO. AMODEO uses the group contribution (GC) method developed by Marrero and Gani (Marrero & Gani 2001; 2002) as its primary tool for evaluating properties from structures. The basic form of this group contribution method is as follows:

$$p_k = f\left(\sum_{i\in F} c_i^k n_i + \sum_{i\in S} c_i^k n_i + \sum_{i\in T} c_i^k n_i\right) \tag{2.1}$$

Here, F is the set of first-order groups, S is the set of second-order groups, and T is the set of third-order groups. These groups are further detailed in Marrero & Gani (2001). c_i^k is a coefficient fitted to the number of occurrences of each group (n_i) for the estimation of the k-th property in a vector of thermophysical properties p. AMODEO's speed in its search of the entire chemical space is enabled by a few key elements in its approach: (1) An optimization formulation exploits the linearity of a set of property prediction models, which allows for quick estimates over a large pool of molecules; (2) The molecular compositions (a raw summation of various molecular sub-groups) are first calculated as an optimization problem, which greatly reduces the dimension of the overall problem from a high-dimensional problem in chemical structure space to a lower-dimensional problem in group composition space to an even lower-dimensional problem in the space of selected groups; (3) Deterministic solution algorithms are used to guarantee optima rather than relying on stochastic algorithms or enumeration of the solutions. To be more specific, AMODEO divides the CAMD problem into two main steps: (1) composition design and (2) structure generation. The composition design phase is solved with an MILP, an abbreviated version of which is shown below:

$$\begin{aligned}
\min_{n} \quad & 0 && (2.2)\\
\text{s.t.} \quad & \kappa_k \leq \sum_{i \in F} c_i^k n_i \leq \pi_k \qquad \forall k \in \mathcal{K} && (2.3)\\
& n \in \mathcal{S} && (2.4)\\
& \vdots
\end{aligned}$$

Here, we solve a feasibility problem to determine which combinations of first-order groups fall within the specified property ranges $[p_k^L, p_k^U]$ for each property k in a set of all properties of interest, $\mathcal{K}$. In other words, we design for every possible set of groups such that these groups satisfy property bounds and structural feasibility. In addition, the traditional group contribution formulation is truncated after the first-order groups and then inverted to yield upper and lower bounds for each $\sum_{i \in F} c_i^k n_i$. These upper and lower bounds are $\pi_k = \max_{p_k \in [p_k^L, p_k^U]} f_k^{-1}(p_k)$ and $\kappa_k = \min_{p_k \in [p_k^L, p_k^U]} f_k^{-1}(p_k)$, respectively. It should be noted that π_k and κ_k can be calculated directly as the functions f_k are always monotonic in the models we use.

In the above model, equation (2.2) indicates a feasibility problem. Equation (2.3) represents the inverted group contribution equation used to determine which vectors of group occurrences, n, fall into the specified property ranges. Finally, equation (2.4) abbreviates a number of valence constraints (Odele & Macchietto 1993; Sahinidis et al. 2003) and simply requires that all groups in a solution can assemble into feasible structures.

The composition design phase relies on a relaxation of p^L and p^U to account for error in the coarser first-order estimates as discussed in Samudra & Sahinidis (2013b). The necessary relaxation is often small as the first-order groups alone nearly always provide estimates within 10% of the full Marrero-Gani model. Once a number of compositions have been determined, each must be assembled into a molecular structure. This is done by solving for entries in an adjacency matrix, solved as an MILP shown in abbreviated form below:

$$\begin{aligned}
\min_{y} \quad & 0 \\
\text{s.t.} \quad & \sum_{\substack{j' \in \mathcal{J} \\ j' \neq j}} y_{jj'} = v_j \qquad \forall j \in \mathcal{J} && (2.5) \\
& y_{j'j} = y_{jj'} \qquad \forall j, j' \in \mathcal{J} && (2.6) \\
& y_{jj'} \leq 1 \qquad \forall j, j' \in \mathcal{J} \setminus \mathcal{J}_0, j' \neq j && (2.7) \\
& y_{jj'} \leq v_j - 1 \qquad \forall j \in \mathcal{J} \setminus \mathcal{J}_0,\ j' \in \mathcal{J} && (2.8) \\
& \vdots
\end{aligned}$$

In this formulation, each composition is represented as a set of nodes in a graph whose arcs are unknown. Each node j corresponds to group j which has a valence v_j, which is the node's required degree for the molecule to be chemically feasible. $y_{jj'}$ is a binary variable equal to 1 if group j is connected to group j', i.e., if there exists an arc between nodes j and j'. $\mathcal{J}$ is the set of nodes determined from the composition design phase and $\mathcal{J}_0$ represents the set of nodes with a valence of 1. Since we aim to find all feasible structures for a set of nodes, we solve a feasibility problem as indicated by the objective. Equation (2.5) ensures that every node has its valence requirements satisfied. Equation (2.6) enforces symmetry in the adjacency matrix. Equation (2.7) means that each pair of nodes can have at most one connection. Finally, Equation (2.8) prevents some simple disconnected subgraphs from being solutions. Beyond these more basic equations, the AMODEO formulation also incorporates efficient constraints for enforcing completely connected graphs, uniqueness cuts to ensure each molecule is only designed once, and redundancy cuts to account for multiple occurrences of the same group. A pictorial representation of the AMODEO approach is provided in Figure 2.1. In short, the algorithm transforms the specified search region into a composition of molecular fragments that are finally assembled into molecular structures for which property estimates are further refined. For more information, see Samudra & Sahinidis (2013b).

2.2.2 *Derivative-free optimization*

Our approach to mixture design will involve optimization on top of product design subproblems which can be solved with existing CAMD methodology. While this decomposition facilitates the use of highly efficient molecular design software, optimization on top of AMODEO-like approaches must be done in the absence of derivatives or even an algebraic objective function. This is because AMODEO applies a highly complex

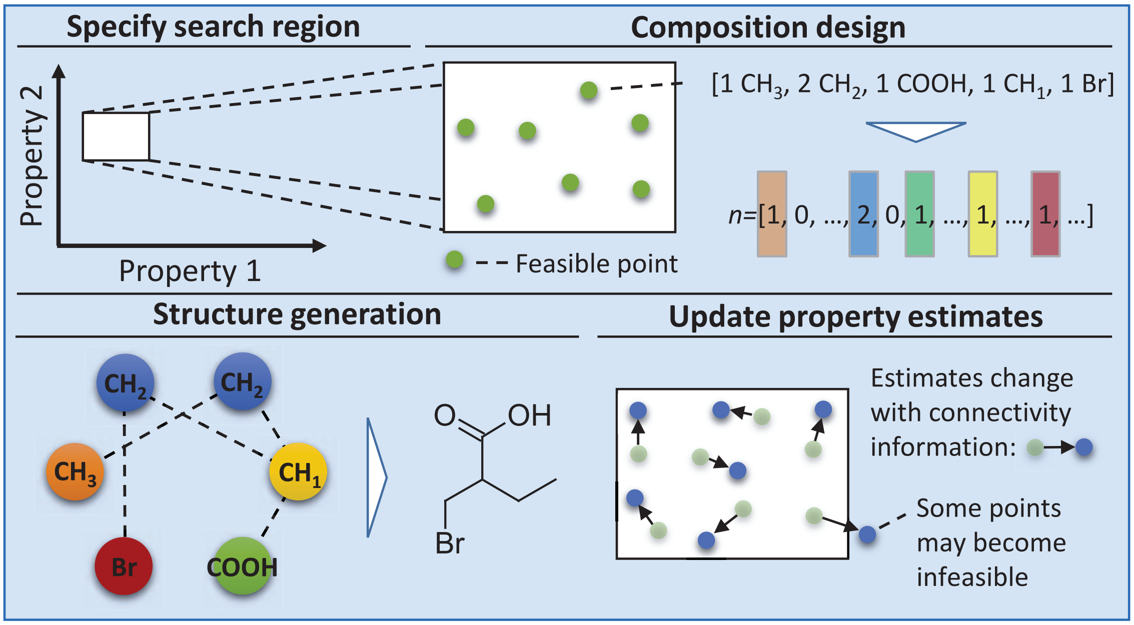

Specify search region
Composition design
Property 2
Property 1
Feature point
Feasible point
[1 CH$_3$, 2 CH$_2$, 1 COOH, 1 CH$_1$, 1 Br]
n=[1, 0, ..., 2, 0, 1, ..., 1, ..., 1, ...]
Structure generation
Update property estimates
CH$_2$
CH$_2$
CH$_3$
CH$_1$
Br
COOH
O
OH
Br
Estimates change with connectivity information:
Some points may become infeasible

Figure 2.1: AMODEO framework for molecular design

sequence of operations, many of which involve the solution of optimization subproblems themselves. For this reason, we will make use of derivative-free optimization (DFO) algorithms. DFO defines a group of algorithms which seek to optimize a function for which no derivative information is available, a function or derivative evaluation which may be computationally expensive, or a function which is not available in algebraic form. Often, the function in question is assumed to be deterministic. DFO algorithms evaluate one or several points in the space of the independent variables, interpret the results based on a wide variety of approaches, and choose one or more new points to evaluate or terminate because some convergence criterion was reached or a time limit was exceeded. There are a number of DFO algorithms, and they can broadly be divided into a few categories:

1. Local vs. global. Local algorithms aim to determine a point of local optimality. These algorithms often search in a small area around a current trial solution, attempting to find an improving direction. Global algorithms, on the other hand, attempt to search the entire feasible space for a point of global optimality. These algorithms typically require more function evaluations than local search, as they try to check objective values in many areas of the feasible region.

2. Deterministic vs. stochastic. When run multiple times from the same starting point, deterministic algorithms will repeatedly evaluate the same set of points and arrive at the same solution. Stochastic algorithms incorporate some element of randomness, often in the form of probabilities of taking a new solution over a previous one. Deterministic algorithms assume a fixed set of operations will lead to good solutions, while stochastic algorithms rely on an often large number of trial points to ensure a good solution.

3. Model-based vs. direct. Model-based algorithms try to fit some functional form to objective values collected over the space of the independent variables. These models can be quite simple and computed from a few points, or they can be very complicated and require the evaluation of many trial points. Direct algorithms assume no underlying model form and simply evaluate the objective function based on some pattern. These patterns are often very straightforward: evaluating the objective at all centroids of the hypercubes surrounding a current trial point, or dividing the entire feasible region into sections and evaluating the objective in a point of each section.

In our approach, the mixture design problems can be posed with few independent variables, which many DFO algorithms have been shown to be efficient at solving (Rios & Sahinidis 2013). For our study, we will consider 27 DFO solvers, summarized in Table 2.1. These solvers apply a variety of algorithms, including spatial search (Hooke & Jeeves 1961; Nelder & Mead 1965; Audet & Dennis Jr. 2006), trust region methods (Powell 2002; Conn et al. 1997), surrogate model building (Booker et al. 1999; Huyer & Neumaier 2008), genetic algorithms (Holland 1975), and hit-and-run methods (Boneh & Golan 1979; Smith 1984). For more information, see Rios & Sahinidis (2013). While Rios & Sahinidis (2013) provides a comparison of these algorithms on over 500 test problems, all these problems are algebraic. The literature currently lacks systematic comparisons of these algorithms on true black-box problems. We aim to fill part of this gap in the computational results section of this study.

Table 2.1: DFO solvers tested in a performance comparison on 4 CAMxD test problems

	Global	Local	
Deterministic	MCS (Neumaier 2011) SNOBFIT (Huyer & Neumaier 2008) TOMLAB/CGO (Holmström et al. 2011), TOMLAB/GLB (Holmström et al. 2011) TOMLAB/RBF (Holmström et al. 2011)	DFO (Scheinberg 2003), BOBYQA (Powell 2009) IMFIL (Kelley 2011), NEWUOA (Powell 2006)	Model-based
	DAKOTA/DIR (Sandia National Laboratories 2011) TOMLAB/GLC (Holmström et al. 2011), TOMLAB/LGO (Pintér et al. 2006) TOMLAB/MM (Holmström et al. 2011)	NOMAD (Abramson et al. 2017; Le Digabel 2009), DAKOTA/PAT (Sandia National Laboratories 2011), FMINSEARCH (Lagarias et al. 1998), SID-PSM (Custódio & Vicente 2008) HOPSPACK (Plantenga 2009), TOMLAB/MSNLP (Holmström et al. 2011) TOMLAB/GLCC (Holmström et al. 2011)	Direct
Stochastic	CMA-ES (Hansen 2017)	n/a	Model-based
	ASA Ingber (2011), DAKOTA/EA, DAKOTA/S-W (Sandia National Laboratories 2011), GLOBAL (Csendes et al. 2008), PSWARM (Vaz 2011)	PRAXIS (Brent 1973)	Direct

2.3 MIXTURE DESIGN USING DERIVATIVE-FREE OPTIMIZATION

With the tools of AMODEO and DFO briefly described, we are now in a position to present our DFO/AMODEO-driven mixture design algorithm. We begin by stating the mixture design problem and presenting a general formulation for it. We consider the problem of designing a K-component mixture. Some of the components may be predetermined but at least one is unknown and its structure and mole fraction in the mixture must be determined in a way that optimizes a function of mixture properties. In addition, it may be necessary to enforce constraints on various pure component and mixture properties.

The indices i, j, and k will, respectively, denote components in the mixture $(i = 1, \ldots, K)$, pure component properties $(j = 1, \ldots, C)$, and mixture properties of interest $(k = 1, \ldots, N)$. For component i, x_i will denote its mole fraction in the mixture. Let p_{ij} denote the value of property j for pure component i. We will assume that we can estimate all these properties via some family of functions f from their corresponding chemical structures, which are determined by specifying a vector n of molecular descriptors. Component structures must satisfy molecular bonding and connectivity constraints, denoted here by requiring that $n \in \mathcal{S}$. The mixture itself possesses properties, q_k, $k = 1, \ldots, N$, that are functions of the pure component properties and mole fractions, i.e., $q_k = g_k(x, p)$, $k = 1, \ldots, N$. The mixture design problem is to determine the components and their mole fractions so that a certain performance criterion $C(q)$ is optimized. We can therefore formulate this problem as follows:

$$\text{(CAMxD)} \qquad \min_{n,x} \quad C(q) \tag{2.9}$$

$$\text{s.t.} \quad q = g(x,p) \tag{2.10}$$

$$p = f(n) \tag{2.11}$$

$$h(x,p,q) \leq 0 \tag{2.12}$$

$$l(x,p,q) = 0 \tag{2.13}$$

$$\sum_i x_i = 1 \tag{2.14}$$

$$p^L \leq p \leq p^U \tag{2.15}$$

$$q^L \leq q \leq q^U \tag{2.16}$$

$$n \in \mathcal{S} \tag{2.17}$$

The problem involves a combinatorial aspect (variables n) to determine the molecular structure of each component as well as a continuous part (variables x, p, and q) involving mole fractions and properties. In (2.9), some function C of mixture properties q is minimized over the continuous variables, x, and discrete variables, n, of the problem. Equation (2.10) transforms individual component properties and mole fractions into mixture properties. Equation (2.11) encompasses the functions which estimate individual component values from each molecule's constitutive subgroups (Marrero & Gani 2001; 2002). Constraints (2.12) and (2.13) are inequality and equality constraints imposed on mixture and component properties. Constraint (2.14) simply requires all mole fractions to sum to 1. Constraints (2.15) and (2.16) represent the bounds placed on individual component and mixture properties. Finally, constraint (2.17) requires that the designed compositions must assemble somehow into a chemically feasible structure, as given by a class of valence balance constraints (Odele & Macchietto 1993; Sahinidis et al. 2003).

Molecular property prediction can be approached through GC methods. Molecular and mixture properties are nonlinear functions of molecular structure and mole fractions (Joback & Stephanopoulos 1990; 1995; Odele & Macchietto 1993; Marrero & Gani 2001). As a result, model (CAMxD) can be viewed as a mixed-integer nonlinear optimization formulation (MINLP).

Approached from an MINLP point of view, CAMxD is a challenging problem to solve given its extremely large size and non-linearity. Even models for pure molecular structure design result in very difficult MINLPs that have required the use of decomposition techniques to solve them. To wit, Samudra & Sahinidis (2013b) has demonstrated that decomposition techniques can reduce the time required to solve CAMD problems to $< 1\%$ of the time required to solve the monolithic MINLP, and even more so if many groups are considered. In its simplest form ($K = 1$), CAMxD reduces to pure component design. Therefore, CAMxD can be expected to be much harder, in general, than the pure component design problem of Samudra & Sahinidis (2013b). To address this challenge, we project CAMxD onto the space of pure component properties, thus facilitating a natural decomposition scheme that can capitalize on existing molecular design methodology. In the space of pure component properties, we no longer need to model molecular structures explicitly, but we still must determine feasible values for variables in the CAMxD problem. Because the objective function of the problem is an implicit function of x and n, we must relate the individual component property space to feasible values for structures (n) and then use these to determine optimal mole fractions (x). More formally, the optimization of CAMxD can be approached as follows:

1. Given a candidate property vector p_T, find n with corresponding $f(n)$ that is as close to p_T as possible; this is a CAMD problem.

2. Use $f(n)$ to find optimal mole fractions and solve the continuous part of the problem; we will show how to address this problem via algebraic optimization.

3. Interpret the objective value and choose a new p if necessary; we address this problem via DFO.

The algorithm is shown pictorially in Figure 2.2. Note the decomposition of the CAMxD problem into two stages. Once p_T has been specified, Step 1 relies on executing a highly complex process that involves the solution of several optimization problems. Therefore, computation of $C(q)$ in Step 3 is no longer an explicit function of p_T after projection. For this reason, we will utilize DFO algorithms.

Our algorithm employs a decomposition strategy similar to those mentioned from the literature, but designs compositions successively without having to enumerate every possible structure for every component. More specifically, the algorithm exploits DFO as a tool to probe pure component property space, so every iteration either produces a set of molecules which does satisfy some specified mixture criteria or determines that no molecules in the search region are compatible. This means that we can search over small property ranges for each component and consider every type of molecule while avoiding the computational burden of enumerating every possible structure for every component.

As a DFO algorithm guides the search through molecular property space, Step 1 of the above algorithm calls for the design of molecules for a given property. In our implementation, this problem is solved using the molecular design methodology AMODEO (Samudra & Sahinidis 2013b) to find molecules in a small neighborhood in the vicinity of the DFO trial point in property space. AMODEO is highly efficient for this task, thus mak-

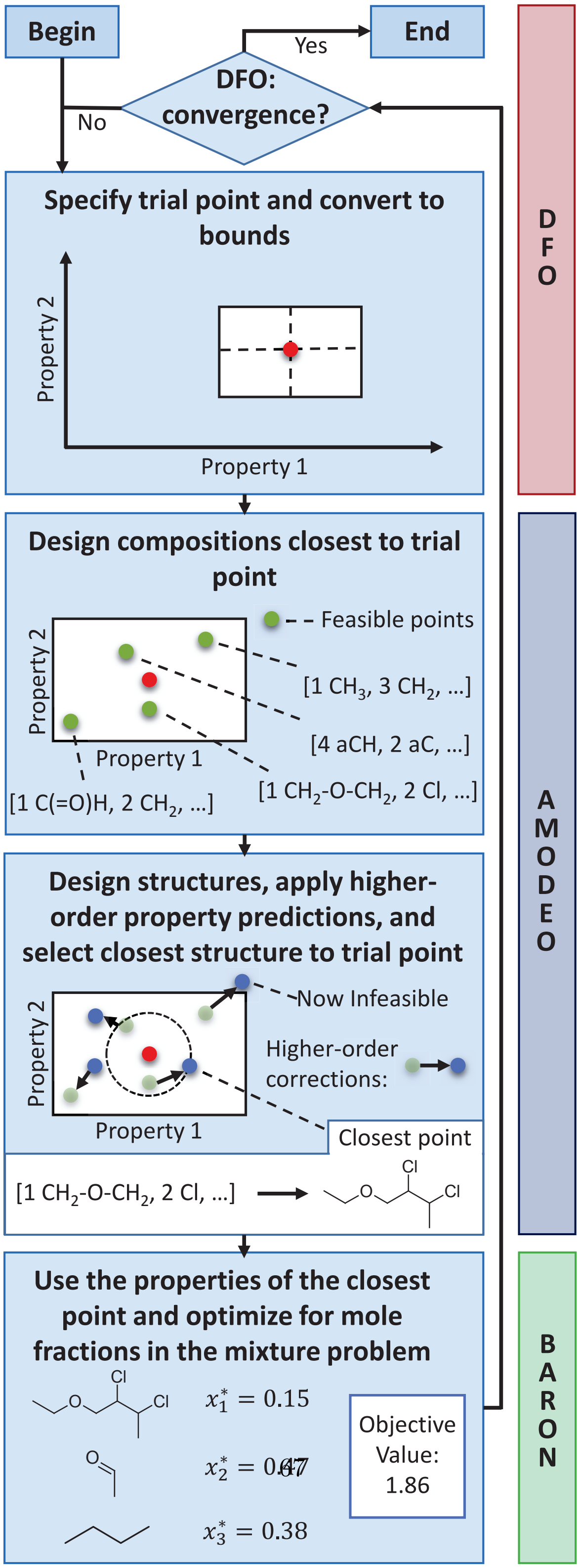

Begin
End
Yes
DFO: convergence?
No
Specify trial point and convert to bounds
Property 2
Property 1
Design compositions closest to trial point
Feasible points
[1 CH$_3$, 3 CH$_2$, ...]
[4 aCH, 2 aC, ...]
[1 CH$_2$-O-CH$_2$, 2 Cl, ...]
[1 C(=O)H, 2 CH$_2$, ...]
Property 2
Property 1
Design structures, apply higher-order property predictions, and select closest structure to trial point
Now Infeasible
Higher-order corrections:
Property 2
Property 1
Closest point
[1 CH$_2$-O-CH$_2$, 2 Cl, ...]
Cl
O
Cl
Use the properties of the closest point and optimize for mole fractions in the mixture problem
Cl
O
Cl
$x_1^* = 0.15$
O
$x_2^* = 0.47$
$x_3^* = 0.38$
Objective Value: 1.86
DFO
AMODEO
BARON

Figure 2.2: Mixture design algorithm

ing it possible for the overall approach to design optimal mixtures while considering a much larger search space than previously possible. One change in the AMODEO formulation is introduced in order to make the approach more amenable to mixture design. Since we aim to design structures with properties as close as possible to the DFO trial point, we alter the Phase 1 problem to minimize the distance between first-order group estimates and the DFO point, p_T:

$$\begin{aligned}
\min_n \quad & \sum_k w_k \left[\frac{d_k^+ + d_k^-}{\pi_k - \kappa_k} \right] && (2.18) \\
\text{s.t.} \quad & d_k^+ - d_k^- = \sum_{i \in F} c_i^k n_i - f_k^{-1}(p_T^k) \quad \forall k \in \mathcal{K} && (2.19) \\
& \kappa_k \leq \sum_{i \in F} c_i^k n_i \leq \pi_k \quad \forall k \in \mathcal{K} && (2.20) \\
& n \in \mathcal{S} && (2.21) \\
& \vdots
\end{aligned}$$

In this modified formulation, we introduce two positive continuous variables, d_k^+ and d_k^-, to account for positive and negative deviations of the transformed property estimate from the transformed DFO target point,$f_k^{-1}(p_T^k)$, for each property k. These positive and negative deviations are captured in the new constraint Equation (2.19). d_k^+ and d_k^- are scaled by the transformed property ranges and minimized in the objective, Equation (2.18). Also, we introduce a weighting parameter, w_k, to give preference to minimizing certain properties over others. For our problems, w_k will always take a value of 1 or 0, depending on whether a property is incorporated into our property search space. This alteration of the composition design stage of AMODEO is well-suited to our algorithm as it aids in selecting molecular structures closest to the DFO trial point. Furthermore, we can specify a maximum number of compositions ($C_{\max}$) to design in

Phase 1. With the reformulated Phase 1 problem and a reasonable value for $C_{\max}$, the modified composition design stage can glean accurate information about the neighborhood of the DFO trial point while avoiding the cumbersome design of every structure within the property bounds.

The algorithm begins with the DFO solver providing a trial point. The trial point signifies relevant property values for each component in the mixture. We choose pure component properties which are expected to exhibit some relationship with the objective function value. For example, the trial point $p_T = (335, 1.2, 5, 421, 1.0, 2)$ may represent a two-component mixture with three trial property values for each component, say melting point, viscosity, and number of oxygens in the structure. The property values are then each transformed as follows:

$$
\begin{aligned}
&p^L = p_T - \tau(P^U - P^L)\\
&p^U = p_T + \tau(P^U - P^L)\\
&\tau \in (0, 1]
\end{aligned}
$$

where p_T is the value of the property generated by the DFO solver, p^L and p^U are the lower and upper bounds, respectively, of the property ranges we will use in the CAMD problem, P^L and P^U are the lower and upper bounds for the property in the entire search space, and τ is a multiplier that tells us how much of the feasible property range to check. With these property bounds, AMODEO generates a small number of structures for each component, based on minimizing the distance of each structure's transformed first-order property estimates to $f^{-1}(p_T)$.

After solving the molecular design problem for each component, we select the compounds whose calculated properties (now using the full Marrero-Gani model (Marrero & Gani 2001; 2002)) are closest to the trial point by summation of percent error. These selected structures can be solutions of the molecular design problem or can come from a list of all previously found, feasible solutions. In Step 2, the estimated property values of the molecular structures are converted into parameters for an optimization problem. The optimization problem is typically nonconvex and solved with the global solver BARON (Tawarmalani & Sahinidis 2004) to determine the mole fractions of each component in the mixture, and the objective function value is reported to the DFO solver. If the overall problem is infeasible or if no structures are found in the area around the trial point, a large objective function value is reported to the DFO solver. Then, in Step 3, the DFO algorithm uses the previously collected information to assess termination criteria and possibly provide a new trial point.

DFO solvers often perform better if given a good starting point. Occasionally, the CAMxD formulation lends itself to a pre-screening stage designed to produce a promising starting point. This is done by solving (CAMxD) while disregarding the equation $p = f(n)$ and discarding the combinatorial part of the problem. This means that we solve the continuous part of the problem to determine the ideal pure component properties for every component, regardless of whether a molecular structure exists for those property values. Given these ideal pure component properties, we solve a CAMD problem for each component to determine feasible structures closest in property space to this ideal point. This pre-screening stage can sometimes provide a good trial point for DFO solvers, but depending on the nature of the problem, this can also lead to highly unrealistic property targets for which no molecular structures exist in the surrounding area. In the context of our mixture design algorithm, this pre-screening stage can only be performed if every pure component property in the problem can be estimated by Marrero-Gani groups. Otherwise, the ideal property values may represent properties that are not present in our design space. If the continuous part of the problem relies on more complex property models, we can provide DFO with a set of known molecular structures as a starting point. For example, in attempting to find an alternative to a certain material in an industrial process, we can supply a starting point that represents the properties of that material in our design space. For these starting point procedures, we both introduce a property vector starting point for DFO as well as add the corresponding molecular structures to our list of feasible solutions. In some cases, a DFO solver which ignores a given starting point can still exhibit an improved solution with these starting point procedures due to the presence of additional molecular structures in the list of feasible solutions. We will investigate these starting point generation procedures in the following sections.

Finally, we note that there are guarantees of global optimality for some of the DFO algorithms we investigate. These are often dependent on certain problem conditions and are discussed in more detail in Conn et al. (2009). In all of the following case studies, we report CPU times with a 2.84 GHz processor.

2.4 ILLUSTRATIVE EXAMPLES

We will first motivate and provide background to the approach with illustrative examples. For the purposes of these examples, we will first assume that the objective function of the mixture design problem is some algebraic expression $a(q)$ of the mixture properties.

Further, we do not consider any additional constraints involving mixture and component properties:

$$
\text{(CAMxD)} \qquad \begin{aligned}
\min \quad & a(q) && (2.22)\\
\text{s.t.} \quad & q = g(x,p) && (2.23)\\
& p = f(n) && (2.24)\\
& \sum_i x_i = 1 && (2.25)\\
& p^L \leq p \leq p^U && (2.26)\\
& q^L \leq p \leq q^U && (2.27)
\end{aligned}
$$

The function g now represents a set of mixing rules used to estimate mixture properties from pure component properties and mole fractions. Below are two examples of mixing rules we will use for this purpose:

- Boiling Point: $T_b^{\text{mix}} = \sum_j x_j T_b^j$ (linear mixing rule).
- Kinematic Viscosity: $\eta^{\text{mix}} = \exp\left(\sum_j \ln(\eta_j)\, x_j\right)$ (nonlinear mixing rule).

2.4.0.1 *First illustrative example*

An example was done to demonstrate the concept on the following problem:

> Given three mixture components (THF, acetone, and cyclohexane), design a fourth component to maximize the ratio of the mixture's kinematic viscosity to its boiling point. Each existing component must make up at least 20 mol % of the mixture.

Here, the DFO trial points will be two-dimensional and will represent the kinematic viscosity and boiling point of the fourth component in pure component property space.

We design molecules which are aliphatic and have no rings. We include structures with up to ten carbons, two oxygens, and two nitrogens. Since the molecule we design should be part of a liquid mixture, we will make sure it is a liquid at operating temperature (300 K) by bounding the fourth component's melting and boiling points. We further assume that we do not want to work with a compound that is too viscous, so we add a constraint to ensure the viscosity of this compound does not exceed 2.4 cP. We impose a 10 minute time limit on each DFO solver, which is strict given the large molecular search space. It should be noted that the limits we impose on the molecular structures are much more constraining than our algorithm requires. This is only done so that a comparison to a decomposition and enumeration approach is possible. The specifications of the problem are outlined in Table 2.2.

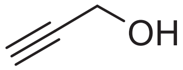
For the sake of comparison, we followed the enumeration procedure of Karunanithi et al. (2005) to solve this problem to global optimality. Specifically, we designed every structure ($\sim 120,000$ molecules) in the given feasible region using AMODEO and evaluated the objective for each of those molecules. This process took over 2 days of computer time and produced the same optimal solution as was found by many DFO solvers in 10 minutes. In both cases, we solve the continuous part of the problem with BARON (Tawarmalani & Sahinidis 2004) to global optimality. It should be noted that a few structures exist that do have values close to the theoretical optimum and do produce objective values of approximately 0.002996 with a theoretical optimal η_4 and T_b^4 of—quite unsurprisingly—2.4 cP and 300 K. However, these structures usually did not have appropriate groups available for the predicted properties or, in a few cases, produced estimates widely deviant from reality. As such, these solutions were removed from consideration. The best structure found had an objective value of 0.00276, and is shown in Table 2.3.

Table 2.2: Summary of important values for illustrative examples 1 and 2

Parameter	Value/Range	Additional Information
Time limit	600 s	Maximum allowable time for the algorithm
Iteration limit	300	Maximum number of steps the algorithm can perform
DFO inputs	η_4, T_b^4	Viscosity and boiling point of component 4
τ	10%	Property bounds relaxation around DFO trial point
$C_{\max}$	10	Maximum number of compositions determined during each iteration
η_4	[0 cP, 2.4 cP]	Range for kinematic viscosity of component 4
T_b^4	[300 K, 600 K]	Range for boiling point of component 4
T_m^4	[0 K, 290 K]	Range for melting point component 4
Carbons	10	Maximum number of carbons in the designed component
Oxygens	2	Maximum number of oxygens in the designed component
Nitrogens	2	Maximum number of nitrogens in the designed component
Double bonds	2	Maximum number of double bonds in the designed component
Enumeration Time	> 2 days	Time to solve the problem with enumeration

Each DFO algorithm was tested with five different randomly generated starting points. For each run, we calculated the % error between the objective function value of the optimal solution and that returned by the DFO solver. The results are shown in Figure 2.3 for 27 different DFO algorithms and compare average percent error over five runs with random starting points, best percent error over the five runs, and percent error of a single run from a favorable starting point (denoted by SP in this and subsequent figures). The starting point was generated by removing the $p = f(n)$ constraint as well

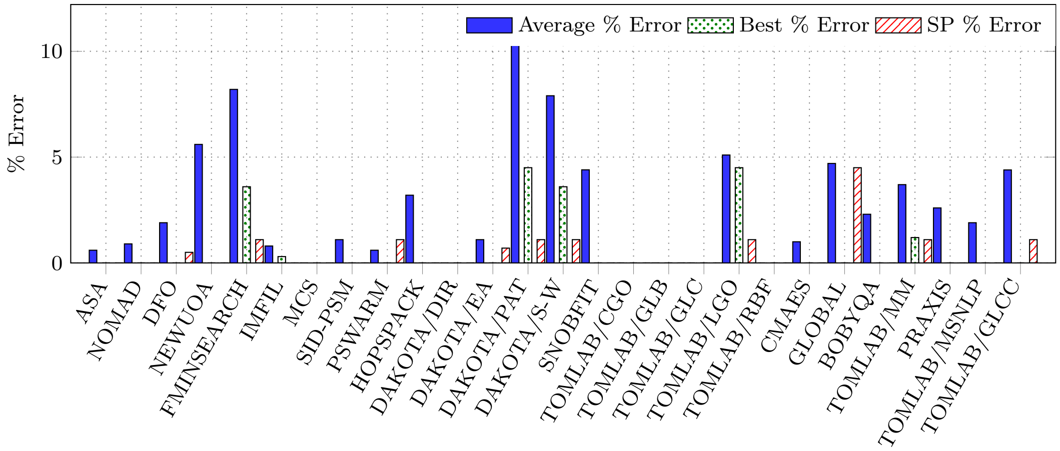
Table 2.3: Optimal structure for illustrative example 1

Optimal structure	**Properties**	
 **propargyl alcohol**	Objective value:	0.00276
	Molar mass:	42.04 g/mol
	Melting point:	237.79 K
	Boiling point:	326.09 K
	Viscosity:	1.88 cP

as all discrete variables n. This means that the objective was optimized over the feasible property space of component 4, regardless of whether a molecular structure existed for the optimal properties. An optimization problem was then solved to determine the closest feasible structure to the ideal properties. The resultant starting point structure was suboptimal but very near to the properties of the globally optimal solution. With this favorable starting point, almost every DFO solver was able to find the global optimum. The success of many solvers when given a starting point suggests that this methodology is greatly benefited by a good initial value, and that the approach can also be successfully applied to refine known solutions.

We also consider how much of the feasible molecular space was probed by each DFO solver. These numbers provide some idea of the efficiency of the DFO approach to mixture design. In this example, the solvers never explored more than $\sim 60\%$ of the projected property space. A value of 60% does not necessarily indicate that every structure within that 60% of feasible space was designed. Due to the limit placed on the number of feasible compositions ($C_{\max}$) to design at each step, the algorithm actually designs a much smaller number of structures than theoretically feasible. For the first illustrative example, the algorithm typically designed about 0.5% of possible solutions,

Figure 2.3: Comparison of DFO solvers for illustrative example 1



depending on the DFO method used, and only a small number of these solutions were evaluated for their objective value. We observe a clear positive correlation between DFO codes that find better solutions and codes that explore more of the feasible space. Only five DFO codes do not find the best solution from any of the five starting points, although they obtain solutions within less than a 5% error.

2.4.0.2 *Second illustrative example*

A highly non-convex problem is now addressed to demonstrate the potential of the algorithm in more challenging applications. This example retains the constraints of the previous problem but substitutes the previous objective with a much more difficult one.

The objective in this case could represent some value metric for the final properties of some formulated product. It is shown below.

$$\begin{aligned} \min \quad & 0.05(T_b^{\text{mix}} - 335)^2(-200(\eta^{\text{mix}} - 1.1)^2 + 2100(\eta^{\text{mix}} - 1.7)^2 - 2000\eta^{\text{mix}}) \\ & - 8300000\exp(-120(\eta^{\text{mix}} - 0.57)^2 - 0.0005(T_b^{\text{mix}} - 335)^2) \\ & - 8100000\exp(-120(\eta^{\text{mix}} - 0.8)^2 - 0.0005(T_b^{\text{mix}} - 350)^2) \end{aligned}$$

The results from applying the 27 different DFO algorithms to this problem are summarized in Figure 2.4. Many solvers found the best solution obtained by the enumeration procedure, shown in Table 2.4. The results of both examples indicate that many solvers can successfully be applied to these types of problems. The second example exhibited slightly worse performance in general, but we would expect an improvement with relaxing our stringent iteration (300) and time (600 s) limits. As shown in Figure 2.4, providing the algorithms with a good starting point led to the globally optimal solution in every DFO solver instance. Note again that the starting point structure was sub-optimal. This demonstrates that even harder, highly non-linear problems can be addressed successfully by all DFO solvers if given a good starting feasible solution. Only seven DFO codes do not find the best solution from any of the five starting points, although they obtain solutions within less than a 2% error.

The solvers explored at most $\sim$ 60% of the theoretically feasible property space. However, every solver in this example designed and tested on the order of hundreds of molecules. This is a clear efficiency improvement when contrasted with the $\sim 120,000$ structures designed for the enumeration procedure.

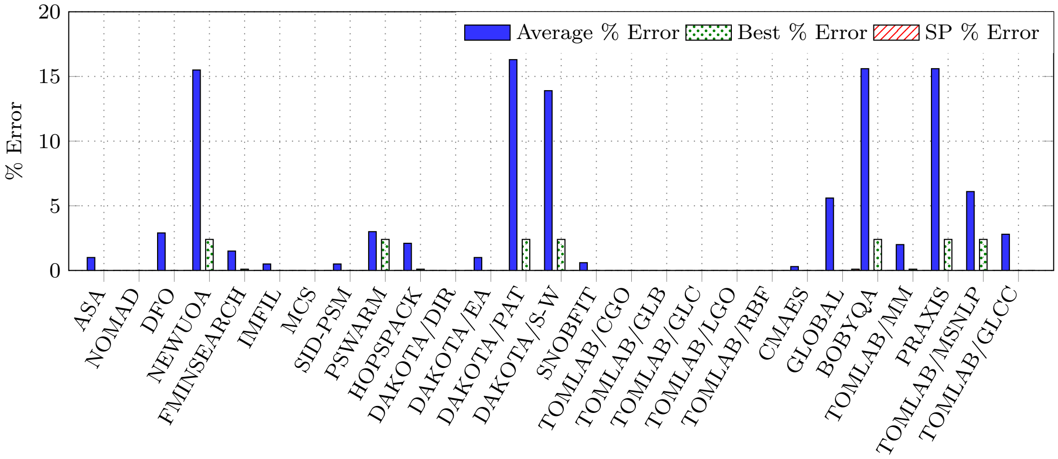


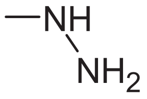
Figure 2.4: Comparison of DFO solvers for illustrative example 2

2.5 CASE STUDIES

2.5.1 *Case study 1: Cooling crystallization for ibuprofen*

We consider the problem posed in Karunanithi et al. (2006) for the design of an optimal solvent for purification via cooling crystallization. The process proceeds by dissolving a

Table 2.4: Optimal structure for illustrative example 2

Optimal solvent	**Properties**	
	Objective value:	8.313
	Molar mass:	46.07 g/mol
	Melting point:	257.95 K
	Boiling point:	323.72 K
methylhydrazine	Viscosity:	1.01 cP

solid compound of interest in a solvent at a high temperature (at which it should be very soluble) and then letting the solution cool until it reaches a certain lower temperature at which the solute is much less soluble. The solute crystallizes out of solution, leaving many impurities behind in the liquid phase. Our approach begins by specifying five independent variables: the three Hansen solubility parameters (δ_h, δ_p, δ_d), the octanol-water partition coefficient (K_{ow}), and aqueous solubility (C_w, given in mg/L). These properties were chosen because they are most closely related with solubility, the focus of this problem. The mixture design optimization formulation for cooling crystallization is as follows:

$$\text{(CAMxD-CC)} \qquad \max \quad \frac{100}{1-X_1}(1 - X_1/X_2) \tag{2.28}$$

$$\text{s.t.} \quad \ln(x_{1j}) = \frac{\Delta H_{fus}^{\text{Ibu}}}{RT_m^{\text{Ibu}}}\left(1 - \frac{T_m^{\text{Ibu}}}{\text{Temp}_j}\right) - \ln(\gamma_{1j}) \quad \forall j \tag{2.29}$$

$$\gamma_{1j} = \text{UNIFAC}(n, x_1, \text{Temp}_j) \quad \forall j \tag{2.30}$$

$$\left.\begin{array}{l} \delta_h \geq 8,\ T_f \geq 323,\ -\log(\text{LC}_{50}) \leq 3.3 \\ T_m \leq 270,\ T_b \geq 340,\ \mu \leq 1 \end{array}\right\} \tag{2.31}$$

$$X_j = \frac{M_1 x_{1j}}{\sum\limits_{i=1}^{2} x_{ij} M_i} \quad \forall j \tag{2.32}$$

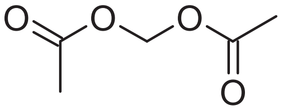
$$260 \leq \text{Temp}_j \leq 320 \quad \forall j \tag{2.33}$$

$$\sum_i x_{ij} = 1 \quad \forall j \tag{2.34}$$

$$n \in \mathcal{S} \tag{2.35}$$

In this formulation, the index j denotes the process at high and low temperatures for $j = 1$ and $j = 2$, respectively. The index i denotes ibuprofen ($i = 1$) and the solvent to be designed ($i = 2$). X_j represents the weight fraction solubilities of ibuprofen at temperature j. The X_j's are calculated using the molar masses M_i of each component. The x_{ij} variables are the mole fractions of each component i at each process condition j. $\Delta H_{fus}^{\text{Ibu}}$ and T_m^{Ibu} are the enthalpy of fusion and melting point of ibuprofen, respectively. Temp_j is the temperature of the solution for process condition j, R is the gas constant, and γ_{ij} is the activity coefficient of component i at temperature Temp_j. T_m and T_b represent the solvent's melting and boiling points, respectively. These are so constrained to ensure the designed solvent remains a liquid at all possible temperatures for this process. T_f is the flash point of the designed solvent, and it is constrained to be above 323 K for safety reasons. All of the variables and parameters representing temperature are given in degrees Kelvin. μ is the solvent's viscosity. This is constrained to be 1 cP at maximum to ensure its ease of use in the process. δ_h is the Hansen solubility parameter for hydrogen bonding. In keeping with Karunanithi et al. (2006), this is so constrained because Gordon & Amin (1984) report that a $\delta_h \geq 8$ is characteristic of good solvents for the cooling crystallization of ibuprofen. Solvents with this property produce crystals with larger particle size, drop in bulk volume, excellent manufacturability, and various other favorable properties. LC_{50} is a measure of toxicity and is estimated with the Martin-Young model (Martin & Young 2001), which is a group contribution method designed to estimate a chemical's 96-h LC_{50} for the fathead minnow (*Pimephales promelas*). This is constrained to be below a threshold of 3.3 in keeping with the original constraints of Karunanithi et al. (2006). The activity coefficients are calculated in (2.30) using the UNIFAC group contribution method (Fredenslund et al. 1975) with some additional parameters (Hansen et al. 1991; Wittig et al. 2003; Balslev & Abildskov 2002). Equation (2.29) represents a solid-liquid equilibrium

Table 2.5: Optimal structure for case study 1

Optimal solvent	**Properties**	
	Percent recovery:	97.94%
	Molar mass:	132.11 g/mol
	Melting point:	225.54 K
	Boiling point:	429.08 K
	Flash point:	362.66 K
methanediyl diacetate	$-\log(\mathrm{LC}_{50})$:	3.29
	Viscosity:	0.97 cP

condition (Gmehling et al. 1978) and is used for predicting the solubility of ibuprofen in a solvent. In (2.28), we optimize percent recovery for this process. A percent recovery of 100% would indicate that all of the ibuprofen put in to the process at the high temperature was recovered at the low temperature. As in Karunanithi et al. (2006), we constrain these process temperatures to a practical range in (2.33). Constraints (2.31) represent individual component requirements and are satisfied in the AMODEO phase of the algorithm, along with structural feasibility constraints (2.35) for molecular design. Our algorithm's decomposition of the problem allows us to address the remaining constraints as an NLP, which is solved with BARON (Tawarmalani & Sahinidis 2004) to global optimality.

Unlike the Karunanithi et al. (2006) formulation, we do not place any constraints on the Hildebrand solubility parameter. This is done because we are already working in the space of the Hansen solubility parameters, an alternative way to quantify solvent properties. Moreover, the Hildebrand solubility parameter cannot capture hydrogen bonding and polarizability effects, both of which play an important role in this particular solvation problem. In Table 2.6, we summarize the important parameters and constraints used in this case study.

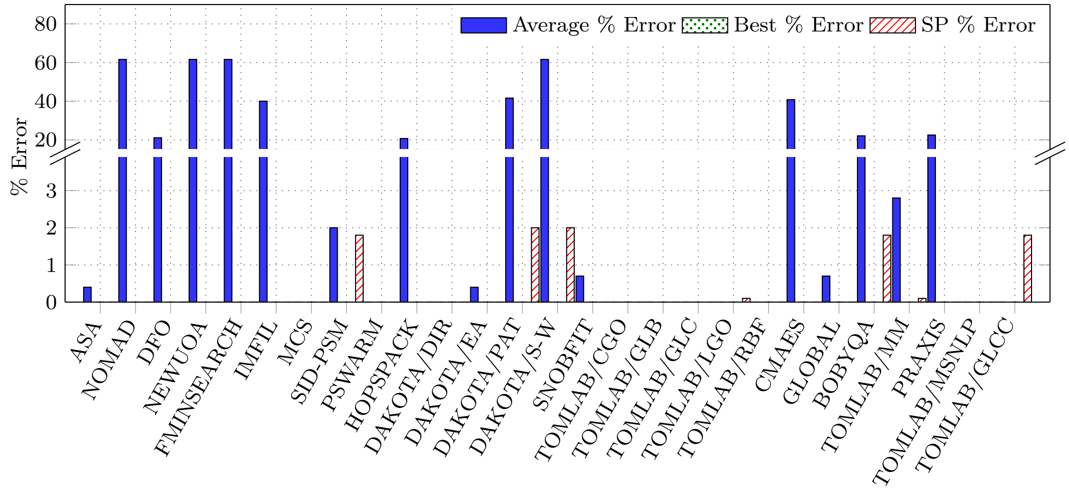
In Figure 2.5, we compare the results for this problem using the 27 DFO solvers. All solvers are able to find the globally optimal solution from one of the five random starting points. The optimal solvent for this problem as given by enumeration is shown in Table 2.5. Using our implementation of UNIFAC, this solvent provides a percent recovery of 97.94%. The current industrial standard for this process is *n*-hexane, which provides a percent recovery of 98.33% according to our UNIFAC implementation. However, *n*-hexane is not feasible in the above formulation because it violates the hydrogen bonding solubility parameter constraint ($\delta_H \geq 8$). Our optimal solvent, on the other hand, satisfies the hydrogen bonding parameter criteria specified by Gordon & Amin (1984), meaning a lower percent recovery in our case may still provide a preferable solvent to the industrial standard. If the solubility constraint is eliminated from the formulation, we obtain *n*-hexane as a solution. This observation validates the design potential of our approach. Furthermore, our approach finds a better solution than previously reported in the literature for this problem. Karunanithi et al. report a solvent that has a percent recovery of 89.93% (Karunanithi et al. 2007), while our designed solvent provides a higher percent recovery of 97.94%, very similar to that of the industrial standard.

Finally, we also tested the performance of each DFO solver when given a good starting point. This starting point was determined by searching the space of the five pure component property variables in the problem and then calculating the closest feasible structure to the industrial standard for this process, *n*-hexane. The closest feasible structure was non-optimal but led all but seven tested solvers to the globally optimal solution.

Table 2.6: Summary of important values for case study 1

Parameter	Value/Range	Additional Information
Time limit	600 s	Maximum allowable time for the algorithm
Iteration limit	1000	Maximum number of steps the algorithm can perform
DFO inputs	δ_h, δ_p, δ_d, $\log K_{ow}$, $\log C_w$	Solubility parameters of component 2
τ	20%	Property bounds relaxation around DFO trial point
$C_{\max}$	30	Maximum number of compositions determined during each iteration
δ_h	$[8, 14]$	Range for the hydrogen bonding solubility parameter
δ_p	$[4, 14]$	Range for the polarizability solubility parameter
δ_d	$[11, 18]$	Range for the dispersion solubility parameter
$\log K_{ow}$	$[-1, 2]$	Range for the octanol-water partion coefficient
$\log C_w$	$[2, 7]$	Range for the aqueous solubility coefficient
Carbons	10	Maximum number of carbons in the designed component
Oxygens	4	Maximum number of oxygens in the designed component
Enumeration time	3890 s	Time to solve the problem with enumeration

Figure 2.5: Comparison of DFO solvers for case study 1



2.5.2 *Case study 2: Drowning out crystallization*

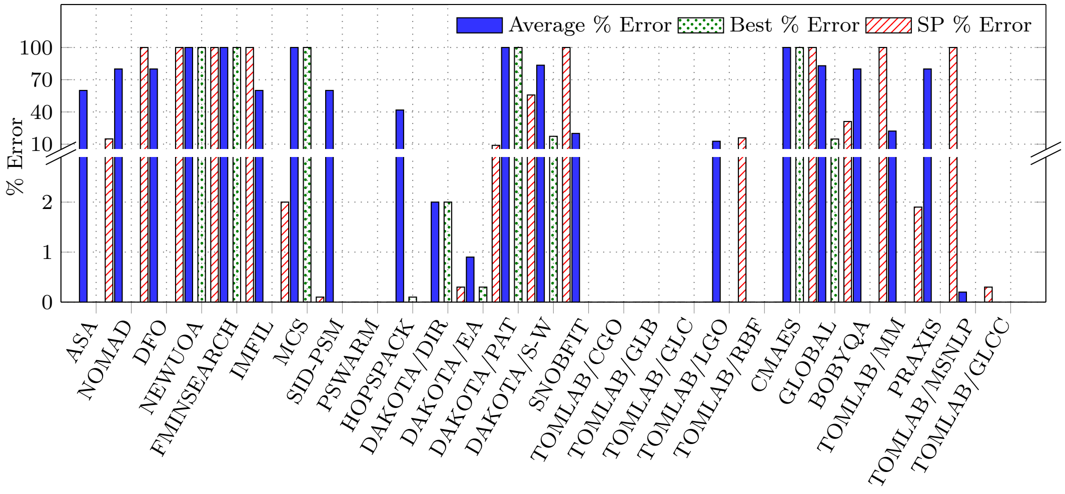
Cooling crystallization may not always be a feasible purification method. For example, temperature-sensitive compounds may decompose or react at higher temperatures, and some compounds may not demonstrate an appreciable change in solubility over a practical temperature range. For these reasons and others, pharmaceutical purification must often turn to drowning out crystallization for purification. Drowning out crystallization works by simply taking a solute/solvent mixture and adding an anti-solvent which reduces the solubility of the solute in the resulting mixture. The objective function in this case is slightly altered in the following formulation:

$$
\begin{aligned}
\text{(CAMxD-DOC)} \qquad & \max \quad \frac{100}{1-X_1}\left(1-\frac{X_1}{X_2}\left(1+\frac{M_{as}}{M_T}\right)\right) \\
& \text{s.t.} \quad \ln(x_{1j}) = \frac{\Delta H_{fus}^{\text{Ibu}}}{RT_m^{\text{Ibu}}}\left(1-\frac{T_m^{\text{Ibu}}}{298\text{ K}}\right) - \ln(\gamma_{1j}) \quad \forall j \\
& \frac{1}{x_{2j}} + \frac{\delta \ln(\gamma_{2j})}{\delta x_{2j}} \geq 0 \quad \forall j \qquad (2.36) \\
& \gamma_{ij} = \text{UNIFAC}(n, x_{ij}, 298\text{K}) \quad \forall j \\
& \left.\begin{aligned}
& \delta_h \geq 8,\ \mu \leq 1 \quad \text{for } i = 2 \\
& T_f \geq 323,\ -\log(\text{LC}_{50}) \leq 3.3 \quad \text{for } i \in \{2,3\} \\
& T_m \leq 270,\ T_b \geq 340 \quad \text{for } i \in \{2,3\}
\end{aligned}\right\} \qquad (2.37) \\
& X_j = \frac{M_1 x_{1j}}{\sum\limits_{i=1}^{3} x_{ij} M_i} \quad \forall j \\
& x_{31} = 0 \qquad (2.38) \\
& \sum_i x_{ij} = 1 \quad \forall j \\
& n \in \mathcal{S}
\end{aligned}
$$

There are three components in this problem, as indicated by index i: ibuprofen ($i = 1$), solvent ($i = 2$) and anti-solvent ($i = 3$). The index j represents the two situations: (1) the solvation of ibuprofen in just the solvent and (2) the presence of a ternary mixture of all three components. M_{as} and M_T represent the mass of the anti-solvent and total mass of the ternary mixture, respectively. Constraint (2.36) ensures that the solvent and anti-solvent are miscible, as given by Bernard et al. (1967). Constraint (2.38) guarantees that there is no anti-solvent in situation 1. We apply constraint (2.37) to the individual component properties of both solvent and anti-solvent in keeping with Karunanithi et al. (2006) that first posed this problem.

We address this problem in the space of six pure component property variables: the three Hansen solubility parameters (δ_d, δ_p, and δ_h) for both the solvent and anti-solvent. We removed K_{ow} and C_w from the design space because this problem is ultimately one of solvent/anti-solvent interactions, which were assumed to be more related to the solubility parameters. We compare the 27 DFO algorithms in Figure 2.6 given the problem conditions in Table 2.7. As shown, many of the algorithms can find the global optimum. Other solvers terminate in local minima due to the sparsity of feasible structures which provide good % recoveries for this design problem. Nonetheless, seventeen DFO solvers are able to find the global optimum from at least one of five different random starting points. The optimal solvent/antisolvent shown in Table 2.8 provide a recovery of 91.33%. Among other considerations, this pair likely works by reducing hydrogen bonding potential of ibuprofen with its original solvent upon the addition of the antisolvent. Again, our consideration of a larger molecular search space produces a better solution than previously reported. Specifically, the enumeration-based approach of Karunanithi et al. produces an optimal pair of molecules with a reported percent recovery of 69% (Karunanithi et al. 2006). We can also generate a favorable starting point based on finding the closest feasible solvent and anti-solvent from methanol and water, respectively. These structures are identified as a favorable solvent pair for this process in Filippa & Gasull (2013). Again, providing each DFO solver with a good starting point improves the results for many solvers, although not to the extent previously observed. In this case, the starting points were not close to the properties of the optimal structures in the projected space.

Figure 2.6: Comparison of DFO solvers for case study 2



2.5.3 *Extended case studies 1 and 2: Considering a larger feasible region*

In the two case studies, we looked at problems for which an enumeration approach was viable, although it often proved very time consuming. This was done to determine a global optimum for each problem so the DFO solvers could be accurately benchmarked. To demonstrate the full potential of the algorithm, we now re-examine case studies 1 and 2 in a much larger molecular search space. This large search space is enabled both by our DFO-decomposition approach to mixture design and our decomposition and optimization approach to molecular design. As shown in Table 2.9, this problem considers a large part of the molecular search space and is prohibitively large for a straightforward decomposition and enumeration approach.

We use the TOMLAB/CGO algorithm on each of these two larger problems, as it was able to solve every problem considered in the illustrative examples and prior case studies to global optimality. We provide no starting point to the solver. For the cool-

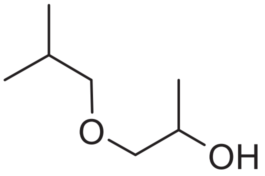
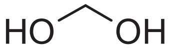
Table 2.7: Summary of important values for case study 2

Parameter	Value/Range	Additional Information
Time limit	3600 s	Maximum allowable time for the algorithm
Iteration limit	1000	Maximum number of steps the algorithm can perform
DFO inputs	δ_h, δ_p, δ_d	Solubility parameters of components 2 and 3
τ	20%	Property bounds relaxation around DFO trial point
$C_{\max}$	30	Maximum number of compositions determined during each iteration
δ_h	[8,30]	Range for the hydrogen bonding solubility parameter
δ_p	[0,20]	Range for the polarizability solubility parameter
δ_d	[0,20]	Range for the dispersion solubility parameter
Carbons	10	Maximum number of carbons in the designed components
Oxygens	3	Maximum number of oxygens in the designed components
Enumeration time	> 6 days	Time to solve the problem with enumeration

ing crystallization problem, our methodology with a larger feasible region was able to find a molecule which provides a 99.45% recovery, better than the current industrial standard of n-hexane, which provides only a 98.33% recovery. This structure is shown in Table 2.10. Furthermore, our designed solvent satisfies toxicity, viscosity, flash point, melting point, and boiling point constraints as well as the hydrogen bonding solubility parameter constraint given by Gordon & Amin (1984).

In the case of drowning out crystallization, our consideration of a larger molecular search space yields a solution with a 97.00% recovery, shown in Table 2.11. This outperforms the solvent pair given by Filippa & Gasull (2013), which only provides a 96.52% recovery. The identification of these two solutions by our algorithm speaks well to

Table 2.8: Optimal structures for case study 2

Optimal solvent	**Properties**	
	Molar mass:	120.15 g/mol
	Melting point:	157 K
	Boiling point:	430.68 K
	Flash point:	323.048 K
	$-\log(\mathrm{LC}_{50})$:	1.59
1-Isobutoxy-2-propanol	Viscosity:	0.86 cP
Optimal antisolvent	**Properties**	
	Molar mass:	48.04 g/mol
	Melting point:	259.93 K
	Boiling point:	393.04 K
	$-\log(\mathrm{LC}_{50})$:	-0.443
methanediol	Viscosity:	16.84 cP
Percent recovery: 91.33%		

its use in a general mixture design context. Clearly, considering a larger part of the molecular search space has great potential to provide better solutions. Such a large feasible region is no longer intractable and can be efficiently searched through with our DFO/AMODEO-driven approach.

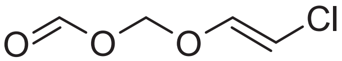
2.6 DISCUSSION

While many solvers were able to find the global optimum in every problem, some insight can be gleaned regarding which solvers can best be applied to these types of problems.

Table 2.9: Summary of important values for extended case studies 1 and 2

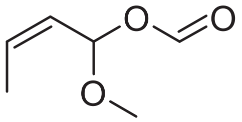
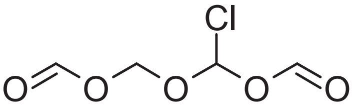
Parameter	Value/Range	Additional information
Time limit	N/A s	Maximum allowable time for the algorithm
Iteration limit	2000	Maximum number of steps the algorithm can perform
DFO inputs	δ_h, δ_p, δ_d	Solubility parameters of components 1 and 2
τ	20%	Property bounds relaxation around DFO trial point
$C_{\max}$	100	Maximum number of compositions determined during each iteration
δ_h	[0,40]	Range for the hydrogen bonding solubility parameter
δ_p	[0,30]	Range for the polarizability solubility parameter
δ_d	[0,30]	Range for the dispersion solubility parameter
$\log K_{ow}$	[-4,4]	Range for the octanol-water partion coefficient
$\log C_w$	[0,9]	Range for the aqueous solubility coefficient
Aliphatic chain carbons	15	Maximum number in the designed components
Aliphatic ring carbons	10	Maximum number in the designed components
Aromatic carbons	6	Maximum number in the designed components
Oxygens	5	Maximum number of carbons in the designed components
Chlorines	1	Maximum number of chlorine atoms in the designed components
Bromines	1	Maximum number of bromine atoms in the designed components
Double bonds	1	Maximum number of aliphatic double bonds in the compounds

Table 2.10: Optimal structure for cooling crystallization in extended case study 1

Optimal Solvent	**Properties**	
 **[2-chloroethenyl]oxymethyl formate**	Percent recovery:	99.45%
	Molar mass:	136.53 g/mol
	Melting point:	248.30 K
	Boiling point:	433.32 K
	Flash point:	333.30 K
	$-\log(\text{LC}_{50})$:	2.93
	Viscosity:	0.99 cP

To that end, we will perform a brief analysis on each solver's performance on all the problems. We will use z-scores to quantify a solver's performance on each problem. A z-score is defined as $z = (x - \mu)/\sigma$, where x is a data value, μ is the average of the whole range of data, and σ is its standard deviation. Thus, a z-score provides some idea of how much above or below the average a certain piece of data is when scaled by the standard deviation. We calculate z-scores based on average % error from the optimal solution using 5 randomly-generated starting points and report the average over the four problems considered in Figure 2.7. In our case, a lower % error indicates a better solution, so the best z-score will be the most negative. In Figure 2.7, it is shown that the algorithms TOMLAB/CGO, TOMLAB/GLB, TOMLAB/GLC, and TOMLAB/RBF outperform all the other solvers on these test problems. All of these solvers were able to find the globally optimal solution for all of the test problems considered. On the other end of the spectrum, many solvers with low scores are local and likely terminated in local minima or did not explore a large enough percentage of the feasible property space to find feasible structures. Most of the algorithms on top in the comparison

Table 2.11: Optimal structures for drowning out crystallization in extended case study 2

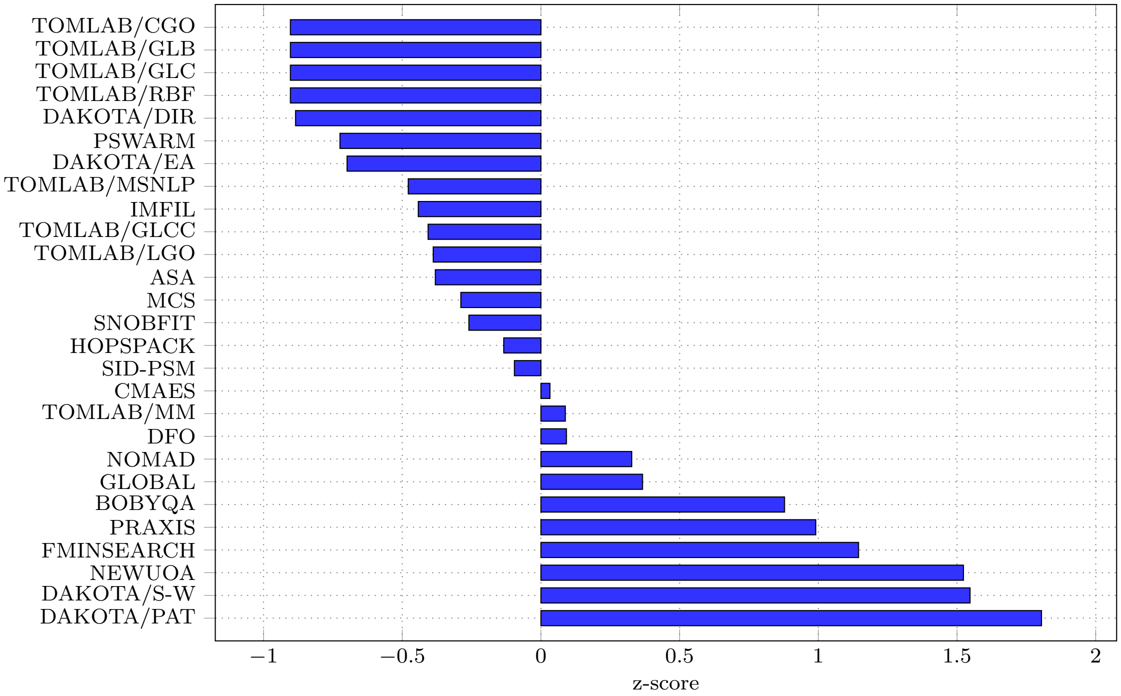
Optimal Solvent	**Properties**	
	Molar mass:	130.14 g/mol
	Melting point:	220.42 K
	Boiling point:	436.09 K
	Flash point:	326.28 K
	$-\log(\mathrm{LC}_{50})$:	3.28
1-methoxybut-2-en-1-yl formate	Viscosity:	0.66 cP
Optimal Antisolvent	**Properties**	
	Molar mass:	168.53 g/mol
	Melting point:	262.72 K
	Boiling point:	477.71 K
	Flash point:	374.50 K
[chloro(formyloxy)methoxy]methyl formate	$-\log(\mathrm{LC}_{50})$:	2.54
	Viscosity:	3.37 cP

Percent recovery: 97.00%

could find the optimal solution from every randomly generated starting point in many of the cases. However, the algorithms DAKOTA/EA and ASA did not find the optimal solution in any of the examples for any randomly generated starting point. These stochastic solvers, though usually able to provide good solutions, cannot reliably produce globally optimal solutions. Figure 2.8 shows the same comparison of z-scores when considering the best solution from the five randomly-generated starting points. Now, the algorithms TOMLAB/GLCC, TOMLAB/RBF, TOMLAB/GLC, TOMLAB/GLB, TOMLAB/CGO, SID-PSM, DFO, and NOMAD are able to solve all problems to global optimality with at least one of the five randomly generated starting points. Of note in this figure is the improved performance of a few stochastic solvers (GLOBAL, ASA, DAKOTA/EA). When given different starting points and more opportunities to solve the same problem, stochastic solvers can produce good solutions. Furthermore, a few local solvers (TOMLAB/GLCC, DFO, NOMAD, SID-PSM, HOPSPACK, IMFIL) showed improved performance. These solvers were likely given at least one starting point in a good region of property space. Figure 2.9 provides the same comparison when each solver is provided with a favorable starting point, as determined from one of the two starting point generation procedures outlined above. Again, many solvers (TOMLAB/RBF, TOMLAB/GLC, TOMLAB/GLB, TOMLAB/CGO, SID-PSM, SNOBFIT, HOPSPACK, MCS, TOMLAB/MSNLP, DAKOTA/DIR) were able to solve every problem to global optimality. A few model-based solvers (MCS, SNOBFIT) showed improved performance when utilized in conjunction with our initialization strategy. Furthermore, the local solvers TOMLAB/MSNLP, IMFIL, HOPSPACK, and SID-PSM demonstrated good performance at refining these favorable starting points.

In Figure 2.10, we compare the z-scores of each of the categories of DFO solvers we discussed in Section 2.2. In this case, we include z-scores for the average of five runs with randomly generated starting points, the best of those five runs, and the result after a starting point was generated based on methods discussed above. In the comparison of average of five runs, it is clear that the global solvers far outperform local solvers. This is expected as three of the four problems considered were highly non-convex. Next, model-based algorithms perform better than direct ones in general. This observation supports our assumption that the problem could be effectively considered in property space. The variables specified as inputs to DFO should have some bearing on the objective values, and model-based algorithms are able to deduce some relationship with each problem's variables and objective. Furthermore, deterministic algorithms perform better than stochastic ones on average. This is likely due to the fact that many

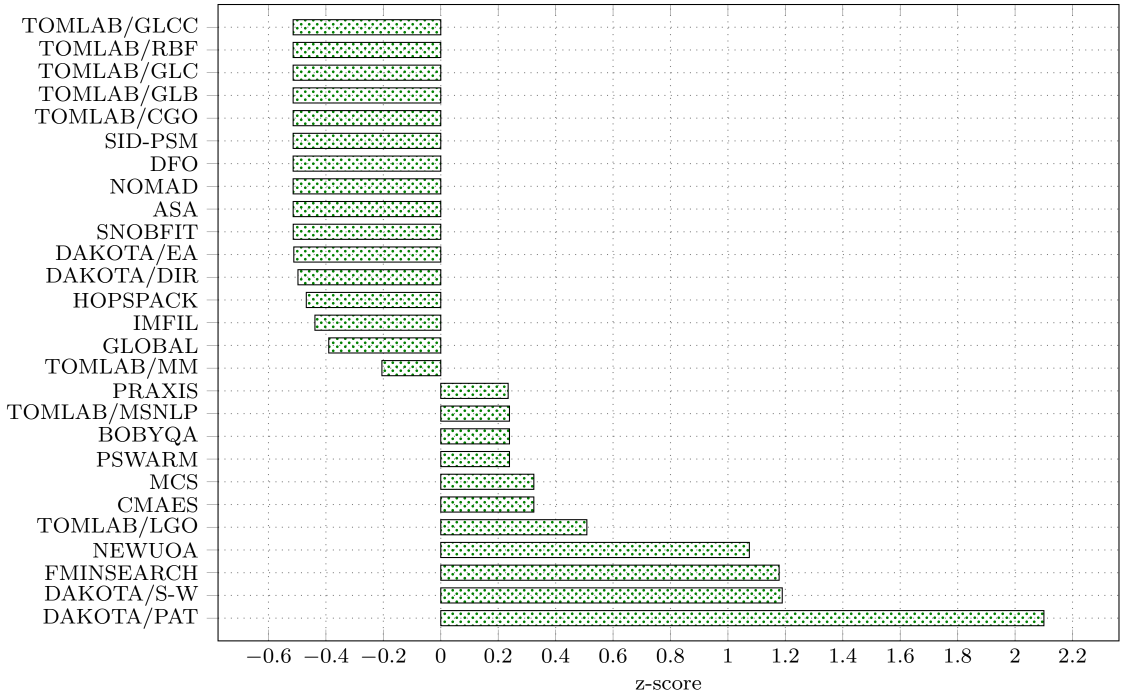
Figure 2.7: z-scores for the average objective value with randomly-generated starting points



deterministic solvers search the feasible region in a methodical way. Stochastic solvers can often miss a potentially good area of the search space due to the random nature of these algorithms.

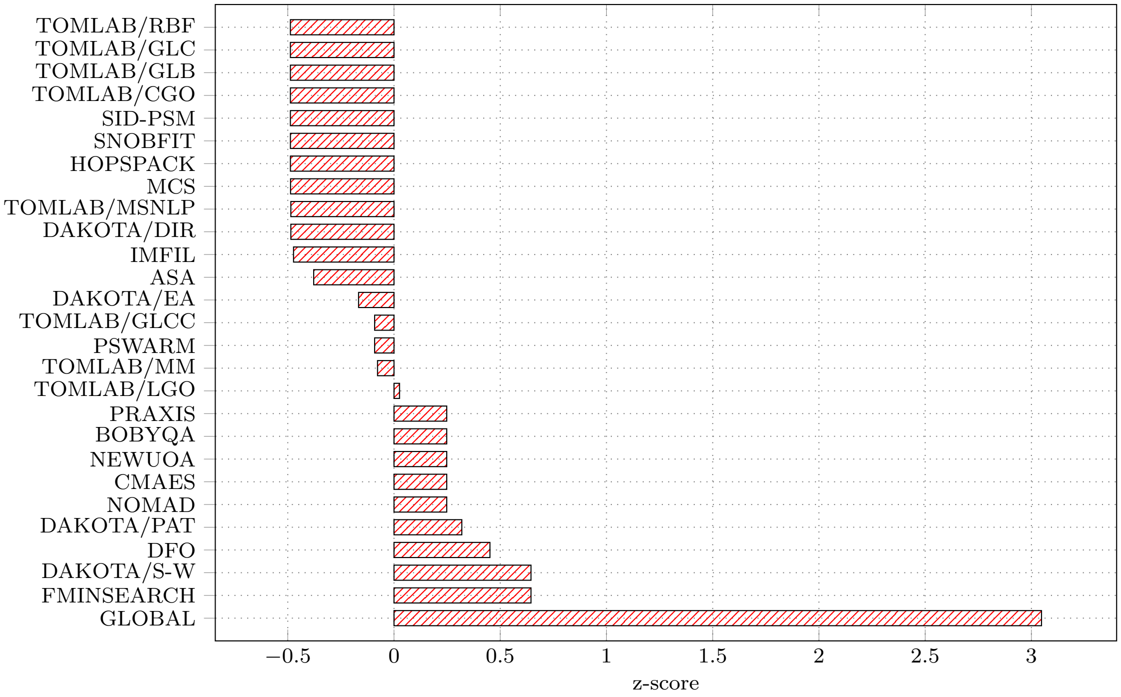
The comparison of the best value returned of the five runs illustrates a smaller but still significant performance gap between global and local solvers. In this case, global solvers are still ahead, but to a lesser extent. This is likely due to the fact that some of the randomly generated starting points placed local solvers in a favorable area of the feasible space. Stochastic solvers also improve marginally in comparison to deterministic ones. This is likely due to the fact that some of these starting points provide reasonable starting points and fewer stochastic solvers terminate due to lack of feasible

Figure 2.8: z-scores for the best objective value with randomly-generated starting points



solutions. Finally, model-based algorithms demonstrate improved performance as compared to direct ones. This can be explained by the fact that model-based algorithms can conceivably glean more information from a favorable starting point. The final comparison is between z-scores when a good starting point is generated based on one of the two methods discussed above. Here, local solvers show even more improvement, likely because these starting points are close to the global optimum in many cases. Deterministic algorithms, when provided a good starting point, do much better than their stochastic counterparts. Again, stochastic algorithms are still largely dependent on chance, and deterministic algorithms can use a good starting point for a more methodical search. Finally, model-based algorithms gain even more ground over direct ones in this com-

Figure 2.9: z-scores for the objective value with a good starting point



parison. This results from model-based algorithms attempting to deduce a relationship between the variables and the objective, a process greatly benefited if in the region around the global optimum.

Finally, in Figure 2.11, we remove all local solvers from consideration. The same trends are observed here, with model-based solvers outperforming direct ones and deterministic solvers outperforming stochastic ones.

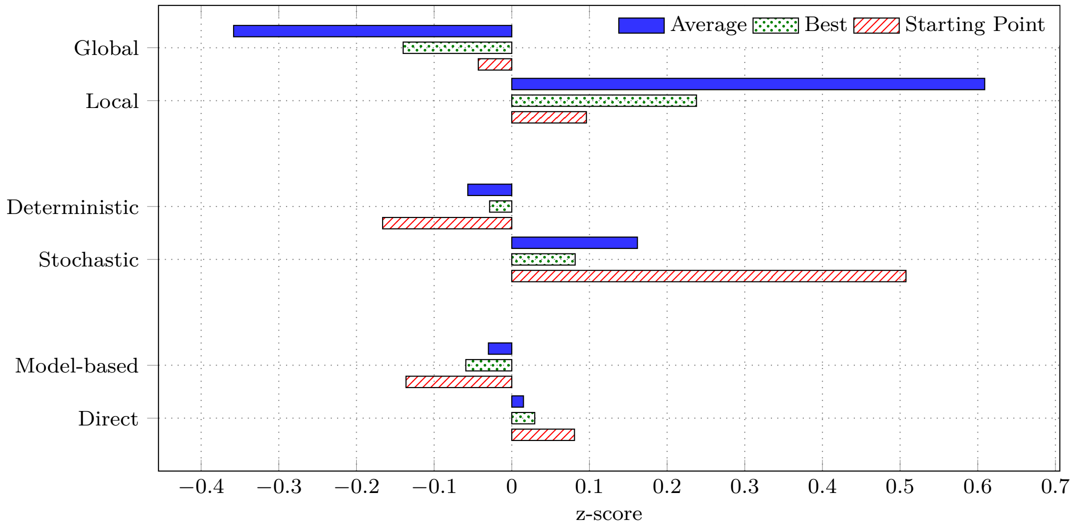


Figure 2.10: z-scores by solver category

2.7 CONCLUSIONS

A new, general-purpose methodology was developed to solve mixture design problems. In our approach, the large CAMxD MINLP is first projected onto the low-dimensional component property space. This projection leads to a natural way to decompose the mixture design problem into molecular design and mole fraction optimization problems. The search through component property space was performed with derivative-free optimization algorithms, and our computational results demonstrate that a portfolio of DFO algorithms is efficient at solving mixture design problems. Of note were global DFO algorithms and also those which work via some surrogate model building. The global algorithms search a larger area and have more stringent convergence criteria than local algorithms and, as a result, find better solutions. The good performance of model-building algorithms is consistent with our assumption that some underlying rela-

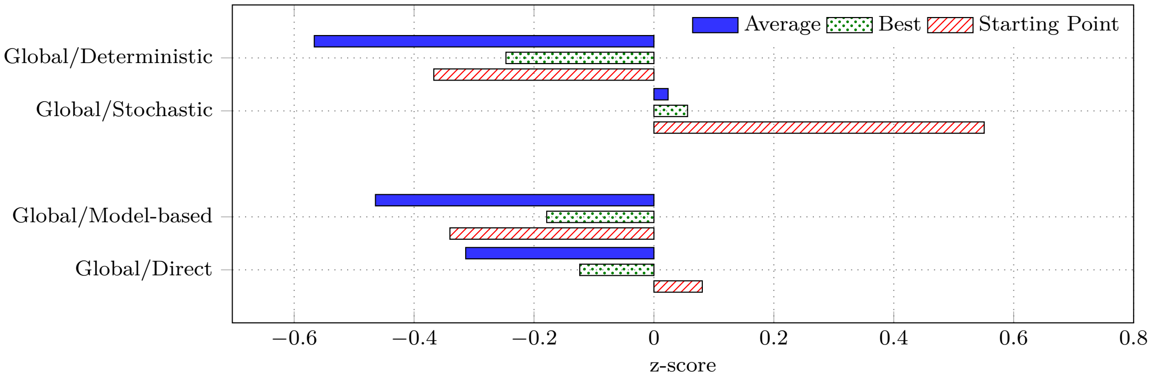


Figure 2.11: z-scores by solver category considering only global solvers

tionship should exist between the pure component properties and design objectives. One key advantage of our mixture design algorithm is that it is able to consider a very large molecular search space. The importance of such a large search space is underscored by the fact that our algorithm found better solutions to two problems from the literature than previously reported. This lends credence to the utility of the proposed approach in solving otherwise intractable CAMxD MINLPs. Finally, the reliable performance of DFO algorithms on these problems suggests that the projection onto pure component property space captures much of the relevant problem information and is a promising strategy for the solution of CAMxD problems in general.

3

MIXTURE DESIGN BASED ON COSMO-RS AND -SAC THERMODYNAMICS

3.1 INTRODUCTION

One distinguishing feature of CAMxD problems is that they simultaneously consider many different types of compounds and must be able to predict relevant mixture (typically solution-phase) properties. Given the complex nature of many mixture properties, it is often necessary to incorporate mixture thermodynamics equations directly into CAMxD problems. To this end, a very large number of CAMxD approaches have applied the UNIFAC (Fredenslund et al. 1975) method to calculating solubilities, phase equilibrium, partition coefficients, and various other properties. For example, early work from Gani & Brignole (1983) proposed the use of the UNIFAC method as a way to calculate activity coefficients in the design of an extraction solvent. Odele & Macchietto (1993) also applied UNIFAC to calculate mixture thermodynamics in a few solvent design problems. Many other approaches facilitated the use of UNIFAC in the context of molecular and mixture design, including multi-stage optimization strategies (Naser & Fournier 1991) and decomposition approaches (Gani & Fredenslund 1993; Klein et al.

1992; Conte et al. 2012; 2011b). Furthermore, the applications have been numerous, ranging from integrated design of compounds and processes (Papadopoulos & Linke 2006c) to calculating phase equilibria in designing crystallization solvents (Karunanithi et al. 2006) to designing solvents and solvent blends to reduce the environmental impact of industrial processes (Buxton et al. 1999; Pistikopoulos & Stefanis 1998). The solution strategies to UNIFAC-based molecular and mixture design are sometimes heuristical. For example, Ourique & Telles (1998) used simulated annealing to reduce the complexity of the problem so as to best apply UNIFAC. Other approaches (Benavides et al. 2015; Dyk & Nieuwoudt 2000) have also combined heuristic optimization techniques with UNIFAC to reduce the difficulty of searching through a large feasible region. This list of works is by no means exhaustive and is only intended to provide some idea of the diversity of applications of UNIFAC in the CAMxD literature.

More recently, there has been growing interest in using the SAFT equation of state (Chapman et al. 1989) to solve CAMD/CAMxD problems. For example, Pereira et al. (2011b) addressed the design of separations solvents using SAFT. Lampe et al. (2014) used SAFT to solve fluid selection and process optimization problems to design an organic Rankine cycle, and Lampe et al. (2015) utilized group contribution methods to better incorporate CAMD methodologies into these problems. SAFT is quickly becoming a useable model in a CAMD/CAMxD context due to the development of group contribution methods like SAFT-γ (Lymperiadis et al. 2007), application to the prediction of mixture properties (Papaioannou et al. 2011), and the use of the group contribution models in various design problems (Burger et al. 2015).

UNIFAC and SAFT-γ are natural choices for CAMD/CAMxD as they use group contribution methods, and groups often represent the design space of these problems. Furthermore, both of these methods have been demonstrated to be accurate and useful in a molecular and mixture design context. However, one significant issue with both is that they rely on binary interaction parameters for every pair of groups in solution. Estimating these parameters requires large data sets of thermodynamic properties, and such data sets often lack enough chemical diversity to make robust parameter estimates for many types of molecular structures. Consequently, many of these binary interaction parameters are simply not available. In this way, the design space of any CAMxD problem using UNIFAC or SAFT-γ is inherently limited to the chemical space represented by the available binary interaction parameters. An alternative way of estimating the thermodynamics of mixtures is through using one of several post-processing methods for the COSMO solvation model (Klamt & Schüürmann 1993), a relative of continuum solvation models used in quantum chemistry calculations. COSMO-RS (Klamt 1995) and COSMO-SAC (Lin & Sandler 2002) are two of these post-processing methods that are continuing to gain popularity. What distinguishes COSMO-RS and -SAC and makes them particularly attractive in a CAMxD context is that they do not involve binary interaction parameters. Using only molecular volumes and composition-independent charge density distributions called sigma profiles, COSMO-RS and -SAC are able to make accurate mixture thermodynamics estimates. In a CAMxD context, COSMO-based thermodynamics enable a much larger search space in that we are free to consider any molecular species so long as we can estimate its sigma profile and molecular volume.

A COSMO-based mixture design approach incorporates accurate *ab initio* quantum chemical information for any species that is fixed in the mixture (i.e., not in the design space). This means that this COSMO-based approach can be applied to thermodynamics calculations for non-standard species like transition states, radicals, and ionic liquids. Other methods cannot capture the complexities present in these systems as they are often parameterized for neutral, ground state structures. This ability to incorporate quantum chemical information greatly expands the classes of problems that can be addressed by CAMxD.

In this chapter, we propose the use of group contribution methods to estimate sigma profiles and molecular volumes. This enables the use of numerous established mixture and molecular design strategies. Furthermore, we solve the CAMxD problem by (1) decomposition into constituent molecular design and mole fraction problems and (2) projection of the design variables on a lower-dimensional space, namely that of the sigma moments (analogous to statistical moments of the sigma profiles) of each compound in solution. We probe the search space defined by the sigma moments with derivative-free optimization algorithms, which enable an efficient search through our design variables without the computational burden of calculating mixture thermodynamics for a large number of solutions.

In the next section, we provide an introduction to the COSMO solvation model and the COSMO-RS and COSMO-SAC post-processing steps. Furthermore, we provide more detail for sigma profiles and sigma moments. In Sec. 3.3, we discuss our group contribution models for estimating sigma profiles, sigma moments, and molecular volumes. Then, in Sec. 3.4, we integrate COSMO-based thermodynamics into the CAMxD problem. In Sec. 3.5, we investigate two case studies: a separation solvent design problem and a reaction rates optimization problem. Finally, in Sec. 3.6, we provide a summary of the work and draw conclusions about COSMO-based thermodynamics as applied to mixture design problems.

3.2 AN OVERVIEW OF COSMO AND COSMO-BASED THERMODYNAMICS

3.2.1 *Sigma profiles and sigma moments*

The COSMO solvation model (Klamt & Schüürmann 1993) is a variant of quantum chemistry continuum solvation models. While a standard quantum chemistry calculation provides molecular geometries and energies in the gas phase, COSMO—like other continuum solvation models—describes a molecule in the solution phase with an approximate representation of its surroundings as a continuum. In the specific case of COSMO, the continuum is assumed to be an ideal conducting medium, meaning it has a dielectric constant of infinity ($\epsilon = \infty$). In performing a COSMO calculation, a large number of point charges are first placed on the surface of the molecule. The molecule is embedded in an ideal conductor, and the energies of the point charges on the surface are calculated accordingly. The result of a COSMO calculation describes a discretized surface of a molecule i, where each point charge m has a three-dimensional coordinate, a surface area A_{im}, a screening charge density (charge/area) σ_{im}, and a few other properties. To simplify the thermodynamics, this three-dimensional charge distribution is projected onto a two-dimensional probability distribution function called a sigma profile, P_i. For consistency with later parts of this study, we deviate slightly from the definition of Klamt (1995) and define a sigma profile for a molecule i as:

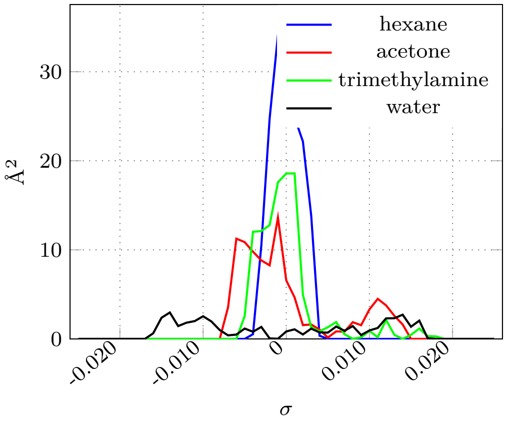
$$P_i(\hat{\sigma}) = \sum_{m, \sigma_{im} \in H(\hat{\sigma})} A_{im}$$

where $\hat{\sigma}$ is a discrete set of σ values used to approximate the sigma profile. $H(\hat{\sigma})$ represents a subset of the point charges, m, and can be defined as

$$H(\hat{\sigma}) = [\hat{\sigma} - \Delta/2, \hat{\sigma} + \Delta/2)$$

with Δ representing a certain margin around the discrete values $\hat{\sigma}$. Naturally, the distance between two of these neighboring $\hat{\sigma}$ values should be equal to Δ. Represented this way, a sigma profile is essentially a histogram that plots surface area according to the $\hat{\sigma}$ value that every point charge is closest to. In other terms, a sigma profile

Figure 3.1: Example sigma profiles



represents the probability of finding a certain screening charge density on the molecular surface. The sigma profiles of a few common structures are given in Fig. 3.1 and come from the Virginia Tech sigma profile database (Mullins & Oldland 2007). The sigma profile of the mixture P_S is simply a linear combination of the sigma profiles of the individual components, weighted by their mole fraction and normalized by their total surface areas:

$$P_S(\sigma) = \frac{\sum\limits_i P_i x_i}{\sum\limits_i A_i x_i}$$

where A_i is the total surface area of molecule i. Therefore, the sigma profile of a solution of only one compound is simply that compound's sigma profile normalized to unity, or the sigma profile divided by the total surface area.

While sigma profiles accurately capture detailed, high-dimensional information about the surface charges on a structure, a more general, low-dimensional view of a molecule's behavior in an ideal conductor is often advantageous. One well-studied way to consolidate the information in a sigma profile into lower-dimensional descriptors is via the use of sigma moments. The n-th sigma moment M_n for $n \in \{0, 1, 2, 3\}$ of a sigma profile P is given by:

$$M_n = \sum_{\hat{\sigma}} P(\hat{\sigma})\hat{\sigma}^n$$

Some of these sigma moments represent physical properties of a structure. The 0-th sigma moment, M_0, is equal to the total surface area of a molecule. M_1 is the total COSMO polarization charge on the surface of the molecule. M_2 is highly correlated with the total COSMO polarization energy, meaning it represents the capacity of a solute molecule to interact with a polarizable continuum. Finally, M_3 does not express any easily-understandable physical property, but according to Klamt (2005) it "represents a kind of skewness in the σ-profile." Furthermore, Klamt defines two hydrogen bonding sigma moments, called the acceptor, M_{acc}, and donor, M_{don}, moments. These are given by:

$$M_{\text{acc/don}} = \sum_{\hat{\sigma}} P(\hat{\sigma}) f^{hb}_{\text{acc/don}}(\hat{\sigma})$$

with

$$f^{hb}_{\text{acc/don}}(\hat{\sigma}) = \begin{cases} 0 & \text{if} \quad \pm\hat{\sigma} < \sigma'_{hb} \\ \pm\hat{\sigma} - \sigma'_{hb} & \text{if} \quad \pm\hat{\sigma} \geq \sigma'_{hb} \end{cases}$$

where σ'_{hb} defines a hydrogen bonding cutoff value.

Sigma moments are useful in a modeling and design context for a few reasons. First, this lower-dimensional space aids in the development of quantitative structure property relationships (QSPRs) to calculate properties not directly available from COSMO-RS and COSMO-SAC calculations (Klamt 2005). The development of QSPRs from sigma moments will not be used in this study but is nonetheless worthwhile to mention as a potential application for COSMO-based CAMD and CAMxD problems. More importantly, sigma moments are particularly useful in our approach to mixture design problems as the low-dimensional space defined by the sigma moments enables the efficient use of derivative-free optimization (DFO) algorithms. From previous studies (Rios & Sahinidis 2013; Austin et al. 2016), we know that DFO algorithms are reliable at solving problems with a small number of degrees of freedom. This approach will be discussed in more detail in a later section.

3.2.2 *COSMO-RS and COSMO-SAC*

The energy of a COSMO calculation represents the total energy of the molecule in an ideal conductor. The liquid phase is considered to be a closely-packed group of molecular structures, while neighboring molecules in an ideal conductor can exhibit surface charges in close contact that are not balanced. Clearly, this is not the situation in a real solvent. In order to calculate the energy of the molecule in an arbitrary solvent, the surface charges must be adjusted to account for the removal of the ideal conductor. This is done with one of a few COSMO post-processing steps, one of which is known as COSMO-RS (Klamt 1995). Given two surface charge densities of segments in close contact, $\hat{\sigma}$ and $\hat{\sigma}'$, the electrostatic energy of interaction per unit area—or "misfit" energy—in a real solvent is expressed as the following:

$$E_{\text{misfit}}(\hat{\sigma}, \hat{\sigma}') = a_{\text{eff}} \frac{\alpha'}{2} (\hat{\sigma} + \hat{\sigma}')^2 \tag{3.1}$$

where a_{eff} represents the area of contact and α' is constant. Both are adjustable parameters. COSMO-RS also predicts the interaction energy of two surfaces due to hydrogen bonding as follows:

$$E_{\text{HB}}(\hat{\sigma}, \hat{\sigma}') = a_{\text{eff}} c_{\text{HB}} \min(0, \hat{\sigma}_{\textbf{donor}} + \sigma_{\text{HB}}) \max(0, \hat{\sigma}_{\textbf{acceptor}} - \sigma_{\text{HB}}) \tag{3.2}$$

where σ_{HB} is the hydrogen bonding cutoff radius beyond which segments with surface charges $\hat{\sigma} > \sigma_{HB}$ or $-\hat{\sigma} > \sigma_{HB}$ are assumed to participate in hydrogen bonding. c_{HB} is an adjustable parameter accounting for the energy of a hydrogen bond. We can now calculate the sigma potential, or the chemical potential of a surface segment of screening charge density $\hat{\sigma}$ in some solvent S. This is given by the following:

$$\mu_S(\hat{\sigma}) = -RT \ln \left(\sum_{\hat{\sigma}'} P_S(\hat{\sigma}') \exp \left(\frac{\mu_S(\hat{\sigma}') - E_{\text{misfit}}(\hat{\sigma}, \hat{\sigma}') - E_{\text{HB}}(\hat{\sigma}, \hat{\sigma}')}{RT} \right) \right) \tag{3.3}$$

With all this information, we can finally calculate the chemical potential of a molecule i in some solvent environment S. This expression provides the chemical potential as defined from the reference state of the ideally screened molecule i:

$$\mu_i = \sum_{\hat{\sigma}} P_i(\hat{\sigma}) \mu_S(\hat{\sigma}) + \mu_i^C$$

where μ_i^C is the combinatorial contribution to the chemical potential. This is essentially an entropic term to account for size and shape differences among molecules in the solution. There are a number of empirical estimates for this term, and those used in this document will be discussed in the appropriate sections. For more information on the COSMO-RS method, see Klamt (1995); Klamt et al. (1998); Eckert & Klamt (2002). For a complete review of the COSMO-RS method and discussion of implementations and applications, readers are directed to Klamt (2005).

Another post-processing method for COSMO, COSMO-SAC (Lin & Sandler 2002), applies a very similar approach. COSMO-SAC defines a term called the segment exchange energy, ΔW, which is analogous to the sum of COSMO-RS's misfit energy and hydrogen bonding energy:

$$\Delta W(\hat{\sigma}, \hat{\sigma}') = \frac{\alpha'}{2}(\hat{\sigma} + \hat{\sigma}')^2 + c_{\text{hb}} \min(0, \hat{\sigma}_{\mathbf{donor}} + \sigma_{\text{hb}}) \max(0, \hat{\sigma}_{\mathbf{acceptor}} - \sigma_{\text{hb}}) \tag{3.4}$$

COSMO-SAC then calculates the activity coefficient of some surface segment $\hat{\sigma}$ in the some solution S:

$$\ln \Gamma_S(\hat{\sigma}) = -\ln\left(\sum_{\hat{\sigma}'} P_S(\sigma')\Gamma_S(\hat{\sigma}') \exp\left(\frac{-\Delta W(\hat{\sigma}, \hat{\sigma}')}{RT}\right)\right) \tag{3.5}$$

Note again that these expressions have been adapted slightly from those given in the original papers.

Finally, COSMO-SAC defines a restoring free energy term, $\Delta G*^{\text{res}}_{i/S}$, to calculate the energy change in moving one of the mixture components i from the ideal conductor to some solvent environment S. This again bears a strong resemblance to the COSMO-RS calculation of the property:

$$\frac{\Delta G*^{\text{res}}_{i/S}}{RT} = X_i^{num}\left(\sum_{\hat{\sigma}} P_i(\hat{\sigma}) \ln \Gamma_S(\hat{\sigma})\right) \tag{3.6}$$

where X_i^{num} represents the total number of surface segments for molecule i. Note that there is also a combinatorial term (not shown) involved in this model. For more discussion of the COSMO-SAC method as well as a series of modifications to the model, readers are directed to Wang et al. (2007); Hsieh et al. (2010); Xiong et al. (2014).

3.3 GROUP CONTRIBUTION METHOD FOR CALCULATING SIGMA PROFILES

Disregarding the combinatorial term, the sigma profile of each mixture component is all that is required to make mixture thermodynamics calculations with COSMO-RS and COSMO-SAC. Though highly-accurate, these calculations are time-consuming and not suitable for the solution of CAMxD problems. For this reason, we propose the use of group contribution methods to estimate sigma profiles. To derive these group contribution methods, we use Virginia Tech's sigma profile database (Mullins & Oldland 2007), the only free, publicly available database of its kind. The profiles contained in this database are calculated with parameters based on the COSMO-SAC model.

We follow the example of Mu et al. (2007) and discretize the σ values along the profile. Now, a sigma profile contains 51 discretized intervals $s = [1, \ldots, 51]$, each Δ=0.001 $e/\text{Å}^2$ wide. The corresponding $\hat{\sigma}$ values will go from -0.025 $e/\text{Å}^2$ for s=1 to +0.025 $e/\text{Å}^2$ for s=51. Now, a sigma profile is represented as the total area of segments whose surface charge density is within $\Delta/2$ of $-0.026 + 0.001s\ e/\text{Å}^2$.

Given the above discretization, we now estimate a sigma profile P_i with the following group contribution formulation

$$P_i = \sum_g c_g n_g \tag{3.7}$$

where n_g defines the number of occurrences of group g and c_g is a vector of dimension $|s|$ that quantifies how much each group contributes to each interval s of the sigma profile. The coefficients were obtained by fitting our group contribution method to all the non-ionic compounds contained in the VT sigma profile database. We list the average deviation of the group contribution method by $\hat{\sigma}$ interval in Fig. 3.2. We list the correlation coefficient, r^2, for each interval in Fig. 3.3. As shown, the group contribution method is fairly good at predicting the sigma profiles, exhibiting an $r^2 > 0.7$ for most of the intervals.

Furthermore, we can estimate the sigma moments with another group contribution method. These sigma moments can be estimated very accurately with group contribution (GC) methods. Using a different set of groups, the regression produces an $r^2 > 0.9$ for five of the six moments. These r^2 values are shown in Fig. 3.4.

Finally, the volumes of the COSMO cavity for each molecule were fit to a final group contribution method. Volumes are needed to calculate the combinatorial contribution to chemical potential. As volumes are an additive property, group contribution methods were able to predict these quite well, exhibiting an $r^2 > 0.999$. In Fig. 3.5, we plot the GC-estimated values for COSMO volumes against those taken from the VT database (Mullins & Oldland 2007). As seen in this figure, the proposed GC technique estimate this data set nearly perfectly. Finally, we again note that the groups to estimate sigma profiles, sigma moments, and molecular volumes are different. These sets will be denoted as G_P, G_M, and G_V, respectively, and the corresponding group contribution methods will be denoted as f_P, f_M, and f_V. The groups and their coefficients for volumes and sigma moments are provided at Austin & Sahinidis (2016).

As shown, the information captured in the σ moments and molecular volumes can be estimated accurately with group contribution methods. The sigma profile estimation, on the other hand, is not as reliable. We emphasize that the need to estimate sigma profiles with group contribution methods is the main shortcoming of such a COSMO-based approach to mixture design. For many types of molecules, this is very accurate, but there are some classes of structures for which group contribution methods have proven inadequate to estimate sigma profiles (Mu et al. 2007). Overall, COSMO-based mixture design has the advantages of no binary interaction parameters and easy integration with quantum chemistry calculations. These advantages come at the occasional price of

Figure 3.2: Average deviation by interval

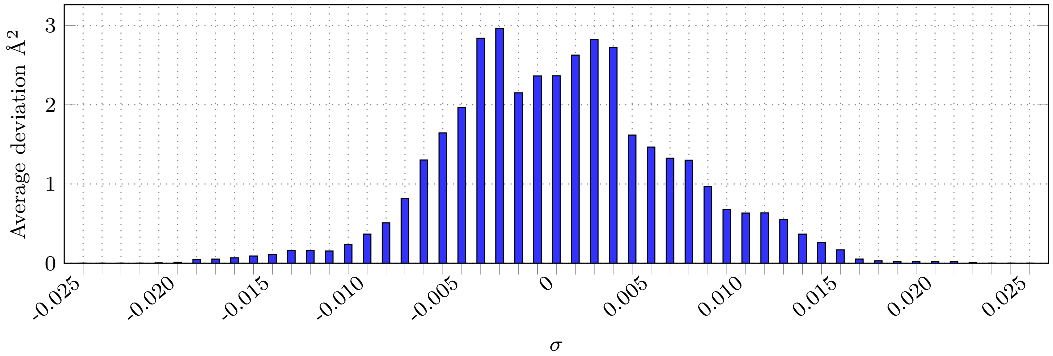
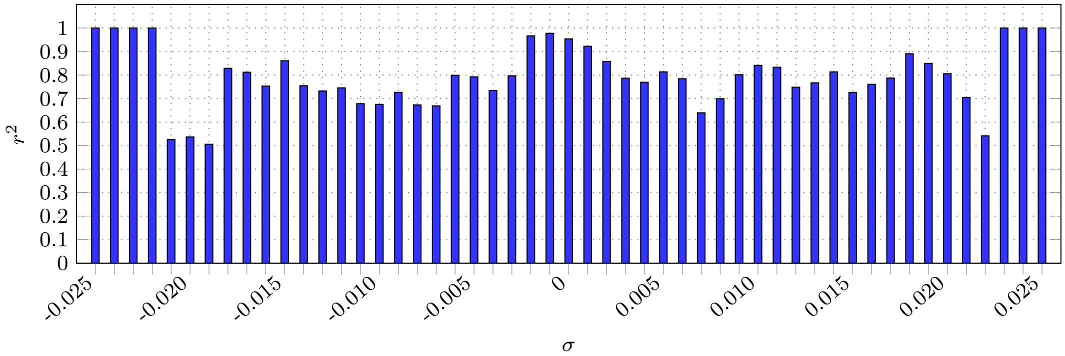
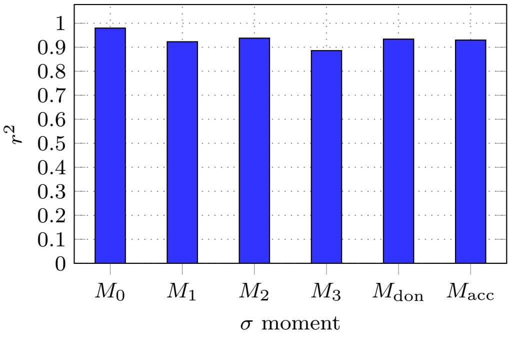


Figure 3.3: Correlation coefficient by interval



accuracy. However, the development of non-standard group contribution methods for sigma profile estimation would alleviate this issue to some degree. Such methods may involve non-linear terms, interaction parameters, and inclusion of non-group descriptors, for example. Finally, it should be noted that, like any group contribution method, the proposed method for sigma profile estimation would benefit from a larger and more diverse training set.

Figure 3.4: Correlation coefficient by moment

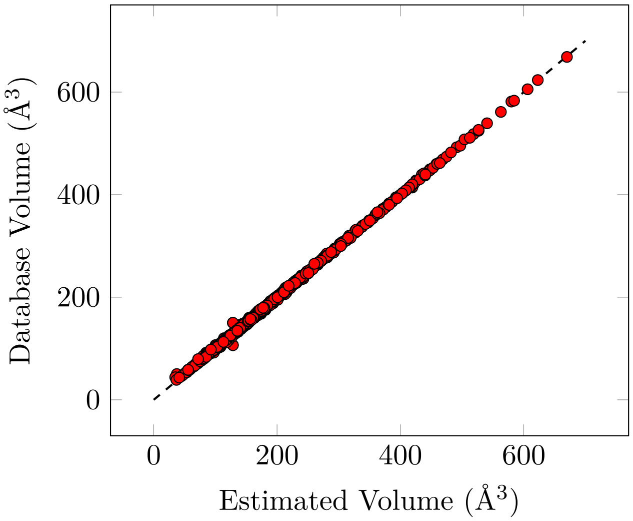


3.4 THE COSMO MIXTURE DESIGN PROBLEM

We will first remind readers of the general form of the mixture design problem and then present a COSMO-based approach. The goal of the problem is to design some K-component mixture such that the mixture properties q optimize some function $C(q)$. Some of the components may be predetermined but at least one is assumed to be unknown. This problem requires determination of the molecular structures of every unknown component as well as optimization over mole fractions for every species in the mixture. Finally, there may be constraints on component and mixture properties.

The indices i, j, and k will, respectively, denote components in the mixture $(i = 1, \ldots, K)$, pure component properties $(j = 1, \ldots, C)$, and mixture properties of interest $(k = 1, \ldots, N)$. For component i, x_i will denote its mole fraction in the mixture. Let p_{ij} denote the value of property j for pure component i. These p's can be estimated by a family of functions $f(n)$, where n is a vector of occurrences of various molecular subgroups. In the majority of CAMxD problems, these f's are group contribution methods. Furthermore, this vector n must capture molecular subgroup information for

Figure 3.5: GC-estimated volume vs. database volume from VT's sigma profile database (Mullins & Oldland 2007)



every unknown component of the mixture, meaning that n will be indexed over the set of unknown components. There are also properties of the mixture, q_k, $k = 1, \ldots, N$, that are functions of the pure component properties and mole fractions, i.e., $q_k = g_k(x, p)$, $k = 1, \ldots, N$. The mixture design problem is to determine the components and their mole fractions so that a certain performance criterion $C(q)$ is optimized. We can therefore formulate this problem as follows:

$$
\begin{aligned}
\text{(CAMxD)} \qquad \min_{n,x} \quad & C(q) && (3.8)\\
\text{s.t.} \quad & q = g(x,p) && (3.9)\\
& p = f(n) && (3.10)\\
& h(x,p,q) \leq 0 && (3.11)\\
& l(x,p,q) = 0 && (3.12)\\
& \sum_i x_i = 1 && (3.13)\\
& p^L \leq p \leq p^U && (3.14)\\
& q^L \leq q \leq q^U && (3.15)\\
& n \in \mathcal{S} && (3.16)
\end{aligned}
$$

First, equation (3.8) optimizes some function C of mixture properties. Equation (3.10) represents group contribution methods used to estimate pure component properties from each unknown component's n vector. Constraints (3.11) and (3.12) are equality and inequality constraints imposed on mixture and component properties. Constraint (3.13) simply requires all mole fractions to sum to 1. Constraints (3.14) and (3.15) represent the bounds placed on individual component and mixture properties. Constraint (3.16) represents a number of group valence constraints (Odele & Macchietto 1993; Sahinidis et al. 2003) used to ensure the chosen n vector contains groups that will assemble into a chemically feasible structure. Finally, equation (3.9) is a number of functions—here collectively identified as g—used to transform individual component properties and mole fractions into mixture properties. These functions are key in enabling the use of single molecule design methodologies in mixture design.

The choice of these g functions has important consequences for how a mixture design problem is solved and what design space it can effectively consider. Using the COSMO-RS and -SAC methods for the g functions, we obtain a modified CAMxD formulation. First, we replace eq. (3.9) with the following:

$$q = \text{COSMO-RS/-SAC}(x, P, V) \tag{3.17}$$

Now, we calculate mixture thermodynamics using the COSMO-RS and -SAC methods. COSMO-RS and -SAC are functions of the mole fractions x, sigma profiles P, and cavity volumes V of every species in solution. Though technically subsumed in eq. (3.10) in the previous formulation, we note the addition of our group contribution estimation methods for P and V to the formulation for clarity:

$$P = f_P(n) \tag{3.18}$$

$$V = f_V(n) \tag{3.19}$$

where P and V are estimated by the group contribution methods f_P and f_V discussed above.

In the above formulation, we are optimizing over the design space of both discrete variables n and continuous variables x. Given that there are a number of unknown components we are designing for, the discrete space, which is defined in terms of the n variables, can be quite large. Additionally, optimizing over the mole fractions, x, is typically challenging as the mixture thermodynamics models introduce many non-linearities and non-convexities. These features make CAMxD a challenging MINLP that cannot easily be solved using both an appreciable number of possible groups and appropriately complex thermodynamic models. To address this difficulty, in Austin et al. (2016) (discussed in Chapter 2) we introduced a strategy whereby we projected the CAMxD problem onto the space of the properties of the individual components in the mixture, p. This projection facilitates a natural decomposition of the CAMxD problem, enabling efficient single-molecule design methodologies to be used with a more straightforward mole fractions optimization problem. In the specific case of our COSMO-based mixture design problems, we will adopt a similar strategy and project our problem onto the space of the sigma moments of each unknown mixture component. The objective of the problem is now an implicit function of x and n, meaning algebraic optimization techniques can no longer be directly applied. Furthermore, the design space is significantly lower dimensional. Due to these features, derivative-free optimization (DFO) algorithms can be effectively applied to optimize over the space of the component properties as many DFO algorithms have been shown to be highly efficient at solving problems with few degrees of freedom (Rios & Sahinidis 2013). Our optimization strategy follows the approach outlined in Chapter 2 but optimizes over the space of sigma moments, M. This DFO-based strategy is detailed below:

1. Given a candidate property vector M_T, find n with corresponding $f_M(n)$ that is as close to M_T as possible; this is a single molecule design problem done for each unknown component in the mixture.

2. Use group contribution methods, $f_P(n)$ and $f_V(n)$, to generate sigma profiles and molecular volumes for each compound in solution. Fix the values of n in the original CAMxD problem and solve the continuous part of the problem.

3. Interpret the objective value and choose new M_T's if necessary; we address this problem via DFO.

We begin by specifying a sigma moments target M_T for each unknown species in solution. In a hypothetical mixture design problem where two components are unknown, a possible sigma moments vector for the first component may be $M_T^1 = [M_1^1, M_2^1, M_{\text{don}}^1]$, representing the first, second, and hydrogen bond donor moments, respectively. A similar sigma moments vector M_T^2 would also exist for the second component. Because these sigma moments are molecular properties, we can exploit efficient optimization techniques for single-molecule design to quickly determine a molecular structure with sigma moments closest to the target values. For each of these components, we design a molecular structure using the AMODEO methodology (Samudra & Sahinidis 2013b). The AMODEO approach optimizes over the space of groups in our sigma moments group contribution method, minimizing the distance between M_T and the group contribution estimates. The optimum of the problem, n_i^*, represents the number of occurrences of each group in component i. Furthermore, we only consider solutions that fall within a certain range $[M_k^L, M_k^U]$ as given for each sigma moment, k, in the design space. This range is determined by:

$$M_k^L = M_k - \tau({M_k^{\text{all}}}^U - {M_k^{\text{all}}}^L) \tag{3.20}$$

$$M_k^U = M_k + \tau({M_k^{\text{all}}}^U - {M_k^{\text{all}}}^L) \tag{3.21}$$

$$\tau \in (0, 1] \tag{3.22}$$

where τ represents a multiplier to quantify a fraction of the entire feasible range of a specific moment. In short, M_k^L and M_k^U define lower and upper bounds around a particular property target point M. ${M_k^{\text{all}}}^L$ and ${M_k^{\text{all}}}^U$ define lower and upper bounds over the entire sigma moments design space. The molecular design formulation for determining molecular structures from moments is shown below. We note that this formulation is almost identical to the formulation discussed in Eqs. (2.18)—(2.20).

$$\min_{n} \quad \sum_{k} \left[\frac{d_k^+ + d_k^-}{M_k^U - M_k^L} \right] \tag{3.23}$$

$$\text{s.t.} \quad d_k^+ - d_k^- = \sum_{g \in G_M} c_g^k n_g - M_k \quad \forall k \tag{3.24}$$

$$M_k^L \leq \sum_{g \in G_M} c_g^k n_g \leq M_k^U \quad \forall k \tag{3.25}$$

$$n \in \mathcal{S}$$

In the above, d_k^+ and d_k^- are positive continuous variables that quantify positive and negative differences between group contribution estimates and target values for each sigma moment k. In Eq. (3.23), we minimize the sum of these differences normalized by the target sigma moments value. Eq. (3.24) calculates these differences. Eq. (3.25) ensures that the groups selected produce an estimate that falls within our specified property bounds for each sigma moment.

After this problem is optimized, the optimal groups are connected to produce actual molecular structures. This is achieved either by an enumeration procedure or by a graph theory optimization approach in AMODEO. More information on the graph theory approach can be found in Samudra & Sahinidis (2013b). Once we have molecular structures for each unknown component of the mixture, we apply group contribution methods to quickly generate sigma profiles and COSMO volumes for each of these components. Now the original CAMxD problem can be solved, fixing the molecular compositions, n_i, of each unknown component. The CAMxD problem thus reduces to an NLP, which is solved over the mole fractions in solution. Note that this NLP does not produce a globally optimal solution to the CAMxD problem. It only provides an optimal solution in the mole fractions space for a set of compounds corresponding to a certain σ moments vector.

We utilize derivative-free optimization (DFO) algorithms as an optimization strategy in the lower-dimensional projected space of σ moments. DFO algorithms in this case supply a particular target value M_T for each component in solution. This value is used to determine molecular structures and optimize some mixture objective function in the process described above. The objective value is reported back to the DFO solver and it either supplies a new target point M_T or determines that convergence has been achieved. If any of the component design subproblems or mole fractions problem is infeasible, a large value is reported back to the DFO solver. This algorithm is summarized in Fig. 3.6. Again, we emphasize the similarity between this algorithm and the one detailed in Fig. 2.2.

Like many examples in the literature, this approach solves the mixture design problem by decomposing it into single molecule design and mole fractions optimization subproblems. However, many of these approaches require designing every possible molecular structure for every component. While this can be done fairly efficiently, the mole fractions problem presents far more difficulty. With these approaches, every combination of possible molecular structures must be evaluated for optimal mole fractions. Because the thermodynamics of these mixture design problems often lead to non-trivial mole fractions problems, only a limited number of these feasible combinations can be practically considered. In the above algorithm, however, every iteration either produces an objective function value corresponding to some region of component property space or determines that no molecular structures exist in that area that are feasible for the mixture design problem. This feature enables a much more efficient and thorough search through the feasible design space.

We next apply this algorithm to two case studies. For these case studies, we use the TOMLAB/CGO solver (Holmström et al. 2007) as the DFO algorithm, and we use a 2.84 GHz processor.

3.5 CASE STUDIES

3.5.1 *Liquid-liquid extraction solvent*

In a liquid-liquid extraction, two liquid components, A and B, are separated by the addition of a third component called an extractant, identified here as C. Assuming the intent is to produce pure B, a good extractant should solvate A much more favorably than B solvates A. Additionally, liquid-liquid extractions should produce two distinct liquid phases, so it is desirable that C is only partially miscible or completely immiscible in B. As an industrial process, we assume that a feed of A+B and a feed of C are mixed together in a single-stage extraction unit. One of the liquid phases that is produced is called the extract and should contain mostly A and C. The other phase is called the raffinate and should contain mostly B.

We investigate the problem originally posed by Seader & Henley (1998) and studied in a mixture design context using UNIFAC by Karunanithi et al. (2005). The problem involves recovering acetic acid from a mixture of acetic acid and water. We assume one feed to the extraction unit contains 8 wt % acetic acid in water and flows into the unit at 13500 kg/h. Another feed containing the extractant feed has a flow rate of 16300 kg/h. The extract and raffinate phases are removed with flow rates F_E and F_R, respectively. The mixture design problem determines the optimal extractant and optimal flow rates for the extract and raffinate phases. The optimal extractant in this case should lead to the least amount of acetic acid loss. Furthermore, we have to ensure mass balances

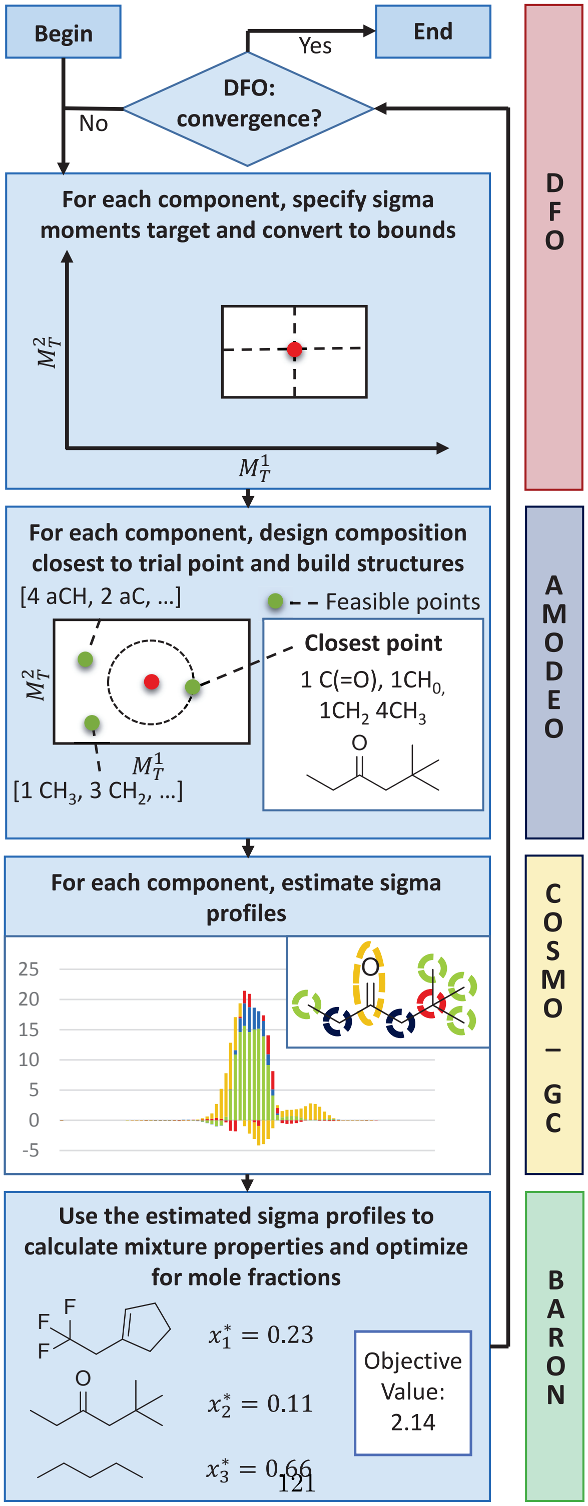

Begin
End
Yes
DFO:
convergence?
No
For each component, specify sigma moments target and convert to bounds
M_T^2
M_T^1
For each component, design composition closest to trial point and build structures
[4 aCH, 2 aC, ...]
Feasible points
Closest point
1 C(=O), 1CH$_0$, 1CH$_2$ 4CH$_3$
O
M_T^2
M_T^1
[1 CH$_3$, 3 CH$_2$, ...]
For each component, estimate sigma profiles
25
20
15
10
5
0
-5
O
Use the estimated sigma profiles to calculate mixture properties and optimize for mole fractions
F
F
F
$x_1^* = 0.23$
O
$x_2^* = 0.11$
$x_3^* = 0.66$
Objective Value: 2.14
DFO
AMODEO
COSMO – GC
BARON

Figure 3.6: A pictorial representation of the COSMO-based mixture design algorithm

and liquid-liquid equilibrium constraints as well as constrain solvent properties to be favorable for this process. As approached with our mixture design algorithm, a candidate structure is generated at each iteration and then a subproblem must be solved to quantify the process performance of the candidate structure. This is done by optimizing over the mole fractions and process conditions. The formulation of the subproblem is presented below.

$$\begin{aligned}
\min_{x,F_R,F_E} \quad & X_A^R F_R && (3.26)\\
\text{s.t.} \quad & P = f(n) && (3.27)\\
& V = f(n) && (3.28)\\
& \gamma = COSMO - SAC(P, V, x) && (3.29)\\
& \gamma_i^E x_i^E - \gamma_i^R x_i^R = 0 \qquad \forall i && (3.30)\\
& X_i^E F_E + X_i^R F_R = X_i^F F_F \qquad \forall i && (3.31)\\
& \sum_i x_i^R = 1 && (3.32)\\
& \sum_i x_i^E = 1 && (3.33)\\
& m = \frac{\gamma_{A,B}^{\infty}}{\gamma_{A,B}^{\infty}} \frac{MW_A}{MW_B} \geq 0.49 && (3.34)\\
& SL = \frac{1}{\gamma_{C,B}^{\infty}} \leq 0.0038 && (3.35)\\
& \beta = \frac{\gamma_{B,C}^{\infty}}{\gamma_{A,C}^{\infty}} \geq 11 && (3.36)\\
& SP = \frac{1}{\gamma_{A,C}^{\infty}} \geq 0.778 && (3.37)
\end{aligned}$$

In the above formulation, we optimize over the mole fractions of each component i and the flow rates of the two phases F_R and F_E. Because there are two liquid phases in this extraction process, we further index the mole fractions over two phases. x_i^R indicates mole fractions in the raffinate phase and x_i^E indicates mole fractions in the extract phase. The capital X's represent mass fractions. MW_i is the molar weight of molecule i. Values m, SL, β, and SP represent the distribution coefficient, solvent loss, selectivity, and solvent power and are constrained in keeping with Karunanithi et al. (2005). $\gamma_{i,i'}^{\infty}$ is the infinite dilution activity coefficient of compound i in a solution of compound i'. The objective, Eq. (3.26), minimizes the mass of acetic acid lost to the raffinate phase, again in keeping with Karunanithi et al. (2005). Eq. (3.29) uses the information from the sigma profiles, mole fractions, and molecular volumes to calculate activity coefficients γ_i^R and γ_i^E for each component in each of the phases. These activity coefficients are calculated with the COSMO-SAC parameters and combinatorial term that are given in Mullins et al. (2006). The liquid-liquid phase equilibrium condition is captured in Eq. (3.30) and Eq. (3.31). Eqs. (3.32) and (3.33) ensure that all mole fractions add to one in both phases. Eqs. (3.34), (3.35), (3.36), and (3.37) place constraints on the properties of the system to ensure favorable characteristics for extraction. The resulting problem is a nonlinear and nonconvex optimization problem that is solved for a fixed input of species C. This problem is solved with BARON (Tawarmalani & Sahinidis 2004) in each iteration of the algorithm.

It is also important to note that the sigma profiles of acetic acid and water, P_A and P_B, come directly from the Virginia Tech sigma profile database. This means that the profiles of these two species reflect full quantum chemical accuracy. The same is true of the molecular volumes, which are also taken directly from the database. The sigma profile and molecular volume of the extractant, P_C and V_C, are the only properties estimated with our group contribution methods. We advocate the approach of including as many quantum-chemistry-calculated sigma profiles as possible in solving COSMO-based mixture design problems. In general, doing so provides the mixture de-

Table 3.1: Summary of important values for the liquid-liquid extraction solvent case study

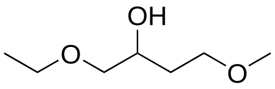
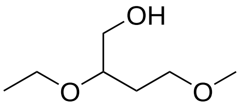
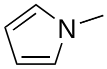
Parameter	Value/Range	Additional Information
Time limit	30 min	Maximum allowable time for the algorithm
Iteration limit	2000	Maximum number of iterations the algorithm can perform
DFO inputs	M_0, M_1, M_2, M_3, M_{acc}, M_{don},	Lower-dimensional design space for the solvent
τ	20%	Property bounds relaxation around DFO trial point
$C_{\max}$	10	Maximum number of compositions determined during each iteration
Carbons	12	Maximum number of carbons in the designed component
Triple bonds	1	Maximum number of triple bonds in the designed component
Double bonds	2	Maximum number of double bonds in the designed component
Non-carbons	3	Maximum number of non-carbons in the designed component

sign problem with the maximum amount of accurate information, relying on the group contribution estimates of sigma profiles only for molecules in the design space of the problem.

In their analysis of this problem, Karunanithi et al. (2005) employ the UNIFAC method and report 2-hexanone as the optimal solvent for this process. They also indicate that this is the industrial standard used for this separation. Using the problem specification from above and taking the sigma profile and molecular volume of 2-hexanone directly from the Virginia Tech Database (Mullins & Oldland 2007), the optimal objective value is found to be 202.27 kg/h. The search space for this problem is defined by the six σ-moments discussed above. We set bounds on the moments based on slightly widening the range of values observed in the VT sigma profile database. Furthermore, we constrain the chemical structure of the designed compound to have fewer than twelve carbons, no more than three non-carbons, and up to one triple bond and two aliphatic double bonds. We also account for aromatic structures. The relevant problem data is summarized in Table 3.1. We note that these constraints on the design space are far more stringent than they need to be. We reduce our search to this region for the purposes of producing solutions which can be practically considered as synthetically viable alternatives to this process.

The solution of our COSMO-based mixture design problem resulted in several promising molecules, a small subset of which is shown in Table 3.2. Using our group contribution methods to estimate sigma profiles and molecular volumes, 1-ethoxy-4-methoxybutan-2-ol, the first entry in the table and best compound found, was determined to have an optimal objective value of 137.70 kg/h. This is a better solution than the industrial standard and solution found by Karunanithi et al. (2005) by roughly 40%. The second molecule we report, 2-ethoxy-4-methoxybutan-1-ol, is another solution to the problem similar in structure to the optimal molecule. Interestingly, both of these structures represent part of this problem space for which UNIFAC would have a missing interaction parameter, specifically that between the UNIFAC main groups -COOH and -OCCOH (Dortmund Data Bank Software and Separation Technology GmbH 2014). The fact that the Karunanithi et al. study did not discover our optimal structure is likely the result of this missing interaction parameter. For the sake of completeness, we acknowledge that this structure has an alternative UNIFAC representation using simply the ether and alcohol main groups. Such a representation would have allowed for our optimal molecule to be in the feasible region defined by Karunanithi et al. In this

Table 3.2: Representative structures for the liquid-liquid extraction case study

Solvents	**Properties**	
	Objective value:	137.70
	x_A^R:	0.0101
	x_A^E:	0.0264
	F_R:	4619.96 kg/h
	F_E:	25180.04 kg/h
	Objective value:	289.78
	x_A^R:	0.0107
	x_A^E:	0.0354
	F_R:	8732.22 kg/h
	F_E:	21067.78 kg/h
	Objective value:	774.11
	x_A^R:	0.0183
	x_A^E:	0.0246
	F_R:	13888.94 kg/h
	F_E:	15911.06 kg/h

case, the sub-optimality of this structure in their approach would likely result from the alternative UNIFAC representation not accounting for the full behavior of the molecule. The different results may also be due to differences between UNIFAC and COSMO-SAC, but the effect is not likely to be so significant.

Coincidentally, a molecule appears in the VT sigma profile database which is very similar to these first two designed structures. This is 2,2-ethoxyethoxyethanol, and it contains two ethers and an alcohol, like both of the first listed structures. Using the sigma profile and molecular volume from the database, we solve the subproblem above and obtain an objective value of 120.43 kg/h, again better than the objective value of the industrial standard. It is noteworthy that this molecule also contains the -OCCOH UNIFAC group and thus could not be considered in a UNIFAC-based design problem.

We also include a third structure in the list to convey an idea of the molecular diversity explored in the algorithm. The third molecule, *N*-methylpyrrole, is not a good solvent for this process, but it again represents part of the design space that could not be explored with UNIFAC. Despite its poor performance in this design problem, its inclusion in the design space could be more significant given a different objective and altered process conditions.

Finally, we note that the algorithm also generated a large number of ketones similar to and including 2-hexanone, all with similar objective values to the industrial standard. This observation illustrates that this COSMO-based approach can reproduce results obtained with UNIFAC. However, this case study primarily underscores the potential of a COSMO-based mixture design approach to determine solutions that are simply not part of the search space in other methods. In this example, some of these solutions are better than those attainable by using a more constrained search space.

3.5.2 *Reaction rates optimization solvent*

The reaction medium plays a critical role in determining the success of a particular reaction, the rate at which it proceeds, and whether any undesirable side-products are formed. Furthermore, there is limited customization in solvent choice as many reactions are performed in one of a handful of common laboratory solvents or in a simple blend of these solvents. For this reason, designing a solvent to optimize some function of reaction rates has considerable application potential in liquid-phase chemistry.

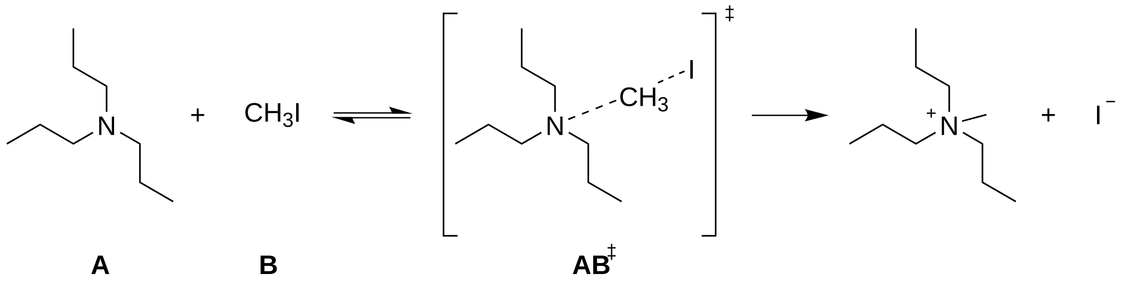
Because a rigorous modeling of reaction rates involves some knowledge of transition states, this class of solvent design problems cannot be approached using methods like UNIFAC. UNIFAC, though powerful in its own domain, is parameterized for neutral, ground state molecules and poorly predicts the electronic complexities of transition states. COSMO-based methods, on the other hand, are particularly suitable for this application as the transition states of relevant reaction pathways can be modeled accurately using quantum chemistry techniques. In addition, COSMO-RS and -SAC directly calculate chemical potential, which trivially yields the free energy of solvation of a molecular species.

In this case study, we design a solvent to maximize the reaction rate of a particular Menschutkin reaction. The Menschutkin reaction is the reaction of a tertiary amine with an alkylhalide to form a quaternary ammonium salt. In our example, the Menschutkin reaction we investigate is the reaction of tripropylamine with methyl iodide. This reaction is shown in Fig. 3.7. The Menschutkin reaction provides a particularly valuable case study for solvent design as its reaction rate is known to be sensitive to the choice of solvent (Lassau & Jungers 1968). In addition, the specific Menschutkin reaction we investigate proceeds via a simple S_N2 pathway, so the calculation of the reaction rates is straightforward. Referring to transition state theory (TST), we calculate the reaction rate constants as a function of the energy differences between the reactants (A and B) and the transition state ($AB^\ddagger$). Specifically, we use the following equation.

$$k = \kappa \frac{k_B T}{h} \exp\left(-\frac{\Delta G^\ddagger_{\text{gas}} + \Delta G^{AB^\ddagger}_{\text{solv}} - \Delta G^{A}_{\text{solv}} - \Delta G^{B}_{\text{solv}}}{RT} \right) \tag{3.38}$$

In the above, k is the reaction rate constant, k_B is Boltzmann's constant, h is Planck's constant, and T is the temperature. $\Delta G^\ddagger_{\text{gas}}$ represents the free energy of activation in the gas phase. $\Delta G^{AB^\ddagger}_{\text{solv}}$, $\Delta G^{A}_{\text{solv}}$, and $\Delta G^{B}_{\text{solv}}$ represent the free energy of solvation

Figure 3.7: The Menschutkin reaction between tripropylamine and methyl iodide



for the transition state, tripropylamine, and methyl iodide, respectively. Finally, κ is a proportionality constant to account for the fact that not every vibration of the transition state leads to the products. In this case study, we set κ to 1.

To estimate reaction rates, we first optimize the gas phase geometries of A, B, and TS on Gaussian09 (Frisch & et al. 2009) using the B3LYP functional (Becke 1993; Stephens et al. 1994) and a 6-311g(d,p) basis set. For iodine, we use the parameters from Glukhovstev et al. (1995). Using these energies, we can calculate $\Delta G^{\ddagger}_{\text{gas}}$. Next, we perform a single point calculation on the optimized geometries using COSMO (Barone & Cossi 1998) in Gaussian. This calculation provides energies in the conductor phase as well as sigma profiles for each of the species. Finally, we estimate the sigma profiles and molecular volumes of the designed solvent molecules using the group contribution method of Mu et al. (2007). This method is chosen rather than ours because we have observed sensitivity in the COSMO-based thermodynamics calculations to the choice of quantum mechanics software, calculation method for sigma profiles, basis sets, and functionals. As we were only able to calculate sigma profiles with Gaussian software, we opted for a method that estimated sigma profiles calculated on Gaussian software.

We are consistent with Mu et al. (2007) and calculate the sigma profiles of A, B, and $AB^{\ddagger}$ on Gaussian with a B3LYP functional using a 6-311g(d,p) basis set.

Using the COSMO-RS model with the parameters and combinatorial term taken from Klamt & Eckert (2000; 2003), we estimate $\Delta G_{\text{solv}}^{AB^{\ddagger}}$, $\Delta G_{\text{solv}}^{A}$, and $\Delta G_{\text{solv}}^{B}$. To test the accuracy of this approach, we compare our estimates to the reaction rate data for this specific reaction in 59 different solvents as given in Folic et al. (2008); Lassau & Jungers (1968). We compare deviation of estimated $\log(k)$ (base 10) from experimental $\log(k)$ in Fig. 3.8. As shown, only 1 of the 59 estimated values differs by more than one log deviation (one order of magnitude) from experimental values. These one-log-deviation lines are shown in red. This is a very promising result as full quantum calculations for every species in solution typically only estimate reaction rates to within about one order of magnitude of accuracy. Furthermore, we note that several of the structures in the dataset cannot be fully described using the group contribution method of Mu et al. (2007). It is likely that a large part of the error would be resolved if there were more groups available in this method.

Another measure of accuracy for our estimates is given by average absolute percent error (AAPE), defined as:

$$AAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|\log(k)_i^{\text{est}} - \log(k)_i^{\text{exp}}|}{|\log(k)_i^{\text{exp}}|} \times 100\% \tag{3.39}$$

where i represents a single solvent molecule in our data set of $N = 59$ molecules, $\log(k)_i^{\text{est}}$ is our estimated reaction rate, and $\log(k)_i^{\text{exp}}$ is the experimental reaction rate taken from Lassau & Jungers (1968). In the regression-based approach of Folic et al. (2008), the authors report an $AAPE$ of 18.77% for this data set. For the sake of comparison, our approach, using a group contribution method to estimate the sigma profiles of the solvents, produces an $AAPE$ of 12.50% on the same 59 data points. Furthermore, in the study of Folic et al. (2008), the authors used some data points to build their regression model. Our model, on the other hand, only contained one adjustable parameter to account for differences in standard states. This comparison suggests that COSMO-based methodologies are capable of increased accuracy over traditional CAMD/CAMxD techniques.

To solve this reaction rates optimization problem, we calculate the free energies in solution of each species with the process described above. We optimize in the space of the sigma moments of the unknown solvent molecule, estimating the sigma profile at each feasible iteration using a group contribution method. The important parameters used in this problem are summarized in Table 3.3. Furthermore, we remove sulfoxides, amines, and carboxylic acids from consideration to ensure the solvent is inert with respect to the reactants. We additionally remove any groups that could lead to protic solvents as these would tend to stabilize the nucleophile. The result of this optimization yields several different types of molecules, a representative list of the highest-performing ones given in Fig. 3.4. The first entry in the list, 3-nitro-2-(nitromethyl)propanenitrile, has the highest predicted reaction rate of all molecules considered. It is a relatively small molecule containing two nitro groups and a cyanide group. Its predicted reaction rate is 0.45 $\log(k)$, with k in units of mol/s.

Another significant advantage of COSMO-based mixture design is the ability to easily integrate full quantum chemistry calculations for designed structures and thus predict properties using structures which are all optimized at the quantum chemistry level. Specifically, we use Gaussian09 (Frisch & et al. 2009) to optimize the geometry of the first solvent molecule in the table and then perform a single point COSMO calculation, again using the B3LYP functional at the 6-311g(d,p) level of theory. Processing the result of the COSMO calculation yields a sigma profile which is accurate to the quantum level. Using this sigma profile, the predicted reaction rate constant for the first solvent becomes

$4.11 \log(k)$, nearly six orders of magnitude improvement over the best solvent given in the dataset (Lassau & Jungers 1968), which had a value of $-1.74 \log(k)$. Though this appears to be a significantly better solvent, the estimate of $4.11 \log(k)$ for this structure is likely high. This may be the result of many COSMO surface segments on the nitro groups having very high or very low σ values. In the post-processing COSMO-RS step, some of these segments with extreme σ values are perhaps erroneously considered to participate in H-bonding, which has a large impact on the solvation free energy. Removing H-bonding from the COSMO-RS calculation, we obtain a different value for the reaction rate constant of $0.49 \log(k)$, much closer to our GC estimate. We also note that since there is likely to be some H-bonding in this system, $0.49 \log(k)$ may be a low estimate. Future work will consider only allowing certain segments to participate in H-bonding.

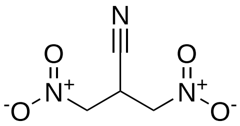
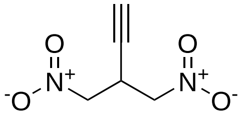
The second solvent, 4-nitro-3-(nitromethyl)but-1-yne, represents another di-nitro compound, although this compound contains a carbon-carbon triple bond and no longer has a cyanide group. It has a slightly lower predicted reaction rate of $-0.11 \log(k)$. Interestingly, this molecule is a di-nitro compound with unsaturated carbon-carbon bonds, defining features of the optimal structure determined by Folic et al. (2008). We also calculate the sigma profile of this structure using Gaussian. This yields a reaction rate constant of $2.88 \log(k)$, representing over four orders of magnitude improvement to the best reported solvent from Lassau & Jungers (1968). As the same issue with H-bonding may exist, we also report a reaction rate constant of $-0.32 \log(k)$ for the same system but without H-bonding. The discrepancy between our estimated reaction rate and the Gaussian-calculated reaction rate is likely due to the inability of the group contribution method to capture the chemical complexities of these structures. In this case, the optimal structure has three very polar groups in close proximity, and the group contribution method treats all of their effects additively. Furthermore, we are not aware of the data

Table 3.3: Summary of important values for the reaction rates solvent case study

Parameter	Value/Range	Additional Information
Time limit	2 hours	Maximum allowable time for the algorithm
Iteration limit	2000	Maximum number of iterations the algorithm can perform
DFO inputs	M_0, M_1, M_2, M_3, M_{acc}, M_{don}	Sigma moments of the solvent to be designed
δ	20%	Property bounds relaxation around DFO trial point
$C_{\max}$	10	Maximum number of compositions determined during each iteration
Carbons	15	Maximum number of carbons in the designed component
Non-carbons	7	Maximum number of non-carbons in the designed component
Triple bonds	2	Maximum number of triple bonds in the designed component
Double bonds	2	Maximum number of double bonds in the designed component

set used by Mu et al. (2007) to create this group contribution method. If the data set contained no di-nitro compounds like the structures shown above, it would be unlikely to predict the sigma profiles of these compounds well. As mentioned, much of the error may also result from an overestimation of the degree of H-bonding in these systems.

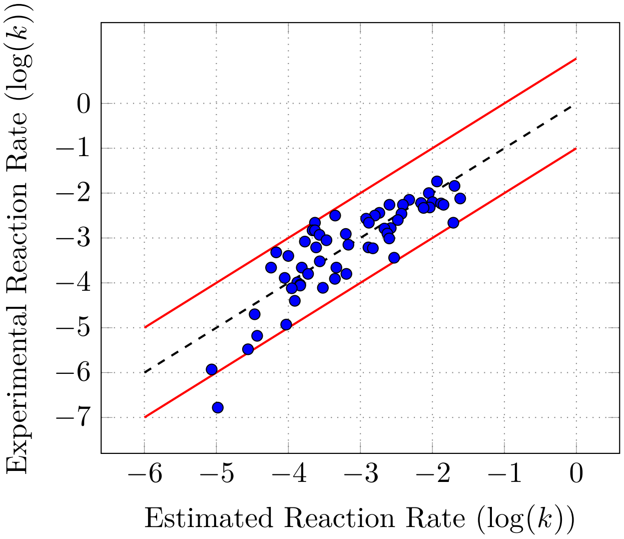
Table 3.4: Representative structures for the reaction rates solvent case study

Solvents	Properties	
	Molar mass:	159.10 g/mol
	Rate constant (GC-COSMO):	$0.45 \log(k)$
	Rate constant (QM-COSMO):	$4.11 \log(k)$
	Rate constant (QM/no H-bonding):	$0.49 \log(k)$
	Molar mass:	158.11 g/mol
	Rate constant (GC-COSMO):	$-0.11 \log(k)$
	Rate constant (QM-COSMO):	$2.88 \log(k)$
	Rate constant (QM/no H-bonding):	$-0.32 \log(k)$

3.5.3 *Reaction rates optimization with mixed solvent*

Finally, we extend the previous case study to investigate binary solvent systems which maximize the reaction rate of the Menschutkin reaction given in Fig. 3.7. To do this, we perform the same optimization algorithm as before but we now consider 12 inputs to our DFO algorithm. These 12 inputs represent two molecular structures, with each structure being defined by the same 6-dimensional vector of σ moments as used above. Upon each iteration of the algorithm, the generated structures are tested for every possible pair of solvents. $C_{\max}$ is a parameter to determine the number of structures designed for each component per iteration, and the number of tested pairs is usually kept small using this parameter. Determining the ratio of the solvents to minimize the activation barrier for a given system is best addressed as a simulation in the mole fraction space.

Figure 3.8: GC-estimated reaction rate vs. experimental reaction rate



Optimization approaches to this problem are possible, but simply scanning through the range of possible mole fraction values proved highly efficient.

We maintain the same parameters for this optimization as discussed in Table 3.3. The mixture design problem in this case can still be solved efficiently with our algorithm. The computational complexity of increasing the number of designed components is somewhat hard to quantify and is highly dependent on the DFO algorithm chosen. Of course, the number of independent variables in the lower-dimensional σ moments space scales linearly with the number of components. For many problems, it has been shown that DFO algorithms are very efficient at determining good and often globally optimal solutions for input-space dimensions up to twenty (Rios & Sahinidis 2013). It is also clear from our analysis of σ moments that all six are often not necessary to capture a solvent's properties with regards to a certain process or chemical objective. Using this observation, we can reduce the dimension of the search space by limiting the number of σ moments included in the design space. For example, recall that the first σ moment represents the total charge of a molecule. Most molecules are very unlikely to deviate much from a value of 0 for this particular σ moment. It is therefore possible to exclude the first moment from many design problems. As with most mixture design algorithms, there is a limit to the number of components our algorithm can simultaneously design. However, we are free to change the inputs to accommodate more components by reducing the dimension of the σ moments space specific to each component.

Nonetheless, optimizing for a binary mixture proved possible using our algorithm. Reaction rates optimization may not be the most practical application of binary solvent design as practitioners are unlikely to synthesize two novel compounds simply to improve a reaction rate. As such, we consider this case study merely as a demonstration of the algorithm and do not account for solvent miscibility. Interestingly, we discovered no composite solvent which outperformed the single component optimum. Our algorithm again found the optimal single component structure and reported it in a ratio of 100:0 with another solvent. This result is perhaps not so surprising. Given the large degree of charge separation in the transition state of the Menschutkin reaction, the rate constant will increase with solvent polarity. Since our single-component optimum, 3-nitro-2-(nitromethyl)propanenitrile, represents one of the most polar (aprotic) species in our feasible region, it is unlikely that a solvent pair would evince higher polarity.

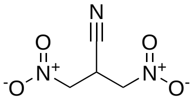
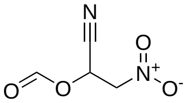
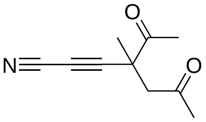
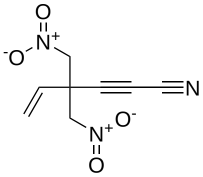
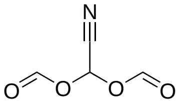
Several other high-performing binary mixtures were identified, a small sample of which is shown in Table 3.5. We note that these molecules contain many of the structural motifs present in the best single-component molecules. For example, prominent substructures include nitro groups, cyanides, and unsaturated carbon-carbon bonds. The first solvent pair in the table has a predicted rate constant of -0.05 $\log(k)$. Optimizing these structures at the quantum chemistry level and subsequently calculating sigma profiles and fixing the mole fraction ratio to 79:21, we obtain a reaction rate constant of 3.64 $\log(k)$. Again, we postulate that this may be unrealistically high due to overemphasizing the H-bonding of the nitro group, so we report a value without considering H-bonding of -0.34 $\log(k)$. A similar result is shown for the second solvent pair reported in the table. The second solution is reported to provide an idea of other structures explored by the algorithm.

Though designing a solvent pair did not provide a better solution, there are many more suitable applications for multi-component mixture design. In fact, there are many reaction rates optimization problems that would likely have better multi-solvent solutions than single-solvent solutions. These will be investigated in a subsequent publication. We primarily demonstrate in this case study that optimization of multi-component mixtures can be efficiently approached using our mixture design algorithm

3.6 CONCLUSIONS

The UNIFAC and SAFT-γ methods for calculating mixture thermodynamics in computer-aided molecular and mixture design problems require binary interaction parameters for every pair of groups in solution, which inherently limits the design space according to which parameters are available. We circumvented this issue by utilizing COSMO-RS and COSMO-SAC methodologies, both of which are capable of accurately calculating mixture thermodynamics without the need for binary interaction parameters. We illustrated that COSMO-based methods can consider chemical systems which cannot be approached with other techniques. By virtue of involving quantum theory, our COSMO-based approach can consider structures with non-standard electronics, deriving the full benefit of quantum chemical accuracy. For example, using quantum chemistry tech-

Table 3.5: Representative structures for the reaction rates solvent case study with mixtures

Solvents		**Properties**	
	–	Ratio (entry 1:entry 2):	100:–
		Rate constant (GC-COSMO):	$0.45\,\log(k)$
		Rate constant (QM-COSMO):	$4.11\,\log(k)$
		Rate constant (QM/no H-bonding):	$0.49\,\log(k)$
		Ratio (entry 1:entry 2):	79:21
		Rate constant (GC-COSMO):	$-0.05\,\log(k)$
		Rate constant (QM-COSMO):	$3.64\,\log(k)$
		Rate constant (QM/no H-bonding):	$-0.34\,\log(k)$
		Ratio (entry 1:entry 2):	56:44
		Rate constant (GC-COSMO):	$-0.10\,\log(k)$
		Rate constant (QM-COSMO):	$0.94\,\log(k)$
		Rate constant (QM/no H-bonding):	$-0.91\,\log(k)$

niques, we modeled the transition state of a reaction and incorporated the complex electronics of this structure directly into a mixture design problem. To best integrate these COSMO methodologies into a CAMxD framework, we developed group contribution methods to estimate sigma profiles, sigma moments, and molecular volumes. The mixture design MINLP was projected onto the space of the sigma moments of each unknown component of the mixture, and derivative-free optimization was used to efficiently optimize over the lower-dimensional projected space.

We demonstrated this COSMO-based mixture design framework on two solvent design problems. Our liquid-liquid extraction solvent design problem resulted in a better solution than the current industrial standard. This solution could only be found by COSMO-based methods as a UNIFAC approach would lack the necessary interaction parameters to make an estimate. Finally, we applied this methodology to a problem that cannot be rigorously approached with UNIFAC and SAFT-γ. We designed a solvent to maximize the reaction rate of a Menschutkin reaction that exhibited a predicted reaction rate increase by nearly six orders of magnitude as compared to the best solvent reported in the experimental study. Removing potentially erroneous H-bonding from the problem, we still obtain a result over two orders of magnitude higher. Finally, we apply the algorithm to design binary mixtures of solvents and discover many solutions better than the best solvent from the experimental study.

Our results suggest that the projection of mixture design onto the lower-dimensional space of the sigma moments results in an effective search space in which to consider CAMxD problems. As the proposed COSMO-based approach can be easily integrated with full-fledged quantum chemistry calculations, it can address a much broader array of problems than previously possible in CAMD and CAMxD.

4

REACTION SOLVENT DESIGN

4.1 INTRODUCTION

In liquid-phase chemistry, the solvent plays a critical role in determining the success of a particular reaction. Altering the solvent can accelerate or diminish a reaction rate, control chemo- or regioselectivity, and influence the formation of any undesirable side-products (Reichardt & Welton 2011). As a result, solvent design/selection can be one of the most crucial considerations in the wider context of reaction design. The importance of the solvent for reaction design is further underscored by its accessibility: solvent choice and mixed solvent mole fraction ratios are not as tightly constrained as are temperatures, pressures, catalysts, reactant structure modifications, and other such variables in reaction design. Despite its impact on reactions, solvent selection is often empirical or based on somewhat rudimentary properties (H-bond donor/acceptor abilities, dielectric constant, solubility parameters, etc.). High-throughput solvent screening is problematic as it can face severe combinatorial difficulties even in the case of binary or ternary solvent mixtures. For these reasons, the solvent selection/design problem for reactions chemistry stands to benefit greatly from leveraging combinatorial optimization approaches in conjunction with more sophisticated thermodynamic and quantum chemical models.

The general solvent design problem is well-studied in the fields of computer-aided molecular design (CAMD) and computer-aided mixture design (CAMxD). As modeling a solvent typically requires some idea of its interaction with its solutes, mixture thermodynamics models often have to be utilized in solving these problems. In the history of CAMD/CAMxD, a very popular choice to calculate interactions with the solvent has been the UNIFAC group contribution method (Fredenslund et al. 1975). There has been significant work in incorporating UNIFAC into solvent design problems. A small sample of some of these studies demonstrates their diversity: extraction solvents (Odele & Macchietto 1993; Pretel et al. 1994), low-environmental-impact solvents (Pistikopoulos & Stefanis 1998; Buxton et al. 1999), crystallization solvents (Karunanithi et al. 2006), and solvents for various consumer products (Conte et al. 2011b). More recent work has focused on solving CAMD/CAMxD problems with the SAFT equation of state (Chapman et al. 1989). Again, applications are numerous, but many have focused on carbon capture solvents (Pereira et al. 2011b; Burger et al. 2015) and fluid design in an organic Rankine cycle (Lampe et al. 2015).

The more specific problem of solvent design for reactions has also been investigated using CAMD techniques, though there are limited examples of such studies. For example, Gani et al. (2005) proposed a rules-based solvent selection/design strategy for reactions based on assigning solvents certain values—so-called "R-indices"—which captured their suitability for a certain reaction in a reduced-dimension space. This method was very successful in optimizing the solvent for a few common reactions as well as for a more complicated problem from the pharmaceutical industry. However, solving problems with this strategy requires somewhat extensive information about specific reaction properties and reaction/solvent relationships. Furthermore, this methodology suggests no straightforward extension to mixed solvents without relying on oversimplifying mixing rules.

A few other approaches have been developed. For example, Folić, Adjiman, and Pistikopoulos (Folić et al. 2007; Folic et al. 2008) designed solvents to maximize the reaction rate of a Menschutkin reaction. Their approach relied on fitting a few experimental data points to the solvatochromic equation (Abraham et al. 1987), a linear solvation free energy relationship, and then solving a CAMD problem with the resultant model. Additional work from Struebing et al. (2013a) proposed an iterative algorithm for reaction solvent design, using DFT and implicit solvation models to estimate reaction rate constants and the solvatochromic equation as a surrogate model to provide lower bounds. This method succeeded in designing solvents with a small number of costly DFT calculations, but the approach is ultimately dependent on the quality of the surrogate model. A limited number of other approaches have integrated quantum chemical calculations with CAMD problems (Stanescu & Achenie 2006).

COSMO-RS (Klamt 1995; Klamt et al. 1998) and COSMO-SAC (Lin & Sandler 2002) are two alternatives for calculating mixture thermodynamics that have some advantages over UNIFAC and SAFT. For example, these COSMO-based methods require no binary interaction parameters, which can place significant limitations on the chemical search space in UNIFAC and SAFT-based group contribution methods like SAFT-γ (Lymperiadis et al. 2007). Furthermore, these COSMO methods are post-processing steps to full-fledged DFT calculations, meaning the use of these methods reflects the accuracy of quantum chemical calculations and can be applied to arbitrary systems. Only very recent work in the CAMD community has focused on integrating these COSMO-based methods into reaction solvent design problems. For example, Zhou et al. (2015) proposed fitting a group contribution method to sections of σ-profiles, building a reduced-order model from this information, and then choosing groups to optimize reaction selectivity. This approach is successful in relating a reduced-dimension σ-profile space to reaction properties but does not use full-order COSMO-RS thermodynamics. This means that, for any solvent design problem, a considerable amount of reaction-specific experimental data is required. Furthermore, reactive species are not considered at the quantum chemistry level, limiting the level of detail this approach can consider. Similar approaches have been applied to screening solvents (Zhou et al. 2014).

The purpose of this work is to build upon previously developed COSMO-based molecular/mixture design methodology (Austin et al. 2016a) (discussed in Chapter 3) and to investigate the utility of COSMO-based CAMD for reaction solvent design. In particular, we aim to make our design methodology amenable to design pure and mixed solvents for several industrially relevant reactive systems. Primarily, we focus on three new additions to the previous methodology: (1) altering the group contribution method to estimate hydrogen-bonding and non-hydrogen-bonding sigma profiles; (2) explicit, *ab initio* modeling of strong solute/solvent interactions such as H-bonding or coordinate bonding and incorporating this information directly into the design problems; and (3) solving mixture design problems limited to common laboratory and industrial solvents. Extensions (1) and (2) lead to considerable improvement in the accuracy of our predictions. Extension (3) re-frames the mixed solvent design problem in a more practical search space, meaning solutions to this problem can be readily implemented without the need to synthesize new compounds. We first apply this methodology to design a solvent to maximize the reaction rate of a Menschutkin reaction, a simple S_N2 reaction. We next consider two systems which are significantly more complicated than systems explored previously in the reaction solvent design literature. The first of these more complicated design problems is optimizing the chemoselectivity of a lithiation reaction. The final design problem involves controlling chemoselectivity in an intramolecular nucleophilic aromatic substitution (S_NAr) reaction to produce substituted xanthones.

Our approach is distinguished from others in the literature because we include quantum-mechanics-derived σ-profiles directly into the CAMD problem for every reactive species. This allows us to consider complicated reaction phenomena at the quantum mechanics level of accuracy. Additionally, our approach requires very few parameters, making it very generalizable.

In the next section, we provide a brief discussion of some of the advantages of using COSMO-RS in the context of reaction solvent design. In Sec. 4.3, we detail three changes to our existing COSMO-based mixture design methodology. Then, in Sec. 4.4, we apply the methodology to the three case studies discussed above. Finally, in Sec. 4.5, we provide a few conclusions about the work and discuss the suitability of COSMO-based mixture design for arbitrary industrial reaction design problems.

4.2 THE UTILITY OF COSMO-RS FOR CAMXD

COSMO-RS is especially useful in a CAMD/CAMxD context. One major advantage is that COSMO-based models are able to calculate chemical potentials (and, trivially, free energies of solvation) without binary interaction parameters. These binary interaction parameters are necessary in models like UNIFAC and SAFT-γ and must be present for every pair of groups in a system in order for these methods to calculate chemical potentials. Due to the limited availability of thermodynamic data for many types of compounds, many of these interaction parameters simply do not exist. This can have significant consequences for CAMD/CAMxD problems using UNIFAC or SAFT-γ as the chemical search space is inherently limited to the portion of the chemical design space for which every binary interaction parameter is available.

Furthermore, COSMO-RS allows for easy integration of quantum chemistry calculations into CAMD/CAMxD problems. This greatly expands the envelope of possible molecular design problems as many previously inaccessible systems can now be considered at a high level of accuracy. For example, we are now able to consider species with complex electronics such as transition states, ionic liquids, radicals, and zwitterions. This methodology can also be extended to organo-metallic chemistry and perhaps general reaction design using transition metal catalysts. One qualification to this claim is that we can only use quantum-accurate σ-profiles for the species in the mixture that are fixed. The σ-profiles and cavity volumes of molecules in our design space must be estimated using lower-order models, which in our case are group contribution methods. For example, in designing an optimal solvent for the simple hypothetical reaction $A + B \rightarrow C$, we can model species A, B, and C as well as the transition state using *ab initio* methods. Though we do have to rely on lower-order methods, most solvent design problems are not intended to produce solutions with complex electronics or other features which would require a quantum chemical treatment.

4.3 EXTENSIONS TO THE EXISTING FRAMEWORK

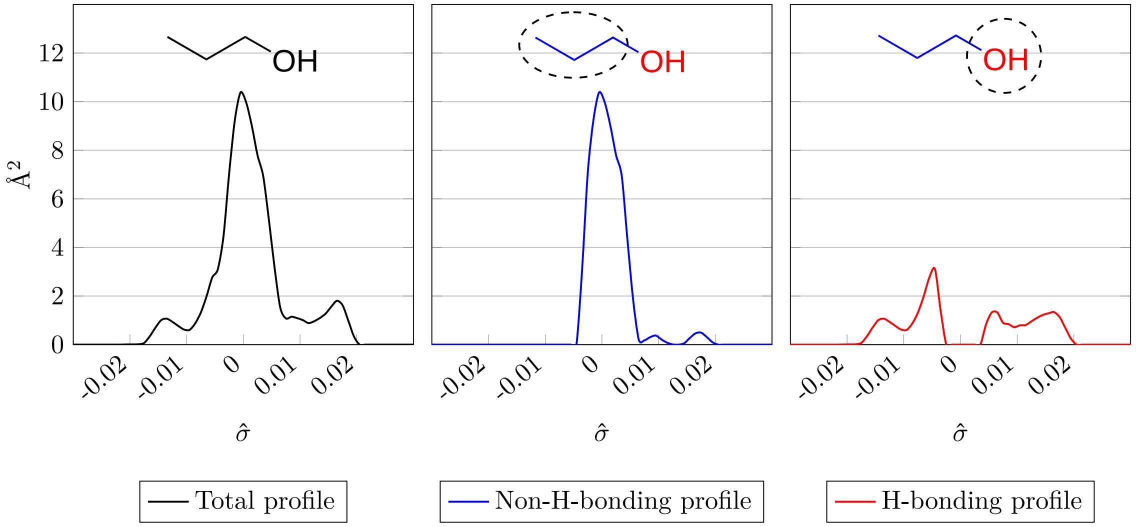
4.3.1 *Splitting the σ-profile into H-bonding and non-H-bonding profiles*

Hydrogen-bonding is one of the strongest intermolecular interactions, and it is given an appropriately large weight in COSMO-RS thermodynamics. However, the classical COSMO-RS (Klamt 1995; Klamt et al. 1998; Klamt & Eckert 2000; Eckert & Klamt 2002) assumes two interacting surface segments (or $\hat{\sigma}$ values in our case) will always participate in hydrogen-bonding if they are beyond a certain threshold σ value, σ'_{hb}. While this assumption may simplify some of the thermodynamics, it can also result in attributing an erroneously large or small interaction energy to certain surface elements. For example, it is unlikely that iodomethane participates in H-bonding as a result of iodine's insufficient electronegativity, but it is assumed to accept H-bonds in the classical COSMO-RS view. The reason for this is due simply to a disproportionate allocation of electrons in the carbon-iodine bond, giving iodomethane's σ-profile non-zero areas for σ values beyond the H-bonding threshold.

For this study, we use the group contribution method of Mu et al. (2007) to predict σ-profiles and COSMO cavity volumes. We opt for this method in lieu of creating our own for compatibility and data accessibility reasons outlined previously (Austin et al. 2016a). In this group contribution method, the authors already distinguish between H-bonding and non-H-bonding parts of the σ-profile. Rather than determine H-bonding using σ cutoff values, Mu et al. (2007) appeal to the traditional definition: only highly electronegative atoms (N, O, and F) and any H's attached to these atoms can participate in H-bonding. This gives rise to two σ-profiles, where groups containing an N, O, F or an H attached to one of these atoms contribute accordingly to an H-bonding σ-profile, and all other groups contribute to a non-H-bonding σ-profile. For example, 1-propanol is a short molecule containing an aliphatic chain and a hydroxyl functional group. Using this definition of H-bonding, all of the groups from the aliphatic chain contribute to the non-H-bonding σ-profile. Since both atoms of the hydroxyl group qualify for the H-bonding definition, these contribute to the H-bonding σ-profile. A more detailed picture of this for 1-propanol is given in Fig. 4.1.

The division into two σ-profiles changes the procedure to calculate chemical potentials slightly. First, we define the two new mixture σ-profiles corresponding to the H-bonding

Figure 4.1: The total, H-bonding, and non-H-bonding σ-profiles of 1-propanol



$(P_S^{HB}(\hat{\sigma}))$ case and the non-H-bonding $(P_S^{NHB}(\hat{\sigma}))$ case. These follow simply from the definitions above:

$$P_S^{HB}(\hat{\sigma}) = \frac{\sum_i P_i^{HB}(\hat{\sigma})x_i}{\sum_i A_i x_i}$$

$$P_S^{NHB}(\hat{\sigma}) = \frac{\sum_i P_i^{NHB}(\hat{\sigma})x_i}{\sum_i A_i x_i}$$

where $P_i^{HB}(\hat{\sigma})$ represents the H-bonding profile and $P_i^{NHB}(\hat{\sigma})$ represents the non-H-bonding profile for a compound i. Note here that the total area of the H-bonding profile and the non-H-bonding profile together is equal to 1. These two profiles lead to two different σ potentials, $\mu_S^{HB}(\hat{\sigma})$ for H-bonding and $\mu_S^{NHB}(\hat{\sigma})$ for non-H-bonding:

$$\begin{aligned}
\mu_S^{HB}(\hat{\sigma}) =& -RT\ln\sum_{\hat{\sigma}'} P_S^{NHB}(\hat{\sigma}')\exp\left(\frac{\mu_S^{HB}(\hat{\sigma}') - E_{\text{misfit}}(\hat{\sigma},\hat{\sigma}')}{RT}\right) \\
&+ P_S^{HB}(\hat{\sigma}')\exp\left(\frac{\mu_S^{HB}(\hat{\sigma}') - E_{\text{misfit}}(\hat{\sigma},\hat{\sigma}') - E_{\text{HB}}(\hat{\sigma},\hat{\sigma}')}{RT}\right) \\
\mu_S^{NHB}(\hat{\sigma}) =& -RT\ln\sum_{\hat{\sigma}'}(P_S^{HB}(\hat{\sigma}') + P_S^{NHB}(\hat{\sigma}'))\exp\left(\frac{\mu_S^{HB}(\hat{\sigma}') - E_{\text{misfit}}(\hat{\sigma},\hat{\sigma}')}{RT}\right)
\end{aligned} \tag{4.1}$$

Finally, we can calculate the chemical potential of a molecule i in the solution S:

$$\mu_i = \sum_{\hat{\sigma}}\left\{P_i^{HB}(\hat{\sigma})\mu_S^{HB}(\hat{\sigma}) + P_i^{NHB}(\hat{\sigma})\mu_S^{NHB}(\hat{\sigma})\right\} + \mu_i^C$$

Note that if the species i is unable to form H-bonds, it will have $P_i^{HB}(\hat{\sigma}) = 0$ for every $\hat{\sigma}$ value. This means it will only interact with the solvent's non-H-bonding profile. Conversely, if compound i has the capability to form H-bonds, it will interact with the solvent's H-bonding profile. This interaction will be weighted by the respective surface areas of the H-bonding and non-H-bonding profiles of all species.

4.3.2 *Explicit treatment of strong intermolecular forces*

For many systems, important reaction characteristics (reaction rate, selectivity, purity, conversion, etc.) are dictated by the presence of strong intra- or intermolecular forces. These can include: H-bonding, coordinate bonding, ionic bonding, ion-dipole interactions, and steric effects. For this reason, we advocate including as much system-specific information in the quantum chemistry calculations as possible. For example, a certain class of solvents may have a strong interaction (say, H-bonding) with one of the reactants, a transition state, or a short-lived intermediate. Since we know H-bonding geometry can vary widely with the system (Desiraju & Steiner 2001; Steiner 1998; Taylor & Kennard 1984; H. Guo & Karplus 1994), it is advantageous to model H-bonding explicitly at the quantum chemistry level. This is also possible as reaction design problems, at least in these examples, necessitate modeling all relevant reactive species at the quantum chemistry level. These structures can be examined for H-bonding potential and re-optimized with an explicit H-bond with a solvent molecule.

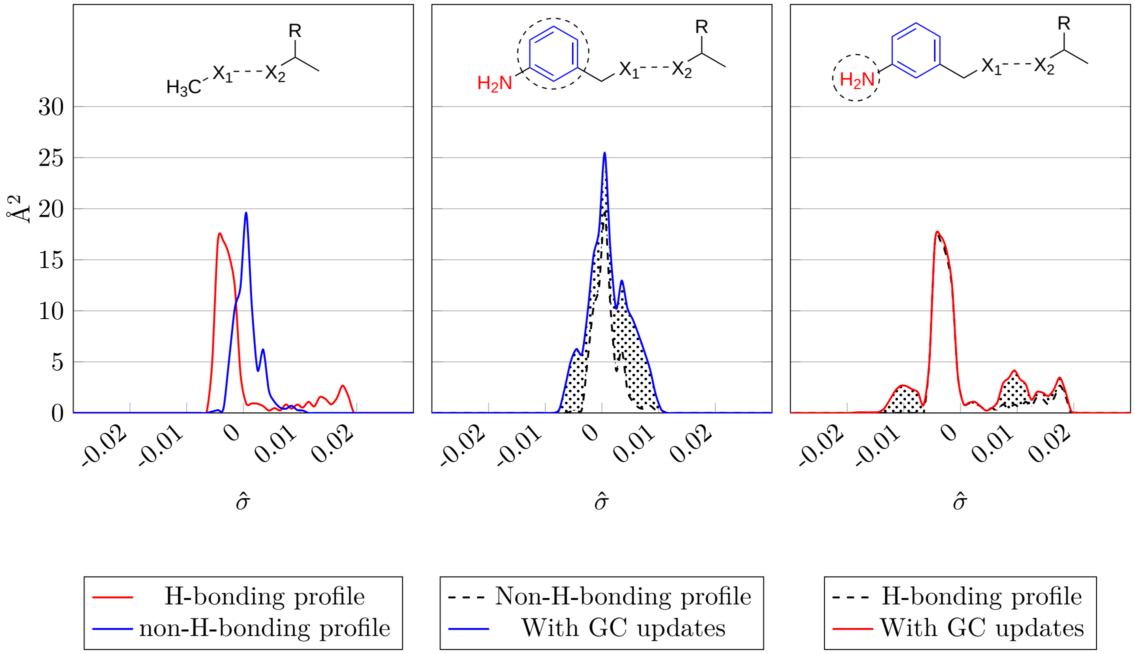
However, our algorithm designs a very large number of possible solvent molecules, meaning we cannot practically consider every designed structure at the quantum level. This necessitates modeling only a representative structure at a high level. To continue with the H-bonding example, if we wish to include alcohols in our design space, we may want our representative structure to be a methanol molecule explicitly H-bonded to any important reactive species. Doing so allows us to capture the particular geometry of any alcohol-donated H-bonds present in the system. Finally, the H-bonding and non-H-bonding σ-profiles of our representative structure only capture the H-bonding of methanol and not all solvent characteristics for other potential alcohol solvents (e.g. bifunctional alcohols) in our design space. To account for this, we modify these σ-profiles for every alcohol solvent in the design space, adding the group contribution estimates of all atoms not present in the representative structure.

In general, we optimize the geometries and obtain the quantum-level σ-profiles for every reaction-relevant species in solution. We next investigate these species and determine if there are any potential strong interactions with any type of solvent we are considering. If so, we model a representative system which captures that interaction for every possible pair of reactive species and solvent class. We optimize the geometries of these representative systems and again obtain the quantum-level σ-profiles. In considering a particular type of solvent in the design problem that falls into one of these classes, we update the corresponding pre-computed σ-profile, using the modified GC method of Mu et al. (2007) to account for all atoms not in the representative structure.

A simple example of this is given in Fig. 4.2. In this example, we assume there is some strong interaction between groups X_1 and X_2. X_2 may be a strong proton donor like a carboxylic acid, and X_1 may be the reactive part of a nucleophile which can act as a H-bond acceptor. This interaction would be important to model as the nucleophilicity of X_1 would be strongly influenced by its H-bond formation in solution. This will, in turn, affect the reaction rate. In this case, we can imagine approximating this interaction with a simple representative system of the nucleophile of interest H-bonded to $X_1 - CH_3$. This simple system now represents a large number of solvents, and many of these may be examined at each iteration of our high-throughput algorithm. In the example in the figure, one such solvent may be $2 - (3 - \text{aminophenyl}) - X_2$, an aniline derivative shown in Fig. 4.2. We first optimize the geometry and obtain the quantum-level σ-profiles from the representative system. Then, we add to those σ-profiles, using the aromatic carbons to contribute to the non-H-bonding profile and the amine to contribute to the H-bonding profile.

For completeness, we note that COSMO-RS does account for H-bonding, but it does so using the statistical thermodynamics that underlie the method rather than a more detailed depiction of two specific interacting atoms. This is done for clear reasons—COSMO-RS is intended to be and is a very general method. However, in these reaction design problems, we are often afforded specific information about important reactive species and, as a result, can determine exact geometries of important interactions. Furthermore, there are a number of systems for which COSMO-RS has yet to be parameterized (e.g. organometallics). Modeling all of the important interactions for these systems at the quantum level means we can still reasonably consider these problems

Figure 4.2: GC updates to the σ-profiles of an example system



using COSMO-RS as solvation has been addressed with explicit representative systems. Overall, this approach is advantageous as it minimizes our reliance on GC methods, allowing us to incorporate much accurate information directly into CAMD/CAMxD problems.

4.3.3 *Mixture design with common solvents*

Though we have investigated the mixture design problem for reaction optimization in the past (Austin et al. 2016a), the solutions to this problem may not always be the most practical. There are two primary reasons for this. First, optimizing for mixed solvent systems requires using a GC method for every unknown component in the solvent blend.

Though GC methods can be trusted in many domains, their errors have the potential to compound when applied in multiple parts of a problem. More specifically, the GC-predicted σ-profiles for each component of the solvent blend may each be close to their respective QM-predicted profiles, but the two σ-profiles of the mixture may deviate more significantly. The second reason for the impracticality of mixed solvent design concerns the likelihood that these compounds will even be synthesized. Assuming even a considerable improvement in reaction performance in a designed multi-component solvent system, the time and expense of synthesizing each component of this mixed solvent can easily outweigh whatever reaction/process advantages it provides.

It is then perhaps more practical to consider the problem of selecting a mixed solvent system from a list of common laboratory and industrial solvents. Solving this problem in favor of the design problem has several advantages. First, solutions can be readily implemented. The assumption is that the solutions will draw from a set of common solvents, so the mixture components should all be easily accessible. Second, unlike the design problem, there is no cost associated with making these solvent blends. This has the advantage in industry of improving the performance without driving up the cost. Finally, considering these problems with COSMO-RS, we are in position to optimize the geometries of all of the solvent structures at the quantum level a priori and then obtain highly accurate σ-profiles for every solvent we wish to consider. This approach is particularly amenable with COSMO-RS as the σ-profiles of a mixed solvent are simply a linear combination of the respective H-bonding and non-H-bonding σ-profiles of the components and normalized by the average surface area. This means we only have to perform one quantum chemistry calculation for every solvent to be considered. These calculations only need to be done once, and the resultant σ-profiles can be used in any number of subsequent design problems.

Table 4.1: Common industrial and laboratory solvents used for mixture design/selection problems

Entry	Solvent name	Entry	Solvent name
1	acetic acid	15	DMSO
2	acetone	16	ethanol
3	benzene	17	ethyl acetate
4	carbon tetrachloride	18	formamide
5	2,2,2-trifluoroethanol	19	isopropanol
6	chloroform	20	acetonitrile
7	cyclohexane	21	methanol
8	dichloromethane	22	*n*-hexane
9	diethyl ether	23	nitromethane
10	diglyme	24	pyradine
11	dimethyl ether	25	THF
12	dioxane	26	toluene
13	DMF	27	trichloroethylene
14	DMPU	28	water

Though high-throughput screening approaches are possible for this problem in industry, designing multi-component solvent systems can quickly face combinatorial issues. Given 100 possible solvents, there are over 4 million combinations of up to 4 components. This number also doesn't capture a more difficult aspect of mixture design: determining mole fractions. Mole fractions, of course, represent a continuous space, so the number of possible solvent systems is actually infinite. In the following case studies, we will consider designing 4-component mixtures from the set of solvents given in Table 4.1. There are 28 solvents listed here, making for over 23,000 possible combinations of up to size 4.

For the solvent selection problem, there is no need to incorporate groups as we already have quantum-level H-bonding and non-H-bonding σ-profiles and COSMO cavity volumes for every potential solvent. The removal of groups from the design space greatly reduces the dimensionality of the overall mixture design problem and means that this problem can be directly approached with numerical optimization techniques. This problem calls for one to make two main determinations: (1) the identities of the solvents making up the optimal mixture and (2) their mole fractions. The choice whether to include a particular solvent in the optimal mixture can be captured with a binary variable, and the COSMO-RS will serve to calculate the thermodynamics of the system.

These features make this problem a mixed-integer nonlinear program (MINLP). The optimization formulation is given below:

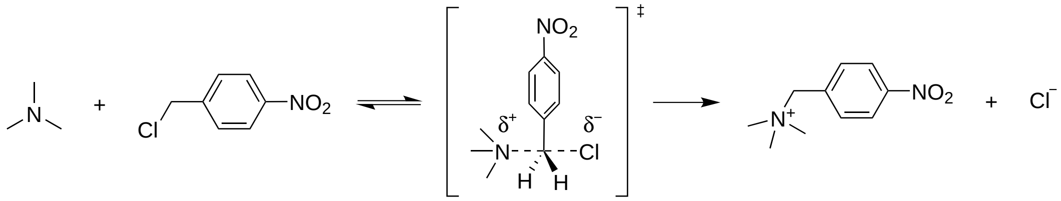
$$
\begin{aligned}
\min \quad & p_1^T \mu_S^{HB} + p_2^T \mu_S^{NHB} + p_3^T \mu^C && (4.2)\\
\text{s.t.} \quad & \mu_S^{HB}(\hat{\sigma}) = -RT \ln \sum_{\hat{\sigma}'} \left\{ P_S^{NHB}(\hat{\sigma}') \exp\left(\frac{\mu_S^{HB}(\hat{\sigma}') - E_{\text{misfit}}(\hat{\sigma},\hat{\sigma}')}{RT}\right)\right. \\
& \left. + P_S^{HB}(\hat{\sigma}') \exp\left(\frac{\mu_S^{HB}(\hat{\sigma}') - E_{\text{misfit}}(\hat{\sigma},\hat{\sigma}') - E_{\text{HB}}(\hat{\sigma},\hat{\sigma}')}{RT}\right)\right\} \quad \forall \hat{\sigma} && (4.3)\\
& \mu_S^{NHB}(\hat{\sigma}) = -RT \ln \sum_{\hat{\sigma}'} \left\{ \vphantom{\frac{a}{b}} \right. \\
& \left. (P_S^{HB}(\hat{\sigma}') + P_S^{NHB}(\hat{\sigma}')) \exp\left(\frac{\mu_S^{HB}(\hat{\sigma}') - E_{\text{misfit}}(\hat{\sigma},\hat{\sigma}')}{RT}\right)\right\} \quad \forall \hat{\sigma} && (4.4)\\
& P_S^{HB}(\hat{\sigma}) = \frac{\sum_i P_i^{HB}(\hat{\sigma}) x_i}{\sum_i A_i x_i} \quad \forall \hat{\sigma} && (4.5)\\
& P_S^{NHB}(\hat{\sigma}) = \frac{\sum_i P_i^{NHB}(\hat{\sigma}) x_i}{\sum_i A_i x_i} \quad \forall \hat{\sigma} && (4.6)\\
& \sum_i y_i \leq K && (4.7)\\
& x_i \leq y_i \quad \forall i && (4.8)\\
& \sum_i x_i = 1 && (4.9)\\
& By \leq 0 && (4.10)
\end{aligned}
$$

In this model, the set $i = \{1, \ldots, I\}$ represents an index over the possible solvents. For our problems, these solvents come from Table 4.1, so $I = 28$. The binary variable y_i is equal to 1 if solvent i is chosen to be in the mixture and is equal to 0 if not. The positive continuous variable x_i represents the mole fraction in solution of solvent i. Parameter vectors p_1 and p_2 represent problem-specific σ-profile information for the reactants, transition states, and important intermediates that affect the reaction objective in (4.2). Parameter vector p_3 is again problem-specific and simply ensures the combinatorial terms are added and subtracted correctly. Though used in our mixture design problems, the calculation of the combinatorial terms for each solute are not included in this general formulation as there are several that have been used with COSMO-RS (Eckert & Klamt 2002; Eberhart & Kennedy 1995; Klamt & Eckert 2000; Klamt 2005). Eqs. (4.3) and (4.4) calculate the σ potential for every $\hat{\sigma}$ value. These potentials are functions of P_S^{HB} and P_S^{NHB}, which are shown in Eqs. (4.5) and (4.6) for clarity but, in our true formulation, are incorporated directly into Eqs. (4.3) and (4.4), respectively. Constraint (4.7) limits our solvent blends to contain at most K components, where for the following case studies $K = 4$. Constraint (4.8) ensures that, if a solvent is not chosen, it must have a mole fraction equal to 0. Eq. (4.9) constrains all mole fractions to sum to 1. Finally, (4.10) defines a number of integer cuts which prohibit immiscible solvents combinations from being a solution. These cuts can also be used to remove single or mixed solvents for any other reason (toxicity, cost, availability, etc.). We solve this model using BARON (Sahinidis et al. 2003). All of the default options are used with the exception of turning on deltaterm and setting deltat to 50.

4.4 CASE STUDIES

For the case studies in this paper, we use the TOMLAB/CGO solver (Holmström et al. 2007) as the DFO algorithm, and we perform all runs on a 2.84 GHz processor. Furthermore, the melting and boiling points of all designed solvents are estimated with the Marrero-Gani GC method (Marrero & Gani 2001) and an extension for missing

Figure 4.3: The Menschutkin reaction between trimethylamine and *p*-nitrobenzyl chloride



groups (Gani et al. 2005). These are constrained in each of the case studies to ensure that each designed solvent is a liquid at the reaction temperature.

4.4.1 *Reaction rates optimization solvent*

The Menschutkin reaction defines a class of reactions in which a tertiary amine reacts with an alkylhalide to form a quaternary ammonium salt. This reaction is a popular choice (Folić et al. 2007; Folic et al. 2008; Austin et al. 2016a) for reaction solvent design problems in CAMD largely due to its simplicity. It proceeds via an S_N2 mechanism and has no competitive pathways. Its transition state has a large degree of charge separation, meaning that the solvent has a dramatic and predictable effect on the reaction rate.

This solvent design problem will focus on maximizing the reaction rate of a Menschutkin reaction. The particular Menschutkin reaction used here is the reaction between trimethylamine and *p*-nitrobenzyl chloride, which is shown in Fig. 4.3. We begin by optimizing the geometries of species **1**, **2**, and **TS1** with Gaussian09 (Frisch & et al. 2009) using the B3LYP functional (Becke 1993; Stephens et al. 1994) and a 6-311g(d,p) basis set. We perform a subsequent COSMO calculation on each of these structures and then obtain their σ-profiles. We retain a few energy values from these calculations.

First, $\Delta G^{\ddagger}_{\text{COSMO}}$ represents the Gibbs energy differences between the transition state and compounds **1** and **2** in the COSMO phase. We additionally assume that vibrational energy differences between the transition state and compounds **1** and **2** are constant in different solvents. For every solvent we investigate, we also calculate a ΔG^{1}_{RS}, ΔG^{2}_{RS}, and $\Delta G^{\text{TS1}}_{\text{RS}}$. These represent the Gibbs energy of transfer from the COSMO phase to a particular solvent. These values are calculated with the COSMO-RS model, parameters, combinatorial term taken from Klamt & Eckert (2000; 2003).

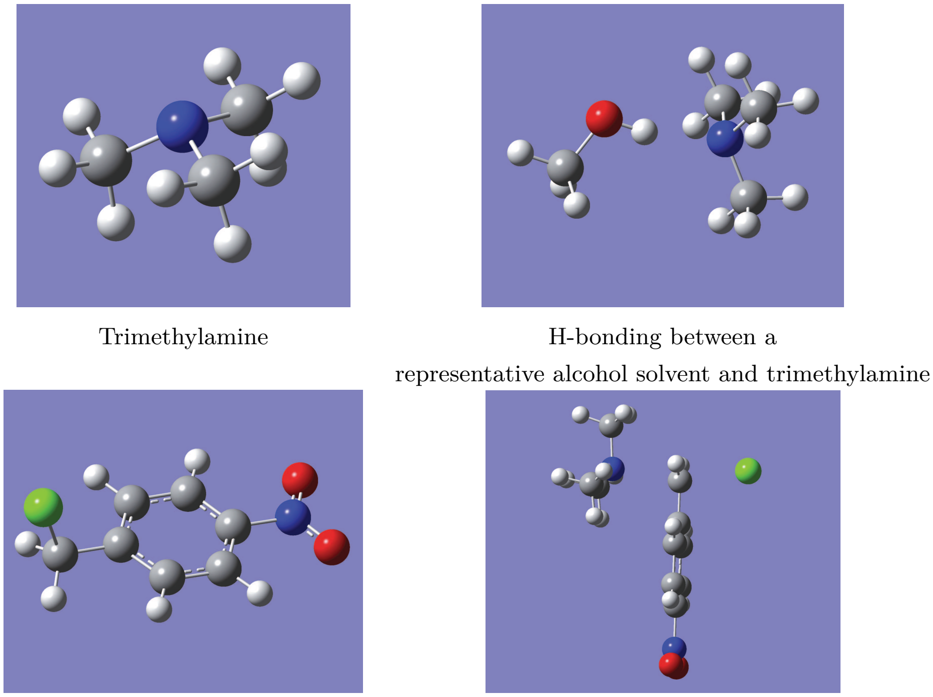
Additionally, we consider H-bonding with the amine nucleophile explicitly at the quantum level. H-bond donating solvents are known to reduce the nucleophilicity of certain reactants, so modeling the effect will serve to increase the accuracy of reaction rate constant predictions. To account for these H-bonding effects, we model the amine nucleophile participating in a H-bond with methanol, which will serve as a representative alcohol. Using the methods detailed above, we update the sigma profile of the methanol-amine system using the GC method of Mu et al. (2007) to account for any atoms not present in methanol. This updated sigma profile is then used when the solvent is an alcohol. The geometries of all of the relevant species for this reaction are given in Fig. 4.2.

Referring to transition state theory (TST), we calculate the reaction rate constants as a function of Gibbs energy terms. Specifically, we use the following equation:

$$k = \kappa \frac{k_B T}{h} \exp\left(- \frac{\Delta G^{\ddagger}_{\text{COSMO}} + (\Delta G^{\text{TS1}}_{\text{RS}} - \Delta G^{1}_{\text{RS}} - \Delta G^{2}_{\text{RS}})}{RT} \right) \tag{4.11}$$

In the above, k is the reaction rate constant, k_B is Boltzmann's constant, h is Planck's constant, T is the temperature, R is the gas constant, and κ is a proportionality constant to account for the fact that not every vibration of the transition state leads to the products. In this case study, we set κ to 1.

Table 4.2: Optimized geometries of the reactants, explicit H-bonding between the nucleophile and the amine, and the transition state used in the reaction rates solvent optimization problem.



Trimethylamine

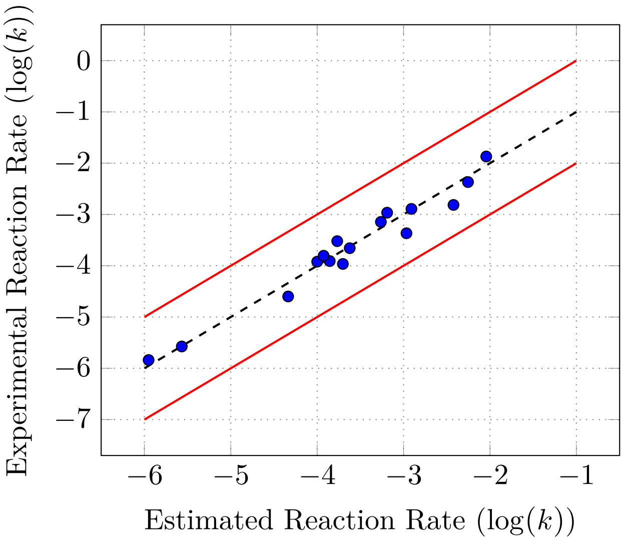
H-bonding between a representative alcohol solvent and trimethylamine

p-nitrobenzyl chloride

Transition state for the S_N2 reaction

Next, we benchmark our approach against experimental reaction rate data for this particular Menschutkin reaction taken from Abraham (1971). For each of the 16 solvents in the data, we estimate the σ-profiles of that solvent using the modified GC method of Mu et al. (2007). We then calculate all of the relevant ΔG terms using the COSMO-RS methodology described above. Note that two of the original 18 solvents were removed from the comparison because the group contribution method could not completely describe them. We compare the experimental and GC-estimated reaction rates for this system in Fig. 4.4. The dashed line in the center signifies the equality

Figure 4.4: GC-estimated reaction rate vs. experimental reaction rate



line between estimated and experimental values for the reaction rate constant k, given here in units of mol/s. The solid red lines on either side represent one $\log(k)$ (base 10) deviation from the experimental values. As shown, none of the solvents falls outside of the one-log-deviation lines. In fact, the largest error for this data is a deviation of 0.39 log units, or about 2.5 mol/s. This is a very small error given the complex nature of estimating reaction rates. In the plot shown, we obtain a coefficient of determination R^2 of 0.96. Again, we investigate the accuracy of our model in terms of AAPE (Eq. (3.39)) for our data set of $N = 16$ molecules from Abraham (1971). Our previous study of a different Menschutkin reaction had an $AAPE$ of 12.50% Austin et al. (2016a) on 59 data points. For the purposes of comparing our modified solvent design methodology, the $AAPE$ for this set of compounds was 5.52%. We note that this comparison is done for two different reactions, so it should not be thought to represent exact, quantitative differences between the two methods. This should only provide some idea of the scale of improvement observed.

To solve this reaction rates optimization problem, we follow the algorithm described above, optimizing in the reduced-dimension space of 5 of the 6 σ moments. Additional optimization parameters for this problem are given in Table 4.3. We calculate the reaction rate constant, k, using COSMO-RS and the TST approach discussed. We additionally remove sulfoxides, amines, and carboxylic acids from the solvent search space to remove potentially reactive solvents. We found a large variety of structures for this problem. A representative list of several of the highest-performing structures is given in Table 4.4. The first entry in the list has the highest predicted reaction rate of all of the structures found as solutions. Its predicted reaction rate constant of $-0.54 \ \log(k)$ is higher than the best-performing solvent in the experimental data by over an order of magnitude. This compound is highly polar and aprotic, making it a very likely candidate for increasing the rate of an S_N2 reaction. The second is a di-nitro compound with unsaturated C-C bonds, a type of compound which has been predicted as a high-performing solvent for the Menschutkin reaction in previous studies (Folić et al. 2007; Folic et al. 2008; Austin et al. 2016a). This particular di-nitro compound, representative of a large number which were found, was predicted to have a reaction rate constant of $-0.88 \ \log(k)$. The third entry represents a family of furans with nitroalkyl groups. These were also high-performers, attaining a predicted reaction rate constant of $-1.03 \ \log(k)$ in the case of the third entry in the table.

Solvent design using this algorithm has one additional advantage. For any molecule determined to be a good solvent, we can optimize the geometry of that molecule at the quantum level, perform a single point calculation using the COSMO solvation model, and extract a quantum level σ-profile for that molecule. This σ-profile can then be used to estimate reaction properties with the σ-profiles of the reactive species (also already at the quantum level). In this way, the solvent design process can be completed with an estimate that is accurate to the full level of COSMO-RS.

We report reaction rate estimates using the QM σ-profiles in Table 4.4. Unfortunately, estimating the reaction rates using the full COSMO-RS does not appear to agree well with the GC-estimates. We have encountered difficulties with nitro groups using COSMO-RS before. Surface charges on nitro groups, containing all H-bond-capable elements, are assigned to the H-bonding profile in our COSMO post-processing step. This is perhaps erroneous as nitro groups are unlikely to participate in H-bonding to such an extent. To remove this effect, we also report reaction rates estimates where we have removed the H-bonding term from the COSMO-RS model. Doing this, we observe a much better agreement with the GC-estimated reaction rates. Interestingly, nitro groups in the GC method of Mu et al. (2007) contribute only partially to the H-bonding profile. This is a more physically-realistic representation and is likely why we observe such good agreement with GC-estimated reaction rates and QM-estimated reaction rates without H-bonding.

Finally, we consider the mixture design problem from the perspective of obtaining practical, easily-implementable solutions. We optimize to find optimal mixtures of up to 4 components, determining both which set of common solvents constitute an optimal solution as well as the mole fraction for each component. This means that all solutions are some combination of the solvents given previously in Table 4.1. The top 10 solutions for this problem are listed in Table 4.5. Pure nitromethane was the best solvent found. This result is not surprising: nitromethane is the most polar, aprotic compound on the list, and polarity should dictate the reaction rate in a reaction with high charge separation in its transition state like the Menschutkin reaction. However, nitromethane may not be a desired solvent for cost or toxicity reasons. Looking at the second solution, a mixture of chloroform and DMSO, we observe a higher reaction rate constant using

Table 4.3: Summary of important values for the reaction rates solvent case study

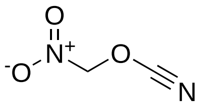
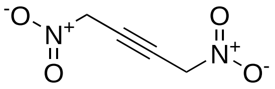
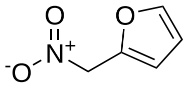
Parameter	Value/Range	Additional Information
Time limit	2 hours	Maximum allowable time for the algorithm
Iteration limit	2000	Maximum number of iterations the algorithm can perform
DFO inputs	M_0, M_2, M_3, M_{acc}, M_{don}	Sigma moments of the solvent to be designed
δ	20%	Property bounds relaxation around DFO trial point
C_{max}	10	Maximum number of compositions determined during each iteration
Carbons	15	Maximum number of carbons in the designed component
Non-carbons	7	Maximum number of non-carbons in the designed component
Triple bonds	2	Maximum number of triple bonds in the designed component
Double bonds	2	Maximum number of double bonds in the designed component

mixed solvents than chloroform or DMSO alone. This solution is also better than every other experimental solvent from Abraham (1971) other than nitromethane.

4.4.2 *Solvent-controlled selectivity of a lithiation reaction*

Lithiation and the lithium-halogen exchange are two powerful and versatile classes of reactions in modern synthetic chemistry (Wakefield 2013). They are especially useful reactions for syntheses encountered in the pharmaceutical industry due to their C-C

Table 4.4: Representative structures for the reaction rates solvent case study

Solvents	**Properties**	
	Rate constant (GC-COSMO):	$-0.54\ \log(k)$
	Rate constant (QM-COSMO):	$3.39\ \log(k)$
	Rate constant (QM/no H-bonding):	$-0.46\ \log(k)$
	Rate constant (GC-COSMO):	$-0.88\ \log(k)$
	Rate constant (QM-COSMO):	$0.69\ \log(k)$
	Rate constant (QM/no H-bonding):	$-0.92\ \log(k)$
	Rate constant (GC-COSMO):	$-1.03\ \log(k)$
	Rate constant (QM-COSMO):	$-1.99\ \log(k)$
	Rate constant (QM/no H-bonding):	$-2.46\ \log(k)$

bond formation and ability to control a reaction's chemoselectivity and regioselectivity. To illustrate that our COSMO-based molecular/mixture design methodology can be applied to more difficult and synthetically-relevant chemical systems, we choose a lithiation reaction to model as our second case study. In particular, we attempt to optimize a reaction solvent to provide maximum selectivity for one product versus the other and a second solvent to promote the reverse selectivity.

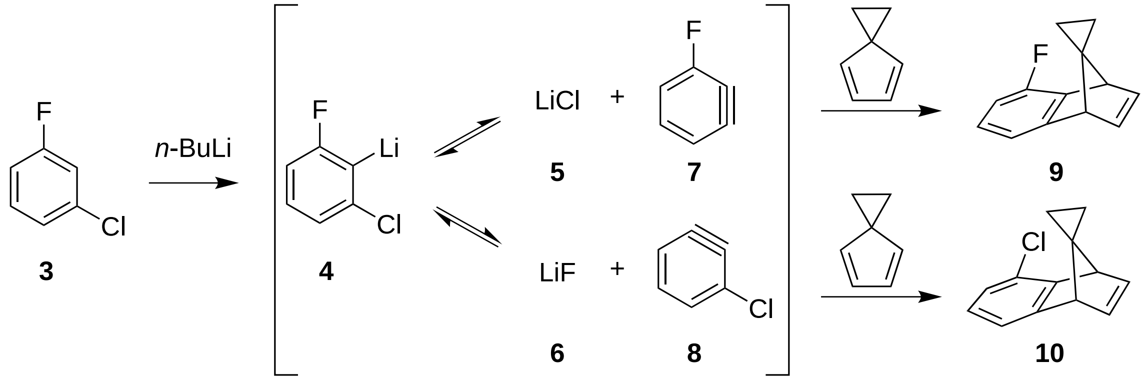
The particular reaction we investigate for this case study comes from Coe et al. (2004). This is a lithiation of 1-chloro-3-fluorobenzene which leads to a solvent-dependent ratio of two products. Coe et al. (2004) suggest that this reaction begins by a lithium substitution at the 2 position (ortho to both fluorine and chlorine). This lithiated species then forms a benzyne, shedding either lithium fluoride or lithium chloride as a salt. The benzyne, now containing a single chlorine or fluorine substitution, is a highly

Table 4.5: Top solvent blends using common solvents for the reaction rates solvent case study

Solution rank	Solvents	Mole fractions	Estimated reaction rate ($\log(k)$)
1	nitromethane	1.000	-1.578
2	chloroform DMSO	0.093 0.907	-2.099
3	dichloromethane DMSO	0.048 0.952	-2.112
4	DMSO	1.000	-2.115
5	acetonitrile	1.000	-2.480
6	DMF pyridine	0.445 0.555	-3.199
7	pyridine	1.000	-3.216
8	dichloromethane DMF	0.137 0.863	-3.219
9	DMF	1.000	-3.223
10	dichloromethane	1.000	-3.354

reactive species which undergoes a Diels-Alder reaction with 1,1-cyclopropylcyclopenta-di-ene. The Diels-Alder adduct of this reaction is afforded in a ratio of chlorinated and fluorinated products. The reaction is shown in Fig. 4.5. As shown in the figure, we model this reaction as an equilibrium involving species **5**, **6**, **7**, and **8**. We note that the main assumption in the subsequent modeling of this reaction is that the reaction's selectivity is a function of the equilibrium between reactive intermediates rather than the kinetics of lithium halide salt formation and benzyne reacting with the diene species.

Figure 4.5: A solvent-controlled chemoselective lithiation reaction which likely proceeds via a benzyne intermediate



We make this assumption in accordance with observations from Coe et al. (2004) in their prior investigation of a reaction involving magnesium halide formation. Though the equilibrium assumption may be erroneous, we found that it fit the experimental data of Coe et al. reasonably well, and we did not see it as our place in this paper to postulate alternative mechanisms.

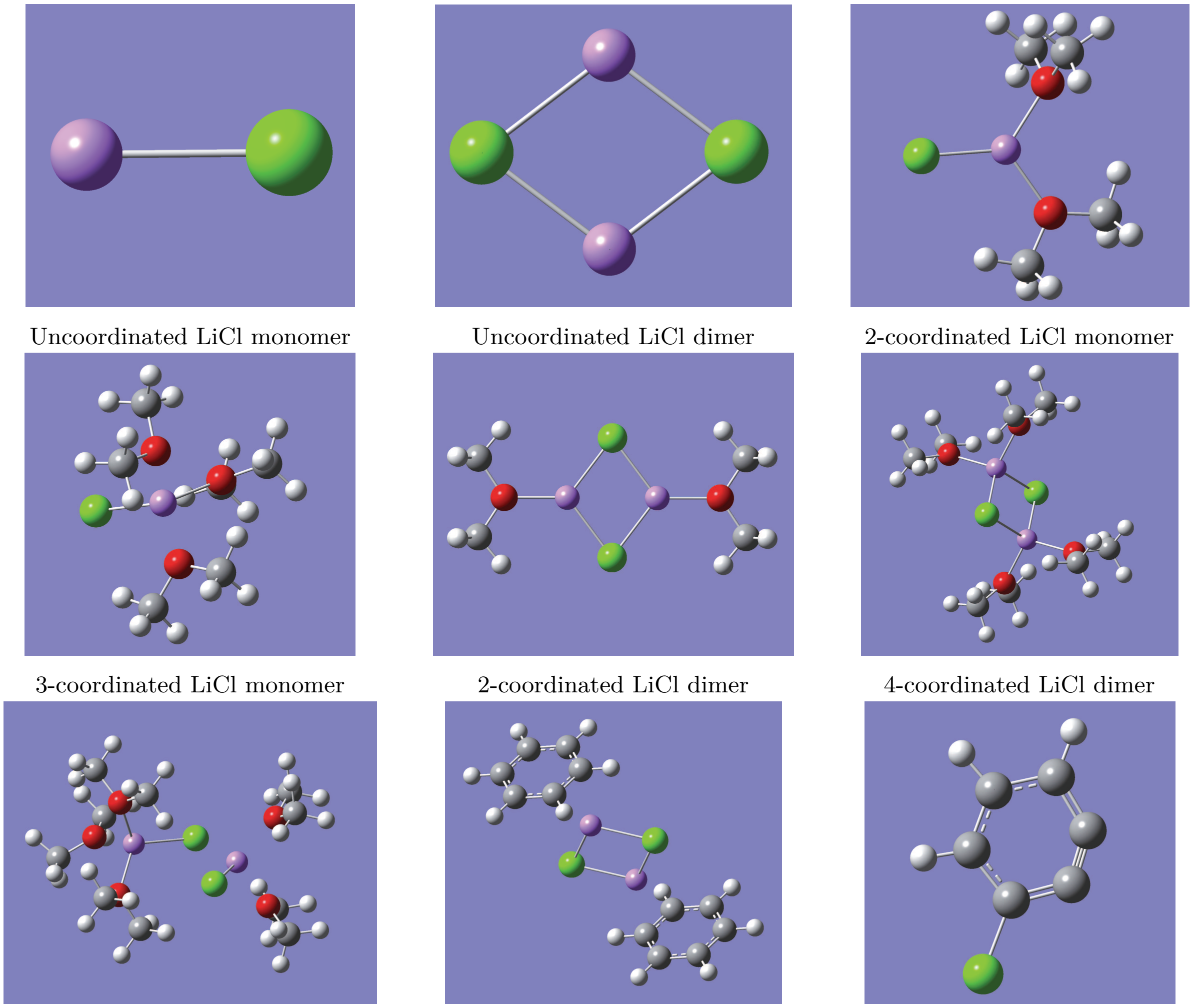
Coe et al. provide some insight into the effects on selectivity of different solvents in this reaction, writing “We suspect that solvent association with intermediates and the departing lithium halide greatly impacts the reaction course.” Their statement is in keeping with the popular belief that the chemistry of lithiations and lithium-halogen exchanges are dictated by coordinating effects of the solvent. Though this belief is grounded in ample evidence, other studies (Jedlicka et al. 1997) have found that other properties of the solvent also have some effect. For this solvent design problem, we choose to model solvent coordination explicitly at the quantum level and leave it to COSMO-RS to predict all other solvent effects on the reaction.

To model this reaction, we first optimize the geometries of the benzyne intermediates (**7** and **8**) and generate their σ profiles. It is assumed that the benzyne intermediates have consistent geometries in all solvents. The geometries of the lithium salts, however, are very likely solvent-dependent. We account for the affect of coordination on LiCl and LiF by explicitly modeling coordinate bonds with lithium atoms, using dimethyl ether as a representative solvent. Furthermore, since there is one aromatic solvent in the dataset of Coe et al. (2004), we include an explicit representation of cation-pi interactions, using benzene as a representative solvent. These σ profiles of these representative solvent systems are updated based on the procedure described above. In all, we consider 8 geometries/coordination complexes for both LiCl and LiF: (1) the linear, uncoordinated lithium halide; (2) the dimeric, uncoordinated lithium halide; (3) the 2-coordinated lithium halide monomer; (4) the 3-coordinated lithium halide monomer; (5) the 2-coordinated lithium halide dimer; (6) the 4-coordinated lithium halide dimer; (7) the 5-coordinated lithium halide dimer; (8) the lithium halide dimer coordinated with an aromatic ring. These structures, along with the chlorinated benzyne are shown in Fig. 4.6.

We note also that the electronic properties of the reactive intermediates necessitated slightly altering the basis set used. All geometries were again optimized using the B3LYP functional, but in this case study we used the 6-311+g(d,p) basis set. We include diffuse functions in this case to account for the electronegativity of fluorine and the partially ionic character of organolithium bonds.

The data set of Coe et al. (2004) contains selectivity data for the lithiation reaction shown above in seven different solvents. To estimate selectivity, we first obtain the free energies for each of the lithium salt species in the COSMO conductor phase. These free energies are then updated using COSMO-RS to reflect the Gibbs energy in each species in different solvents. Note again that we then generate the σ profiles of each solvent

Figure 4.6: Optimized geometries of various forms of LiCl considered in the lithiation selectivity case study. Note that in the problem there are analogous structures for LiF and the fluorinated benzyne.



Uncoordinated LiCl monomer

Uncoordinated LiCl dimer

2-coordinated LiCl monomer

3-coordinated LiCl monomer

2-coordinated LiCl dimer

4-coordinated LiCl dimer

5-coordinated LiCl dimer

Aromatic-cation interactions with LiCl dimer

Chlorinated benzyne

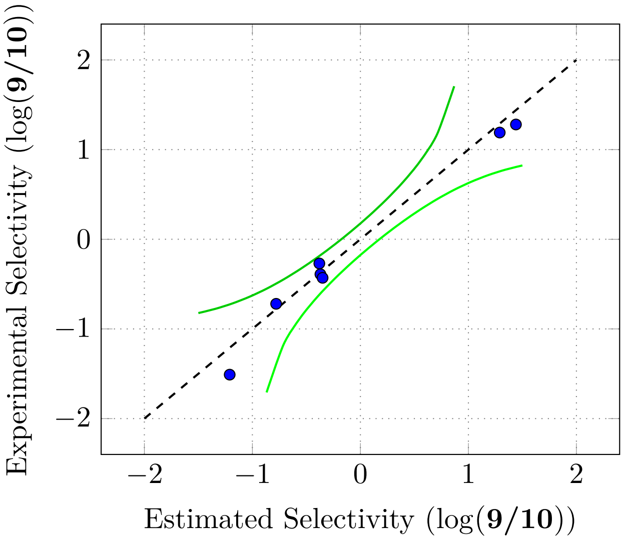
using the GC method of Mu et al. (2007) and use the quantum chemical σ profiles of the benzynes and lithium salts, updating the representative solvents to reflect the actual solvents used. Finally, the lowest energy coordination complex of each lithium salt is taken. These will be denoted G^{*}_{LiCl} and G^{*}_{LiF} for lithium chloride and lithium fluoride, respectively. The selectivity can be estimated as a ratio of the free energies:

$$S = \frac{\text{Amount of } \mathbf{9}}{\text{Amount of } \mathbf{10}} = \exp\left(-\left[(G_{\text{Cl-benzyne}} + G^{*}_{\text{LiF}}) - (G_{\text{F-benzyne}} + G^{*}_{\text{LiCl}})\right]\right) \quad (4.12)$$

where $G_{\text{F-benzyne}}$ and $G_{\text{Cl-benzyne}}$ represent the free energies in solution of the fluorinated and chlorinated benzynes, respectively. We disallow some coordination complexes for certain solvents. Most obviously, we disallow dimethyl ether-complexed salts for solvents which contain no oxygen atoms. We also disallow the aromatic-cation interactions for solvents without an aromatic ring. We incorporate steric effects into the problem by disallowing certain types of solvents to be highly coordinated with lithium. For example, sterically-hindered solvents which contain only one oxygen atom are considered unable to produce the 5-coordinated dimers, the 4-coordinated dimers, and the 3-coordinated monomers. We validate our model on the seven solvents provided by Coe et al. (2004). In Fig. 4.7, we plot the experimental selectivity and the GC-estimated selectivity. Both of these are plotted as the logarithm of the ratio of products **9** to **10**. The green lines in the figure represent 10% error in predicted amount of the products (if the ratio of products is experimentally 70:30, the green lines at this point will indicate predicted values of 80:20 and 60:40). As shown, Fig. 4.7 suggests that our model agrees well with the experimental data, attaining an R^2 of 0.98.

Next, we apply this model to design solvents to optimize the selectivity of this reaction. First, we design a solvent to maximize the ratio of **9** to **10**. Since the trend in the data set suggests more coordinating solvents increase this ratio, we suspect the solution to

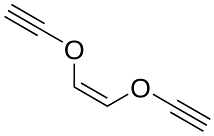
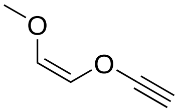
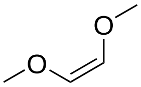
Figure 4.7: GC-estimated selectivity vs. experimental selectivity



contain a number of oxygen atoms. Furthermore, we consider all solvents with only one oxygen to be too sterically hindered to produce the 5-coordinated dimers, the 4-coordinated dimers, and the 3-coordinated monomers. Solvents without any oxygens cannot form any of these coordinated species, and solvents with two or more oxygens are assumed to be able to form all of them. We note that this is an assumption made for simplicity as classifying steric effects into different regimes is a difficult problem.

A few representative solvents for maximizing the ratio of the fluorinated product to the chlorinated product (**9** to **10**) are shown in Fig. 4.8. The best solvent found by our algorithm is shown on top. This is a diether with a double bond and two triple bonds with a predicted ratio of **9:10** of 98.2:1.8. This selectivity is higher than both the experimental and estimated selectivity reported for the best solvent in the data set of Coe et al. (2004). Furthermore, this structure is likely to be a strong coordinating

Figure 4.8: Representative structures for the lithiation selectivity case study

Solvents	Properties
	Predicted ratio **9/10** (GC-COSMO): 98.2/1.8 Predicted ratio **9/10** (QM-COSMO): 97.3/2.7
	Predicted ratio **9/10** (GC-COSMO): 98.0/2.0 Predicted ratio **9/10** (QM-COSMO): 97.5/2.5
	Predicted ratio **9/10** (GC-COSMO): 97.2/2.8 Predicted ratio **9/10** (QM-COSMO): 97.4/2.6

solvent with lithium due to the presence of two oxygens to donate electrons as well as an adjacent double bond for donating electron density via cation-pi interactions (although the cation-pi interaction was not modeled with an explicit solvent as we had no explicit solvent with oxygens and a pi bond). The other two structures shown in Fig. 4.8 would likely behave in a similar way. We again report the estimates from the QM-derived σ-profiles for each of the representative solvents. These demonstrate good agreement.

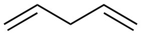
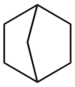
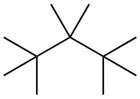
We also optimize selectivity in the opposite direction, maximizing the ratio of the chlorinated product to the fluorinated product (**10**:**9**). A representative list of results of this solvent design problem are shown in Table 4.6. As shown, no oxygen-containing solvents appear in this list as they participate coordination complexes and lead to additional fluorinated product. The first solvent shown is a diene and has the highest predicted selectivity of **10**:**9** at 3.0:97.0. However, the selectivity of this solvent is likely not as high as it is predicted to be. The accuracy of the prediction is subject to question because we do not include a representative solvent model of this system using an explicit alkene representative solvent with cation-pi interactions. We do not do this because the Coe et al. (2004) data set contains no alkene solvents, so we have no experimental data to which we could compare the estimated selectivity values. However, the other two solvents in Table 4.6 are alkane structures. Both demonstrate structural motifs found in the solution pool of high-performing solvents. Specifically, many solutions contained rings and/or a high degree of branching. These two features lead to more compact solvent structures and may have some effect on the selectivity of this reaction. We also note that many solutions were found—including the three shown in Table 4.6—which had a predicted selectivity higher than the predicted selectivity for any solvent in the experimental data set. In the experimental data set, the solvent *n*-hexane had an experimental selectivity of 3.0:90.0 (**10**:**9**) but a predicted selectivity of 5.8:94.2. The selectivity of the solvents in Table 4.6—also alkanes like *n*-hexane—may be similarly underpredicted.

Finally, we note that we do not consider mixture design for this case study. Due to the complex nature of coordination in mixed solvents, it would be challenging to correctly predict the lithium halide coordination complexes for arbitrary solvent systems. Without confidence in the correct complexes, the accuracy of our mixture design predictions would be questionable.

4.4.3 *Optimizing the selectivity of an intramolecular S_NAr reaction*

Nucleophilic aromatic substitution (S_NAr) is a well-known and useful reaction to alter the substituents on aromatic rings (Bunnett & Zahler 1951). This reaction can be applied to fairly diverse syntheses as it works for many types of arenes and het-

Table 4.6: Representative structures for the lithiation selectivity case study

Solvents	**Properties**	
	Predicted ratio **10/9** (GC-COSMO):	97.0/3.0
	Predicted ratio **10/9** (QM-COSMO):	96.8/3.2
	Predicted ratio **10/9** (GC-COSMO):	96.3/3.7
	Predicted ratio **10/9** (QM-COSMO):	84.9/15.1
	Predicted ratio **10/9** (GC-COSMO):	96.0/4.0
	Predicted ratio **10/9** (QM-COSMO):	95.0/5.0

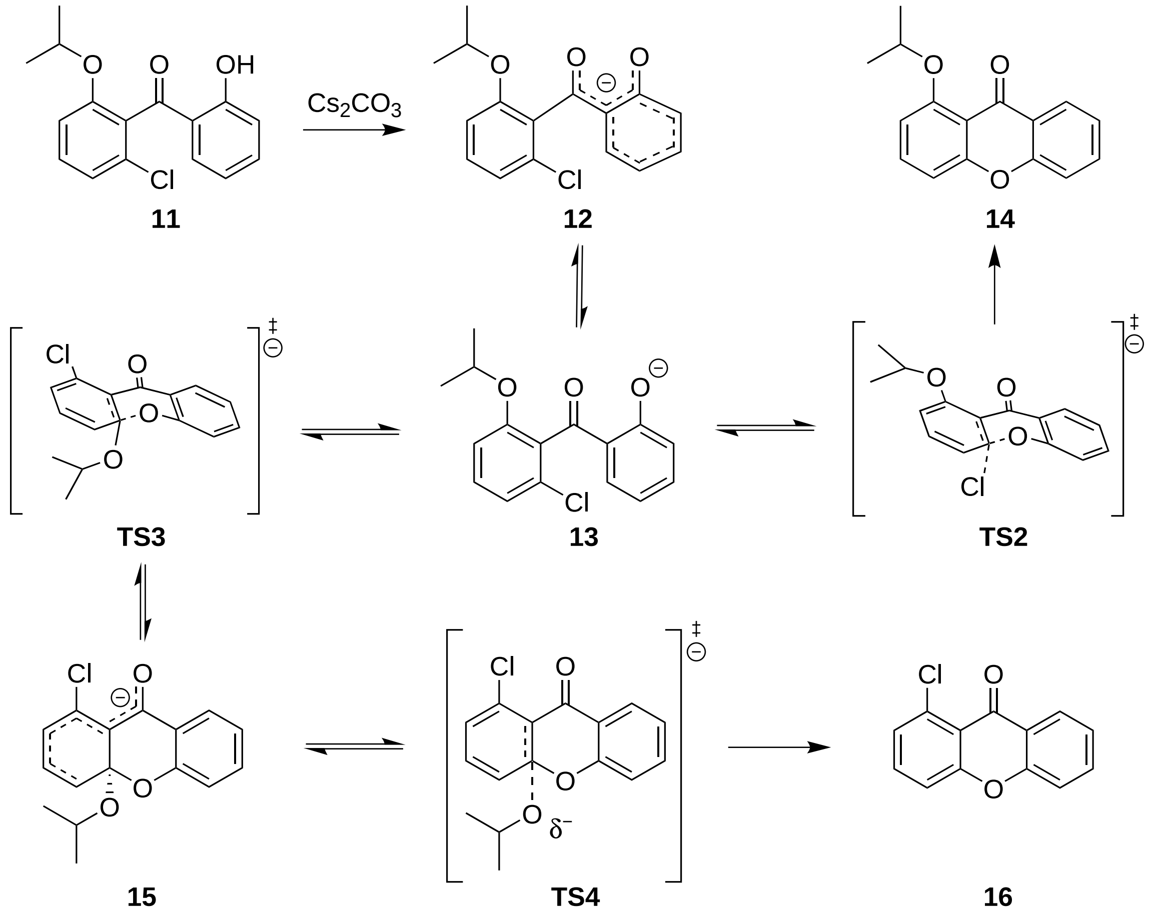
eroarenes. $S_N Ar$ is typically considered to proceed via a substitution-elimination mechanism, although other mechanisms also exist (Bunnett 1958). Overall, the reaction replaces an electron-withdrawing leaving group attached to an aromatic carbon with another substituent. The first step in the substitution-elimination mechanism involves a strong, usually negatively-charged nucleophile attacking the electrophilic aromatic carbon. This carbon—now tetravalent—becomes sp^3 hybridized, destroying the ring's aromaticity, and electron-withdrawing groups ortho- and/or para- to the sp^3 carbon stabilize the negative charge. This charge-stabilized intermediate species is known as a Meisenheimer complex. Finally, the leaving group on the sp^3 carbon dissociates from the ring, restoring aromaticity.

Although $S_N Ar$ reactions are often straightforward and lead to only one product, in some cases there are multiple suitable aromatic carbons for a nucleophilic attack. One such case comes from the study of Hintermann et al. (2008), who proposed a synthetic route towards substituted xanthones, an important class of bioactive natural products. In their study, the authors investigated an intramolecular S_NAr reaction using a reactant with two possible sites for nucleophilic attack. Interestingly, they reported that the selectivity of this attack was highly solvent-dependent. As this S_NAr reaction involves competitive pathways and a rate-determining step that is solvent-dependent, we thought it would demonstrate the ability of our solvent design method to capture more complex chemistry. Again, we address two solvent design problems to maximize the selectivity in both directions.

The specific reaction of Hintermann et al. (2008) uses a benzophenone as the reactant. The benzophenone has an alcohol group on one of its rings. This alcohol is deprotonated and the resultant alkoxide anion becomes nucleophilic. This attacks either of two positions on the neighboring aromatic ring, displacing either a chloride anion or an isopropoxide anion as a leaving group. The authors propose the usual addition-elimination mechanism for this reaction. However, in modeling the reaction at the quantum level, we did not locate the Meisenheimer complex for the case with the chloride leaving group. This resulted in a concerted S_NAr mechanism for the chloride case, a variant of the usual S_NAr mechanism which has been observed in the literature (Neumann et al. 2016). Other than this alteration, we modeled the reaction exactly as proposed by Hintermann et al. (2008). The reaction mechanism is shown in Fig. 4.9.

Again, we used the 6-311+g(d,p) basis set in order to capture the anionic species. The optimized geometries of the relevant reactive species are shown in Fig. 4.10. We estimate the selectivity of this reaction as a ratio of relevant rates of reaction. As pointed out by Hintermann et al. (2008), the solvent likely controls which of the two transition states is rate-determining in this addition-elimination mechanism. Since we found only one

Figure 4.9: A solvent-controlled chemoselective nucleophilic aromatic substitution (S_NAr) reaction



transition state for the chloride leaving group case, estimating its reaction rate constant, k_{Cl}, is a simple matter of applying TST:

$$k_{Cl} = \kappa \frac{k_B T}{h} \exp\left(-\frac{G^{\mathbf{TS2}} - G^{\mathbf{13}}}{RT}\right) \tag{4.13}$$

where now $G^{\mathbf{TS2}} - G^{\mathbf{13}}$ represents the Gibbs energy difference in the solvent phase for species **TS2** and **13**. $G^{\mathbf{TS2}}$ and $G^{\mathbf{13}}$ are the sum of the free energies in the COSMO phase and the updates to Gibbs energy given by COSMO-RS. To calculate the reaction rate constant for the isopropoxide leaving group case, k_{oipr}, we apply TST using the higher of the two transition state energies:

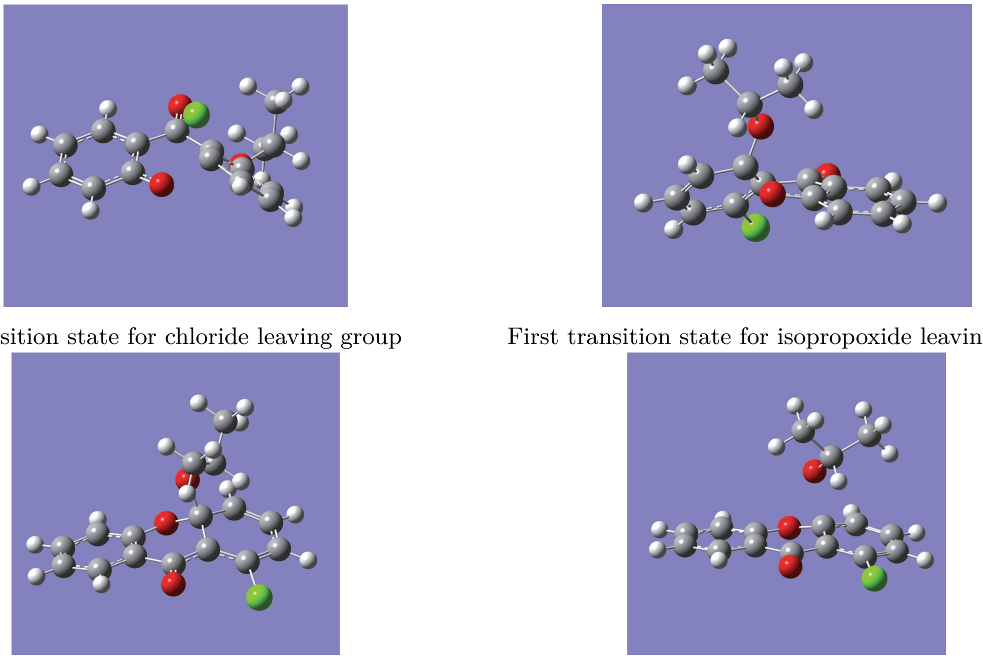
$$k_{oipr} = \begin{cases} \kappa \frac{k_B T}{h} \exp\left(-\frac{G^{\mathbf{TS3}} - G^{\mathbf{13}}}{RT}\right), & \text{if } G^{\mathbf{TS3}} \geq G^{\mathbf{TS4}} \\ \kappa \frac{k_B T}{h} \exp\left(-\frac{G^{\mathbf{TS4}} - G^{\mathbf{13}}}{RT}\right), & \text{if } G^{\mathbf{TS3}} < G^{\mathbf{TS4}} \end{cases} \tag{4.14}$$

where the G terms are defined similarly to the previous case. Finally, we can calculate the selectivity, S, as:

$$S = \frac{\text{Amount of } \mathbf{16}}{\text{Amount of } \mathbf{14}} = \frac{k_{oipr}}{k_{Cl}} \tag{4.15}$$

To validate our approach, we again compare our estimated selectivity to the experimental selectivity given for 5 solvents in Hintermann et al. (2008). This comparison is shown in Fig. 4.11. Again, the solid green lines in this graph represent a absolute deviation of 10% in either product. As shown, selectivity is predicted well, reflecting a correct determination of the rate determining step as well as the solvent properties in general. The R^2 for this parity plot is 0.97.

Figure 4.10: Optimized geometries of various forms of LiCl considered in the lithiation selectivity case study. Note that in the problem there are analogous structures for LiF and the fluorinated benzyne.



Transition state for chloride leaving group

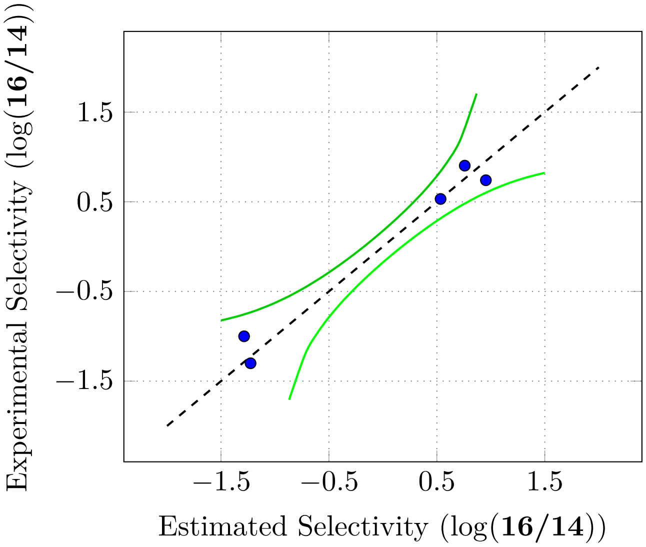
First transition state for isopropoxide leaving group

Meisenheimer complex for isopropoxide leaving group

Second transition state for isopropoxide leaving group

Next, we apply our algorithm to the design of a solvent to maximize the ratio of **14** to **16**. A few solvents which represent the structural motifs in high-performing solutions are given in Table 4.7. The first entry in this table returned the highest objective value for this problem. It is a highly polar and aprotic compound, similar to the best-performing solvents for this selectivity from the experimental solvents. Interestingly, it also appears as a representative solvent in the reaction rates optimization case study. Entry 1 on the table, along with all of the other entries, has an estimated selectivity higher than the estimated selectivity of any of the 5 solvents in the experimental data set. Finally, the last solvent on the list represents a large number of alkanes found in this optimization.

Figure 4.11: GC-estimated selectivity vs. experimental selectivity



As shown, this solvent predicts a selectivity of 91.5:8.5. This corresponds well to the selectivity of alkane solvents in Hintermann et al. (2008).

Next, we consider the inverse problem, determining solvents to maximize the ratio of **16** to **14**. We provide representative solvents for this study in Table 4.8. All of the solvents on the list are polar and protic, much like the top solvents in the experimental data. Furthermore, many solutions were found with higher predicted selectivity than any of the solvents given in the experimental data set. Three such solvents are given in Table 4.8. There were many aprotic molecules found in this solvent design problem, and these often demonstrated comparably high selectivity to proton-donating species. This observation is in keeping with the transition states suggested by Hintermann et al. (2008), where solvent polarity alone accounts for the difference in selectivity (given a particular rate-determining step). However, we postulate that proton-transfer to the

Table 4.7: Representative structures for maximizing the ratio of **14/16** in the $S_N Ar$ selectivity case study

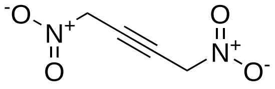
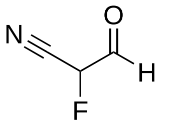
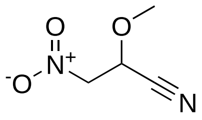
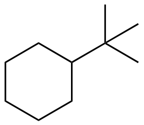
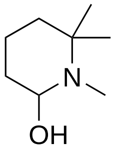
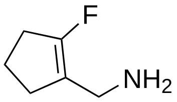
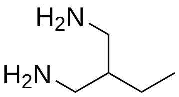
Solvents	Properties	
	Predicted ratio **14/16** (GC-COSMO):	95.0/5.0
	Predicted ratio **14/16** (QM-COSMO):	18.7/81.3
	Predicted ratio **14/16** (QM-COSMO/no H-bonding):	94.6/5.4
	Predicted ratio **14/16** (GC-COSMO):	95.0/5.0
	Predicted ratio **14/16** (QM-COSMO):	92.9/7.1
	Predicted ratio **14/16** (QM-COSMO/no H-bonding):	93.1/6.9
	Predicted ratio **14/16** (GC-COSMO):	94.2/5.8
	Predicted ratio **14/16** (QM-COSMO):	11.9/88.1
	Predicted ratio **14/16** (QM-COSMO/no H-bonding):	93.5/6.5
	Predicted ratio **14/16** (GC-COSMO):	91.5/8.5
	Predicted ratio **14/16** (QM-COSMO):	92.2/7.8
	Predicted ratio **14/16** (QM-COSMO/no H-bonding):	92.2/7.8

Table 4.8: Representative structures for maximizing the ratio of **16/14** in the S_NAr selectivity case study

Solvents	Properties	

	Predicted ratio **16/14** (GC-COSMO):
	Predicted ratio **16/14** (QM-COSMO):
	Predicted ratio **16/14** (QM-COSMO/no H-bonding):
		91.8/8.2
		95.6/4.4
		95.6/4.4

	Predicted ratio **16/14** (GC-COSMO):
	Predicted ratio **16/14** (QM-COSMO):
	Predicted ratio **16/14** (QM-COSMO/no H-bonding):
		90.6/9.4
		95.8/4.2
		95.8/4.2

	Predicted ratio **16/14** (GC-COSMO):
	Predicted ratio **16/14** (QM-COSMO):
	Predicted ratio **16/14** (QM-COSMO/no H-bonding):
		89.9/10.1
		95.7/4.3
		95.7/4.3

isopropoxide leaving group may have a pronounced effect on the kinetics. This effect was not modeled in our problem as we attempted to keep our reaction mechanism as close to the Hintermann et al. (2008) study as possible. Again, we did not see it as our place in this paper to pursue an alternative mechanism, especially given the paucity of experimental data. However, this underscores the importance of modeling the correct mechanism. Had we included proton transfer as a mechanistic step (or simply modeled the equilibrium), many aprotic solvents may have been disfavored in the solution pool. Additionally, we may have found solvents with low-pK_a protons as high-performing solutions.

Next, we perform a mixture design problem to maximize both selectivity ratios at 50° C. We begin with the ratio of **14/16**. Again, we consider a mixture of up to four of a fixed set of common laboratory and industrial solvents. The top ten solutions to this mixture design problem are given in Table 4.9. Interestingly, all of the solvents selected are pure, single-component solvents. This makes sense as this selectivity is usually favored if **TS4** is the rate-determining step. All of the solvents listed here are polar, aprotic species, which would tend to stabilize **TS3**. We note that we do not observe a large range of selectivity. However, polar, aprotic solvents demonstrated fairly extreme selectivity in the experimental data set, so this small range of selectivity is perhaps not surprising. It is possible that scaling the energies or using different COSMO-RS parameters would lead to a wider range of selectivities. We did not pursue this as it would be difficult to justify one set of parameters over another with such limited data.

Our last mixture design problem will determine a solvent to maximize the ratio of **16/14** at 100° C. The results are shown in Table 4.10. There were many solutions found with a selectivity of 96/4. In an effort to communicate more diverse solutions, we only list the top 5 mixtures as the first 5 entries of Table 4.10. We note that these solutions always contain cyclohexane and isopropanol. There may be some merit to this mixture for this particular problem. Pure cyclohexane, a non-polar solvent, is a very poor solvent for maximizing this ratio, having a predicted selectivity of 20.7/79.3. In this case, **TS4** is the predicted rate-limiting step for cyclohexane. However, cyclohexane is predicted to stabilize TS3 over **TS2**. This means that, if the rate-determining step is shifted to TS3, cyclohexane would prefer the desired selectivity. Adding isopropanol, a polar, protic solvent, to the mixture seems to have this effect. Isopropanol stabilizes **TS4** and leads to high selectivity. The first five mixtures listed in Table 4.10 all have a very high predicted selectivity, higher than any solvent in the experimental data set.

Table 4.9: Top solvent blends using common solvents for maximizing the ratio of **14/16** at 50° C in the $S_N Ar$ selectivity case study

Solution rank	**Solvents**	**Mole fractions**	**Estimated ratio of 14 to 16**
1	acetonitrile	1.000	95.4/4.6
2	dioxane	1.000	94.5/5.5
3	DMSO	1.000	94.3/3.7
4	acetone	1.000	94.1/5.9
5	DMF	1.000	94.1/5.9
6	diglyme	1.000	94.0/6.0
7	ethyl acetate	1.000	94.0/6.0
8	dimethyl ether	1.000	93.9/6.1
9	DMPU	1.000	93.8/6.2
10	pyridine	1.000	92.9/7.1

The next five entries represent the solution to the problem with increasingly stringent conditions for the solvent mixture. For the first restriction, we consider solvent mixtures without cyclohexane or *n*-hexane (entry 6). Again, we observe a solution of non-polar solvents combined with a polar, protic solvent. These diverse solvents, combined in the right ratio, are predicted to promote isopropoxide leaving over chloride leaving. The remaining solvent mixtures in Table 4.10 demonstrate the same trend, even with placing increasingly strict limitations on the design space. We note again that all of these solutions are dependent on modeling the correct mechanism and modeling every potential rate-determining step.

Table 4.10: Top solvent blends using common solvents for maximizing the ratio of **16/14** at 100° C in the S_NAr selectivity case study

Solution rank	**Solvents**	**Mole fractions**	**Estimated ratio of 16/14**	**Solvent restrictions**
1	cyclohexane *n*-hexane isopropanol	0.758 0.222 0.020	96.1/3.9	-
2	cyclohexane isopropanol diethyl ether	0.958 0.022 0.020	96.1/3.9	-
3	cyclohexane isopropanol dimethyl ether	0.958 0.022 0.020	96.1/3.9	-
4	cyclohexane isopropanol dioxane	0.956 0.024 0.020	96.0/4.0	-
5	cyclohexane isopropanol diglyme	0.956 0.024 0.020	96.0/4.0	-
-	diethyl ether carbon tetrachloride 2,2,2-trifluoroethanol	0.844 0.136 0.020	95.6/4.4	no cyclohexane or *n*-hexane
-	carbon tetrachloride DMPU 2,2,2-trifluoroethanol ethanol	0.645 0.315 0.020 0.020	95.4/4.6	no ethers
-	toluene ethyl acetate methanol benzene	0.854 0.106 0.020 0.020	94.9/5.1	no chlorine-containing solvents
-	DMPU 2,2,2-trifluoroethanol DMF DMSO	0.880 0.080 0.020 0.020	94.9/5.1	no aromatics
-	ethyl acetate 2,2,2-trifluoroethanol nitromethane	0.959 0.021 0.020	94.5/5.5	no DMPU, DMF, or DMSO

4.5 CONCLUSIONS

In this chapter, we provided three additions to our previously-developed (Austin et al. 2016a) COSMO-based mixture design algorithm. First, we divided σ profiles into separate H-bonding and non-H-bonding profiles. Next, we introduced a method to include explicitly-modeled intermolecular interactions between solute and solvent in molecular/mixture design problems. Finally, we re-framed the mixture design problem as a mixture selection problem with a list of common laboratory and industrial solvents. These three additions serve to increase the accuracy of our predictions, allow for more complex species to be considered, and restrict the mixture design problem to producing practical, readily-implementable solutions.

Directly incorporating quantum mechanics calculations into our design problems, we applied our algorithm to three solvent design problems: (1) maximizing the reaction rate of a Menschutkin reaction; (2) maximizing the selectivity of a lithiation reaction; and (3) maximizing the selectivity of an intramolecular nucleophilic aromatic substitution reaction. The second case study modeled multiple coordination complexes of lithium salts and choose appropriate species based on energy. Using this approach, we observed a very good agreement with experimental data and were able to find improved solvents which performed better than any solvents in the experimental data. The third case study modeled a reaction with multiple competitive pathways and two possible rate-determining steps. Insofar as we observed good agreement with experimental data, our approach chose the appropriate rate determining step. We were also able to design single-component and mixed solvents which had higher predicted selectivities than any solvent in the experimental data.

To our knowledge, the complexity of the systems considered in case studies two and three is unprecedented in CAMD-based reactions solvent design. The successful modeling of such complex systems using COSMO-based methods lends credence to the incorporation of quantum mechanics calculations in CAMD. Using these techniques, arbitrary chemical systems can be considered at the level of quantum mechanical accuracy, making possible a wide diversity of solvent design problems which were previously untenable.

5

CONCLUSIONS AND FUTURE WORK

5.1 CONCLUSIONS AND CONTRIBUTIONS MADE

This work has detailed methods to overcome several limitations in CAMD and expanded the range of problems which can be addressed using CAMD techniques. We first proposed a projection and decomposition technique to solve mixture design problems as well as difficult single-molecule design problems (e.g. a molecular design problem as part of a larger process optimization). This technique requires identifying a lower-dimensional space of individual component properties for every component in the to-be-designed mixture. This lower-dimensional property space serves as the search space for the optimization problem and is expected to have some relevance to the higher-dimensional objective. We note here that the selection of the particular variables in this lower-dimensional property space represents one of the most critical decisions to be made when applying the algorithm. The search through this low-dimensional space is then done using a combination of Derivative-Free Optimization (DFO) techniques to guide the search and efficient MILP optimization problems to relate the projected variables to the space of chemical structures.

We have applied this algorithm throughout this document and have found better solutions than previously reported for all of the case studies we considered. This success was largely the result of our algorithm's ability to efficiently optimize over an extensive region of the chemical design space. Other approaches—typically being limited to a design space with 10-20 groups—have been unable to consider such a large search space and as a result have not been able to find solutions of the quality we report. The high-performing solutions we have obtained confirm the utility of our decomposition-and-projection approach to mixture design. Furthermore, these results demonstrate that the lower-dimensional projected property spaces do contain meaningful information for addressing these complex problems.

Apart from the issue of search space, many mixture design problems also face difficulties in incorporating accurate mixture thermodynamics models. These models are necessary considerations for a straightforward reason: without a very accurate picture of the behavior of designed components in a mixture, very little information can be gleaned from these design problems. UNIFAC (Fredenslund et al. 1975) has been applied extensively as the mixture thermodynamics model used in these problems. However, UNIFAC requires binary interaction parameters for every pair of groups which occur in the mixture in order to make in estimate for activity coefficients. These binary interaction parameters are difficult to estimate for many groups, meaning that many binary interaction parameters simply do not exist. The lack of certain binary interaction parameters can have severe consequences for mixture design problems: problems using UNIFAC are inherently limited to a design space defined by the available binary interaction parameters.

To address this challenge, we have incorporated COSMO-RS- and -SAC-based thermodynamics into mixture design problems. These methods do not require binary interaction parameters and are not restricted to certain design spaces as a result. Using these methods as the thermodynamics underpinning of mixture design problems, we can estimate relevant mixture properties for any mixture so long as we can estimate σ-profiles and molecular volumes for every *individual* component of the mixture. We discovered that σ-profiles and molecular volumes can be accurately estimated using group contribution methods, and this allowed for the complete incorporation of these COSMO-based methods into our mixture design framework. Applying COSMO-based thermodynamics in a few case studies, we obtained better solutions than had previously been reported. In many cases, the solutions generated using COSMO-based thermodynamics were simply unattainable with UNIFAC-based methods as they would have lacked a necessary binary interaction parameter.

Finally, because COSMO-RS and -SAC are methods which are based on quantum chemistry calculations, molecular and mixture design problems using COSMO methods are able to address a much wider set of problems and applications. The simple reason for this is that σ-profiles are typically derived from a molecular structure after a full-fledged geometry optimization at the quantum level. Because the σ-profiles are used for every component in a mixture design problem, we are free to include quantum-level information about any molecular species in molecular/mixture design problems. We note for clarity that quantum-level information is typically only included for species which are fixed in the design problem. The σ-profiles of molecules in the design space are estimated using accurate lower-order methods. Nonetheless, this accessibility of QM information about molecular structures means that a much larger variety of chemical species (e.g., transition states, ionic liquids, radicals, etc.) can now be considered directly in CAMD problems. This extends CAMD applications to many new domains, and we have demonstrated the utility of this approach for one of these new domains: reaction solvent design.

In summary, this work encompasses a new technique to overcome the difficulty of optimizing over the massive chemical design space. We have demonstrated that this algorithm can find superior solutions and is very well-suited to the mixture design problem. Furthermore, we have demonstrated a way to directly incorporate quantum-level information into molecular and mixture design problems. These contributions have allowed for much larger problems to be considered and have greatly expanded the range of problems which are tenable using CAMD.

5.2 FUTURE WORK

5.2.1 *Determining lower-dimensional search spaces using principal components analysis*

In its current incarnation, our DFO-based mixture design algorithm uses a space of pure component properties as its projected search space. As discussed, this search space is novel in CAMD and has provided a meaningful variable space with which to capture the behavior of many CAMD problems. One possible extension to the current framework would be to use functions of some of these properties (or σ moments, group occurrences, etc.) as descriptors. Though many design problems can be modeled effectively using our current approach, there may be some problems in which individual component properties may not be the descriptors with the most information. Using principal component analysis (PCA) would allow us to consider a large number of descriptors at once and at the same time keep the dimensionality of the search space for DFO solvers small. Furthermore, using this technique would remove the subjectivity associated with selecting the descriptors for a given problem. One potential complication is the need for a large dataset of molecules and corresponding objective function values. For many problems, this could be generated without too much difficulty. PCA may not be well-suited for problems requiring difficult simulations or other computationally expensive steps. Partial least squares (PLS) could be similarly applied to these problems.

5.2.2 *Extension to integrated product/process design*

One of the main challenges of solving integrated process/product design problems is that the process variables and objective are often very sensitive to the choice of designed molecule(s). Additionally, even when the product design variables are fixed (i.e., a particular molecule(s) is selected), the resultant process design problem can still be very difficult to solve. Our mixture design algorithm—originally conceived to solve difficult mole fractions subproblems—would be well-suited to the task of exploring a large molecular search space even while solving difficult process optimization subproblems. The efficiency and available design space of our DFO-based algorithm would likely result in better solutions to many of these product/process design problems. The applications are numerous: CO_2 capture, organic Rankine cycle design, crystallization solvent/process design, etc. Again, we would expect the projected design space to be properties of the designed component or some reduced-dimensionality space resulting from PCA/PLS.

5.2.3 *Modeling more complex reactions in reaction solvent design problems*

The reactions provided as case studies in this document are relatively simple compared to some reactions used in laboratories/industry. The solvent design algorithm could be extended to more complex case studies including polymerization reaction solvents or reactions relevant to the pharmaceutical industry. An interesting case study may also be to design a mixed solvent for optimizing battery performance.

5.2.4 *Designing custom ionic liquids*

Ionic liquids are becoming increasingly common in industrial applications. Requiring both a cation and an anion, ionic liquids are by nature mixtures of multiple components. For this reason, the same combinatorial difficulties which exist in typical mixture design problems also make these ionic liquids design problems difficult. Fortunately, our mixture design algorithm is able to efficiently explore a large space of multiple chemical components. Additionally, COSMO-based methods have been applied extensively for modeling ionic liquids. These considerations make ionic liquids design a natural evolution of our current implementation. There are many applications in gas purification, nuclear waste treatment, battery solvents, and biological reactions solvents. Furthermore, ionic liquids have potential to be fully reusable solvents, and a hypothetical design problem considering reusability could also optimize some industrial process while simultaneously determining the optimal ionic liquid.

5.2.5 *Small molecule design for the pharmaceutical industry*

Pharmaceutical drug design represents a unique opportunity for CAMD. The pharmaceutical design space is massive and many drug compounds are complex structures, requiring many steps to synthesize. We have discussed algorithms which are capable of exploring large design spaces with efficiency, meaning that the size of the pharmaceutical design space should not be too daunting. Additionally, pharmaceutical companies are distinguished from commodity chemical companies in that they often invest a large amount of financial resources in the development of a certain product. This means that solutions determined by CAMD approaches are less likely to be discounted on the basis of expense and/or complexity and that solutions to these CAMD problems have the potential to be implemented on the industrial scale. A natural extension of the work described in this document would be to propose a low-dimensional design space related to some pharmaceutical property (bioactivity, toxicity, enzyme binding affinity, etc.) or a function of these properties. As the molecules are successively designed, each can be outsourced to a subproblem whereby these properties can be estimated. These subproblems can be difficult (i.e., quantum chemistry calculations, binding energy calculations, etc.), but the use of DFO may have a significant impact on the number of iterations required for this design procedure. A DFO-based design procedure could effectively explore the large search space and converge on a good solution which could have been difficult to find without such an optimization procedure.

5.2.6 *Custom group contribution methods for modeling*

The applicability of CAMD is limited by the quality of QSPR models used to estimate properties. Traditionally, many group contribution methods have been made using a fairly consistent set of groups as descriptors (the UNIFAC groups). More accurate models could be obtained if the descriptor space were customized for every property of interest. Making these models requires optimizing over the parameter space as well as a discrete space of all possible molecular groups. Though this particular problem is not directly related to anything discussed in this document, it is of general applicability to CAMD and would improve the quality of all of the solutions obtained using the algorithms discussed as well as the efficiency and reliability of the DFO-based search procedure. This group optimization problem is currently under investigation.

BIBLIOGRAPHY

Abraham, M. H. (1971). Substitution at saturated carbon. Part VIII. Solvent effects on the free energy of trimethylamine, the nitrobenzyl chlorides, and the trimethylamine–nitrobenzyl chloride transition states. *Journal of the Chemical Society B: Physical Organic*, 299–308.

Abraham, M. H., Doherty, R. M., Kamlet, M. J., Harris, J. M., & Taft, R. W. (1987). Linear solvation energy relationships. Part 37. An analysis of contributions of dipolarity–polarisability, nucleophilic assistance, electrophilic assistance, and cavity terms to solvent effects on t-butyl halide solvolysis rates. *Journal of the Chemical Society, Perkin Transactions 2*, 913–920.

Abrams, D. S. & Prausnitz, J. M. (1975). Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems. *AIChE Journal*, *21*, 116–128.

Abramson, M. A., Audet, C., Couture, G., Dennis, Jr., J. E., & Le Digabel, S. (2017). The NOMAD project. `http://www.gerad.ca/nomad/`.

American Chemical Society (2017). CAS Registry. Available at `https://www.cas.org/content/chemical-substances`.

Ashbya, M. F., Bréchet, Y. J. M., Cebona, D., & Salvoc, L. (2004). Selection strategies for materials and processes. *Materials and Design*, *25*, 51–67.

Audet, C. & Dennis Jr., J. E. (2006). Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on Optimization*, *17*, 188–217.

Austin, N. D. & Sahinidis, N. V. (2016). Sigma Profile Group Contribution Model. Available at `http://archimedes.cheme.cmu.edu/?q=sigmaprof`.

Austin, N. D., Sahinidis, N. V., & Trahan, D. W. (2016a). A COSMO-based approach to computer-aided mixture design. *Chemical Engineering Science*. DOI 10.1016/j.ces.2016.05.025.

Austin, N. D., Sahinidis, N. V., & Trahan, D. W. (2016b). Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques. *Chemical Engineering Research and Design*, *116*, 2–26.

Austin, N. D., Samudra, A. P., Sahinidis, N. V., & Trahan, D. W. (2016). Mixture design using derivative-free optimization in the space of individual component properties. *AIChE Journal*, *62*, 1514–1530.

Bajaj, S., Sambi, S. S., & Madan, A. K. (2005). Prediction of anti-inflammatory activity of N-arylanthranilic acids: Computational approach using refined Zagreb indices. *Croatica Chemica Acta*, *78*, 165–174.

Balas, E. (1979). Disjunctive programming. *Annals of Discrete Mathematics*, *5*, 3–51.

Balslev, K. & Abildskov, J. (2002). UNIFAC parameters for four new groups. *Industrial & Engineering Chemistry Research*, *41*, 2047–2057.

Bardow, A., Steur, K., & Gross, J. (2010). Continuous-molecular targeting for integrated solvent and process design. *Industrial & Engineering Chemistry Research*, *49*, 2834–2840.

Barone, V. & Cossi, M. (1998). Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *The Journal of Physical Chemistry A*, *102*, 1995–2001.

Basak, S. C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach. *Medical Science Research*, *15*, 605–609.

Basak, S. C., Gieschen, D. P., & Magnuson, V. R. (1984). A quantitative correlation of the LC50 values of esters in Pimephales promelas using physicochemical and topological parameters. *Environmental Toxicology and Chemistry*, *3*, 191–199.

Becke, A. D. (1993). Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, *98*, 5648–5652.

Benavides, P. T., Gebreslassie, B. H., & Diwekar, U. M. (2015). Optimal design of adsorbents for NORM removal from produced water in natural gas fracking. Part 2: CAMD for adsorption of radium and barium. *Chemical Engineering Science*, *137*, 977–985.

Benson, S. W. (1999). New methods for estimating the heats of formation, heat capacities, and entropies of liquids and gases. *The Journal of Physical Chemistry A*, *103*, 11481–11485.

Benson, S. W. & Buss, J. H. (1958). Additivity rules for the estimation of molecular properties. Thermodynamic properties. *The Journal of Chemical Physics*, *29*, 546–572.

Benson, S. W., Cruickshank, F. R., Golden, D. M., Haugen, G. R., O'neal, H. E., Rodgers, A. S., Shaw, R., & Walsh, R. (1969). Additivity rules for the estimation of thermochemical properties. *Chemical Reviews*, *69*, 279–324.

Bernard, G., Hocine, R., & Lupis, C. H. P. (1967). Thermodynamic conditions for spinodal decomposition in a multicomponent system. *Transactions of the Metallurgical Society of AIME*, *239*, 1600.

Boethling, R. S. (1986). Application of molecular topology to quantitative structure-biodegradability relationships. *Environmental Toxicology and Chemistry*, *5*, 797–806.

Bohacek, R. S., McMartin, C., & Guida, W. C. (1996). The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal research reviews*, *16*, 3–50.

Bommareddy, S., Chemmangattuvalappil, N. G., Solvason, C. C., & Eden, M. R. (2010). Simultaneous solution of process and molecular design problems using an algebraic approach. *Computers & Chemical Engineering*, *34*, 1481–1486.

Bonchev, D. (1991). *Chemical graph theory: Introduction and fundamentals*, volume 1. CRC Press.

Boneh, A. & Golan, A. (1979). Constraints' redundancy and feasible region boundedness by random feasible point generator (RFPG). In *Third European Congress on Operations Research (EURO III)*, Amsterdam.

Booker, A. J., Dennis Jr., J., Frank, P. D., Serafini, D. B., Torczon, V. J., & Trosset, M. W. (1999). A rigorous framework for optimization of expensive functions by surrogates. *Structural Optimization*, *17*, 1–13.

Brent, R. P. (1973). *Algorithms for Minimization without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall.

Brignole, E. A., Bottini, S. B., & Gani, R. (1986). A strategy for the design and selection of solvents for separation processes. *Fluid Phase Equilibria*, *29*, 125–132.

Brown, W. M., Martin, S., Rintoul, M. D., & Faulon, J.-L. (2006). Designing novel polymers with targeted properties using the signature molecular descriptor. *Journal of chemical information and modeling*, *46*, 826–835.

Bunnett, J. F. (1958). Mechanism and reactivity in aromatic nucleophilic substitution reactions. *12*, 1–16.

Bunnett, J. F. & Zahler, R. E. (1951). Aromatic nucleophilic substitution reactions. *49*, 273–412.

Burger, J., Papaioannou, V., Gopinath, S., Jackson, G., Galindo, A., & Adjiman, C. S. (2015). A hierarchical method to integrated solvent and process design of physical CO2 absorption using the SAFT-γ Mie approach. *AIChE Journal*, *61*, 3249–3269.

Buxton, A., Livingston, A. G., & Pistikopoulos, E. N. (1999). Optimal design of solvent blends for environmental impact minimization. *AIChE Journal*, *45*, 817–843.

Camarda, K. V. & Maranas, C. D. (1999). Optimization in polymer design using connectivity indices. *Industrial & Engineering Chemistry Research*, *38*, 1884–1892.

Camarda, K. V. & Sunderesan, P. (2005). An optimization approach to the design of value-added soybean oil products. *Industrial & Engineering Chemistry Research*, *44*, 4361–4367.

Cao, W., Knudsen, K., Fredenslund, A., & Rasmussen, P. (1993). Group-contribution viscosity predictions of liquid mixtures using UNIFAC-VLE parameters. *Industrial & engineering chemistry research*, *32*, 2088–2092.

Ceriani, R., Gani, R., & Meirelles, A. J. A. (2009). Prediction of heat capacities and heats of vaporization of organic liquids by group contribution methods. *Fluid Phase Equilibria*, *283*, 49–55.

Ceriani, R., Gonçalves, C. B., Rabelo, J., Caruso, M., Cunha, A. C. C., Cavaleri, F. W., Batista, E. A. C., & Meirelles, A. J. A. (2007). Group contribution model for predicting viscosity of fatty compounds. *Journal of Chemical & Engineering Data*, *52*, 965–972.

Chapman, W. G., Gubbins, K. E., Jackson, G., & Radosz, M. (1989). SAFT: Equation-of-state solution model for associating fluids. *Fluid Phase Equilibria*, *52*, 31–38.

Chavali, S., Lin, B., Miller, D. C., & Camarda, K. V. (2004). Environmentally-benign transition metal catalyst design using optimization techniques. *Computers & Chemical Engineering*, *28*, 605–611.

Chemmangattuvalappil, N. G., Solvason, C. C., Bommareddy, S., & Eden, M. R. (2010). Reverse problem formulation approach to molecular design using property operators based on signature descriptors. *Computers & Chemical Engineering*, *34*, 2062–2071.

Churchwell, C. J., Rintoul, M. D., Martin, S., Visco, D. P., Kotu, A., Larson, R. S., Sillerud, L. O., Brown, D. C., & Faulon, J.-L. (2004). The signature molecular descriptor: 3. Inverse-quantitative structure–activity relationship of ICAM-1 inhibitory peptides. *Journal of Molecular Graphics and Modelling*, *22*, 263–273.

Churi, N. & Achenie, L. E. K. (1996). Novel mathematical programming model for computer aided molecular design. *Industrial & Engineering Chemistry Research*, *35*, 3788–3794.

Coe, J. W., Wirtz, M. C., Bashore, C. G., & Candler, J. (2004). Formation of 3-halobenzyne: Solvent effects and cycloaddition adducts. *6*, 1589–1592.

Cohen, N. & Benson, S. W. (1993). Estimation of heats of formation of organic compounds by additivity methods. *Chemical Reviews*, *93*, 2419–2438.

Conn, A. R., Scheinberg, K., & Toint, P. L. (1997). Recent progress in unconstrained nonlinear optimization without derivatives. *Mathematical Programming*, *79*, 397–414.

Conn, A. R., Scheinberg, K., & Vicente, L. N. (2009). *Introduction to derivative-free optimization*. Philadelphia, PA: SIAM.

Constantinou, L. & Gani, R. (1994). New group contribution method for estimating properties of pure compounds. *AIChE Journal*, *40*, 1697–1710.

Conte, E., Gani, R., Cheng, Y. S., & Ng, K. M. (2012). Design of Formulated Products: Experimental Component. *AIChE Journal*, *58*, 173–189.

Conte, E., Gani, R., & Ng, K. M. (2011a). Design of formulated products: A systematic methodology. *AIChE Journal*, *57*, 2431–2449.

Conte, E., Gani, R., & Ng, K. M. (2011b). Design of formulated products: A systematic methodology. *AIChE Journal*, *57*, 2431–2449.

Csendes, T., Pál, L., Sendín, J. O. H., & Banga, J. R. (2008). The GLOBAL optimization method revisited. *Optimization Letters*, *2*, 445–454.

Cussler, E. L. & Moggridge, G. D. (2011). *Chemical Product Design* (2nd ed.). Cambridge University Press.

Custódio, A. L. & Vicente, L. N. (2008). *SID-PSM: A pattern search method guided by simplex derivatives for use in derivative-free optimization.* Coimbra, Portugal: Departamento de Matemática, Universidade de Coimbra.

Desiraju, G. R. & Steiner, T. (2001). *The weak hydrogen bond in structural chemistry and biology*, volume 9. Oxford University Press on Demand.

Devillers, J. & Balaban, A. T. (2000). *Topological indices and related descriptors in QSAR and QSPAR.* CRC Press.

Domalski, E. S. & Hearing, E. D. (1988). Estimation of the thermodynamic properties of hydrocarbons at 298.15 K. *Journal of Physical and Chemical Reference Data*, *17*, 1637–1678.

Dortmund Data Bank Software and Separation Technology GmbH (2014). Parameters of the original UNIFAC model. `http://www.ddbst.com/published-parameters-unifac.html`. Accessed: 2016-02-14.

Duran, M. A. & Grossmann, I. E. (1986). An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical Programming*, *36*, 307–339.

Duvedi, A. P. & Achenie, L. E. K. (1996). Designing environmentally safe refrigerants using mathematical programming. *Chemical Engineering Science*, *51*, 3727–3739.

Duvedi, A. P. & Achenie, L. E. K. (1997). On the design of environmentally benign refrigerant mixtures. A mathematical programming approach. *Computers & Chemical Engineering*, *21*, 915–923.

Dyk, B. V. & Nieuwoudt, I. (2000). Design of solvents for extractive distillation. *Industrial & Engineering Chemistry Research*, *39*, 1423–1429.

Eberhart, R. & Kennedy, J. (1995). A new optimizer using particle swarm theory. In *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, (pp. 39–43)., Nagoya, Japan.

Eckert, F. & Klamt, A. (2002). Fast solvent screening via quantum chemistry: COSMO-RS approach. *AIChE Journal*, *48*, 369–385.

Eden, M. R., Jørgensen, S. B., Gani, R., & El-Halwagi, M. M. (2004). A novel framework for simultaneous separation process and product design. *Chemical Engineering and Processing: Process Intensification*, *43*, 595–608.

Eslick, J. C., Ye, Q., Park, J., Topp, E. M., Spencer, P., & Camarda, K. V. (2009). A computational molecular design framework for crosslinked polymer networks. *Computers & Chemical Engineering*, *33*, 954–963.

Estrada, E. (1995). Edge adjacency relationships and a novel topological index related to molecular volume. *Journal of Chemical Information and Computer Sciences*, *35*, 31–33.

Estrada, E. & Rodríguez, L. (1999). Edge-connectivity indices in QSPR/QSAR studies. 1. Comparison to other topological indices in QSPR studies. *Journal of Chemical Information and Computer Sciences*, *39*, 1037–1041.

Estrada, E. & Uriarte, E. (2001). Recent advances on the role of topological indices in drug discovery research. *Current Medicinal Chemistry*, *8*, 1573–1588.

Faulon, J.-L., Churchwell, C. J., & Visco, D. P. (2003). The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *Journal of Chemical Information and Computer Sciences*, *43*, 721–734.

Faulon, J.-L., Visco, D. P., & Pophale, R. S. (2003). The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *Journal of Chemical Information and Computer Sciences*, *43*, 707–720.

Filippa, M. A. & Gasull, E. I. (2013). Ibuprofen solubility in pure organic solvents and aqueous mixtures of cosolvents: Interactions and thermodynamic parameters relating to the solvation process. *Fluid Phase Equilibria*, *354*, 185–190.

Folić, M., Adjiman, C. S., & Pistikopoulos, E. N. (2007). Design of solvents for optimal reaction rate constants. *AIChE Journal*, *53*, 1240–1256.

Folic, M., Adjiman, C. S., & Pistikopoulos, E. N. (2008). Computer-aided solvent design for reactions: Maximizing product formation. *Industrial & Engineering Chemistry Research*, *47*, 5190–5202.

Fredenslund, A., Jones, R. L., & Prausnitz, J. M. (1975). Group-contribution estimates of activity coefficients in nonideal liquid mixtures. *AIChE Journal*, *21*, 1086.

Frisch, M. J. & et al. (2009). Gaussian 09 Revision D.01. Gaussian Inc., Wallingford CT.

Galvez, J., Garcia, R., Salabert, M. T., & Soler, R. (1994). Charge indexes. New topological descriptors. *Journal of Chemical Information and Computer Sciences*, *34*, 520–525.

Gani, R. & Brignole, E. (1983). Molecular design of solvents for liquid extraction based on UNIFAC. *Fluid Phase Equilibria*, *13*, 331–340.

Gani, R. & Fredenslund, A. (1993). Computer-aided molecular and mixture design with specified constraints. *Fluid Phase Equilibria*, *82*, 39–46.

Gani, R., Harper, P. M., & Hostrup, M. (2005). Automatic creation of missing groups through connectivity index for pure-component property prediction. *Industrial & Engineering Chemistry Research*, *44*, 7262–7269.

Gani, R., Jiménez-González, C., & Constable, D. J. C. (2005). Method for selection of solvents for promotion of organic reactions. *Computers & Chemical Engineering*, *29*, 1661–1676.

Gani, R., Nielsen, B., & Fredenslund, A. (1991). A group contribution approach to computer-aided molecular design. *AIChE Journal*, *37*, 1318–1332.

Gebreslassie, B. H. & Diwekar, U. M. (2015). Efficient ant colony optimization for computer aided molecular design: Case study solvent selection problem. *Computers & Chemical Engineering*, *78*, 1–9.

Gharagheizi, F., Gohar, M. R. S., & Vayeghan, M. G. (2012). A quantitative structure–property relationship for determination of enthalpy of fusion of pure compounds. *Journal of Journal of thermal analysis and calorimetry*, *109*, 501–506.

Glukhovstev, M. N., Pross, A., McGrath, M. P., & Radom, L. (1995). Element, I. *The Journal of Chemical Physics*, *103*, 1878.

Gmehling, J. G., Anderson, T. F., & Prausnitz, J. M. (1978). Solid-liquid equilibria using UNIFAC. *Industrial & Engineering Chemistry Fundamentals*, 269–273.

Gopinath, S., Jackson, G., Galindo, A., & Adjiman, C. S. (2016). Outer approximation algorithm with physical domain reduction for computer-aided molecular and separation process design. *AIChE Journal*.

Gordon, R. E. & Amin, S. I. (1984). Crystallization of ibuprofen. US Patent 4,476,248.

Gupta, S., Singh, M., & Madan, A. K. (2002). Application of graph theory: Relationship of eccentric connectivity index and Wiener's index with anti-inflammatory activity. *Journal of Mathematical Analysis and Applications*, *266*, 259–268.

H. Guo, H. & Karplus, M. (1994). Solvent influence on the stability of the peptide hydrogen bond: A supramolecular cooperative effect. *The Journal of Physical Chemistry*, *98*, 7104–7105.

Hada, S., Solvason, C. C., & Eden, M. R. (2014). Characterization-based molecular design of bio-fuel additives using chemometric and property clustering techniques. *Frontiers in Energy Research*, *2*, 20.

Hall, L. H., Dailey, R. S., & Kier, L. B. (1993). Design of molecules from quantitative structure-activity relationship models. 3. Role of higher order path counts: Path 3. *Journal of Chemical Information and Computer Sciences*, *33*, 598–603.

Hall, L. H. & Kier, L. B. (2007). The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. *Reviews in Computational Chemistry, Volume 2*, 367–422.

Hall, L. H., Kier, L. B., & Murray, W. J. (1975). Molecular connectivity II: Relationship to water solubility and boiling point. *Journal of Pharmaceutical Sciences*, *64*, 1974–1977.

Hall, L. H., Maynard, E. L., & Kier, L. B. (1989a). QSAR investigation of benzene toxicity to fathead minnow using molecular connectivity. *Environmental Toxicology and Chemistry*, *8*, 783–788.

Hall, L. H., Maynard, E. L., & Kier, L. B. (1989b). StructureâĂŤactivity relationship studies on the toxicity of benzene derivatives: III. Predictions and extension to new substituents. *Environmental Toxicology and Chemistry*, *8*, 431–436.

Hall, L. H. & Story, C. T. (1996). Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks. *Journal of Chemical Information and Computer Sciences*, *36*, 1004–1014.

Hansen, H. K., Rasmussen, P., Fredenslund, A., Schiller, M., & Gmehling, J. (1991). Vapor-liquid equilibria by UNIFAC group contribution. 5. Revision and extension. *Industrial & Engineering Chemistry Research*, *30*, 2352–2355.

Hansen, N. (2017). *The CMA Evolution Strategy: A tutorial.* `http://www.lri.fr/~hansen/cmaesintro.html`.

Harini, M., Adhikari, J., & Rani, K. Y. (2013). A Review of Property Estimation Methods and Computational Schemes for Rational Solvent Design: A Focus on Pharmaceuticals. *Industrial & Engineering Chemistry Research*, *52*, 6869–6893.

Harper, P. M. & Gani, R. (2000). A multi-step and multi-level approach for computer aided molecular design. *Computers & Chemical Engineering*, *24*, 677–683.

Harper, P. M., Gani, R., Kolar, P., & Ishikawa, T. (1999). Computer-aided molecular design with combined molecular modeling and group contribution. *Fluid Phase Equilibria*, *158–160*, 337–347.

Herring, R. H. & Eden, M. R. (2015). Evolutionary algorithm for de novo molecular design with multi-dimensional constraints. *Computers & Chemical Engineering*, *83*, 267–277.

Hintermann, L., Masuo, R., & Suzuki, K. (2008). Solvent-controlled leaving-group selectivity in aromatic nucleophilic substitution. *10*, 4859–4862.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems.* The University of Michigan Press.

Holmström, K., Göran, A. O., & Edvall, M. M. (2007). *User's Guide for* `TOMLAB/CGO`. Tomlab Optimization. `http://tomopt.com`.

Holmström, K., Göran, A. O., & Edvall, M. M. (2011). *User's Guide for* `TOMLAB` *7*. Tomlab Optimization. `http://tomopt.com`.

Hooke, R. & Jeeves, T. A. (1961). Direct search solution of numerical and statistical problems. *Journal of the Association for Computing Machinery*, *8*, 212–219.

Hosoya, H. (1971). Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bulletin of the Chemical Society of Japan*, *44*, 2332–2339.

Hosoya, H., Hosoi, K., & Gutman, I. (1975). A topological index for the totalπ-electron energy. *Theoretica Chimica Acta*, *38*, 37–47.

Hostrup, M., M., P., & Gani, R. (1999). Design of environmentally benign processes: Integration of design and separation process synthesis. *Computers & Chemical Engineering*, *23*, 1395–1414.

Hsieh, C.-M., Sandler, S. I., & Lin, S.-T. (2010). Improvements of COSMO-SAC for vapor–liquid and liquid–liquid equilibrium predictions. *Fluid Phase Equilibria*, *297*, 90–97.

Huyer, W. & Neumaier, A. (2008). SNOBFIT – Stable noisy optimization by branch and fit. *ACM Transactions on Mathematical Software*, *35*, 1–25.

III, R. H. H., Haser, J. C., Hada, S., & Eden, M. R. (2013). Structure based design of non-peptide mimetics. *Proceedings of the 23rd European Symposium on Computer Aided Process Engineering, Lappeenrata, Finland*, 175–180.

Ingber, L. (2011). *Adaptive Simulated Annealing (`ASA`)*. `http://www.ingber.com/#ASA`.

Jalowka, J. W. & Daubert, T. E. (1986). Group contribution method to predict critical temperature and pressure of hydrocarbons. *Industrial & Engineering Chemistry Process Design and Development*, *25*, 139–142.

Jedlicka, B., Crabtree, R. H., & Siegbahn, P. E. M. (1997). Origin of solvent acceleration in organolithium metal-halogen exchange reactions. *16*, 6021–6023.

Joback, K. G. (1989). *Designing molecules possessing desired physical property values*. PhD thesis, Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Joback, K. G. & Reid, R. C. (1987). Estimation of pure-component properties from group contributions. *Chemical Engineering Communications*, *57*, 233–243.

Joback, K. G. & Stephanopoulos, G. (1990). Designing molecules possessing desired physical property values. *Proceedings of the 1989 Foundations of Computer-Aided Process Design Conference, Snowmass, CO,* Elsevier, Amsterdam, 195–230.

Joback, K. G. & Stephanopoulos, G. (1995). Searching spaces of discrete solutions: The design of molecules possessing desired physical properties. *Advances in Chemical Engineering*, *21*, 257–311.

Jonuzaj, S. & Adjiman, C. S. (2017). Designing optimal mixtures using generalized disjunctive programming: hull relaxations. *Chemical Engineering Science*, *159*, 106–130.

Jonuzaj, S., Akula, P. T., Kleniati, P.-M., & Adjiman, C. S. (2016). The formulation of optimal mixtures with generalized disjunctive programming: A solvent design case study. *AIChE Journal*.

Karunanithi, A. & Mehrkesh, A. (2013). Computer-aided design of tailor-made ionic liquids. *AIChE Journal*, *59*, 4627–4640.

Karunanithi, A. T., Achenie, L. E. K., & Gani, R. (2005). A new decomposition-based computer-aided molecular/mixture design methodology for the design of optimal solvents and solvent mixtures. *Industrial & Engineering Chemistry Research*, *44*, 4785–4797.

Karunanithi, A. T., Achenie, L. E. K., & Gani, R. (2006). A computer-aided molecular design framework for crystallization solvent design. *Chemical Engineering Science*, *61*, 1247–1260.

Karunanithi, A. T., Acquah, C., Achenie, L. E. K., Sithambaram, S., Suib, S. L., & Gani, R. (2007). An experimental verification of morphology of ibuprofen crystals from CAMD designed solvent. *Chemical Engineering Science*, *62*, 3276–3281.

Katritzky, A. R. & Gordeeva, E. V. (1993). Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *Journal of Chemical Information and Computer Sciences*, *33*, 835–857.

Katritzky, A. R., Wang, Y., Sild, S., Tamm, T., & Karelson, M. (1998). QSPR studies on vapor pressure, aqueous solubility, and the prediction of water-air partition coefficients. *Journal of Chemical Information and Computer Sciences*, *38*, 720–725.

Kauffman, G. W. & Jurs, P. C. (2001). Prediction of surface tension, viscosity, and thermal conductivity for common organic solvents using quantitative structure-property relationships. *Journal of Chemical Information and Computer Sciences*, *41*, 408–418.

Kelley, C. T. (2011). *Users Guide for* `IMFIL` *version 1.0*. `http://www4.ncsu.edu/~ctk/imfil.html`.

Kier, L. B. & Hall, L. H. (1976). Molecular connectivity VII: specific treatment of heteroatoms. *Journal of Pharmaceutical Sciences*, *65*, 1806–1809.

Kier, L. B. & Hall, L. H. (1993). The generation of molecular structures from a graph-based QSAR Equation. *Quantitative Structure-Activity Relationships*, *12*, 383–388.

Kier, L. B., Hall, L. H., & Dailey, R. S. (1993). Design of molecules from quantitative structure-activity relationship models. 3. Role of higher order path counts: Path 3. *Journal of Chemical Information and Computer Sciences*, *33*, 598–603.

Kier, L. B., Hall, L. H., & Frazer, J. W. (1993). Design of molecules from quantitative structure-activity relationship models 1. Information transfer between path and vertex degree counts. *Journal of Chemical Information and Computer Sciences*, *33*, 143–147.

Kier, L. B., Hall, L. H., Murray, W. J., & Randić, M. (1975). Molecular connectivity I: Relationship to nonspecific local anesthesia. *Journal of Pharmaceutical Sciences*, *64*, 1971–1974.

Kier, L. B., Murray, W. J., Randić, M., & Hall, L. H. (1976). Molecular connectivity V: connectivity series concept applied to density. *Journal of Pharmaceutical Sciences*, *65*, 1226–1230.

Kim, K. & Diwekar, U. M. (2002a). Efficient combinatorial optimization under uncertainty. 2. Application to stochastic solvent selection. *Industrial & Engineering Chemistry Research*, *41*, 1285–1296.

Kim, K.-J. & Diwekar, U. M. (2002b). Integrated solvent selection and recycling for continuous processes. *Industrial & Engineering Chemistry Research*, *41*, 4479–4488.

Klamt, A. (1995). Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena. *The Journal of Physical Chemistry A*, *99*, 2224–2235.

Klamt, A. (2005). *COSMO-RS: From quantum chemistry to fluid phase thermodynamics and drug design.* Elsevier.

Klamt, A. & Eckert, F. (2000). COSMO-RS: A novel and efficient method for the a priori prediction of thermophysical data of liquids. *Fluid Phase Equilibria*, *172*, 43–72.

Klamt, A. & Eckert, F. (2003). Erratum to "COSMO-RS: A novel and efficient method for the a priori prediction of thermophysical data of liquids". *Fluid Phase Equilibria*, *205*, 357.

Klamt, A., Jonas, V., Bürger, T., & Lohrenz, J. C. W. (1998). Refinement and parametrization of COSMO-RS. *The Journal of Physical Chemistry A*, *102*, 5074–5085.

Klamt, A. & Schüürmann, G. (1993). COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *Journal of the Chemical Society, Perkin Transactions*, *2*, 799–805.

Klein, J. A., Wu, D. T., & Gani, R. (1992). Computer aided mixture design with specified property constraints. *Computers & Chemical Engineering*, *16*, S229–S236.

Klincewicz, K. M. & Reid, R. C. (1984). Estimation of critical properties with group contribution methods. *AIChE Journal*, *30*, 137–142.

Klopman, G., Li, J.-Y., Wang, S., & Dimayuga, M. (1994). Computer automated log P calculations based on an extended group contribution approach. *Journal of Chemical Information and Computer Sciences*, *34*, 752–781.

Klopman, G. & Zhu, H. (2001). Estimation of the aqueous solubility of organic molecules by the group contribution approach. *Journal of Chemical Information and Computer Sciences*, *41*, 439–445.

Knight, J. P. & McRae, G. J. (1991). A combinatorial optimization approach to molecular design. *Nanotechnology*, *2*, 142–148.

Kolská, Z., Kukal, J., Zábranskỳ, M., & RuzËĞicËĞka, V. (2008). Estimation of the heat capacity of organic liquids as a function of temperature by a three-level group contribution method. *Industrial & Engineering Chemistry Research*, *47*, 2075–2085.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal on Optimization*, *9*, 112–147.

Lampe, M., Stavrou, M., BuÌĹcker, H. M., Gross, J., & Bardow, A. (2014). Simultaneous optimization of working fluid and process for organic Rankine cycles using PC-SAFT. *Industrial & Engineering Chemistry Research*, *53*, 8821–8830.

Lampe, M., Stavrou, M., Schilling, J., Sauer, E., Gross, J., & Bardow, A. (2015). Computer-aided molecular design in the continuous-molecular targeting framework using group-contribution PC-SAFT. *Computers & Chemical Engineering*, *81*, 278–287.

Lassau, C. & Jungers, J. (1968). L'influence du solvant sur la réaction chimique. La quaternation des amines tertiaires par l'iodure de méthyle. *Bulletin de la Société Chimique de France*, *7*, 2678–2685.

Le Digabel, S. (2009). `NOMAD` user guide version 3.3. Technical report, Les Cahiers du GERAD.

Lin, B., Chavali, S., Camarda, K., & Miller, D. C. (2005). Computer-aided molecular design using tabu search. *Computers & Chemical Engineering*, *29*, 337–347.

Lin, S.-T. & Sandler, S. I. (2002). A priori phase equilibrium prediction from a segment contribution solvation model. *Industrial & Engineering Chemistry Research*, *41*, 899–913.

Lymperiadis, A., Adjiman, C. S., Galindo, A., & Jackson, G. (2007). A group contribution method for associating chain molecules based on the statistical associating fluid theory (SAFT-γ). *The Journal of Chemical Physics*, *127*, 234903.

Lymperiadis, A., Adjiman, C. S., Jackson, G., & Galindo, A. (2008). A generalisation of the saft-group contribution method for groups comprising multiple spherical segments. *Fluid Phase Equilibria*, *274*, 85–104.

Macchietto, S., Odele, O., & Omatsone, O. (1990). Design on optimal solvents for liquid-liquid extraction and gas absorption processes. *Chemical Engineering Research and Design*, *68*, 429–433.

Maranas, C. D. (1996). Optimal computer-aided molecular design: A polymer design case study. *Industrial & Engineering Chemistry Research*, *35*, 3403–3414.

Marcoulaki, E. C. & Kokossis, A. C. (1998). Molecular design synthesis using stochastic optimisation as a tool for scoping and screening. *Computers & Chemical Engineering*, *22*, S11–S18.

Marrero, J. & Gani, R. (2001). Group-contribution based estimation of pure component properties. *Fluid Phase Equilibria*, *183–184*, 183–208.

Marrero, J. & Gani, R. (2002). Group-contribution based estimation of octanol/water partition coefficient and aqueous stability. *Industrial & Engineering Chemistry Research*, *41*, 6623–6633.

Martin, T. M. & Young, D. M. (2001). Prediction of the acute toxicity (96-h LC50) of organic compounds to the fathead minnow (Pimephales promelas) using a group contribution method. *Chemical Research in Toxicology*, *14*, 1378–1385.

Matsuda, H., Yamamoto, H., Kurihara, K., & Tochigi, K. (2007). Computer-aided reverse design for ionic liquids by QSPR using descriptors of group contribution type for ionic conductivities and viscosities. *Fluid Phase Equilibria*, *261*, 434–443.

McLeese, S. E., Eslick, J. C., Hoffmann, N. J., Scurto, A. M., & Camarda, K. V. (2010a). Design of ionic liquids via computational molecular design. *Computers & Chemical Engineering*, *34*, 1476–1480.

McLeese, S. E., Eslick, J. C., Hoffmann, N. J., Scurto, A. M., & Camarda, K. V. (2010b). Design of ionic liquids via computational molecular design. *Computers & Chemical Engineering*, *34*, 1476–1480.

Mercader, A., Castro, E. A., & Toropov, A. A. (2000). QSPR modeling of the enthalpy of formation from elements by means of correlation weighting of local invariants of atomic orbital molecular graphs. *Chemical Physics Letters*, *330*, 612–623.

Mitrofanov, I., Sansonetti, S., Abildskov, J., Sin, G., & Gani, R. (1995). The solvent selection framework: Solvents for organic synthesis, separation processes, and ionic-liquids solvents. *Proceedings of the 22nd European Symposium on Computer Aided Process Engineering June, 2012 London*, 257–311.

Mu, T., Rarey, J., & Gmehling, J. (2007). Group contribution prediction of surface charge density profiles for COSMO-RS(Ol). *AIChE Journal*, *53*, 3231–3240.

Mu, T., Rarey, J., & Gmehling, J. (2009). Group contribution prediction of surface charge density distribution of molecules for COSMO-SAC. *AIChE Journal*, *55*, 3298–3300.

Mullins, E. & Oldland, R. (2007). Sigma Profile Database. Available at `http://www.design.che.vt.edu/VT-Databases.html`.

Mullins, E., Oldland, R., Liu, Y. A., Wang, S., Sandler, S. I., Chen, C.-C., Zwolak, M., & Seavey, K. C. (2006). Sigma-profile database for using COSMO-based thermodynamic methods. *Industrial & Engineering Chemistry Research*, *45*, 4389–4415.

Murray, W. J., Hall, L. H., & Kier, L. B. (1975). Molecular connectivity III: Relationship to partition coefficients. *Journal of Pharmaceutical Sciences*, *64*, 1978–1981.

Nannoolal, Y., Rarey, J., & Ramjugernath, D. (2007). Estimation of pure component properties: Part 2. Estimation of critical property data by group contribution. *Fluid Phase Equilibria*, *252*, 1–27.

Nannoolal, Y., Rarey, J., & Ramjugernath, D. (2008). Estimation of pure component properties: Part 3. Estimation of the vapor pressure of non-electrolyte organic compounds via group contributions and group interactions. *Fluid Phase Equilibria*, *269*, 117–133.

Nannoolal, Y., Rarey, J., & Ramjugernath, D. (2009). Estimation of pure component properties. Part 4: Estimation of the saturated liquid viscosity of non-electrolyte organic compounds via group contributions and group interactions. *Fluid Phase Equilibria*, *281*, 97–119.

Nannoolal, Y., Rarey, J., Ramjugernath, D., & Cordes, W. (2004). Estimation of pure component properties: Part 1. Estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions. *Fluid Phase Equilibria*, *226*, 45–63.

Naser, S. F. & Fournier, R. L. (1991). A system for the design of an optimum liquid-liquid extractant molecule. *Computers & Chemical Engineering*, *15*, 397–414.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, *7*, 308–313.

Neumaier, A. (2011). `MCS`: Global Optimization by Multilevel Coordinate Search. `http://www.mat.univie.ac.at/~neum/software/mcs/`.

Neumann, C. N., Hooker, J. M., & Ritter, T. (2016). Concerted nucleophilic aromatic substitution with 19f- and 18f-.

Nuchitprasittichai, A. & Cremaschi, S. (2013). Optimization of CO2 capture process with aqueous amines–A comparison of two simulation-optimization approaches. *Industrial & Engineering Chemistry Research*, *93*, 247–263.

Odele, O. & Macchietto, S. (1993). Computer aided molecular design: A novel method for optimal solvent selection. *Fluid Phase Equilibria*, *82*, 47–54.

Ourique, J. E. & Telles, A. S. (1998). Computer-aided molecular design with simulated annealing and molecular graphs. *Computers & Chemical Engineering*, *22*, S615–S618.

Papadopoulos, A. I. & Linke, P. (2005). A unified framework for integrated process and molecular design. *Chemical Engineering Research and Design*, *83*, 674–678.

Papadopoulos, A. I. & Linke, P. (2006a). Efficient integration of optimal solvent and process design using molecular clustering. *Chemical Engineering Science*, *61*, 6316–6336.

Papadopoulos, A. I. & Linke, P. (2006b). Multiobjective molecular design for integrated process-solvent systems synthesis. *AIChE Journal*, *52*, 1057–1070.

Papadopoulos, A. I. & Linke, P. (2006c). Multiobjective molecular design for integrated process-solvent systems synthesis. *AIChE Journal*, *52*, 1057–1070.

Papadopoulos, A. I., Stijepovic, M., & Linke, P. (2010). On the systematic design and selection of optimal working fluids for organic Rankine cycles. *Applied Thermal Engineering*, *30*, 760 –769.

Papaioannou, V., Adjiman, C. S., Jackson, G., & Galindo, A. (2011). Simultaneous prediction of vapour-liquid and liquid-liquid equilibria (VLE and LLE) of aqueous mixtures with the SAFT-γ group contribution approach. *Fluid Phase Equilibria*, *306*, 82–96. 20 years of the SAFT equation of state–Recent advances and challenges Symposium.

Papaioannou, V., Lafitte, T., Avendaño, C., Adjiman, C. S., Jackson, G., Müller, E. A., & Galindo, A. (2014). Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments. *The Journal of chemical physics*, *140*, 054107.

Partington, J. R. (1970). A history of chemistry.

Patel, S. J., Ng, D., & Mannan, M. S. (2009). QSPR flash point prediction of solvents using topological indices for application in computer aided molecular design. *Industrial & Engineering Chemistry Research*, *48*, 7378–7387.

Pavurala, N. & Achenie, L. E. K. (2013). A mechanistic approach for modeling oral drug delivery. *Computers & Chemical Engineering*, *57*, 196–206.

Peng, Y., Goff, K. D., dos Ramos, M. C., & McCabe, C. (2009). Developing a predictive group-contribution-based SAFT-VR equation of state. *Fluid Phase Equilibria*, *277*, 131–144.

Pereira, F., Keskes, E., Galindo, A., Jackson, G., & Adjiman, C. (2011a). Integrated solvent and process design using a SAFT-VR thermodynamic description: High-pressure separation of carbon dioxide and methane. *Computers & Chemical Engineering*, *35*, 474–491.

Pereira, F. E., Keskes, E., Galindo, A., Jackson, G., & Adjiman, C. S. (2011b). Integrated solvent and process design using a SAFT-VR thermodynamic description: High-pressure separation of carbon dioxide and methane. *Computers & Chemical Engineering*, *35*, 474–491.

Pintér, J. D., Holmström, K., Göran, A. O., & Edvall, M. M. (2006). *User's Guide for* `TOMLAB/LGO`. Tomlab Optimization. `http://tomopt.com`.

Pistikopoulos, E. N. & Stefanis, S. K. (1998). Optimal solvent design for environmental impact minimization. *Computers & Chemical Engineering*, *22*, 717–733.

Plantenga, T. D. (2009). HOPSPACK 2.0 User Manual. Technical Report SAND2009-6265, Sandia National Laboratories, Albuquerque, NM and Livermore, CA.

Platts, J. A., Abraham, M. H., Butina, D., & Hersey, A. (2000). Estimation of molecular linear free energy relationship descriptors by a group contribution approach. 2. Prediction of partition coefficients. *Journal of Chemical Information and Computer Sciences*, *40*, 71–80.

Powell, M. J. D. (2002). UOBYQA: unconstrained optimization by quadratic approximation. *Mathematical Programming*, *92*, 555–582.

Powell, M. J. D. (2006). The `NEWUOA` software for unconstrained optimization without derivatives. In G. Di Pillo and M. Roma (eds.)*, Large-Scale Nonlinear Optimization,* Springer, New York, NY, 255–297.

Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. Technical report, Department of Applied Mathematics and Theoretical Physics, University of Cambridge.

Pretel, E. J., López, P. A., Bottini, S. B., & Brignole, E. A. (1994). Computer-aided molecular design of solvents for separation processes. *AIChE Journal*, *40*, 1349–1360.

Raman, V. S. & Maranas, C. D. (1998). Optimization in product design with properties correlated with topological indices. *Computers & Chemical Engineering*, *22*, 747–763.

Randic, M. (1975). Characterization of molecular branching. *Journal of the American Chemical Society*, *97*, 6609–6615.

Randic, M. & Zupan, J. (2001). On interpretation of well-known topological indices. *Journal of Chemical Information and Computer Sciences*, *41*, 550–560.

Reichardt, C. & Welton, T. (2011). *Solvents and solvent effects in organic chemistry.* John Wiley & Sons.

Rios, L. M. & Sahinidis, N. V. (2013). Derivative-free optimization: A review of algorithms and comparison of software implementations. *Journal of Global Optimization*, *56*, 1247–1293.

Roganov, G. N., Pisarev, P. N., Emel'yanenko, V. N., & Verevkin, S. P. (2005). Measurement and prediction of thermochemical properties. Improved Benson-type increments for the estimation of enthalpies of vaporization and standard enthalpies of formation of aliphatic alcohols. *Journal of Chemical & Engineering Data*, *50*, 1114–1124.

Rose, K., Hall, L. H., & Kier, L. B. (2002). Modeling blood-brain barrier partitioning using the electrotopological state. *Journal of Chemical Information and Computer Sciences*, *42*, 651–666.

Sahinidis, N. V., Tawarmalani, M., & Yu, M. (2003). Design of alternative refrigerants via global optimization. *AIChE Journal*, *49*, 1761–1775.

Samudra, A. & Sahinidis, N. V. (2009). Design of secondary refrigerants: A combined optimization-enumeration approach. In M. M. El-Halwagi and A. A. Linninger (eds.), *Proceedings of the Seventh International Conference on the Foundations of Computer-Aided Process Design,* CRC Press, 879–886.

Samudra, A. & Sahinidis, N. V. (2013a). Design of heat transfer media components for retail food refrigeration. *Industrial & Engineering Chemistry Research*, *52*, 8518–8526.

Samudra, A. & Sahinidis, N. V. (2013b). Optimization-based framework for computer-aided molecular design. *AIChE Journal*, *59*, 3686–3701.

Sandia National Laboratories (2011). The Coliny Project. `https://software.sandia.gov/trac/acro/wiki/Overview/Projects`.

Sastri, S. R. S. & Rao, K. K. (1992). A new group contribution method for predicting viscosity of organic liquids. *The Chemical Engineering Journal*, *50*, 9–25.

Scheffczyk, J., Fleitmann, L., Schwarz, A., Lampe, M., Bardow, A., & Leonhard, K. (2016). Cosmo-camd: A framework for optimization-based computer-aided molecular design using cosmo-rs. *Chemical Engineering Science*.

Scheinberg, K. (2003). *Manual for Fortran Software Package* `DFO` *v2.0.*

Schramke, J. A., Murphy, S. F., Doucette, W. J., & Hintze, W. D. (1999). Prediction of aqueous diffusion coefficients for organic compounds at 25 C. *Chemosphere*, *38*, 2381–2406.

Seader, J. & Henley, E. (1998). *Separation process principles.* New York: Wiley.

Siddhaye, S., Camarda, K., Southard, M., & Topp, E. (2004). Pharmaceutical product design using combinatorial optimization. *Computers & Chemical Engineering*, *28*, 425–434.

Siddhaye, S., Camarda, K. V., Topp, E., & Southard, M. (2000). Design of novel pharmaceutical products via combinatorial optimization. *Computers & Chemical Engineering*, *24*, 701–704.

Sinha, M., Achenie, L. E. K., & Gani, R. (2003). Blanket wash solvent blend design using interval analysis. *Industrial & Engineering Chemistry Research*, *42*, 516–527.

Sinha, M., Achenie, L. E. K., & Ostrovsky, G. M. (1999). Environmentally benign solvent design by global optimization. *Computers & Chemical Engineering*, *23*, 1381–1394.

Smith, R. L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, *32*, 1296–1308.

Stanescu, I. & Achenie, L. E. K. (2006). A theoretical study of solvent effects on Kolbe–Schmitt reaction kinetics. *Chemical Engineering Science*, *61*, 6199–6212.

Stavrou, M., Lampe, M., Bardow, A., & Gross, J. (2014). Continuous molecular targeting–computer-aided molecular design (CoMT–CAMD) for simultaneous process and solvent design for CO2 capture. *Industrial & Engineering Chemistry Research*, *53*, 18029–18041.

Stein, S. E. & Brown, R. L. (1994). Estimation of normal boiling points from group contributions. *Journal of Chemical Information and Computer Sciences*, *34*, 581–587.

Steiner, T. (1998). Hydrogen-bond distances to halide ions in organic and organometallic crystal structures: Up-to-date database study. *Acta Crystallographica Section B: Structural Science*, *54*, 456–463.

Stephens, P. J., Devlin, F. J., Chabalowski, C. F., & Frisch, M. J. (1994). Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of Physical Chemistry*, *98*, 11623–11627.

Struebing, H., Ganase, Z., Karamertzanis, P. G., Siougkrou, E., Haycock, P., Piccione, P. M., Armstrong, A., Galindo, A., & Adjiman, C. S. (2013a). Computer-aided molecular design of solvents for accelerated reaction kinetics. *Nature Chemistry*, *5*, 952–957.

Struebing, H., Ganase, Z., Karamertzanis, P. G., Siougkrou, E., Haycock, P., Piccione, P. M., Armstrong, A., Galindo, A., & Adjiman, C. S. (2013b). Computer-aided molecular design of solvents for accelerated reaction kinetics. *Nature Chemistry*, *5*, 952–957.

Sundaram, A., Ghosh, P., Caruthers, J. M., & Venkatasubramanian, V. (2001). Design of fuel additives using neural networks and evolutionary algorithms. *AIChE Journal*, *47*, 1387–1406.

Tawarmalani, M. & Sahinidis, N. V. (2004). Global optimization of mixed-integer nonlinear programs: A theoretical and computational study. *Mathematical Programming*, *99*, 563–591.

Taylor, R. & Kennard, O. (1984). Hydrogen-bond geometry in organic crystals. *Accounts of chemical research*, *17*, 320–326.

Tihic, A., Kontogeorgis, G. M., von Solms, N., Michelsen, M. L., & Constantinou, L. (2007). A predictive group-contribution simplified PC-SAFT equation of state: Application to polymer systems. *Industrial & Engineering Chemistry Research*, *47*, 5092–5101.

Vaidyanathan, R. & El-Halwagi, M. (1996). Computer-aided synthesis of polymers and blends with target properties. *Industrial & Engineering Chemistry Research*, *35*, 627–634.

Vaz, A. I. F. (2011). `PSwarm` Home Page.
`http://www.norg.uminho.pt/aivaz/pswarm/`.

Venkatasubramanian, V., Chan, K., & Caruthers, J. M. (1994). Computer-aided molecular design using genetic algorithms. *Computers & Chemical Engineering*, *18*, 833–844.

Venkatasubramanian, V., Chan, K., & Caruthers, J. M. (1995). Evolutionary design of molecules with desired properties using the genetic algorithm. *Journal of Chemical Information and Computer Sciences*, *35*, 188–195.

Visco, D. P., Pophale, R. S., Rintoul, M. D., & Faulon, J.-L. (2002). Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor. *Journal of Molecular Graphics and Modelling*, *20*, 429–438.

Wakefield, B. J. (2013). *The chemistry of organolithium compounds*. Elsevier.

Wang, S., Sandler, S. I., & Chen, C.-C. (2007). Refinement of COSMO-SAC and the applications. *Industrial & Engineering Chemistry Research*, *46*, 7275–7288.

Wang, Y. & Achenie, L. E. K. (2002). Computer aided solvent design for extractive fermentation. *Fluid Phase Equilibria*, *201*, 1–18.

Warrier, P., Sathyanarayana, A., Bazdar, S., Joshi, Y., & Teja, A. S. (2012). Selection and evaluation of organosilicon coolants for direct immersion cooling of electronic systems. *Industrial & Engineering Chemistry Research*, *51*, 10517–10523.

Weis, D. C. & Visco, D. P. (2010). Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. *Computers & Chemical Engineering*, *34*, 1018–1029.

Wiener, H. (1947). Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, *69*, 17–20.

Wittig, R., Lohmann, J., & Gmehling, J. (2003). VaporâĹŠLiquid equilibria by UNIFAC group contribution. 6. revision and extension. *Industrial & Engineering Chemistry Research*, *42*, 183–188.

Wu, X., Zhang, C., Goldberg, P., Cohen, D., Pan, Y., Arpin, T., & Bar-Yosef, O. (2012). Early pottery at 20,000 years ago in Xianrendong Cave, China. *Science*, *336*, 1696–1700.

Xiong, R., Sandler, S. I., & Burnett, R. I. (2014). An improvement to COSMO-SAC for predicting thermodynamic properties. *Industrial & Engineering Chemistry Research*, *53*, 8265–8278.

Xu, W. & Diwekar, U. M. (2005). Improved genetic algorithms for deterministic optimization and optimization under uncertainty. part ii. solvent selection under uncertainty. *Industrial & Engineering Chemistry Research*, *44*, 7138–7146.

Yao, X., Fan, B., Doucet, J. P., Panaye, A., Liu, M., Zhang, R., Zhang, X., & Hu, Z. (2003). Quantitative structure property relationship models for the prediction of liquid heat capacity. *QSAR & Combinatorial Science*, *22*, 29–48.

Zhang, L., Cignitti, S., & Gani, R. (2015). Generic mathematical programming formulation and solution for computer-aided molecular design. *Computers & Chemical Engineering*, *78*, 79–84.

Zhou, T., Lyu, Z., Qi, Z., & Sundmacher, K. (2015). Robust design of optimal solvents for chemical reactions–A combined experimental and computational strategy. *Chemical Engineering Science*, *137*, 613–625.

Zhou, T., Qi, Z., & Sundmacher, K. (2014). Model-based method for the screening of solvents for chemical reactions. *Chemical Engineering Science*, *115*, 177–185.

Zhou, T., Wang, J., McBride, K., & Sundmacher, K. (2016). Optimal design of solvents for extractive reaction processes. *AIChE Journal*.

Zhou, T., Zhou, Y., & Sundmacher, K. (2016). A hybrid stochastic–deterministic optimization approach for integrated solvent and process design. *Chemical Engineering Science*.