Physical properties and chemical product design

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Chemical Engineering

Yijia Sun

Bachelor of Science in Engineering, Chemical Engineering, University of Connecticut

Carnegie Mellon University

Pittsburgh, PA

May 2021

"To be patient toward all that is unsolved in your heart and to try to love the questions themselves like locked rooms and like books that are written in a very foreign tongue. Do not now seek the answers, which cannot be given you because you would not be able to live them. And the point is, to live everything. Live the questions now. Perhaps you will then gradually, without noticing it, live along some distant day into the answer." —Rainer Maria Rilke

Acknowledgments

First, I would like to thank my advisor Dr. Nick Sahinidis for his constant support and guidance over the course of my PhD career. It has been a privilege to work with him, to observe and learn from how he gets to the heart of a problem. His extensive knowledge in many diverse research fields has been a constant source of inspiration. Thank you for giving me the freedom to explore different ideas and for offering your acute insight when I needed it. I am very fortunate to have an advisor who's always been supportive of me pursuing my career goals both inside and outside of chemical engineering. Nick is truly a great mentor and role model, and there is so much one can learn from him. Thank you for teaching me about the art of efficiency and the power of intuition. I have also learnt from him about efficient and effective communication, which proves to be extremely helpful especially during the last 13 months of my PhD when working from home becomes the new normal as a result of the pandemic. I am very fortunate to be your student.

I would like to extend my gratitude to the members of my doctoral committee: Prof. Larry Biegler, Prof. Chrysanthos Gounaris, and Prof. Dave Yaron. Thank you all for offering invaluable insights and constructive advice which contribute significantly to this thesis. Thanks in particular to Larry for being my co-advisor in the last year of my PhD and for your valuable comments to my research proposal.

I would like to take this opportunity to thank many collaborators and mentors that I have worked

with over the years. I would like to thank Exxon Mobil Research and Engineering for partially funding my research and thank Dr. Anantha Sundaram and Myun-Seok Cheon from Exxon Mobil for being incredible collaborators. I would like to thank Prof. George Bollas for introducing me to computational research and for being the first one to teaching me to take ownership of my work. I thank Dr. Jin Wang and the colleagues I met at Morgan Stanley for their help and support during my summer internship. Thanks should also go to Prof. Ignacio Grossmann and Prof. Larry Biegler for introducing me to the field of optimization.

I would like to thank the Sahinidis Group. Special thanks to Marissa Engle and Dr. Tong Zhang for being great mentors. I would like to thank all the friends I have made during the course of my time at CMU. I also thank all faculties and staff from the department for your help throughout the years.

Thanks to the Center for Advanced Process Decision-making and the Department of Chemical Engineering for funding my research.

Finally, I owe my deepest appreciation to my family back in China. They are always my pillar of strength during difficult times. I am forever indebted to my parents for their unconditional love and support. I am equally thankful to my grandparents for their love and encouragement throughout my life. Thanks to my friends in the US and back in China for supporting me on this journey. I am finally very thankful to Aiden for being a constant source of inspiration, for always believing in me, and for helping me to strive for the best. And thanks to our lovely dog Marsh for always keeping us company. Thank you all for shaping my life and helping me grow into the person I am today. This thesis is dedicated to you.

Abstract

Chemical product design is the problem of identifying chemical compounds that satisfy a set of previously identified functional properties for a specific application. Throughout the industrial era, chemical product design has played an increasingly important societal role by addressing the growing demand for better products. Naturally, product design has become a focal area for chemical scientists and engineers. This thesis develops new methodologies for chemical product design problems where the design targets are physical properties. We address at problems with dynamic design targets and stationary design targets separately.

For design applications with stationary property targets, we rely on algebraic optimization models to efficiently locate optimal compounds from the vast chemical design space. In particular, we exploit efficient mixed-integer linear programming (MILP) models and solve for optimal pure chemical components for a specific application: identifying better cooling fluids for electronic equipment. We use group contribution (GC) methods as the major property estimation tool along with additional accurate property models. We derive a metric to measure the cooling performance of a two-phase cooling system consisting of micro-channel heat sinks. Additionally, we carry out a sensitivity analysis on property predictions to assess the effect of prediction uncertainty on the final design outcome.

To extend the coolant design space to include silicon containing structures, we develop a new GC

method capable of property prediction for organosilicon compounds. Along the way, we propose a functional group selection method that deterministically decomposes each molecular structure into the smallest number of non-overlapping functional groups while ensuring each group holds a maximum amount of information. The group selection method is applied to construct GC models to predict eight pure component properties. Utilizing the new GC models, we are able to identify organosilicon cooling fluids with considerably improved heat transfer properties.

As for design application with time-varying targets, or functionalities that are difficult to model via algebraic expressions, we leverage the state-of-art derivative-free optimization (DFO) methods to solve for a diverse span of candidate chemicals. We assess DFO algorithms and demonstrate the viability of DFO in a polymer configuration design example. Our computational results suggest that a collection of derivative-free algorithms can successfully search the chemical design space and identify good solutions with high computational efficiency.

Whether the candidate compounds can be put into production depends on their likelihood to be synthesized. We investigate contemporary synthetic planning methodologies and provide an overview on retrosynthesis frameworks. We address potential challenges and opportunities facing automated synthetic planning.

Lastly, we summarize the major contributions from this work and offer future research directions.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The chemical industry has long been one of the fastest growing sectors in the economy [1]. Ranging from consumer products, pharmaceuticals, and commodity chemicals to oil and gas, the chemical industry produces materials that constitute "the world of things", supporting every aspect of our lives. Chemical companies are always on the lookout for new opportunities to accommodate shifts in market trends and meet consumers' desire for improved products. Consequently, an ever-increasing number of chemicals are added to the chemical design space. As of March 2021, more than 178 million unique chemicals have been registered to the Chemical Abstracts Service (CAS) [2].

Chemical product design is the process of translating desirable functional criteria into an array of suitable products for a specific application [3]. In this process, one first needs to understand customers' wants and needs from an application and translate their desire into a list of quantifiable design targets. The design targets of a product provide specifications on its properties, such as physical, chemical, mechanical, and environmental. The next step is to discover product alternatives that satisfy the desired product specifications. After solving the problem of "what to make", the last

step in chemical product design is to solve "how to make it" by developing processes to manufacture and test the products [4]. The problem of "what to make" differentiates chemical product design from chemical process design. The latter deals with a known product, usually commodity chemical, and the goal is to optimize the manufacturing process efficiency to gain a competitive edge in the market. Conversely, the former often focuses on searching for novel/unknown products for a certain application. This problem attempts to locate desirable products from the chemical space by relying on the relationship between the products' properties and their chemical structures. Therefore, good understanding of the property-structure relationship is essential to any chemical product design activity.

The conventional approach to chemical product design relies on a series of trial-and-error experiments. In other words, new products are synthesized using high throughput screening by varying previous formulations or finding structural analogues of existing compounds. Unfortunately, this conventional approach often limits the chemical search space to what can be synthesized and tested based on prior knowledge and experience, which inevitably overlooks structurally novel compounds that could be performing better than existing ones.

Today, computer-based material design approaches have emerged as an attractive alternative to the traditional synthesize-and-test methodology. Among these approaches, computer-aided molecular design (CAMD) is a well-studied problem for designing single chemicals but also serves as the foundation for more complex material design tasks. CAMD relies on physicochemical prediction models and algorithmic frameworks to efficiently explore the diverse chemical design space and identify suitable chemical structures that meet design targets. In CAMD, molecular structures form by combining submolecular functional groups, followed by a screening step to select structures that meet design constraints. The CAMD approach builds on methodologies that solve two problems: the "forward problem" relies on semi-empirical quantitative structure-

property relationships (QSPRs) to predict properties given molecular structures, while the “backward problem" chooses optimal molecules from the space of theoretically possible chemical structures given target properties. These two problems represent different sides of the CAMD puzzle and have attracted significant attention, especially in the context of specific applications, including the design of alternative refrigerants [5–7], heat transfer fluids [8–10], extraction solvents [11–13], polymer repeating units [14–16], mixtures [17–23], reaction solvents [24–27], ionic liquids [28–31], and pharmaceutical solvents [32, 33]. Several works [34–36] review the development, milestones, applications, and challenges of CAMD.

Most of the CAMD applications have utilized QSPRs or semi-empirical models to connect the chemical design space to the space of properties. Group contribution (GC) based methods are among the most commonly used QSPRs to estimate the physicochemical properties of putative structures. GC methods build upon the group additivity principle and work under one key assumption: a molecule’s properties can be estimated by the number of occurrences of sub-molecular structures called groups. Joback and Reid [37] proposed a GC method that includes functional transformations of otherwise linear group contribution summations. Constantinou and Gani [38] provided another extension to the general form of GC models, which introduces multiple levels of groups to better capture the proximity effects. Later, Marrero and Gani [39, 40] proposed one of the most commonly-used GC methods in CAMD, the GC+ method, that utilizes a three-level group contribution model to cover proximity effects and structural features at the molecular scale. Efforts to improve GC methods accuracy include the introduction of connectivity indices [41, 42] and group interaction terms [43–47].

GC methods link molecular structures to property values via an occurrence vector that represents the frequency of appearances of each group. Each molecular structure is broken down into a set of sub-molecular functional groups to characterize the molecular properties. For example, we can

think of acetic acid being formed by connecting 1 $-CH_3$ group to 1 $-COOH$ group. The molecular composition of acetic acid can be represented using an occurrence vector, $n = [1, 1]$, where each element stands for the frequency of occurrences of $-CH_3$ and $-COOH$, respectively. Following the additive assumption of GC methods, physical properties of acetic acid can be estimated using

$$P = \sum_i c_i n_i$$

where n_i is the i^{th} element in the occurrence vector, and c_i represents the contribution of group i to the overall property value. The contributions are obtained through regression on experimental data. The set of functional groups from the GC methods now become features to describe the chemical design space. Using these features, the CAMD problem is now transformed from searches in the vast space of molecular structures into a search over a much smaller set of groups, the combinations of which meet the design targets. This problem can then be translated into mathematical formulations and solved using modern combinatorial optimization techniques.

1.2 Challenges with existing QSPR and CAMD frameworks

There are a few challenges facing the current QSPR and CAMD frameworks.

- QSPRs, being a family of semi-empirical models, sometimes lack the ability to model complex properties to a satisfactory degree. Examples include physical properties from the quantum level, rheological properties, and properties with time-varying behavior. Advancements in computational modeling have enabled the development of more accurate property estimation simulators. These property simulators facilitate the incorporation of more sophisticated design targets into CAMD. However, combining property simulators with algebraic optimization solvers can be a challenging task for CAMD.

- GC methods use a set of previously selected functional groups as prediction variables to regress the GC coefficients, i.e., the contributions. Combining these groups forms a diverse span of molecular structures in the chemical design space. The number of all possible combinations determines the size of the design space CAMD is modeled upon. Consequently, organic families that cannot be modeled using the GC methods reside outside of the feasible CAMD design region. Organosilicon compounds is one of such organic families that current GC models are unable to provide reliable predictions to. Therefore despite their wide use in commercial products, organosilicon compounds are usually excluded from CAMD studies.

- CAMD approaches often rank candidate molecules via a metric which is a composite of several design target properties. As property values estimated by GC methods come with a degree of uncertainty, these uncertainties propagate into the objective function, adding a considerable layer of propagated uncertainty to the final performance ranking. Often times, it is not straightforward to determine whether the increase in performance predicted for a candidate compound is due to better design or an artifact of prediction uncertainties.

- Candidate molecules identified from the CAMD problem often have completely novel chemical structures. Therefore, identifying reaction paths to synthesize candidate molecules from a set of commercially available reactants is essential in solving the question of "how to make it" in chemical product design. The problem of decomposing target molecules into a number of feasible precursors is called *retrosynthesis*. This problem serves as a critical final step in chemical product design to put design compounds into manufacturing. Yet, how to directly incorporate manufacturability as a screening step in CAMD is an open question.

1.3 Research problem statement

This work focuses on chemical product design problems where the design targets are physical properties. We aim to offer solution approaches to design problems with 1) time-varying property targets and 2) fixed property targets. For problems with dynamic property targets, we propose methodologies to track the time series behavior of physical properties and utilize black-box optimization solvers to efficiently identify suitable solutions [48]. For problems with stationary targets, we focus on a specific design application: electronics cooling fluids [10]. We rely on CAMD and GC models to search for optimal coolants in the chemical design space. The CAMD approach used in this work follows a decomposition scheme that solves for molecular compositions and structures in a sequential manner to reduce computational requirements. In situations where the design targets cannot be predicted using existing GC models, we provide a systematic methodology to build property prediction models from experimental measurements. Once a set of candidate coolants are identified, we execute an uncertainty analysis step to account for the effect of prediction uncertainties on the final design outcomes. Finally, we explore the manufacturability of candidate compounds and provide a review on contemporary retrosynthesis approaches.

1.4 Thesis outline

This thesis is organized as follows:

In chapter 2, we introduce the electronics coolant design problem. Our goal is to design cooling fluids that have better heat transfer performance than the industrial coolant HFE7200. In this work, we rely on a CAMD framework, AMODEO [49], and group contribution methods to solve for optimal coolants. We derive a metric to rank the performance of candidate solutions and carry out a sensitivity analysis to assess the effect of prediction uncertainties on the final performance

improvement. As a result of this work, we identify a number of novel cooling fluids that fall into four organic families.

Chapter 3 extends the work of coolant design to organosilicon compounds. We develop a group contribution model that enables prediction of eight pure component properties for silicon-containing structures. The GC models are based on a deterministic functional group selection method we developed. The proposed GC models are subsequently embedded in the AMODEO framework to design organosilicon coolants systems. The candidate compounds demonstrate superior heat transfer performance compared to HFE7200 and non-organosilicon coolants.

In chapter 4, we look at a chemical product design problem with dynamic property targets. We aim to identify optimal polymer melt mixtures that match a moving rheological target during polymerization reaction. We propose to incorporate derivative-free optimization into the design framework. Doing so facilitates the exploitation of property prediction simulators to generate accurate property models and search the complete design space to increase solution diversity.

To have a better understanding of manufacturability test, in Chapter 5, we provide a review on retrosynthesis strategies to break target molecules into precursors. We focus on how data driven models and machine learning techniques contribute to the field. We go back to the coolant design problem and assess the manufacturability of candidate coolants via the synthetic accessibility score (SAScore) [50].

Finally in Chapter 6, we provide conclusions of this thesis and offer future research directions.

Chapter 2

Design of electronics cooling fluids

2.1 Introduction

The design trend of miniaturized electronics has resulted in a significant decrease in transistor size and an increase in power density. As the demand for high performance electronic devices continues to increase, the heat flux generation of many electronic chips has already exceeded their predetermined value [51], creating serious technical challenges in electronic cooling. The increasing need to find robust thermal management solutions is driving the growth of the electronic cooling industry. The market size of thermal management technologies is expected to reach $18 billion by 2024 [52]. Consequently, manufacturers and researchers are focusing on the development of advanced heat removal techniques and effective coolants to meet the ever-increasing cooling needs that traditional cooling approaches fail to satisfy.

Based on the working fluids utilized, traditional cooling techniques can be classified into several categories, including air cooling, liquid cooling, and refrigerant cooling. Air is the primary coolant because it is easy to reproduce and operate. Being cost-effective and highly reliable, forced convection of air is widely used in CPU cooling, while natural convection is commonly used in

low heat flux applications. Liquid cooling emerged in the early 1970s and became the preferred method for high-flux heat removal. Liquid cooling can be characterized by either a single-phase or two-phase cooling system [53]. Single-phase liquid cooling utilizes natural or forced convection and relies on liquid heat capacity. Two-phase cooling uses fluids that evaporate at lower temperatures (50^{o}C – 120^{o}C). Since the heat of vaporization is orders of magnitude higher than liquid heat capacity, two-phase cooling systems can achieve high heat flux while maintaining a safe operating temperature.

Murshed and Nieto de Castro [54] reviewed both traditional and emerging techniques and fluids for electronic cooling. Among all emerging cooling techniques, microchannel-based cooling systems have a much higher heat transfer performance than any traditional heat exchanger. Microchannel heat sinks significantly minimize the package size and can be adapted to on-chip integration. Thus, they are one of the most promising thermal management solutions to high heat generating electronic devices. Figure 2.1 shows a simplified version of a microchannel cooling system. From direct contact with the working fluid, the modified surface fins receive heat from the electronic device through conduction and emit heat to the coolant reservoir. As rising chip temperature generates more heat flux, nucleate boiling of the coolant takes place near the surface fins, which in turn draws more liquid to the surface as bubbles continue to depart. The combined action of heat conduction, latent heat absorption, and forced convection leads to significant cooling effects, allows sufficient heat dissipation, and maintains the chip temperature below the maximum junction temperature.

A variety of fluids are employed in a number of cooling systems. Common industrial coolants in the 20^{th} Century were fully halogenated chlorofluorocarbons (CFCs). CFCs have a boiling point range from -30^oC to 24^{o}C, suitable for a variety of two-phase cooling applications. However, CFCs are banned by the Montreal Protocol [55] due to their ability to destroy the ozone layer. On the other hand, water is used in many cooling applications because of its exceptional heat transfer capabilities

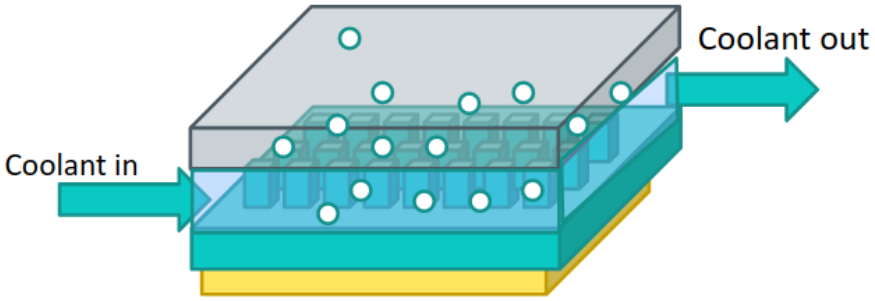


Figure 2.1: Schematic of a microchannel heat sink

and low viscosity. However, the use of water is not permissible in closed loop cooling systems due to its high freezing point and expansion upon freezing. Common microchannel cooling fluids are hydrofluorocarbons (HFCs) and hydrofluoroethers (HFEs) [56]. Although carbon-based refrigerants remain safe and stable when in contact with electronic circuits, greenhouse gases like HFCs have very long atmospheric lifetimes. A recent global agreement signed in Kigali [57] has limited the consumption and production of HFCs due to their high global warming potential (GWP). HFEs have short atmospheric lifetimes and low GWP but unfortunately have low thermal conductivity and low specific heat [58]. Their performance is also limited by low surface tension. Therefore, it is crucial to seek more efficient and environmentally benign cooling fluids for utilization in microchannel systems.

Computer-aided molecular design (CAMD) has been used to design alternative refrigerants and heat transfer fluids in a number of studies [5, 7, 59–63]. These works utilize enumeration methods or mixed-integer nonlinear programming (MINLP) techniques to solve CAMD problems. The development of physiochemical models is a substantial intellectual contribution to these works. These models are suitable for the specific application and are complemented with an algorithmic framework for searching molecules with desired properties.

In our work, we develop an optimization model for designing cooling fluids to identify compounds that have optimal thermal and environmental performance. We solve this model using a recently developed CAMD solution methodology that utilizes mixed-integer linear programming (MILP) techniques and a decomposition framework for fast computation [49]. Additionally, we derive a heat transfer performance metric for microchannel-based systems and execute an uncertainty analysis of the performance criterion. Specifically, we provide a detailed analysis of the candidate molecules that includes explicit models for biodegradability, toxicity characteristics, and kinetic stability. To solve the problem of finding ideal fluids in microchannel cooling systems, we match a set of property targets that are specifically based on properties of existing industrial cooling fluids. We set objectives for minimizing environmental impact and search for novel compounds with superior heat transfer performance.

The remainder of this chapter is structured as follows. In Section 2, we provide background information, including a review of CAMD and the molecular design framework we utilize. In Section 3, we present the formulation of the general design problem based on property constraints and environmental impact. In Section 4, we propose our model for the design of electronic cooling fluids by specifying property prediction methods and performance metrics. In Section 5, we discuss the results and analyze the most promising candidate molecules.

2.2 CAMD framework

Computer-aided molecular design is used in many applications to identify promising molecules that satisfy predefined properties [8, 11, 25, 59, 61, 64–72]. The traditional molecular design approach assumes a molecular structure and solves the forward calculation problem, with the goal to predict properties from the pre-identified molecular structures. CAMD is the reverse problem–its objective is to determine molecular structures that meet the desired property targets. This approach expands

the molecular search space by considering a large diversity of structures and relies heavily on the availability of property estimation models. Group contribution (GC) methods [37, 38, 40, 61, 73, 74] are widely used to estimate physical properties and are suitable for the CAMD approach. The GC methods are mostly based on the assumption that key molecular properties satisfy the group additivity principle. The CAMD problem can be formulated as an optimization problem where the design variables model the molecular structure and the constraints utilize GC-like methods to enforce requirements on physical properties.

From a set of property targets, CAMD will find a combination of molecular substructures that compose a molecule and satisfy the specified properties. To address the combinatorial challenge associated with searching the vast space of molecular structures, Samudra and Sahinidis [49] propose a decomposition scheme in which they determine the solutions to molecular compositions and the structures separately. The complex CAMD problem contains three simpler subproblems that provide a solution pool with increasing property prediction accuracy and finer structure resolution. The series of subproblems are automated and implemented into a software, Automated Molecular Design Using Optimization (AMODEO), which can handle a variety of design problems. Figure 2.2 presents an overview of the framework. The following subsections briefly introduce each subproblem of the framework.

2.2.1 Property estimation models

Group contribution methods are a class of quantitative structure property relationships (QSPRs). A molecule is a collection of groups, where each group has a property-dependent contribution. Properties are estimated by adding the product of the number of occurrences to the contribution of each group in a molecule.

Constantinou and Gani [38] model molecules as a collection of groups in different orders.

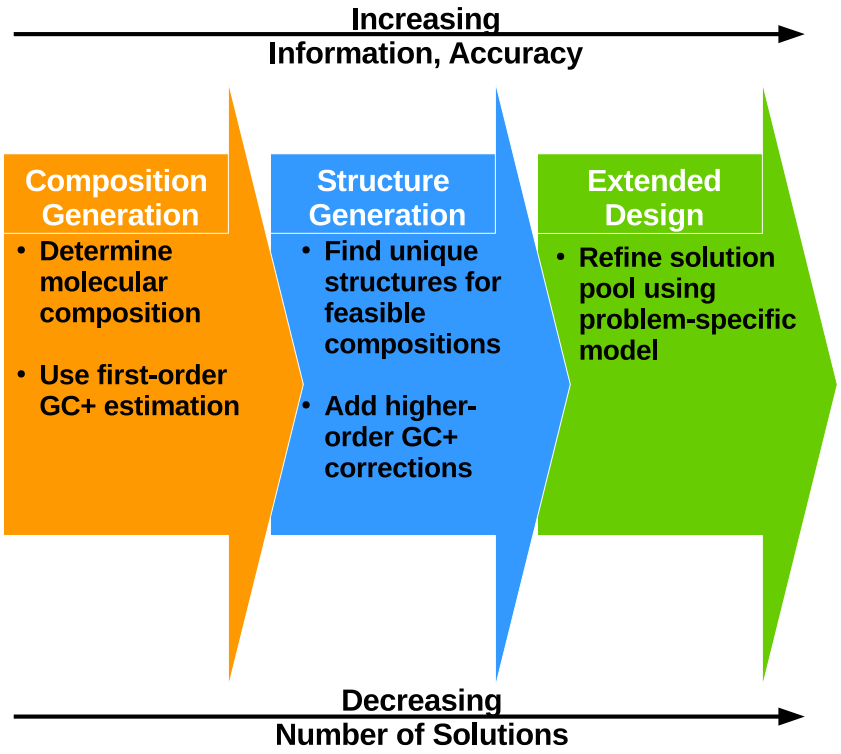


Figure 2.2: CAMD framework

The first-order groups are typically non-overlapping chemical subgroups that compose a molecule. First-order estimation is a method that uses only first-order groups for property estimation. The higher-order groups are combinations of first-order groups, and by capturing the proximity effects of first-order groups, they provide corrections to first-order estimations. Furthermore, each property is a function of additive contributions of individual groups. Functions are chosen to closely match experimental data.

Gani et. al. [41, 42] extended the GC method by using connectivity indices to predict contributions of groups absent in the original GC model. This expanded method is known as the GC+ method, which involves 182 first-order, 122 second-order, and 66 third-order groups. These functional groups increase prediction accuracy in the GC+ method. We use the GC+ method as the main property prediction model in our approach.

2.2.2 Composition design

Composition design seeks diverse molecular compositions that match relaxed design targets. The GC+ method predicts key molecular properties through using only first-order groups as building blocks. This problem can be formulated as follows:

$$\min_{n} \quad 0 \tag{2.1}$$

$$\text{s.t.} \quad p_k^L \leq f\left(\sum_{i \in F} c_{ik} n_i\right) \leq p_k^U, \ \forall k \in K \tag{2.2}$$

$$s_1(n) \leq 0 \tag{2.3}$$

$$s_2(n) = 0 \tag{2.4}$$

$$n \in \mathbb{Z}^N \tag{2.5}$$

In the above model, F is the set of first-order groups that contains a total number of N groups. c_{ik} is the contribution of the i^{th} group for the estimation of the k^{th} property. n_i is the number of occurrences of group i. Each property falls in the range $[p_k^L, p_k^U]$, for a set of K properties. This stage includes structural constraints to enforce acyclic tree structures of the molecules. Rings or multiring clusters are treated as fictitious super-nodes in the tree. We ensure the acyclic tree condition by allowing exactly $V - 1$ acyclic bonds between V groups, rings, and multi-ring clusters. $s_1(n) \leq 0$ and $s_2(n) = 0$ summarize a set of structural constraints [7, 66].

We apply a relaxation on the property bounds $[p_k^L, p_k^U]$ to account for first-order estimation error. Using the relaxed property bounds, the nonlinear property constraints are transformed into linear constraints due to the monotonic nature of the GC+ model. The upper and lower bounds after inversion of f, π_k and κ_k, can be calculated directly from f^{-1}. The nonlinear constraint $p_k^L \leq p_k \leq p_k^U$ in the property space, p_k, is equivalent to the linear constraint in the $f^{-1}(p_k)$ space as shown below:

$$\kappa_k \leq \sum_{i \in F} c_i^k n_i \leq \pi_k \tag{2.6}$$

$$\kappa_k = f^{-1}(p_k^L),\ \pi_k = f^{-1}(p_k^U) \tag{2.7}$$

Property constraints imposed by design targets are enforced by this outer-linearization representation in this stage. By exploiting linearity in the space of functional groups, we can formulate an MILP model to identify all feasible molecular compositions.

2.2.3 Structure generation

Each resulting composition that is obtained through composition design may correspond to a number of distinct isomers with different properties. The structure design problem identifies unique

structures for all feasible compositions and adds higher-order corrections to property prediction. We use a graphical representation of molecules to generate all possible distinct tree graphs representing unique isomers. For each set of functional groups and their frequency determined in composition design, we generate all possible planar graphs where the nodes and edges represent the functional groups and chemical bonds, respectively. Each molecular graph is described by an adjacency matrix to capture all the necessary bond information, the entries of which are binary variables representing the presence of bonds between nodes. Isomeric structures have distinct connectivity relationships, thus characterized by unique adjacency matrices. Higher-order groups are identified by traversing adjacency matrices to account for property differences in isomeric structures. To determine the molecular structures, we solve an optimization problem with binary variables in the adjacency matrix. In this problem, a feasible molecular structure must satisfy the following constraints:

1. Each node must satisfy the valency constraint for the corresponding group.
2. The adjacency matrix must preserve symmetry.
3. The graph must be completely connected.

Interested readers can refer to [49] for a detailed explanation. We solve the MILP repeatedly and add cuts to obtain unique and feasible isomeric structures. We then convert the resulting molecules to simplified molecular line entry specification (SMILES) strings [75], which enables the use of other external computational chemistry tools.

2.2.4 Extended design

This stage accounts for properties that group contribution methods are unable to predict. We incorporate more accurate and complex property prediction models to complement group contribution methods and allow for further screening. The choice of property estimation models are application-dependent, such as empirical correlations, complex performance metrics, or simulation-based

estimates. With the facilitation of SMILES strings, many external property estimation models can be adopted to further refine the solution pool.

2.3 Design targets

When selecting working fluids, it is crucial to study the various properties that affect the safety, stability, and efficiency of the cooling loop. Improving the cooling mechanism can increase the heat transfer efficiency of the cooling device. However, the choice of heat-transfer medium greatly affects the overall efficiency. New cooling fluids must have good thermo-physical properties to obtain high heat transfer coefficients. To guarantee that the newly identified liquids exhibit better heat transfer characteristics than the existing industrial coolants, the properties of a hydrofluoroether compound, HFE 7200, are used as the basis to develop property constraints. All property values are evaluated at 298K. The desired characteristics of an ideal replacement are:

- Heat conductivity (k): Heat transfer occurs at a higher rate across fluids with high thermal conductivity. Thus, high heat conductivity is critical.
- Latent heat of vaporization (h_v): Fluids with high enthalpy of vaporization can absorb more heat during phase change and are desirable in two-phase cooling systems.
- Boiling point (T_b): According to the International Technology Roadmap for Semiconductors[51], the maximum junction temperature must fall below 85^{o}C. Therefore, the boiling point needs to fall in a narrow range between ambient temperature and 85^{o}C to allow phase change.
- Viscosity (μ): Fluids with low viscosity flow faster throughout the channels, increasing the rate of convective heat transfer. Hence, a low viscosity is desirable.
- Flash point (T_f): The fluids need to have high flash point and high auto-ignition temperature to achieve thermal stability.

Table 2.1: Property targets used in the design

Property	Target Range
Viscosity at 300 K	μ ≤0.0025 Pas
Boiling Point	320 K≤ T_b ≤370 K
Melting Point	T_m ≤273 K
Latent Heat of Vaporization	$H_v \geq$ 35 kJ/mol

Additionally, the working fluids must exhibit low corrosivity to metal and polymeric materials, minimal environmental effects, and minimal toxicity. Table 2.1 lists the property targets in this model. Metrics used to estimate toxicity, safety, and environmental impact will be discussed in later sections.

Here, the flash point is used as a ranking criterion instead of a property constraint. Flash point is the lowest temperature at which a volatile compound forms sufficient vapor to ignite into air. Compounds with higher volatility tend to have higher vapor pressure at a given temperature, which corresponds to a lower boiling point. Therefore, there is a trade-off between maintaining a low boiling point and a high flash point. Compounds that satisfy the property targets listed in Table 2.1 are screened and ranked according to their flash points to produce the final solution pool. As the main molecular design algorithm does not guarantee synthesizability of the resulting molecules, Chapter 5 addresses synthesizability questions separately.

2.4 Property models and performance metric

2.4.1 Physical properties

Most of the key physical properties used in the composition design are calculated using the GC+ methods as described by Marrero and Gani [39]. These properties include: melting point (T_m),

boiling point (T_b), critical temperature (T_c), critical pressure (P_c), critical volume (V_c), standard enthalpy of vaporization at 298 K (H_v), surface tension (σ), and liquid viscosity at 300 K (η_l). Properties that are unavailable through the GC+ methods are calculated in the extended design, which include gas and liquid density(ρ_g, ρ_l), heat conductivity (k), gas phase viscosity (η_g), and flash point (T_f).

It is necessary to account for flash point estimation in the design in order to guarantee a safe operating condition. The Catoire and Naudet correlation [76] is used to calculate the flash point:

$$T_f = 1.447 {T_b}^{0.79686} {H_v}^{0.16845} {n_C}^{-0.05948} \tag{2.8}$$

where T_b is the boiling point, H_v is standard enthalpy of vaporization at 298 K, and n_C is the number of carbon atoms in a molecule.

The accentric factor, ω, is determined by the Ambrose and Walton coefficient [77] by defining reduced temperatures $T_{br} = T_b/T_c$ and $T_r = T/T_c$ calculated at 298 K:

$$\begin{aligned} \alpha &= -\ln \frac{P_c}{1.013} - 5.97214 + \frac{6.09648}{T_{br}} && (2.9)\\ &+ 1.28862 \ln T_{br} - 0.169347 T_{br}^6 && (2.10)\\ \beta &= 15.2518 - \frac{15.6875}{T_{br}} - 13.4721 \ln T_{br} + 0.43577 {T_{br}}^6 && (2.11)\\ \omega &= \frac{\alpha}{\beta} && (2.12) \end{aligned}$$

We calculate the liquid density of each molecule with the accentric factor using the Gunn-Yamada

method [78] as follows,

$$M = \sum_{i=1}^{N} n_i M_i \tag{2.13}$$

$$H_{1a} = 0.33593 - 0.33953 T_r + 1.51941 {T_r}^2 \tag{2.14}$$

$$- 2.02512 {T_r}^3 + 1.11422 {T_r}^4 \tag{2.15}$$

$$H_{2a} = 0.29607 - 0.09045 T_r - 0.04842 {T_r}^2 \tag{2.16}$$

$$V_l = \frac{RT_c}{P_c}(0.292 - 0.0967\omega) H_{1a}(1 - \omega H_{2a}) \tag{2.17}$$

$$\rho_l = \frac{M}{V_l} \tag{2.18}$$

and gas density is calculated by the cubic equation-of-state.

The heat conductivity is calculated at 298 K using the Sato-Reidel correlation [77]:

$$k = \frac{1.1051}{M^{0.5}} \left[\frac{3 + 20(1 - T_r)^{2/3}}{3 + 20(1 - T_{br})^{2/3}} \right] \tag{2.19}$$

Gas phase viscosity can be determined by the Lee et al. correlation of hydrocarbon gas viscosity [79]:

$$\eta_g = 10^{-4} K \exp(X \rho_g^Y) \tag{2.20}$$

$$K = \frac{(9.379 + 0.01607 M_w) T^{1.5}}{209.2 + 19.26 M_w + T} \tag{2.21}$$

$$X = 3.448 + \frac{986.4}{T} + 0.01009 M_w \tag{2.22}$$

$$Y = 2.447 - 0.2224 X \tag{2.23}$$

In the extended design, we include estimation of environmental impact by exporting solutions to the Estimation Program Interface (EPI) Suite developed by the U.S. EPA [80]. The EPI Suite contains

several built-in algorithms that are able to estimate a compound's environmental fate, including biodegradability, lethal concentration (LC_{50}), and atmospheric oxidation. Candidate molecules are exported to the estimation tool box as SMILES strings. A compound's biodegradability is estimated using the BioWin software provided by the EPI Suite. A compound is classified as readily degradable if the predicted probability of biodegradation is greater than 0.5. For the compounds that are not readily degradable, the primary and ultimate degradation time are estimated using Models 3 and 4 in BioWin. Henry's LC_{50} values and atmospheric reaction rate are used to measure the feasibility of candidate molecules. Although no constraints are posed on these environmental properties, estimation values are used for additional screening of the candidate molecules.

2.4.2 Performance metric

A two-phase micro-channel cooling device, where microchannel flow requires small coolant charge and easy installation, still tends to produce large pressure drop that is associated with small hydraulic diameter. Large pressure oscillation indicates that phase transition happens in the early section of the channel, which prevents the uniform flow of coolant along the fairly long channel. For the coolant to evaporate and absorb more heat along the later sections, it requires a balance between heat transfer and pressure drop. Therefore, apart from the above-mentioned physical property constraints, another criterion needs to be satisfied: the fluids must have high heat transfer coefficient and mitigate channel pressure drop. The ratio of heat transfer coefficient per unit pressure drop (HT/PD) is then used as a performance metric.

In a two-phase flow system, the measured pressure drop has three components: frictional, contractional, and accelerational pressure change. The frictional pressure drop in a boiling channel makes substantial contributions to the combined impact from contraction and acceleration, and comprises more than 80% of the total pressure drop. A previous study [81] shows that two-phase

flow patterns can be modeled by the separated flow model where each phase formulates separate mass, momentum, and energy balance. The development of the Lockhart and Martinelli frictional pressure drop correlation was based on the separated flow model where a frictional pressure drop multiplier was used [82]:

$$\left(\frac{\Delta P_f}{\Delta Z}\right)_{TP} = \phi_L^2 \left(\frac{\Delta P_f}{\Delta Z}\right)_L \tag{2.24}$$

where $\left(\frac{\Delta P_f}{\Delta Z}\right)_{TP}$ and $\left(\frac{\Delta P_f}{\Delta Z}\right)_L$ represent frictional pressure drop from two-phase flow pattern and liquid-phase only flow pattern, respectively. The two-phase frictional multiplier ϕ_L can be used as a representation of two-phase frictional pressure drop. The multiplier provides a dimensionless ratio of the two-phase frictional pressure drop to the liquid-phase only frictional pressure drop, and can be correlated in terms of the Lockhart-Martinelli parameter X as shown in the following equation,

$$\phi_L^2 = 1 + \frac{C}{X} + \frac{1}{X^2} \tag{2.25}$$

Here, the parameter C measures the interaction between two phases, which is the net result of the interactions between liquid inertia, liquid viscous force, and surface tension [83]. For a two-phase laminar flow, the parameter C is a function of the Reynolds (Re) and Weber (W) numbers

$$C = 2.16 Re^{0.047} We^{0.60} \tag{2.26}$$

$$Re = \frac{G d_h}{\eta_l} \tag{2.27}$$

$$We = \frac{v_l G^2 d_h}{\sigma} \tag{2.28}$$

where G is mass velocity, d_h is hydraulic diameter, η_l is viscosity of saturated liquid, v_l is specific volume of saturated liquid ($v_l = 1/\rho_l$), and σ is surface tension.

Drastic property changes occur in a boiling micro-channel. Therefore, Lee and Mudawar [84]

incorporated two dimensionless numbers: the boiling number, Bo, and the liquid Weber number, We, to capture how property variation affects two-phase heat transfer coefficient. The two-phase heat transfer coefficient, h_{tp}, can be calculated as

$$h_{tp} = 436.48 Bo^{0.522} We^{0.351} X^{0.665} h_{sp,f} \tag{2.29}$$

$$Bo = \frac{q"}{G H_v} \tag{2.30}$$

where q" is heat flux through the heat sink base area, H_v is enthalpy of vaporization, and $h_{sp,f}$ is the single phase liquid heat transfer coefficient, defined as a function of Nusselt number as well as liquid heat conductivity, $h_{sp,f} = \frac{Nuk}{d_h}$. The Nusselt number is a function of the ratio of channel depth to width, dependent only on the geometry of the channel.

At the same flow states, the Lockhart-Martinelli parameter X can be correlated to

$$X = \left(\frac{\eta_l}{\eta_g}\right)^{0.5} \left(\frac{1 - x_e}{x_e}\right)^{0.5} \left(\frac{\rho_g}{\rho_l}\right)^{0.5} \tag{2.31}$$

where x_e is the thermodynamic equilibrium quality, which is assumed to be a constant in the regime.

For a given mass velocity (G), channel diameter (d_h), and heat flux (q"), parameters that affect the performance metric are related to physical properties of the compounds. The metric HT/PD can be calculated as

$$HT/PD = \frac{h_{tp}}{\phi_L^2} = \frac{Bo^{0.522} We^{0.351} X^{0.665} k}{1 + \frac{C}{X} + \frac{1}{X^2}} \tag{2.32}$$

This performance metric can be generalized and adapted to micro-channel based cooling systems with various sizes. A robustness margin is required to account for the differences in channel geometry as well as other hardware properties as shown in Eq. (27). With the uncertainty margin, a robust optimization model can be formulated where maximization on the performance criterion

is sought against uncertainties in the cooling systems. The performance metric calculated using Eq. (30) can be used as the nominal value. In this work, we consider a fixed cooling system with constant micro-channel geometry. Therefore, the dominating factors in estimating heat transfer coefficient involve only physical properties of the cooling fluids. A higher value in the performance metric corresponds to better heat transfer characteristics exhibited by the fluids. This metric is used to rank the cooling performance of all candidate molecules.

2.5 Results and analysis

The goal of this work is to find replacements for the current cooling fluids. Nonfunctional groups such as $-CH_3$ and $-CH_2-$ are allowed to repeat up to 10 times each. Chlorine, phenol, amine, and amide functional group are excluded from the design because of low biodegradability [85]. Due to concerns about toxicity, organic bromine and iodine compounds are not permitted in the design. Organosilicon compounds are also excluded from the design as no accurate group contribution model is presently available to predict their physical properties.

In this work, we generate molecules by combining functional groups that permit either single or double bonds between groups. We allow up to two double bonds to be present in each molecule. The maximum number of functional groups in a molecule is limited to 15 [73]. Property constraints eliminate the feasibility of most of the functional group combinations. A 10% relaxation is applied to the physical property bounds to allow for first-order estimation error, since the GC+ method rarely exceeds an average error of 10% in first-order property estimation. In eight seconds, 308 compositions were identified. For these compositions, 944 structures were generated in 1308 seconds.

After ranking the candidates with our performance metric, we found that 96.5% of the candidate compounds have better performance than the base fluid HFE 7200. Upon further inspection, we

found that 86% of the candidate molecules were fluorinated compounds, indicating that fluorinated compounds outperform the other classes of compounds.

2.5.1 Uncertainty analysis

The GC+ method contains a set of k independent property prediction models each of which has an estimation error. The set of properties are used in the calculation of the performance metric $HT/PD = f(P_1, P_2, ...P_k)$. Property estimation errors are therefore entered in the calculation, determining the propagated error in the performance metric. A robust optimization model can be formulated to account for prediction uncertainties where optimization on the design targets is sought against property estimation errors from the GC models. In this work, we explore the impact of prediction uncertainties in a post-hoc analysis. Let e_{P_i} represent the estimation error of the i^{th} property, P_i. The propagation of error can be calculated in the same way as [86],

$$e_f = \sqrt{\left(\frac{\partial f}{\partial P_1}\right)^2 e_{P_1}^2 + \left(\frac{\partial f}{\partial P_2}\right)^2 e_{P_2}^2 + ... + \left(\frac{\partial f}{\partial P_k}\right)^2 e_{P_k}^2} \tag{2.33}$$

where $\partial f/\partial P_i$ is the local sensitivity, which is calculated by the partial derivative of the function f with respect to P_i. Following this formula, we calculate the propagated error of the performance metric for all candidate molecules. In order to examine whether the change in performance metric surpasses the noise range, we plot the performance metric versus prediction uncertainty in Figure 2.3. 91.8% of candidate molecules have an increase in the performance metric of more than 100%, doubling the cooling performance of HFE 7200, while the corresponding propagated error is lower than 20%. This observation suggests that methodology developed in our work enables the identification of high performance coolants with reliable property estimation, and the proposed performance metric is a valid ranking criterion. For candidates with higher than 20% propagated error, more detailed analysis is required to identify whether the propagated error has a positive

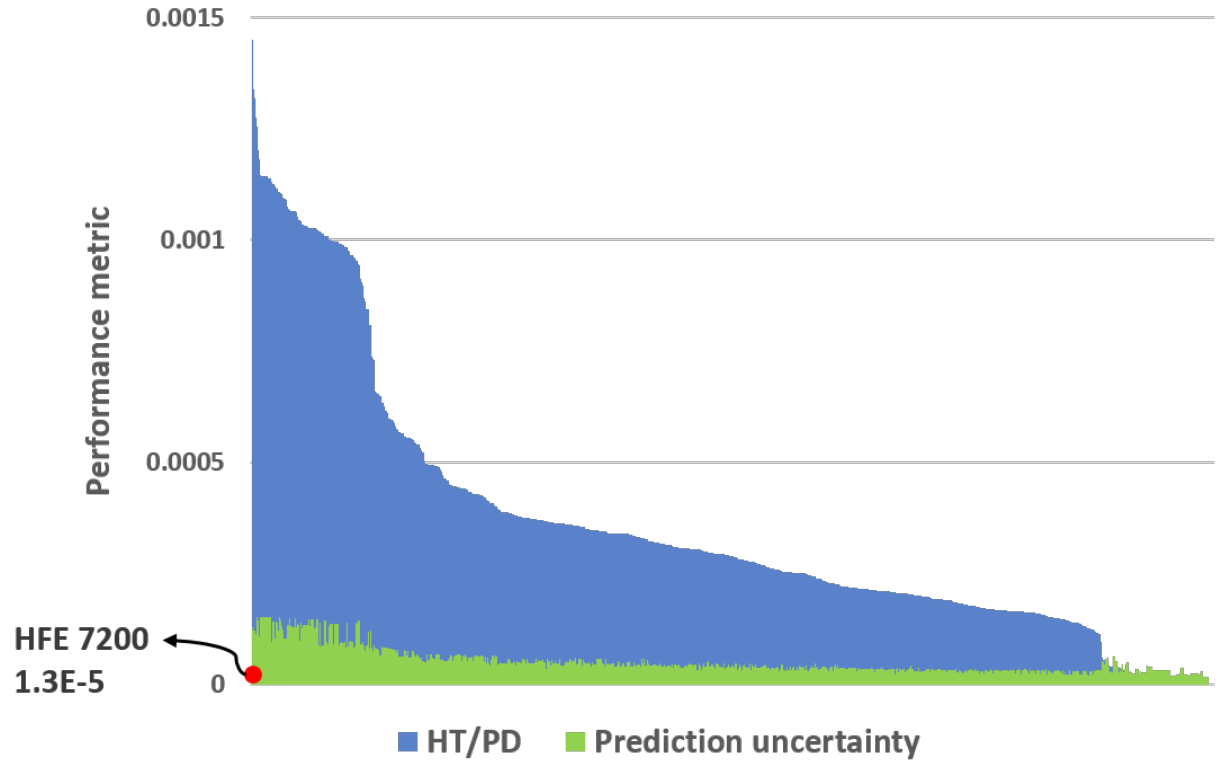


Figure 2.3: Heat transfer performance of all candidate molecules is shown in blue. The prediction uncertainty on the performance is shown in green. 84.5% of candidate molecules have a prediction uncertainty under 20%.

or negative impact on the true performance. In the following sections, analysis is carried out on candidates with propagated error lower than 20%, corresponding to a total of 797 candidate compounds. At this point, all resulting compounds have a significant increase in the predicted performance metric compared with HFE 7200.

2.5.2 Organic families

Further screening is placed on the chemical and thermal stability of the candidate molecules, which eliminates structures consisting of epoxide, peroxide, alkene, and oxygen-fluorine bond. Aromatic compounds are removed from the solution pool because of toxicity concerns. As for

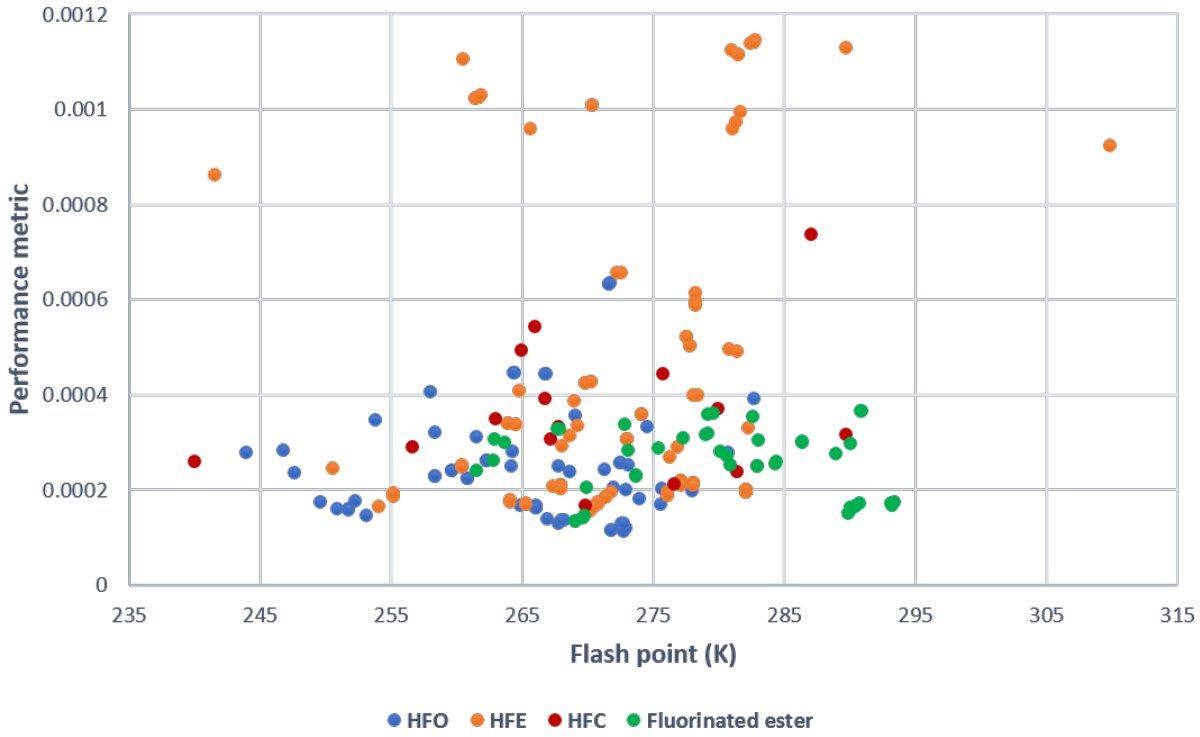


Figure 2.4: Heat transfer performance of the four organic families versus their flash points.

the fluorinated compounds in the presence of unsaturated carbons, we only keep structures with conjugated double bonds and more substituted carbon atoms to ensure stability. We identify four organic families with previous industrial applications. These organic families are: hydrofluoroethers (HFEs), hydrofluoroolefins (HFOs), esters, and hydrofluocarbons (HFCs).

Figure 2.4 plots the heat transfer per unit pressure drop metric versus the flash point of candidates from these four families. All compounds have better performance than HFE 7200. The estimated flash points of these compounds are below 300K except for one molecule, while HFE7200 is known to be non-flammable. We postulate that these compounds should be used with proper safety measures. Operation settings should be specified in the context of their application.

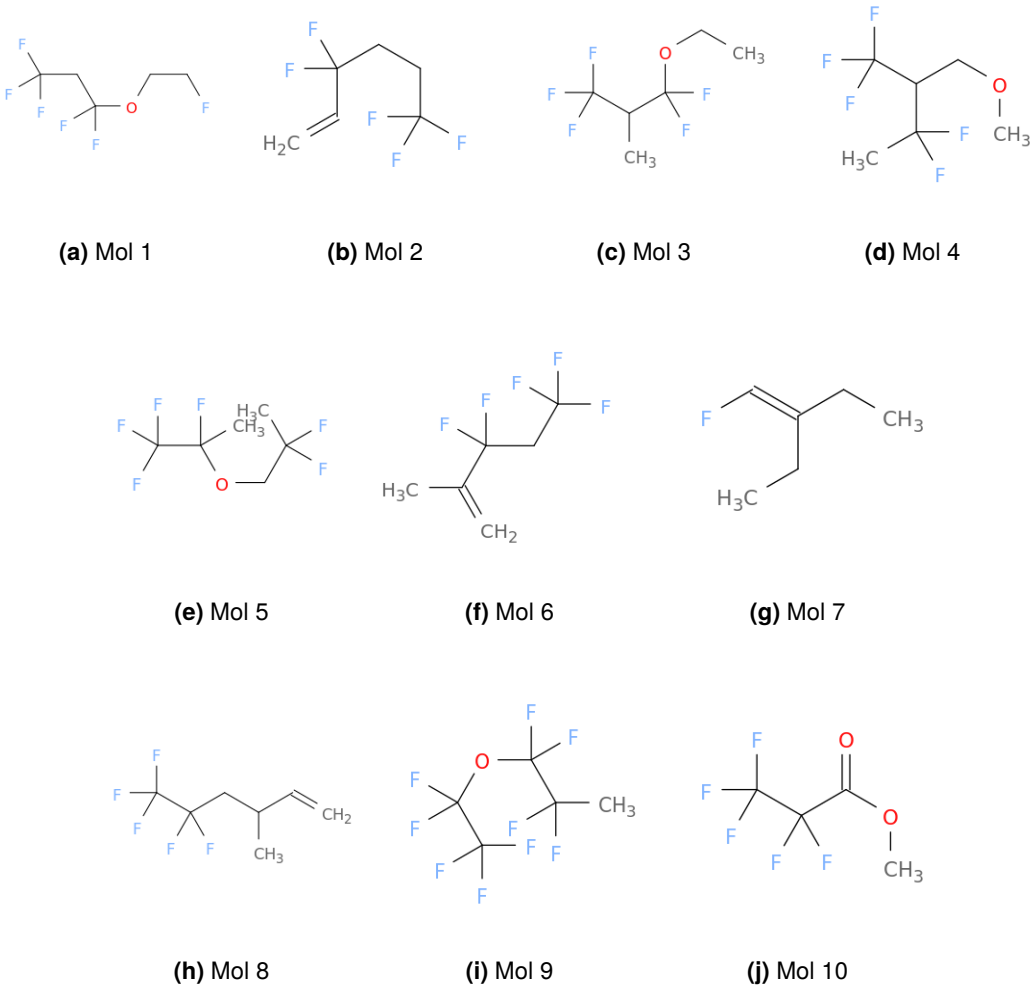
In order to determine whether the resulting molecules can be synthesized from a set of possible

compounds, we search for existing reaction templates on SciFinder [87] for each candidate molecule and the functional groups that appear in the structure. We are able to identify existing reaction templates for compounds that share similar structures to the candidate molecules, which is discussed in later sections.

Organofluorine compounds

The fluorine atom has the highest electronegativity of all elements[88], resulting in significant dipole moment in the carbon-fluorine bond and making the $C-F$ bond one of the strongest single bonds in organic compounds. As a result, organofluorine compounds usually possess high thermal and chemical stability. The $C-F$ bond is relatively short, compared to the bonds of carbon with other halogens. The short bond length, together with the small Van der Waals radius of fluorine substituent, leads to zero steric strain in polyfluorinated compounds, further advancing their thermal stability. A total of 297 fluorinated structures are identified and fall into four organic families: HFEs, HFOs, HFCs, and fluoro-esters. HFEs, HFOs, and HFCs are common heat transfer fluids in the cooling industry. Although less commonly used, fluorinated esters in cooling applications are useful as dielectric fluids in electronic devices and as heat transfer agents. A total of 16 HFCs appeared in the final solution, most of which are novel compounds and their application in the cooling industry has yet to be reported. Further inspection of these HFCs reveals that they exhibit moderate to low performance. Therefore, this work does not give detailed analysis of HFCs or their global warming potential.

For better illustration purpose, a random selection of ten fluorinated molecules as well as their performances are summarized in Table 2.2. Figure 2.5 presents explicit illustrations of their molecular structures.



(a) Mol 1 **(b)** Mol 2 **(c)** Mol 3 **(d)** Mol 4

(e) Mol 5 **(f)** Mol 6 **(g)** Mol 7

(h) Mol 8 **(i)** Mol 9 **(j)** Mol 10

Figure 2.5: Molecular structures of ten randomly selected fluorinated molecules

Table 2.2: Ten randomly selected fluorinated candidates

Organic Family	Solution No.	Canonical SMILES	Flash Point(K)	Metric (10^{-4})
HFE	1	FCCOC(CC(F)(F)F)(F)F	282.1	1.31
HFO	2	C=CC(CCC(F)(F)F)(F)F	253.0	1.70
HFE	3	CCOC(C(C(F)(F)F)C)(F)F	276.7	5.07
Segregated HFE	4	COCC(C(F)(F)F)C(F)(F)C	267.9	1.77
HFE	5	CC(COC(C(F)(F)F)(F)C)(F)F	276.7	1.33
HFO	6	CC(=C)C(CC(F)(F)F)(F)F	252.3	1.33
HFO	7	CCC(=CF)CC	264.4	1.33
HFO	8	C=CC(CC(C(F)(F)F)(F)F)C	266.0	1.33
HFE	9	CC(C(OC(C(F)(F)F)(F)F)(F)F)(F)F	270.1	1.39
Fluoro ester	10	COC(=O)C(C(F)(F)F)(F)F	269.6	2.29

Oxygenated compounds

Oxygenated compounds usually have a low boiling point due to the dipole-dipole interactions created by the polarized $C-O$ bond. Branched chain structures in these compounds, which reduce intermolecular forces, can also lead to low boiling point. Low boiling point invariably causes low flash point as these properties exhibit a linear relationship. One possible way to suppress flammability would involve the addition of water, as oxygenated compounds are usually miscible with water. If safety is of major concern, adding a small amount of water would suppress flammability and make these compounds safer.

The oxygenated compounds fall into two organic families: HFEs and fluorinated esters. Perfluorinated and partially-fluorinated ether compounds comprise the majority of the candidates and generally have good performance than the other organic groups. The top 15 best performing molecules are all stemming from the HFE family which contains a total of 179 compounds. Compared with HFCs, HFEs have zero ozone depletion potential and low global warming potential [89, 90], allowing them to be used as drop-in replacements.

Among all the candidates within the HFE family, we observed a few fluorinated molecules with branched structure exhibited promising performance. Additionally, most of the resulting HFEs are a special type–segregated HFEs–where an oxygen atom separates the fluorinated carbons from the non-fluorinated carbons. Segregated HFEs are known to have even lower environmental impact than ordinary HFEs [91].

In addition to segregated HFEs, we also identified compounds with dioxy groups which are separated by an alkylene group. Although no prior application with these candidate molecules has been reported, several reaction paths have been proposed and patented[92–94] to synthesize hydrofluoroethers with the dioxy groups, suggesting that these compounds can be manufactured. These molecules are identified as novel compounds and can potentially be commercialized as the next generation of cooling fluids.

Fluorinated esters have previously been used as solvents in industrial applications. Our work identifies 46 esters, all of which are in the range of C3-C5. Since lighter esters correspond to lower flash points, application of such coolants would require proper safety apparatus, which might increase the operating costs. Among all the identified fluorinated esters, molecules have better performance where an ester group separates fluorinated carbons from non-fluorinated carbons. The fluorinated ester presented in Table 2.2 illustrates this special structure. Further experimental validation is necessary to determine whether fluorinated esters are good replacements, as the application of fluorinated ester in the cooling industry has yet to be reported.

Non-oxygenated compounds

Our work identifies non-oxygenated compounds, including hydrofluoroolefins (HFOs), HFCs, and non-fluorinated aliphatics. HFOs have zero ozone depletion potentials and low global warming potentials. They are the fourth generation refrigerants replacing HFCs. This work identified 46

HFOs, all of which exhibit moderate to low performance. Their flash points invariably are below 273 K.

We also identified seven straight chain and branched non-fluorinated aliphatic hydrocarbons. These compounds show moderate to good performance, but they are ultimately insufficient due to their combustible nature. Although aliphatics are predicted to have better performance than HFE 7200, they are not ideal replacements due to safety concerns and their environmental impact.

2.5.3 Toxicity estimation

We estimated LC_{50} values by using the EPA's Toxicity Estimation Software Tool (TEST) version 5.1 [95] via a single multi-linear model. The chemical hazard classification category from the OPP (Office of Pesticide Programs) determines the level of hazard. Among all 944 candidate molecules, 69.3% of them are predicted to be in Category IV (non-toxic), 30.5% are predicted to be in Category III (slightly irritating), and 2 compounds are predicted to be Category II (moderately toxic). As an example, predicted LC50 values of the 10 fluorinated compounds from Table 2.2 are shown in Table 2.3. All ten molecules fall into category IV (not hazardous), suggesting safe operating conditions.

2.5.4 Kinetic stability

It is crucial to have kinetically stable fluids in the cooling system to guarantee that the coolants are not subject to chemical transformation. The highest occupied molecular orbital (HOMO)–lowest occupied molecular orbital (LUMO) energy gap is one of the most common indicators of kinetic stability. A high HOMO–LUMO gap indicates that a molecule has high kinetic stability and low chemical reactivity. We used the B3LYP method to optimize molecular geometry, and the 6-31G* basis set for frequency calculations. This choice results in high accuracy of the extensively tested

Table 2.3: 96-h log LC50 value

Solution No.	96-h log(LC50)	Safety
1	3.60	Category IV
2	4.28	Category IV
3	3.52	Category IV
4	3.11	Category IV
5	3.76	Category IV
6	4.32	Category IV
7	4.03	Category IV
8	4.43	Category IV
9	4.13	Category IV
10	3.05	Category IV

DFT methods [96]. We calculated the HOMO–LUMO gaps of the 10 fluorinated candidates from Table 2.2 using Gaussian 09W (G09) [97] as well as Jaguar from Schrodinger [98] to provide further validation. The HOMO–LUMO gap in units of Hartrees of the 10 fluorinated candidates are shown in Table 2.4. Both molecular modeling software lead to similar results. G09 identified that the ninth molecule, a hydrofluoroether compound, has the highest HOMO-LUMO gap. This HFE molecule, with a linear structure and only one non-fluorinated carbon, was among the most promising candidates. The best molecule identified by Jaguar is molecule 1. This HFE molecule is predicted to exhibit both satisfying heat transfer performance and low toxicity. The careful investigation of both HFE molecules will bring great benefits to the cooling industry. Future work could investigate using redox potential as another indicator of chemical stability.

Table 2.4: HOMO-LUMO gap of 10 fluorinated candidates using the B3LYP method and 6-31G* basis set

Software / Solution No.	G09	Jaguar
1	0.3680	**0.39123**
2	0.2675	0.28233
3	0.3609	0.38825
4	0.3389	0.27338
5	0.3600	0.35581
6	0.2634	0.2734
7	0.2661	0.26593
8	0.2693	0.2793
9	**0.3853**	0.38015
10	0.2572	0.25155

Chapter 3

Functional group selection method for GC models

3.1 Introduction

In this chapter, we continue the design of coolants for electronics cooling systems. Efficient removal of high heat flux from a compact area has received great attention to cope with the uprising miniaturization trend of electronic devices. One of the most effective cooling schemes is the microchannel-based system where evaporation of cooling fluids removes high heat flux while maintaining a safe operating temperature [54]. Existing commercial coolants such as Novec fluid HFE7100 and HFE7200 [99] are limited in their ability to remove high heat flux from compact spaces [58, 100]. Therefore, interest emerges in developing new cooling fluids with high heat removal capability. This problem is a perfect application area for CAMD and has been studied extensively in the past three decades [5, 7, 9, 10, 62, 63, 65, 101]. Organosilicon compounds generally demonstrate high heat conductivity, high electrical insulation, and low viscosity. These traits have motivated

us to consider organosilicon compounds as part of the set of candidate replacements for existing coolants.

In our past approaches to coolant design [49], we relied extensively on GC+[102] to identify potential molecular designs. We studied microchannel cooling regimes and identified a number of promising coolants that are environmentally friendly, safe to operate, and more heat transfer efficient than HFE7200 [10]. Unfortunately, GC+[102] was built with only 42 measurements involving organosilicon compounds; most Si groups were involved in no more than two measurements for most properties of interest, which makes use of this GC+ model unreliable for silicon-based structures. Warrier et al.[65] studied organosilicon coolants by developing group contribution models using 44 compounds from DIPPR 801[103]. New groups were defined to represent silicon substructures, while the majority of functional groups and their contributions remained the same as in the GC+ model. This model is limited by the small number of organosilicon measurements used and may favor other types of compounds that are present in larger numbers in the dataset used. In general, organosilicon compounds are usually excluded from CAMD studies, despite their wide use in commercial products.

To facilitate a reliable CAMD study that allows for silicon compounds, in this paper, we develop a GC model using data sets containing 747 measurements from organosilicon compounds. The first step to develop a GC model is to determine a set of functional groups that are chemically important substructures that correlate well with the property values. The individual group contributions are obtained by fitting experimental data to the property model, where the numbers of group occurrences are the predictor variables. In the GC literature, functional groups are often limited to UNIFAC groups[104] in order to facilitate utilization of packages for calculating phase equilibrium and mixture properties [105–107]. To improve predictive accuracy, more recent works are not limited to UNIFAC groups [61, 108–110]. Once selected, certain functional groups can be formed by

combining smaller-sized elementary groups. Common fragments in functional groups can lead to more than one way to decompose a molecular structure and a non-unique molecular model. For example, group —$CH_2C(=O)$— can form by combining —CH_2— and > C = O. If all three of these groups are utilized in a GC method, this would result in different representations of the same molecule and different property prediction values. Another challenge in the development of GC methods is how to develop property models that perform well on a set of unseen data while avoiding unnecessary model complexity that could lead to overfitting. This is important for properties for which there do not exist large amounts of measurements.

To address the above challenges, we propose a new method for developing a GC model. We propose a functional group selection method that deterministically decomposes each molecular structure into the smallest set of non-overlapping functional groups. Each resulting functional group is structurally simple but holds maximum information in a well-defined sense. This method serves as a consistent means to deconstruct molecules into substructures, hence avoiding model building bias. We then use the identified functional groups to build group contribution models that enable property estimation of organosilicon compounds.

We select the optimal property model by minimizing an information criterion, thus preventing overfitting, reducing root mean squared error over the training data, and increasing generalization. The black-box modeling tool ALAMO [111] is used in regression to minimize the Bayesian Information Criterion (BIC). The resulting models are subject to a test set to demonstrate their predictive power. We then apply the GC models to the coolant design problem that motivated this work.

The remainder of this chapter is structured as follows. Section 3.2 proposes our functional group selection algorithm and variable selection process using ALAMO. Section 3.3 details the results of parameter estimation using the proposed GC method. Section 3.4 presents the coolant design

problem and detailed solution analysis.

3.2 Proposed group contribution methodology

3.2.1 Background

CAMD is the problem of identifying molecular structures that satisfy a set of predefined property targets. To formulate a CAMD problem, one needs to specify a collection of building blocks representing sub-molecular descriptors, various property models that correlate these descriptors to properties, and a set of property targets to match. Property models are essential in any CAMD approach as they enable the prediction of molecular properties from structural descriptors that quantify molecular structures. The quality of molecular descriptors determines how much chemical information the descriptors can convey and subsequently be exploited by the overall CAMD approach. Therefore, the property models used in CAMD can influence the solution quality and validity, and the numerical solution techniques used.

Many property models relate structures to properties with various degrees of accuracy. Models based on atom-level *ab initio* quantum mechanics are highly accurate but computationally intensive. On the other hand, empirical models based on molecular mechanics provide fast but rough property estimation. In the middle ground of these two approaches, semi-empirical models are based on statistical regression that fits proposed models to experimental property values. They are usually easy to implement and provide property estimates with an acceptable accuracy level in reasonable time.

Group contribution methods are a class of semi-empirical methods frequently used for property estimation in CAMD. The property of the molecule can be expressed as a function of each group's

frequency in the molecule:

$$f(P) = \sum_i c_i n_i \tag{3.1}$$

Equation (3.1) shows a typical group contribution model, where P is the property value to be determined, c_i is the contribution of group i to the property, and n_i is the frequency of occurrence of group i in the molecule. The contribution parameters c_i and the transformation function f are obtained through regression on experimental property values.

Marrero and Gani proposed a three-level GC model where groups are divided into first-order and higher-order groups. In the Marrero-Gani (MG) model, the first-order groups are non-overlapping substructures that compose a molecule, whereas overlapping higher-order groups can capture the proximity effects. The property model consisting of both compositional terms (first-order) and structural terms (higher-order) has the following form

$$f(P) = \sum_{i \in F} c_i n_i + w \sum_{i \in S} c_i n_i + z \sum_{i \in T} c_i n_i \tag{3.2}$$

where F is the set of first-order groups, S represents the set of second-order groups, and T is the set of third-order groups. In[102], the MG model was updated to involve 220 first-order, 130 second-order, and 74 third-order groups.

In this work, we build a new GC model using the same functional group base as in [102] with the addition of the following nine first-order groups for organosilicon compounds: SiH_3, SiH_2, SiH, Si, SiOH, SiO, cSi, cSiH, $cSiH_2$ (cSi represents a Si group in a cyclic structure). The next section will introduce the functional group selection method that determines groups from the group base that result in the GC model. The following target properties are investigated: normal boiling point (T_b), normal melting point (T_m), critical temperature (T_c), critical pressure (P_c), critical volume (V_c), enthalpy of vaporization at 298K (H_{vap}), surface tension at 298K and atmospheric pressure (σ),

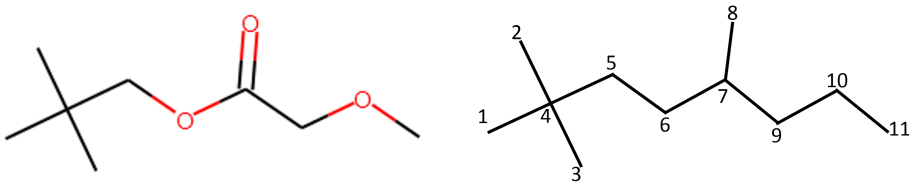


Figure 3.1: Graphical representation of a molecule

and dynamic viscosity at 298K and atmospheric pressure (μ).

3.2.2 Functional group selection method

When selecting functional groups, we assume that large groups with a higher sum of atomic weights hold more information about the molecular structure than smaller groups. Therefore, we will select the functional groups following the rule of thumb that larger/heavier groups take priority over smaller/lighter groups. This assumption is made for mathematical convenience to facilitate calculations and will be shown to provide good results. For each molecular structure in the data set, the group selection method starts with identifying all first-order groups from the group base that appear in the molecular structure and recording each group's number of occurrences. The molecular structure is then delineated as a hydrogen-suppressed graph where non-hydrogen atoms become vertices and bonds become edges. Each atom is assigned a unique number. Multi-covalent bonds are not distinguished from single bonds. Figure 3.1 gives a representation of a molecule in terms of its molecular structure and its corresponding hydrogen-suppressed graph. Using the numbered atoms, we construct a matrix where each column corresponds to a numbered atom and each row corresponds to an occurring first-order group. The (i, j) value of the matrix equals 1 if group i covers atom number j in the hydrogen-suppressed graph; it is 0 otherwise.

The molecular structure in Figure 3.1 corresponds to an atomic occurrence matrix, as shown in

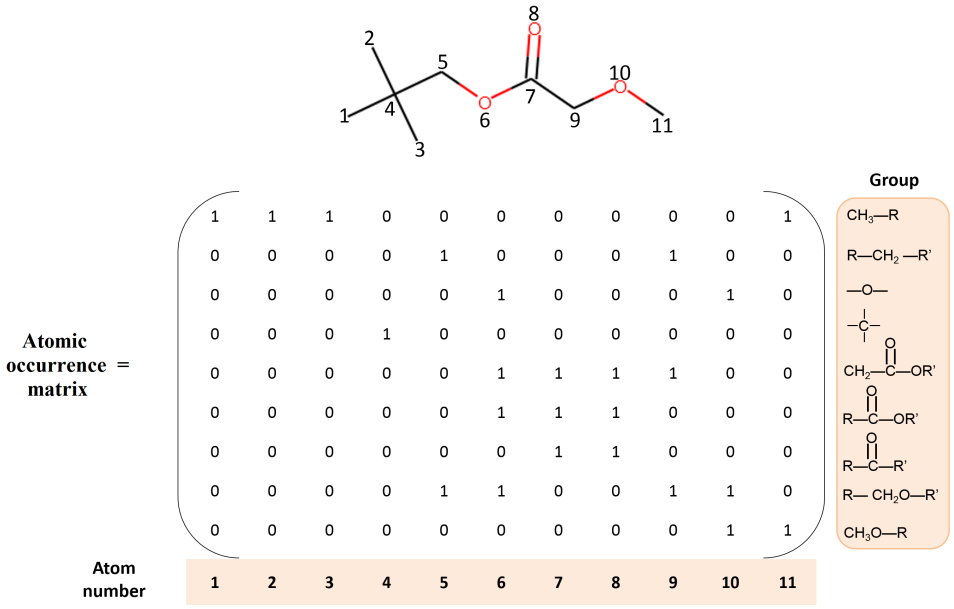


Figure 3.2: Atomic occurrence matrix of a molecule

Figure 3.2. We would like to select functional groups from the list of occurring groups such that the chosen groups are non-overlapping substructures that compose the molecule. The non-overlapping requirement can be satisfied by ensuring each atom in the hydrogen-suppressed graph is covered exactly once. This constraint serves as a checkpoint on the atomic balance and molecular weight balance. Therefore, given a molecular structure and its atomic occurrence matrix with dimension $K \times M$, we need to select a subset of rows such that the summation of each column equals 1. This

problem can be formulated as the following mixed-integer linear program (MILP):

$$\begin{aligned} \max \quad & w_1\alpha - w_2 \sum_{i=1}^{K} n_i x_i \\ \text{s.t.} \quad & \alpha \geq x_i l_i, \ i = 1, ..., K \\ & \sum_{i=1}^{K} z_{ij} x_i = 1, \ j = 1, ..., M \\ & x_i \in \{0, 1\} \end{aligned}$$

In this model, the binary variable x_i denotes the selection of group i in the occurring group list, the corresponding descriptor frequency is represented by n_i, and z_{ij} denotes the (i, j) entry from the atomic occurrence matrix. An auxiliary variable α is introduced to identify the largest individual group. The size of each group is indicated by l_i. In this formulation, we aim to identify the smallest set of functional groups from a molecular structure where each group pertains as much information as possible. The objective function is a weighted sum of two optimization goals: minimize the number of functional groups in a set and maximize the size of each selected group. Using scalars w_1 and w_2, this formulation reduces the two-objective to a single-objective optimization problem. If a functional group exists alongside its fragments, the objective function ensures the group itself is selected instead of the substructure fragments. The weights $w_1 > w_2 > 0$ are tunable hyperaparamters that determine how we prioritize selecting bigger groups over minimizing the number of selected functional groups. Varying w_1 and w_2 will lead to a number of optimal solutions that form a Pareto frontier. For any given choice of w_1 and w_2, this formulation produces a unique selection of groups generated by a specific solver. In this study, w_1 and w_2 are empirically set to be 5 and 1, respectively. The constraint set ensures that the largest group is identified while maintaining that non-overlapping descriptors have complete atom coverage. We use CPLEX 12.9 [112] to solve the MILP and obtain a unique set of groups for each molecule. The example shown in Figure 3.2 results in the final group

Property	f(P)
Boiling point (K)	$\exp(T_b/T_{b0})$
Melting point (K)	$\log(T_m/T_{m0})$
Critical temperature (K)	$\exp(T_c/T_{c0})$
Critical pressure (bar)	$1/\sqrt{P_c/P_{c0}} - P_{c1}$
Critical volume (cc/mol)	$V_c - V_{c0}$
Enthalpy of vaporization (kJ/mol)	$H_{vap} - H_{vap0}$
Surface tension (N/m)	$\sigma - \sigma_0$
Dynamic viscosity (mPa s)	$\log(\mu/\mu_0)$

Table 3.1: Transformation function for each property

set [3 ($-CH_3$), 1 (> C <), 1 ($-CH_2-$), 1 ($-CH_3O$), 1 ($-CH_2COO-$)].

3.2.3 Property model and data pre-processing

The property estimation models $f(P)$ have the following form

$$f(P) = \sum_{i \in F} c_i n_i + \sum_{i \in S} c_i n_i + \sum_{i \in T} c_i n_i \tag{3.3}$$

where first-order groups are determined using the group selection method and higher-order groups are subsequently determined by using first-order groups as building blocks. The left-hand side $f(P)$ is a function of the target property P that provides good extrapolation over a wide range of experimental data. The function $f(P)$ is determined through the monotonic transformation of data to minimize variance and reduce skewness in the data distribution. The eventual transformation functions are selected from a list of potential functions including linear, $\log x$, $\exp x$, $\sqrt{x}$, $\sqrt[3]{x}$, and $\frac{1}{x}$. The target properties and their corresponding transformation functions are listed in Table 3.1.

To estimate property model parameters, we collect experimental data of pure components from various organic families from the following databases: DIPPR801[103], Knovel[113], and the EPI suite[80]. For substances with multiple experimental measurement entries, we utilize the

measurement average. Table 3.2 details each property data set in terms of the number of components in each organic family. Each property data set is subject to the corresponding transformation function in Table 3.1, followed by an outlier detection procedure to filter out data points with a z-score higher than two in the transformed data distribution. Structures with complex ring structures tend to exhibit steric hindrance, adding nonlinear features to properties that we cannot accurately capture using linearly additive GC methods. Therefore, any components with more than three ring structures are eliminated from the data set. As a result, the proposed GC models are most suitable for molecules with no more than two ring structures. Post-processed experimental data is split into ten consecutive folds for cross-validation. Each fold is used once as the test set while the remaining nine folds form the training set. Repeated training and test data split follows a stratified manner. In other words, property data is partitioned into subgroups in advance based on how many standard deviations a data point resides from the mean. Training and test split occur in each subgroup where 9/10 of the population fall into the training set while the rest belong to the test set. Cross-validation is carried out to determine the optimal universal parameters T_{b0}, T_{m0}, T_{c0}, P_{c0}, V_{c0}, H_{vap0}, σ_0, μ_0 in the transformation functions, as shown in Table 3.3. Values are chosen to minimize the average cross-validation error between the estimated and true property values.

Organic family	T_b	T_m	T_c	P_c	V_c	H_{vap}	σ	μ
Hydrocarbons C_nH_m	1459	636	287	287	287	547	217	63
Oxygenated $C_nH_mO_x$	2724	1540	273	273	273	730	700	236
Nitrogenated $C_nH_mN_x$	837	471	74	73	73	287	135	43
Sulfur containing $C_nH_mS_x$	199	116	31	31	31	81	56	7
Halogen								
Fluorinated $C_nH_mF_x$	135	62	9	9	8	42	13	12
Chlorinated $C_nH_mCl_x$	350	180	30	30	30	122	91	31
Brominated $C_nH_mBr_x$	160	102	8	8	8	44	46	10
Iodinated $C_nH_mI_x$	46	44	5	5	5	38	26	0
Silicon containing $C_nH_mSi_x$	269	92	59	55	62	128	36	46
Polyatomic functional groups	2262	2035	73	73	74	462	279	66
Total	8445	5280	860	852	852	2473	1599	514

Table 3.2: Number of measurements by organic family used in parameter estimation of different properties

Constant	Value
T_{b0} (K)	295.51
T_{m0} (K)	233.90
T_{c0} (K)	295.51
P_{c0} (bar)	2.63
P_{c1} (bar$^{-0.5}$)	0.17
V_{c0} (cc/mol)	57.34
H_{vap0} (kJ/mol)	13.49
σ_0 (N/m)	19.97
μ_0 (mPa s)	1.38E-2

Table 3.3: Universal constants used in the property transformation functions

3.2.4 Parameter estimation

Parameter estimation is carried out via regression on the training data set for each property of interest. Using (3.3), we fit the transformed target properties to a linear function of group occurrences. The determination of contribution coefficients (c_i's) in the property model is carried out by the model selection tool ALAMO [111] (Automated Learning of Algebraic Models). ALAMO learns algebraic functions by building a low-complexity surrogate model through best subset selection on regressors. Several model fitness metrics can be employed to balance the bias-variance trade-off. In this work, the Bayesian information criterion (BIC) is used as the optimization objective to find a balance between the goodness of fit and overfitting:

$$BIC = \frac{\sum_{m=1}^{N}(P_m - \hat{P}_m)^2}{\hat{\sigma}^2} + k \log(N) \tag{3.4}$$

In this equation, N is the number of data points in the training set, k denotes the number of non-zero regressors, and P_m and $\hat{P}_m$ respectively represent the actual and predicted property value for measurement m. $\hat{\sigma}^2$ is an estimation of the residual variance calculated by ALAMO. ALAMO assigns a contribution coefficient to each functional group by minimizing BIC. Groups with non-zero

coefficients end up participating in the GC model.

3.3 Validation of derived GC method

As in (3.4), in this section, we will use P_m and $\hat{P}_m$ to represent, respectively, the experimentally measured property values and the values predicted with the derived GC models at the m^{th} measurement point. The property mean will be represented with $\overline{P}$. In order to quantify the goodness of fit of the developed GC models, we will adopt the following metrics: R-squared (R^2), root mean squared error (RMSE), average absolute deviation (AAD), and average relative deviation (ARE%). These metrics are defined as:

$$R^2 = 1 - \frac{\sum_{m=1}^{N} \left(P_m - \hat{P}_m\right)^2}{\sum_{m=1}^{N} \left(P_m - \overline{P}\right)^2} \tag{3.5}$$

$$RMSE = \sqrt{\frac{\sum_{m=1}^{N} \left(\hat{P}_m - P_m\right)^2}{N}} \tag{3.6}$$

$$AAD = \sum_{m=1}^{N} \frac{|P_m - \hat{P}_m|}{N} \tag{3.7}$$

$$ARE\% = \frac{1}{N} \sum_{m=1}^{N} \frac{|P_m - \hat{P}_m|}{P_m} \times 100 \tag{3.8}$$

R^2 measures to what extent the observed variation in the dependent variable can be explained by the model's inputs. An R^2 value of 1 indicates perfect fitting accuracy. RMSE measures how far away the data is around the line of best fit. AAD measures the average distance between the true values and predicted values. ARE% expresses the average of the absolute error of the predictions with respect to the true values in percentage.

We use 411 functional groups to compose the functional group base, including both first-order

and higher-order groups, to consider the complete set of molecular structures. The group selection procedure decomposes molecular structures into a subset of the 411 groups, after which ALAMO determines the final set of "best groups" through best subset selection. Eventually, a total of 315 functional groups are selected to model all target properties. To achieve optimal fitting results, we use the full data set to build the property models. Table 3.4 presents the fitting status of the final property models using the full data set. The model size indicates the number of parameters involved in each GC model. A parity plot of each target property is provided in the Appendix. Most property models achieving a satisfactory correlation coefficient R^2 proves the viability of the proposed GC method for its predictive power. Visual observation on the parity plots indicates good model fitting status where most of the data are fitted to an adequate degree of accuracy. The residual error plot of each property is also provided in the Appendix. All histograms exhibit Gaussian bell shape curves, suggesting evenly distributed deviations of training data about the regression line. The spikes around 0 indicate that the proposed GC models perfectly fit property values. Further analysis shows that most of the presently tested compounds deviated by no more than one RMSE to the true value. Promising model performance is further corroborated by low ARE% values for most target properties. ARE%s are relatively high for melting point and viscosity. Difficulty in melting point modeling is mainly due to the strength of the crystal lattice, which is primarily a function of intermolecular forces, molecular symmetry, and conformational degrees of freedom of a molecule [114], all of which are not explicitly modeled in the proposed linearly additive GC method. On the other hand, dynamic viscosity suffers from a small data size. Therefore, extra care should be taken when using GC models for these two properties. A complete list of participant functional groups and their group contribution coefficients are provided in the supplementary material.

This work aims to develop reliable models with good predictive power. Therefore, it is vitally essential to subject the model to an unseen data set. For this purpose, we performed 10-fold

Property	R^2	ARE%	AAD	RMSE	Model size
T_b (K)	0.96	2.65	11.64	15.69	226
T_m (K)	0.90	7.19	22.26	26.94	186
T_c (K)	0.96	2.02	12.38	20.31	222
P_c (bar)	0.95	4.09	1.53	2.48	136
V_c (cc/mol)	0.99	3.10	11.05	17.45	148
H_{vap} (kJ/mol)	0.95	5.79	3.02	4.31	163
σ (N/m)	0.89	5.48	1.62	2.15	165
μ (mPa s)	0.91	31.05	0.45	0.78	181

Table 3.4: Fitting metrics of the proposed GC model on the full data set

cross-validation, where 1/10 of the data was taken out as a test set while the remaining was used as the training set to fit the property models. Fitted property models were evaluated on the test set. Viscosity was not subject to cross-validation due to the small data size. Table 3.5 summarizes the cross-validation results; each statistical metric is associated with a 95% confidence interval obtained through multiple cross-validation calculations—most properties under investigation exhibit high R^2 and Q^2 on both training and cross-validation sets. The small difference between the two metrics is clear proof of the reliability of the present GC model for its predictive power. It is worth noticing that most of the GC models can be expressed using models with less than 200 variables, as shown in the *Model Size* column in Table 3.4. This observation demonstrates the proposed method's ability to identify a small predictive model.

Property	**Training R^2**	**Training AAD**	**Test R^2 (Q^2)**	**Test AAD**
T_b (K)	0.96 (±0.05E-2)	11.68 (±6.22E-2)	0.95 (±6.33E-2)	11.47 (±0.57)
T_m (K)	0.90 (±0.04E-2)	22.31 (±5.71E-2)	0.88 ±0.56E-2)	21.80 (±0.55)
T_c (K)	0.97 (±0.12E-2)	12.14 (±0.11)	0.79 (±5.92E-2)	29.03 (±5.65)
P_c (bar)	0.95 (±0.33E-2)	1.59 (±2.06E-2)	0.93 (±0.04)	1.54 (±0.20)
V_c (cc/mol)	0.99 (±0.02E-2)	10.79 (±0.11)	0.98 (±0.54E-2)	13.17 (±1.06)
H_{vap} (kJ/mol)	0.93 (±0.15E-2)	3.17 (±2.52E-2)	0.90 (±0.02)	3.25 (±0.24)
σ (N/m)	0.88 (±0.14E-2)	1.66 (±0.92E-2)	0.85 (±0.02)	1.62 (±8.78E-2)

Table 3.5: Cross-validation results

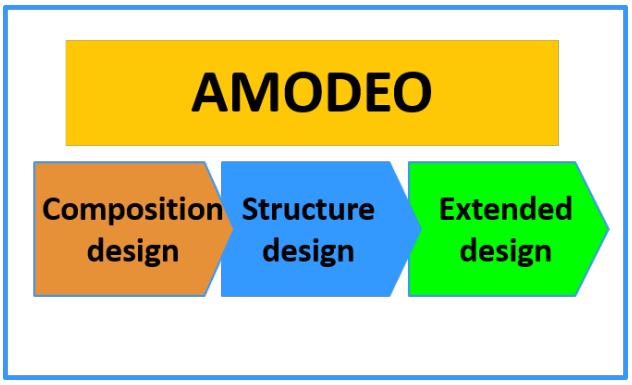


Figure 3.3: AMODEO framework

3.4 Coolant design

This section addresses the problem of designing coolants for electronic devices. We will incorporate the new GC method into the AMODEO CAMD methodology [49]. AMODEO utilizes a decomposition scheme that allows efficient exploration and optimization in the chemical design space in a computationally efficient manner. As Figure 3.3 shows, AMODEO first solves for molecular compositions. This is done by identifying all feasible combinations of first-order groups that match the design targets without establishing chemical bonds between groups to form a complete molecular structure. At this stage, target property ranges are widened in order to compensate for prediction errors in the GC method used. The subsequent structure design stage adds chemical bonds to each feasible composition to identify unique structures and differentiate among isomers. This stage includes higher-order groups as correction terms to property estimation. Finally, the extended design stage incorporates additional property models that are not available through GC to further refine the solution pool.

The design objective for the problem at hand is to identify organosilicon cooling liquids that enhance the heat removal performance in a microchannel two-phase cooling system. We seek replacement coolants with better heat transfer properties than the commercial coolant Novec

Boiling point	320 K $\leq T_b \leq$ 370 K
Melting point	$T_m \leq$ 273 K
Latent heat of vaporization	$H_{vap} \geq$ 35 KJ/mol
Dynamic viscosity at 298 K	$\mu \leq$ 0.0025 Pa s

Table 3.6: Property targets used in the design

HFE7200 and develop a set of property targets listed in Table 3.6. This set of design targets was identified in our previous work [10]. In a two-phase cooling system, coolant vaporizes to absorb sufficient heat from the electronics to maintain a chip temperature below 85°C [51]. Low viscosity allows fluids to flow throughout the microchannel with less resistance to enhance convective heat transfer. Our design targets reflect these desired characteristics.

Once property targets have been identified, we widen property target intervals by 15% in order to compensate for estimation errors from the GC model. Widening bounds allows us to identify structures that reside in the vicinity of design targets. The effect of prediction errors will be further examined using uncertainty analysis on the derived molecules.

The metric developed in Chapter 2 is used here to evaluate the performance of microchannel two-phase cooling fluids:

$$\frac{HT}{PD} = f(H_{vap}, \sigma, \mu, T_{cri}, P_{cri}, T_b) \tag{3.9}$$

The metric is formulated as the ratio of heat transfer coefficient (HT) and pressure drop (PD) to maximize heat transfer rate while minimizing pressure drop across microchannels to guarantee phase change occurs throughout the cooling system.

We limit the chemical design space to solely contain organosilicon compounds. AMODEO identified 656 compositions in the composition design stage. For these compositions, 3568 organosilicon structures were generated and subsequently subjected to the extended design stage for further screening. All structures identified at this stage surpass the heat transfer performance of

Novec HFE7200.

To assess the effect of prediction uncertainty on the cooling performance, an error propagation analysis is carried out in the same way as in Ku [86] using local sensitivities with respect to all properties. The prediction uncertainty is calculated as follows:

$$e_f = \sqrt{\left(\frac{\partial f}{\partial p_1}\right)^2 e_{p_1}^2 + \left(\frac{\partial f}{\partial p_2}\right)^2 e_{p_2}^2 + \ldots + \left(\frac{\partial f}{\partial p_\kappa}\right)^2 e_{p_\kappa}^2} \tag{3.10}$$

Here, we take partial derivatives of the performance metric function f with respect to each target property p_j $(j = 1, \ldots, \kappa)$ involved in the expression and use AAD to represent the estimation error of each property. Using this formula, we calculate the propagated error of the performance metric for all candidate molecules.

Figure 3.4 shows the cooling performance of the candidate molecules (blue area) compared with HFE7200 (red dot). The majority of the resulting compounds have a drastic improvement in heat transfer performance. One of the top-performing compounds achieved a 1500% enhancement in cooling performance compared to the base material. The result from the error propagation is presented in the same plot using green shades to quantify the range of uncertainty of the predicted performance improvement. The range of uncertainty is insignificant compared to the actual performance improvement. This observation suggests that we have identified a pool of organosilicon cooling liquids with performance exceeding one of the market's fluids.

In the extended design step, we carry out an investigation on the operational safety and dielectric properties of the identified candidate coolants. New cooling fluids should exhibit a flash point at least higher than room temperature and low toxicity to ensure a safe operating environment. Using the flash point estimation model from Catoire and Naudet [76], we find that 85% of the candidate liquids have flash points higher than 300K. Toxicity was estimated using the EPA's Toxicity Estimation

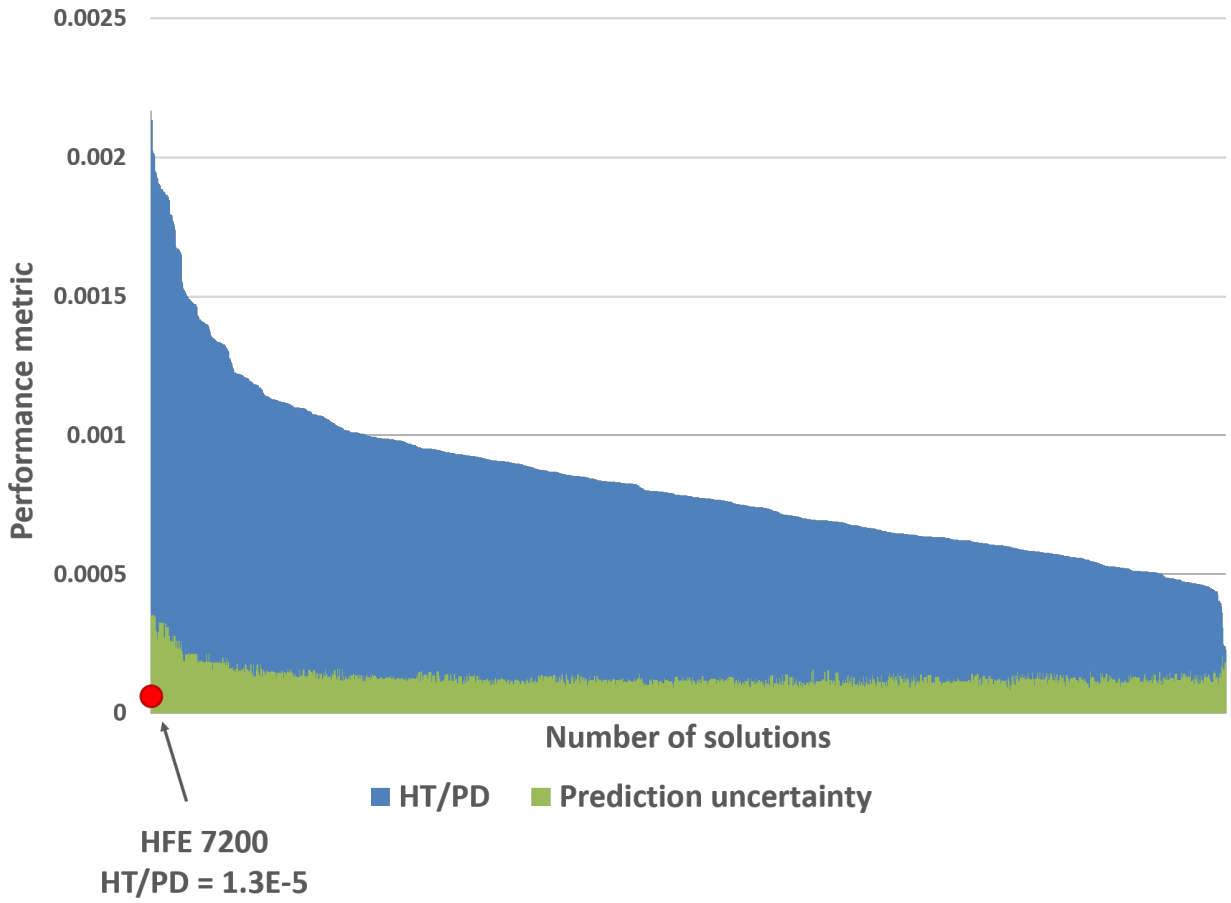


Figure 3.4: Heat transfer performance of candidate molecules is shown in blue. The prediction uncertainty on the performance metric of each compound is shown in green.

Software Tool (TEST), version 5.1 [95]. We use a multilinear single model to predict 96-hour lethal concentration (LC_{50}). Using the chemical hazard classification category from the OPP (Office of Pesticide Programs), we find that 83.85% of the candidates fall into category IV (not hazardous), while the remaining 16.15% of the candidate coolants fall into category III (harmful if inhaled). The results from both investigations demonstrate that the oragnosilicon compounds identified are very likely to result in safe operating conditions.

Even for non-direct-to-chip cooling systems, such as the microchannel cooling system addressed in this study, it is desirable for cooling fluids to possess low dielectric constants in order to maintain electrical resistivity over an extended period of time and minimize potential damage to electronic equipment. Following Kirkwood [115], we evaluate the dielectric constants of our top-10 candidates with the best heat transfer performance. We assume that the refractive indices of all candidate molecules fall within the typical range 1.33-1.50, and utilize the Lorentz-Lorenz equation to calculate molar polarizabilities. Dipole moments are obtained using the B3LYP/6-31+g(d,p) method in Gaussian09 [116]. The dielectric constants of the top-10 molecules range from 1.67 to 2.23, lower than the reported dielectric constant of HFE 7200 [117]. This preliminary analysis suggests that our identified coolants demonstrate good electrical resistivity, adding more confidence to their operational safety. Therefore, they should be considered seriously by industry.

Finally, we compare the performance of organosilicon compounds to non-silicon structures identified from Chapter 2. Figure 3.5 plots the heat transfer performance versus flash point of both silicon containing and non-silicon containing candidates. In general, the addition of a silicon group not only enhances the heat transfer efficiency but also improves the operational safety by providing a higher flash point. Additionally, the toxicity estimation of silicon containing structures yields a higher percentage of benign compounds compared to the candidate coolants from Chapter 2, as summarized in Table 3.7. These observations further consolidate that organosilicons can be an ideal

Organic family	Category I	Category II	Category III	Category IV
Organosilicons	0%	0%	16.2%	83.8%
Non-organosilicons	0%	0.2%	30.5%	69.3%

Table 3.7: Comparison of LC50 values between organosilicons and non-silicon containing compounds

replacement of current commercial cooling fluids.

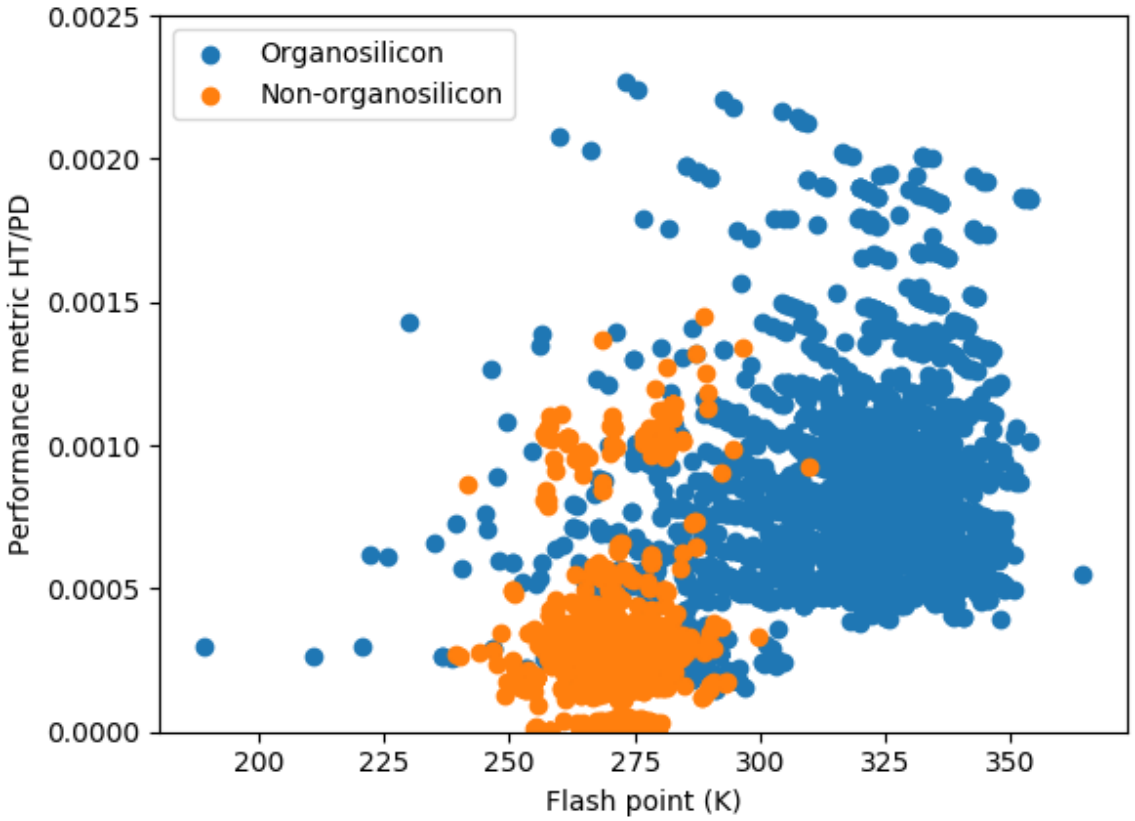


Figure 3.5: Comparison between organosilicon coolant and non-organosilicon coolants. Organosilicons generally have higher flash points and higher heat transfer efficiency compared to non-silicon containing structures.

Chapter 4

Derivative-free optimization for chemical product design

4.1 Introduction

Computer-aided molecular design (CAMD) aims to efficiently explore the diverse chemical design space and identify suitable chemical structures with desirable properties. Efficient search algorithms and accurate property prediction models are the two key components in CAMD frameworks. While search based on algorithms can generate molecular structures from a small set of submolecular building blocks, most CAMD problems utilize optimization techniques and are formulated as mixed-integer nonlinear programming (MINLP) models to explore the complete molecular design space. The complexity of property prediction models together with the vast continuous/discrete molecular solution space can lead to MINLPs with high degree of nonlinearity. These problems are difficult to solve using standard MINLP solvers even for design spaces containing small molecules. Several methodologies have been developed to couple decomposition techniques with MINLP formulations

to solve the complex molecular design problems as a series of relatively easier sub-problems [49]. Harper et al.[60] proposed a structured multi-level generate and test approach where the complexity of property prediction models increases at each stage. A decomposition scheme is utilized in [12, 20] to solve the solvent mixture and crystallization solvent design problem. Samudra and Sahinidis [49] have developed a decomposition scheme that demonstrates high computational efficiency when solving CAMD problems with many functional groups. Austin et al. [118] provide a review of CAMD frameworks and solution methods.

Applications of derivative-free optimization (DFO) algorithms in CAMD appeared as early as Venkatasubramanian et al. [15, 119], who used genetic algorithms (GAs) to design polymers and fuel additives. Other chemical product design applications utilizing GAs can be found in integrated solvent and process design [120], ionic liquid design [121], and solvent design [30, 122, 123]. Marcoulaki and Kokossis utilized simulated annealing to solve the solvent design problem for liquid/liquid extraction processes [124, 125]. Gebreslassie and Diwekar studied the solvent selection problem using ant colony optimization method [126]. A CAMD approach is proposed in [127] for the optimal design of absorbents for adsoprtion of natural gas fracking waste and is solved using efficient ant colony optimization method. All these papers utilize stochastic DFO algorithms. Although stochastic search procedures can be used as a fast means to generate candidate solutions, the recent work of Rios and Sahinidis [128] suggests that deterministic algorithms perform better in practice for a large collection of problems. Additionally, each of these papers utilized a single DFO approach. Recently, Austin et al. [129] utilized many DFO methods to efficiently search the vast space of solvent designs based on the key observation that the number of degrees of freedom is very small in the space of properties important to the application and that, once the degrees of freedom are fixed, numerical simulation/optimization determines values for all other variables. A portfolio (i.e. collection) of DFO algorithms was used to solve the problem. Using a portfolio of

DFO algorithms can increase the possibility of identifying good solutions. The idea of using a portfolio of DFO algorithms has otherwise remained unexplored in chemical product design.

As more reliable and accurate property simulation models become available, employing property simulators in product design will enable engineers to account more accurately for the effect of design variables on product characteristics that might otherwise be difficult or impractical to measure. For example, Monte Carlo simulation is widely used to estimate thermal conductivity and polymeric properties [130], computational fluid dynamic models have been used to accurately represent industrial processes [131], and density functional theory finds increasing applications in the prediction of complex material behavior [132]. While combining these simulators with algebraic optimization solvers is a formidable task, DFO can naturally integrate with complex simulators at a minimal effort from the design engineer. Thus, there exist vast opportunities for the use derivative-free optimization in chemical product design practice.

This chapter provides a short survey on recent developments of derivative-free optimization algorithms and their applications. We aim to show why applications in this area of research are pertinent and why researchers ought to fill the gap in the literature concerning the use of portfolios of DFO algorithms in chemical product design problems. We describe algorithmic approaches for solving DFO models, prior reviews in literature, and a variety of available software for DFO algorithms. We also provide a case study in order to demonstrate the application of a portfolio of DFO solvers in chemical product design.

4.2 Derivative-free optimization

In derivative-free optimization, algorithms seek optimal solutions of an optimization problem where the objective function and/or constraints are not necessarily algebraically available. DFO algorithms search the domain of independent variables, evaluate the objective function at one or several points,

interpret the results based on a wide variety of approaches, and choose one or more new points for evaluation or terminate if a convergence criterion was reached or a time limit exceeded. The objective function and constraints can be computed through a call to a closed-source or very complex code that is treated as a black box. Thus, the literature on derivative-free optimization often uses the term *optimization over black box* interchangeably with *derivative-free optimization*. The closely related term simulation optimization (SO) is typically reserved for derivative-free optimization when noise or variability exists in the simulation outputs [133].

Many DFO algorithms exist and can be classified into the following categories:

- Model-based or direct. Model-based algorithms fit a surrogate model to the objective values collected over the space of independent variables. This surrogate model provides derivative approximations and guides the search. These methods are suitable for computationally expensive applications, as a way to deduce the underlying relationships between the input variables and the objective function without carrying out expensive simulations. The theoretical development of model-based algorithms revolves around the choice of surrogate models to estimate the unknown underlying models as well as the criteria used to determine the next sampling point. A review of model-based box-constrainted DFO problems can be found in Forrester et al. [134]. Direct search algorithms search a set of points around a current trial point, and use the information collected in order to determine search directions. These methods are popular for their simplicity and flexibility. Under certain assumptions on smoothness and differentiability, convergence to a stationary point is guaranteed [135–137]. Kolda et al.[137] provided extensive review on direct-search methods for derivative-free optimization.
- Local or global. Local search algorithms search in the vicinity of a current trial solution, while attempting to find a direction with improved objective value within a local subspace. Direct

search methods that are local in nature include mesh adaptive direct search (MADS) [138], Nelder-Mead simplex based algoritm [139] and pattern search method [136, 137]. Model-based local search method include trust-region methods [140] and implicit filtering [141]. These algorithms aim to identify a point of local optimality. Major disadvantages of local direct search are the high dependency on the initial point, large number of function evaluations and likelihood of getting trapped in local optima. Multi-start approaches can be used to increase the probability to converge to global optimum. On the other hand, global algorithms refer to methods that do not require a starting point. These algorithms explore the entire search domain in ways that balance local refinement with global exploration. Popular direct search methods in the context of global-search include the DIRECT algorithm [142] and multilevel coordinate search [143]. Global model-based search algorithms aim to construct surrogate models for the entire search space. With more function evaluations, global search algorithms can refine the search domain by examining many areas of the feasible region. Both local search and global search can converge to a local stationary point under certain assumptions but global optimality is not guaranteed unless the search is "dense" [135, 143, 144].

- Deterministic or stochastic. When they have the same starting point, deterministic algorithms will follow a fixed set of operations, evaluate the same set of sample points, and arrive at the same final solution. These algorithms include several variants of the MADS algorithm [138, 145], implicit filtering [146], and branching-based algorithms such as SNOBFIT [147]. Stochastic algorithms add randomness to the search, often following a probability distribution that chooses a new solution over the previous one. These methods include algorithms such as CMA-ES [148], particle swarm algorithms, and GAs. Because they lack a deterministic termination criterion, stochastic algorithms may require a large number of evaluations.

The review by Rios and Sahinidis [128] and the textbook by Conn et al. [149] provide more

details on the theory behind derivative-free optimization. The recent increased interest in the DFO field is largely due to advances in related software implementations. Rios and Sahinidis [128] provide a systematic comparison of software implementations on a large collection of test problems. The authors found that no single solver is sufficient to solve all problems. Additionally, all solvers can provide the best solution possible for at least some of the problems tested. Austin et al. [19] recently studied the chemical mixture design problem using derivative-free optimization, and cross-compared 27 DFO solvers on a proposed decomposition formulation. The authors concluded that a portfolio of DFO algorithms is efficient at solving mixture design problems. Among the solvers tested, global model-based methods provided better solutions. Findings from these two papers suggest that using a set of solvers can lead to better solutions in comparison to using a single solver. Recently, the use of a portfolio of DFO solvers has found considerable success in engineering design [150] and algorithm tuning [151].

To encourage more applications of a portfolio of DFO solvers in the CAMD field, we will demonstrate how this approach can be used in a product design problem. In the following section, we introduce a polymer design problem involving a black-box simulator. We use a portfolio of DFO algorithms to search the multi-dimensional design space and identify polymeric configurations that can match a target rheological behavior.

4.3 Design of polymer structure and flow

Every commercially available plastic object goes through the melting stage during manufacturing to be shaped into the final product. Polymer melts are entangled macromolecules that exhibit time-dependent viscoelastic behavior. It is important to analyze and quantify the dynamic viscoelastic behavior in order to control the rheology of the polymeric materials within certain range for smooth process operation. Normally, an oscillatory test is carried out to measure the complete rheological

responses of a polymer melt. This process applies a sinusoidal stress at different frequencies to a polymer melt and creates deformation, after which the material relaxes in response to the external stimuli. The resulting strain is measured during the relaxation response.

The rheology of polymers melt is very sensitive to the polymeric configuration and structure. For example, heavier polymers (higher molecular weight) correspond to higher viscosity, and branching structures affect the elasticity of the polymer melt. In this section, we are interested in the following problem: Given a polymer melt and a quantitative prediction of its rheological behavior, identify all feasible polymer melt architectures of which the rheological time-dependency resembles that of the given target melt.

The motion of a polymer molecule is assumed to follow a tube model which restricts the movement of a polymer within a virtual tube formed by the surrounding entangled polymers. We use a black-box simulator to calculate the rheological responses of the polymer melts using the extended tube model [152, 153]. We consider the following design variables for each component in the polymer blends: mole fraction, weight-averaged mass and polydispersity index (PDI). We focus on binary polymer blends with star structure and lognormal arm length distribution. A total of six input variables are fed into the simulator to calculate the transient response in strong shear and extension of the input test materials. The goal of this case study is to search the design space in order to identify a combination of polymer properties that leads to rheological behavior similar to that of the target polymer melt. Since rheological computation is carried out in a black-box operation, which is considered computationally expensive, derivative-free optimization is a suitable method to solve this problem.

The target rheological response is simulated for a low density polyethylene (LDPE) blend with fixed material properties and polymer structures listed in Table 4.1. The rheological responses of the target polymer melt are shown in Figure 4.1. G' is the shear storage modulus that measures the

Entanglement time	1.15E-7 seconds
Temperature	463.15 K
Mass of a monomer	42.08 atomic units
Number of monomers in an entanglement length	128
Mass-density of the polymer	723.28 kg/m^3
Number of components	2
Dynamic dilation exponent	1
Number of polymers in each component	2000
Polymer type	Star with 3 arms
Arm length distribution	Lognormal

Table 4.1: Material properties and polymer structures used in the simulation

Configuration	Component 1	Component 2
Mole fraction	0.3	0.7
Polydispersity index	1.8	2.5
Weight-averaged mass	25000	30000

Table 4.2: Polymer configurations of the target binary LDPE blend

stored energy due to elasticity, and G'' is the loss modulus that measures the dissipated energy due to viscosity. Polymer configurations of each component in the target blend are presented in Table 4.2.

Physical properties listed in Table 4.1 are treated as fixed parameters, while we search in each variable space as shown in Table 4.2 to identify a set of polymer configurations of which the newly simulated rheological responses match those of the target response curves.

We assume that the given polymer melts exhibit linear viscoelasticity, i.e., there exists a linear relationship between stress and strain at any given time. Rheological curves therefore are evaluated at the same set of angular frequencies. Complex modulus is used to quantify the overall resistance to deformation of a material upon oscillations. It is calculated as a composite of the rheological responses G' and G''

$$G^* = \sqrt{G'^2 + G''^2} \tag{4.1}$$

In order to fit the target rheological responses, we seek to identify values for the simulator inputs

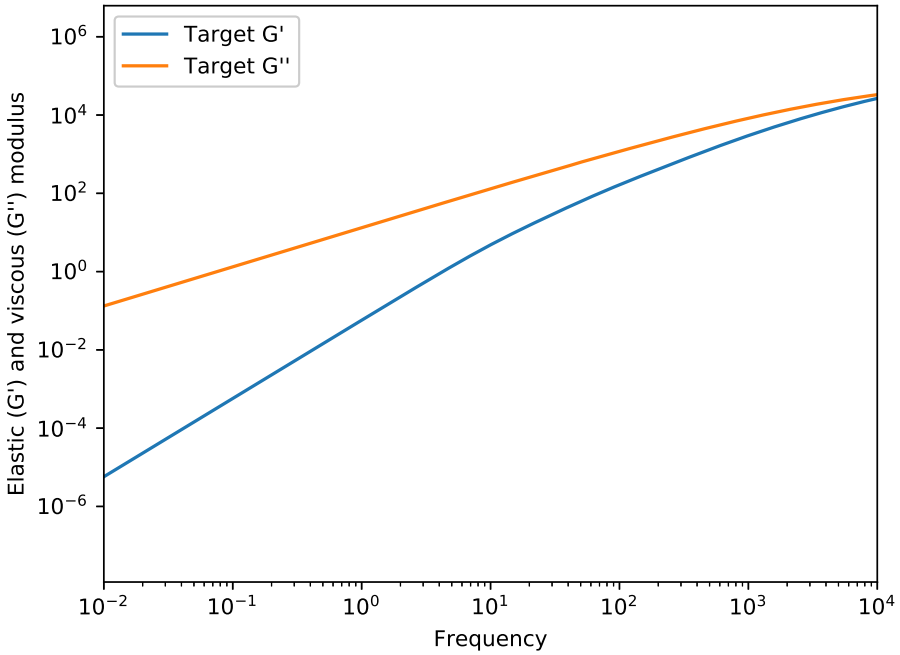


Figure 4.1: Target rheological responses

that minimize the sum of squared error between the complex moduli of the desired and simulated rheological responses. We optimize simultaneously for both responses as they are equally important to guarantee solution consistency. Additionally, our model satisfies the mass balance constraint and physical bounds on all input variables. We experimented with more than 20 DFO solvers. Here, we describe results with the following 11 solvers that performed better than other solvers:

- ASA
- CMAES
- HOPSPACK
- MCS
- NOMAD
- PSWARM
- SID-PSM
- SNOBFIT
- TOMLAB/GLCCLUSTER
- TOMLAB/MULTIMIN
- TOMLAB/LGO

These solvers can handle continuous problems with box-bounded constraints. The theoretical underpinnings and algorithmic descriptions for each solver can be found in their user manuals [143, 147, 148, 154–159].

Some DFO solvers, including CMAES, NOMAD, SID-PSM, and SNOBFIT, utilize the provided starting point to initialize the search process. Other solvers, such as MCS, discard the starting point or rely on one of the bounds. In order to evaluate these solvers for their ability to perform a thorough search of the design space independent of the quality of the starting point, the initial values were chosen to be far away from the configurations that gave the target rheological curves.

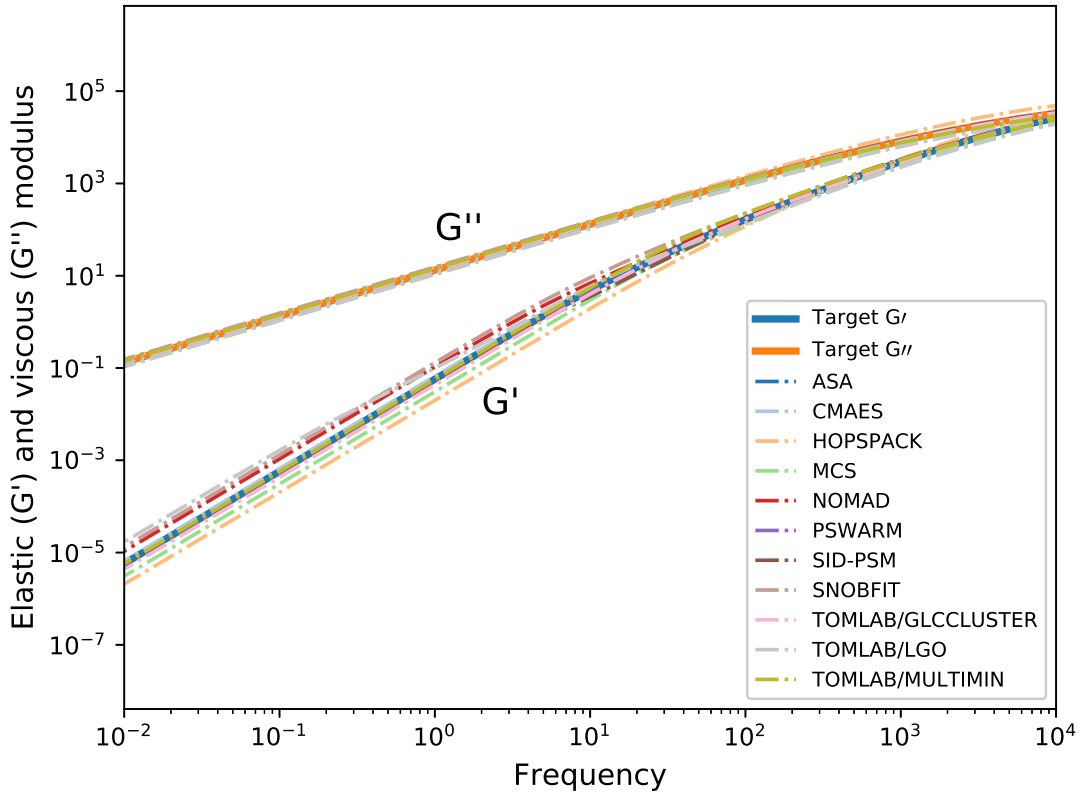


Figure 4.2: Comparison of rheological behavior of the identified polymer configurations

Upon termination, many DFO codes are not guaranteed to provide a local optimum, even though some of the underlying algorithms do so in theory. Thus, by using a collection of DFO solvers, we are increasing the likelihood of identifying good solutions.

A maximum of 1500 function evaluations was imposed on each DFO solver. For each solver, we report the number of function evaluations required to reach a solution with a 2.83 GHz processor.

By running the above DFO solvers with a limit of 1500 calls to the simulator, we obtained a set of 'optimal' configurations. When fed to the simulator, these configurations produced the rheological behavior curves that are shown in Figure 4.2. Clearly, all solves identified polymer configurations that were able to match the rheological targets very closely.

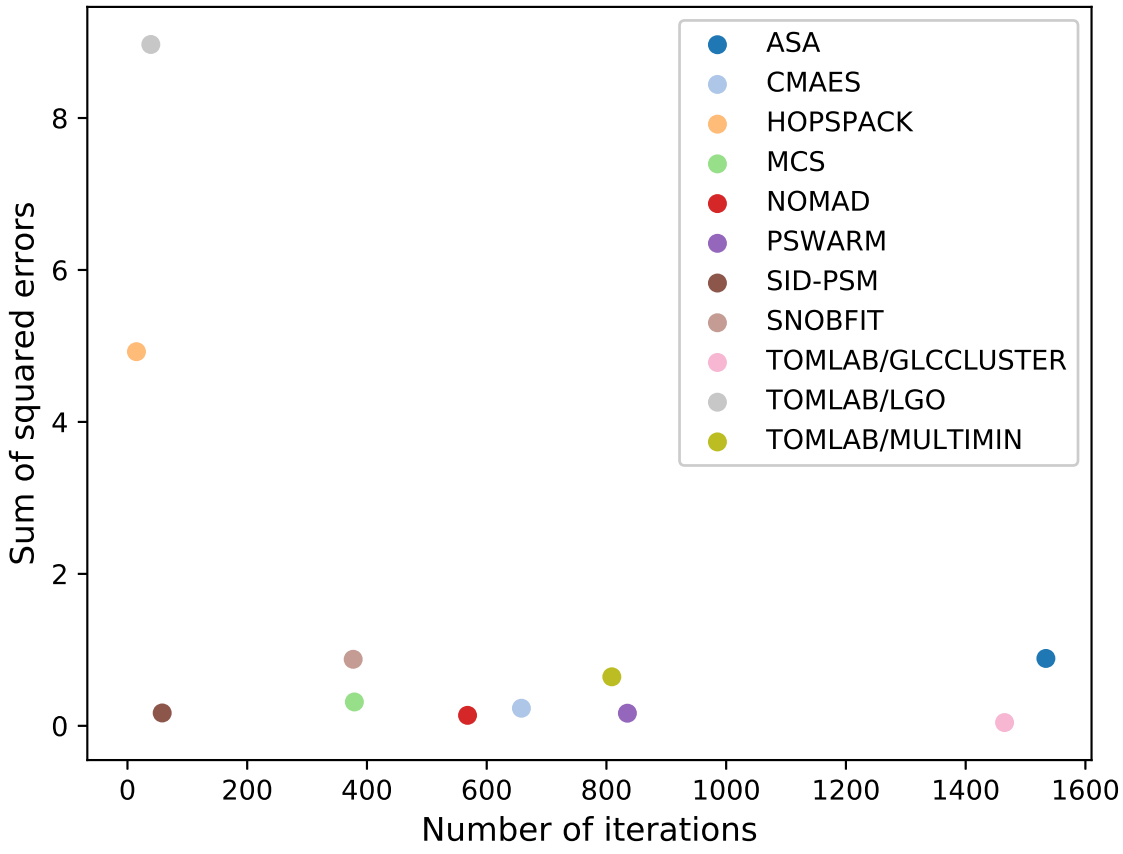


Figure 4.3: Performance of DFO solvers: quality of solution versus number of simulations required to complete the search

Figure 4.3 summarizes the performance of all 11 DFO solvers, including the final objective values and number of iterations required to reach a solution. All DFO solvers reach a solution within 1500 function evaluations. More specifically, SID-PSM was able to reach a satisfying solution within 100 function calls. Solvers like PSWARM, SID-PSM, NOMAD and TOMLAB/GLCCLUSTER were able to fit the target rheological responses even more closely. These solvers demonstrate that a large design space can be efficiently searched using these algorithms.

We further analyze the final design space to determine the distance from the identified solutions

to the polymer configurations as shown in Table 4.2. Each subplot of Figure 4.4 corresponds to a property-algorithm combination. The horizontal and vertical axes of each subplot correspond to the property values of the two components of the binary blend. Green circles denote the point in property space that was identified as best by the corresponding DFO algorithm. The red circles correspond to the configurations that we used to generate the target curves. As seen in this figure, differences in the search algorithms lead to a dispersive distribution of the identified polymer configurations, making them also different from the configuration used to generate the target rheological curves. This observation underscores the importance behind using a portfolio of DFO algorithms, instead of a single DFO algorithm. Using a portfolio of algorithms not only increases the likelihood of identifying good solutions but also creates diverse designs. A diverse solution pool is beneficial when it comes to polymer synthesis as it can potentially increase product diversity while maintaining similar rheological features. When the target polymeric characteristics are hard to maintain or expensive to achieve, the identified configurations can act as drop-in replacements of the target blends, further expanding product variety.

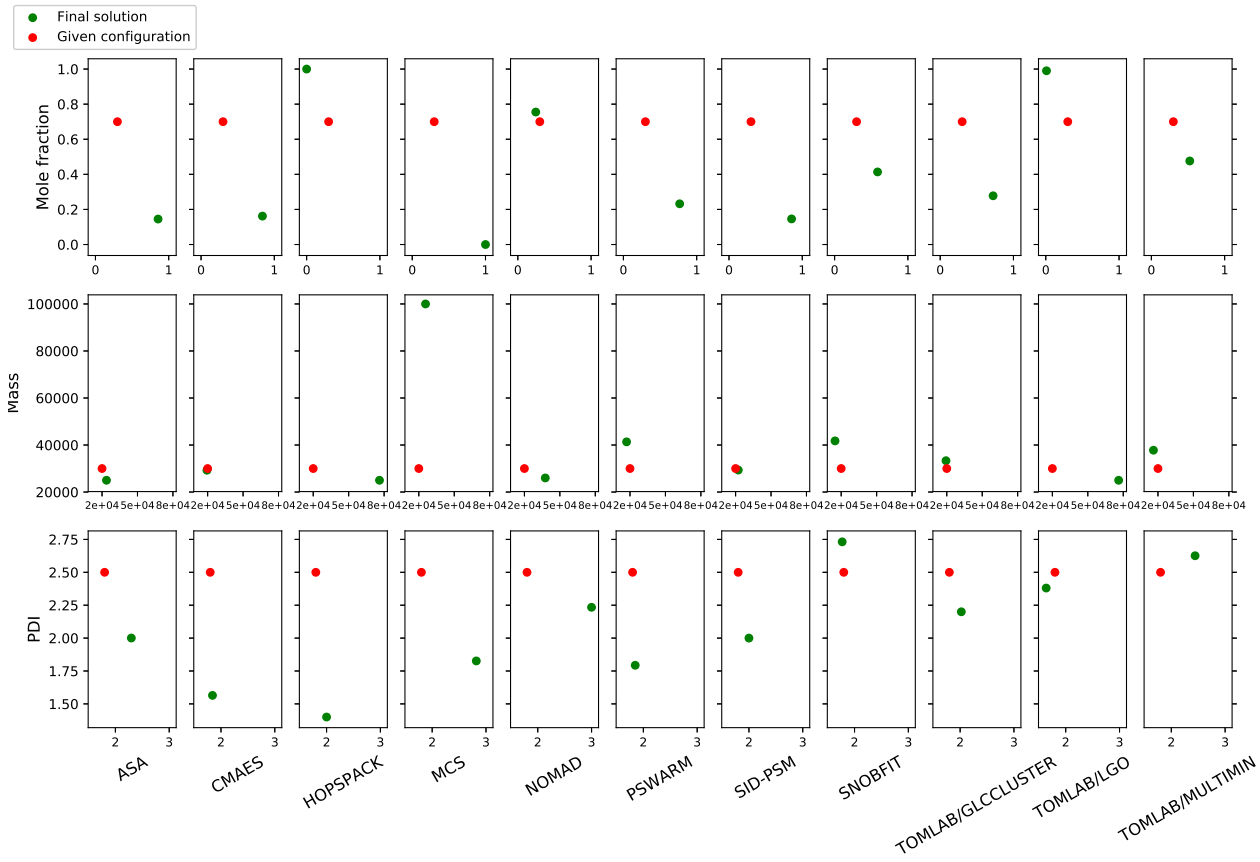


Figure 4.4: Solution diversity obtained by portfolio of DFO solvers

Chapter 5

Computer-aided retrosynthesis

5.1 Introduction

Retrosynthesis, a technique for planning a synthesis route backward, reduces a target organic molecule into a sequence of increasingly simpler precursors along reaction pathways which can ultimately lead to commercially available starting materials. This inherently complex problem requires searching in a large space of possible actions to transform the target molecules, either by the imaginary disconnection of bonds or by converting a functional group into another to match existing reaction templates. Traditionally, retrosynthesis tasks were largely based on existing knowledge, experience and intuition from synthetic chemists. The lack of systematic approaches imposed challenges to the synthesis of complex structures. For example, the synthesis of vitamin B12, a monumental achievement in organic synthetic chemistry, required the collaborative effort of over 100 chemists for nearly 12 years. As structural complexity increases, recognizing available starting materials requires search in a reaction space that grows exponentially with the number of reaction steps. Szymkuć et al. [160] estimated the dimension of the reaction search space is in the order of $10^{30} - 10^{50}$ for long reaction sequences. Computational algorithms can save a tremendous amount

of time and effort in the search of the enormous number of theoretically possible transformations.

Since 1967 when Nobel Laureate E. J. Corey made the first attempt to use computational tools for synthesis design [161], computer-aided retrosynthesis has been evolving rapidly. Early works in retrosynthesis used logic-based synthesis trees where the target molecule is placed at the root node with branches representing alternative or convergent pathways linking to intermediate reagents. Johnson's SYNLMA [162], SYNCHEM from Gelernter [163, 164], and Corey's LHASA [165] resulted from some of the early pioneering efforts in this field. These template-based endeavors tend to be labor-intensive as they require manual encoding of reaction rules. The number of encoded reactions subsequently determines the size of the reaction search space. More detailed perspectives on template-based synthetic planning tools can be found in [160, 166–171].

In the past decade, chemists and computer scientists have made significant improvements in computer-aided retrosynthesis thanks to the rapid and explosive advancement of data-driven decision-making tools and the establishment of a comprehensive reaction database. Data-driven models combined with machine learning (ML) techniques can infer latent relationships from high-dimensional data to extract implicit and potentially meaningful context and have demonstrated expert-level performance in various applications [172–174]. This compelling progress has become increasingly attractive in response to the rapid discovery of novel molecules and chemical information, contributed to template-based synthesis planning, and incubated the development of template-free synthetic planning tools. For example, in template-based approaches, ML techniques provide the means in automatic reaction rule extraction [175, 176], template relevance ranking [176], and candidate reactants scoring [177]. As for template-free methods, natural language processing models that built for machine translation tasks [178, 179] are adopted to translate product SMILES strings [180] to reactant SMILES strings [181]. As a result, the number of publications in retrosynthesis has increased significantly in the past decade (Figure 5.1). These tools have

enabled various applications, including drug design, novel synthetic route discovery, and design of biologically active compounds.

This paper is a survey on contemporary retrosynthesis strategies. We review and evaluate computer-aided retrosynthesis tools developed mainly in the past five years. The review is concise and approachable for beginners in this topic. Interested readers can refer to [171, 182, 183] for more detailed reviews. In the following sections, we cover the two dominant approaches to computer-aided retrosynthesis: template-based and template-free. Lastly, we provide a future outlook and address potential challenges of the field.

5.2 Retrosynthesis planning

A generic framework for retrosynthesis planning consists of a reaction rule library, a chemical database with commercially available starting materials, and strategies that select bond disconnection rules and direct the search toward the chemical database. Corey et al. [161, 166] established the majority of the framework decades ago. In the past decade, the landscape of retrosynthesis has changed significantly due to the establishment of large reaction databases and advancement in data-driven computational tools. Here, we review how ML can improve retrosynthesis performance from the following aspects: reaction rule extraction and synthesis planning strategies.

5.2.1 Reaction templates

Traditionally, reaction rules are expert-defined and manually encoded. Chematica [160] is one of the most well-known, commercially available, and manually encoded reaction libraries that cover most of the known reaction rules. It has demonstrated its ability to design synthetic pathways for high value medicinally relevant compounds [184, 185]. These design pathways were successfully validated in

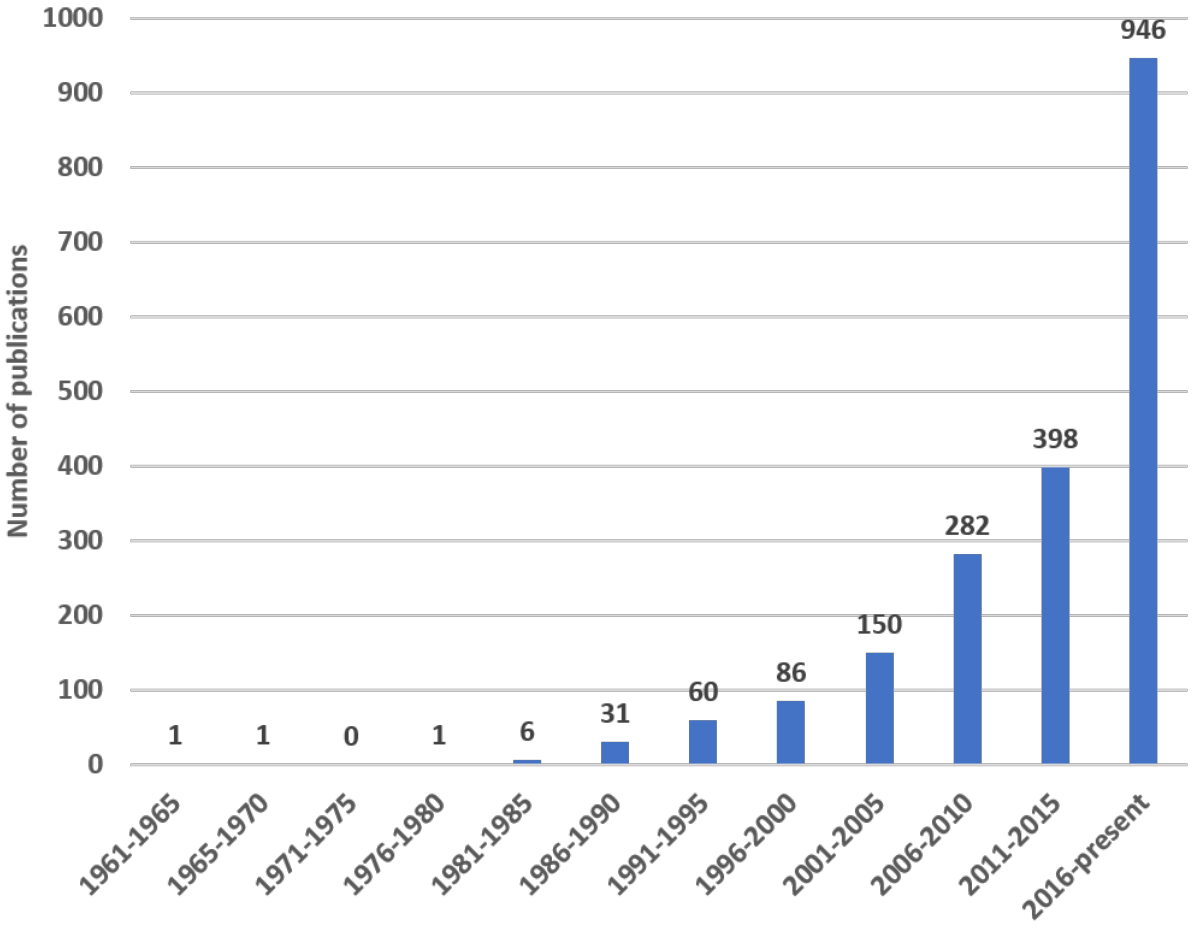


Figure 5.1: Number of publications containing the terms “retrosynthesis” or “synthetic planning.” Source: Google Scholar, 3/26/2021.

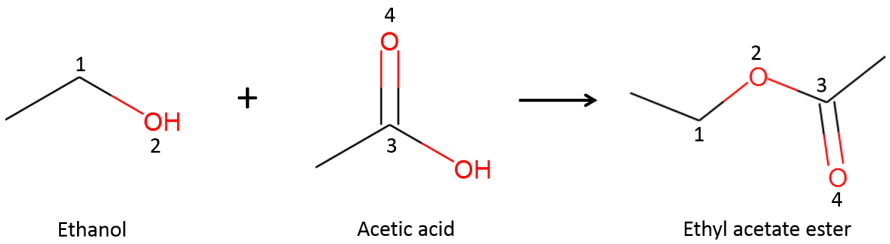
the laboratory. As the chemical space grows exponentially at an annual rate of 4.4% [186], manually annotating all existing chemical knowledge becomes a formidable task. A more contemporary approach to reaction encoding utilizes algorithmic extraction of reaction centers via atom-atom mapping to identify the correspondence between reactant and product atoms. For a given reaction, one can identify the set of atoms that change bond connectivities as the reaction centers. Then the reaction centers and adjacent atoms are algorithmically extracted and generalized to form the corresponding retrosynthesis template.

In general, the retrosynthesis template of a single-outcome reaction with N candidate precursors can be depicted by a subgraph rewriting rule [187]

$$o^T \rightarrow r_1^T + r_2^T + \cdots + r_N^T \tag{5.1}$$

where o^T represents the subgraph pattern extracted from the reaction center on the target molecule, and r_i^T, $i \in 1, 2, \ldots, N$ represents the corresponding subgraph patterns extracted from the i^{th} reactant. Iteratively matching templates to a target molecule equals solving a subgraph isomorphism problem. How far to extend the reaction center to auxiliary atoms highly affects the solution quality of retrosynthesis prediction. Including more neighboring atoms can improve specificity and lead to computationally expensive large subgraphs and poorly generalized databases. On the other hand, including too few neighboring atoms might overlook crucial information, which can generate inaccurate predictions. Heuristics have been developed to balance specificity and efficiency [188–190].

After the extraction, reaction centers must be stored in a machine-readable format. The RDKit [191] reaction SMARTS [192] format is commonly used to encode the reaction core patterns for both the reactants and product. Figure 5.2 shows an example of encoding the esterification reaction into the reaction SMARTS string. The reaction centers, as well as atom-atom mapping,



Reaction SMARTS string with atom matching

[CH2:1][OH:2].[OH][C:3]=[O:4]>>[C:1][O:2][C:3]=[O:4]

Figure 5.2: Reaction SMARTS string of an esterification reaction

are explicitly laid out. The above mentioned template generalization rules can be applied to the reaction template accordingly. This example, and most existing retrosynthesis applications, do not consider reaction conditions in the templates. Although specifying reaction conditions is crucial to the prediction outcome, including reaction conditions would add another search layer to the already enormous reaction search space. Several works attempted to recommend reaction conditions in forward reaction prediction [193–196]. However, to the best of our knowledge, suggesting suitable reaction conditions remains a largely open question in retrosynthesis applications.

Whether a reaction template library is used to convert target molecules into candidate reactants, retrosynthesis strategies can be categorized into template-based and template-free.

5.2.2 Retrosynthesis strategy evaluation

To evaluate the performance of retrosynthesis strategies, we need a common metric to examine if the ground truth precursors, the actual reactants reported in the template library for the corresponding target molecule, are among the top-N highest-ranked precursors suggested by the model. The

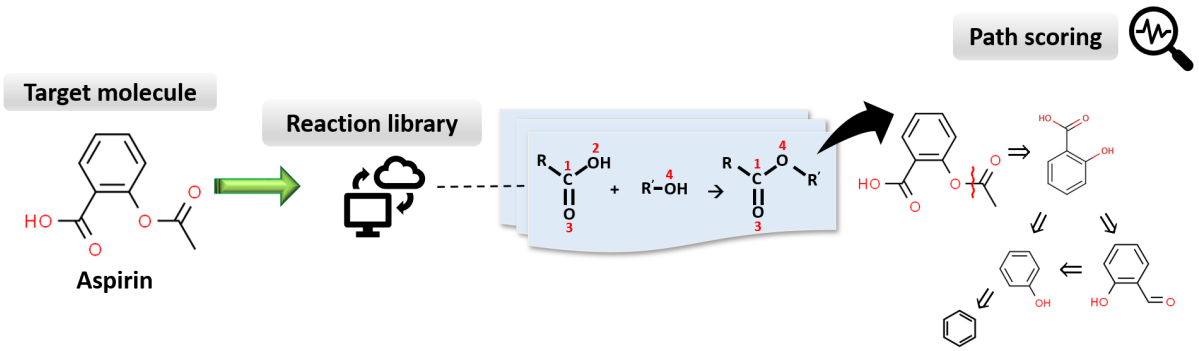


Figure 5.3: Template-based modeling to determine a retrosynthetic path for aspirin.

percentage is often referred to as the top-N accuracy.

5.2.3 Template-based models

Template-based strategies match target molecules to the entire template library by solving a subgraph isomorphism problem to obtain candidate reactants. An example is provided in Figure 5.3. These approaches usually require exhaustive enumeration through the reaction database, therefore are complemented with efficient graph-theoretical algorithms [197, 198] and virtual screening techniques [199].

Precursors not readily purchasable must be recursively expanded until all reactants on the path are commercially available or maximum depth is reached. To circumvent combinatorial explosion, scientists have made efforts to limit recursive expansion to the most promising bond disconnections that lead to easily synthesizable structures. Consequently, several metrics were developed to quantify the synthesizability of molecular structures. The traditional metric relies on the length of the resulting SMILES strings and aims to divide molecules into the smallest possible reactants. Synthetic accessibility score (SA score) utilizes fragment contributions that scale linearly with commonly synthesizable structural features and penalizes the presence of rare and complex structural features [50, 200]. Chematica [160] provides a metric that evaluates synthetic difficulty

as a function of both structural complexity and length of reaction steps, with an extra penalty for reaction conflicts and protection groups. SCScore [201] is based on the premise that the products of a reaction should be synthetically more complex than their reactants. Other synthetic scoring functions include support vector machines-based DRSVM [202] and current complexity [203].

Template-based approaches are useful for their interpretability and ability to provide fully specified chemical precursors. However, these approaches are computationally demanding and have limited generalization outside the template library.

5.2.4 Machine learning in template-based models

Research in template-based models has focused on overcoming the high computational cost resulting from the exhaustive enumeration of reaction templates. To address this challenge, researchers utilize ML to select only the relevant templates instead of using the full template library. This class of template-based models is referred to as "focused template application" [182].

Utilizing the most relevant subset of reaction rules alleviates the computational cost of full library application while preserving chemical interpretability. In the context of focused template applications, molecules are encoded into Extended-Connectivity Fingerprints (ECFP) [204] to better exploit the scalability and expressiveness of neural networks. Although training neural networks is an expensive task, the resulting models can make predictions at virtually no cost.

Segler and Waller [176] proposed one of the first focused-template models. They trained a neural-symbolic model on molecular fingerprints to score template relevance. This model solves a multiclass classification problem that categorizes similar templates into sub-groups. Segler et al. combined this approach with reinforcement learning [205] for fast and robust retrosynthetic pathway design and reaction prediction. Coley et al. [177] proposed a reaction similarity-based [206, 207] method formulated on the premise that similar reactions tend to produce similar compounds.

They computed molecular similarities between the target molecule and precedent reactions to rank templates. This model, known as *retrosim* [189], demonstrated good prediction accuracy and is often used as a benchmark. Baylon et al. [208] presented a multiscale neural network reaction recommendation system to suggest the first retrosynthetic reaction step. This method partitioned the retrosynthesis problem into two tasks. The first neural network predicted which reaction class can produce the target molecule. The second neural network, solely trained on the subset of reactions from the identified reaction group, determined the appropriate transformation to produce the target molecule. Dai et al. [187] proposed a conditional graph model built on top of graph neural networks to directly calculate the conditional joint probability of using a specific template and reactant set. Top-N prediction accuracy of this model outperforms retrosim.

As a template-based method, focused template applications can reduce computational intensity from the original rule-based approaches while retaining the same level of chemical interpretability. Yet, this class of models tends to convey the same limitations stemming from the underlying rules. They cannot predict outside the reaction knowledge base to suggest novel bond disconnections. As a result, they are rarely used to provide insights outside the scope of general chemistry.

5.2.5 Template-free

Recently, template-free approaches have attracted increasing attention as they avoid the computationally-expensive subgraph matching problems. These approaches utilize text representation of molecules (SMILES or InChI [209]) to cast the retrosynthesis problem as a sequence-to-sequence (seq-2-seq) prediction problem. Transforming target molecules to reactants becomes a translation task that converts the SMILES string of a product to that of reactants. This process no longer involves atom-atom mapping to identify reaction centers.

The idea of combining chemistry with natural language processing was first proposed by Cadeddu

et al. [210]. Several studies have since explored the use of seq-2-seq models in *de novo* molecular design [210–214] and forward reaction prediction problems [215–218]. These works used recurrent neural networks (RNNs) to learn the SMILES string's sequential relationships to regenerate SMILES as output. In the context of retrosynthesis, Liu et al. [181] reported a seq-2-seq architecture which involves two RNNs for target molecules and reactants and a beam search procedure to limit the number of candidates at each retrosynthesis step. They compared the top-N prediction accuracy to a template-based expert system on the test data set. Although the seq-2-seq model did not significantly benefit prediction accuracy, the model demonstrated several advantages over the rule-based baseline model. First, the seq-2-seq model can implicitly learn both the reaction rules and candidate ranking metrics, which avoid using stand-alone reaction complexity ranking metrics as in template-based approaches. Second, the seq-2-seq model is easier to scale up than rule-based approaches. The efficiency of template-based approaches depends on the number of reactions stored in the database since every rule needs to be applied to match the target molecules. At the same time, the efficiency of seq-2-seq model is primarily dependent on the width of beam search. Lastly, the seq-2-seq model can learn a latent environment of molecules and propose fundamentally novel bond disconnections.

Adopting template-free models is still relatively new in retrosynthesis. Recent development involves machine translation techniques [219, 220] that have yet to show significant improvement in prediction accuracy over template-based approaches. One common shortcoming of template-free approaches that needs to be addressed is the output of invalid SMILES strings. Further improvement is expected to enhance chemical interpretability and allow the prediction of multi-outcome/multi-step reactions.

5.3 Outlook

ML techniques have contributed to multiple steps in the retrosynthetic planning framework by providing efficient and better means to learn from the rich history of chemistry. However, as a data-driven model, the performance of the ML method largely depends on the data quality. Curating and maintaining a reliable database is challenging, but having access to open-source data with standardized representation and consistent quality would accelerate the development and offer a fair comparison baseline for various methodologies. ML models often suffer from a lack of interpretability. Naturally, template-free strategies are more prone to the lack of interpretability. These approaches, trained on text sequences, can neglect significant chemistry meaning behind bond disconnection, which sometimes can lead to infeasible suggestions. Methods to improve ML interpretability can be a potential solution to this challenge.

For any *in silico* design process, suggested synthesis routes should be validated experimentally to determine the true performance of computer-aided retrosynthesis. High-throughput and parallelized experimentation are commonly used for rapid data generation and experimental validation. Yet, the absence of experimental conditions in retrosynthesis data imposes further constraints on experimental planning. Recent developments exist in automated design of experiments (DoE) and self-optimization utilizes ML algorithms to optimize and identify feasible reaction conditions [221, 222]. Combining automated DoE tools with robotic experimental instruments can facilitate retrosynthesis design validation.

Improvements in our ability to manufacture chemical compounds can bring tremendous social and technological impact. Computer-aided synthesis planning, automatic laboratories, and automated material design tools are the key players in achieving a closed-loop automatic material discovery paradigm. Fueled by digitalization trends, artificial intelligence is anticipated to be a fundamental building block in establishing an automated chemical synthesis system, offering valuable assistance

in material discovery and eventually serving as a future robo-chemist [223]. As retrosynthesis techniques mature, we are likely to witness their integration within techniques for automatic computer-aided molecular design [36], wherein retrosynthesis-based metrics can be used to expedite the search by screening for manufacturability and even cost effectiveness.

5.4 Case study: Manufacturability test for electronic coolants

Here, we propose to incorporate a manufacturability test in the molecular design framework. We go back to the electronic coolant design problem reported in Chapter 2 and Chapter 3 and examine the candidate coolants for their ease of synthesis. Upon inspecting the newly discovered coolants, we find that most of the candidate molecules are novel structures with no previously reported application in the coolant industry. Instead of explicitly planning out synthetic routes for the candidate coolants, we are interested in knowing the likelihood of synthesizing these structures from existing synthetic knowledge, based on the assumption that similar compounds can be produced via similar reactions. Synthetic accessibility score (SAscore) [50] is one of the commonly used metrics to determine the ease of synthesis of a target molecule. It relies on molecular fragment contributions and structural complexity penalty to assign a score ranging from 1 to 10 to a molecular structure. Higher scores indicate compounds easier to synthesize. Molecular fragment contributions represent the frequency of appearance of structural features observed from already synthesized molecules. Complex structural features such as rings and isomers are assigned a complexity penalty. Both fragment contributions and complexity penalties are obtained by training on one million past synthesized molecules from PubChem (https://pubchem.ncbi.nlm.nih.gov), thus capturing historical synthetic knowledge.

We calculate the SAscore for both organosilicon and non-organosilicon candidates. Figure 5.4 illustrates the distribution of their SAscores. Organosilicon candidates correspond to SAscores

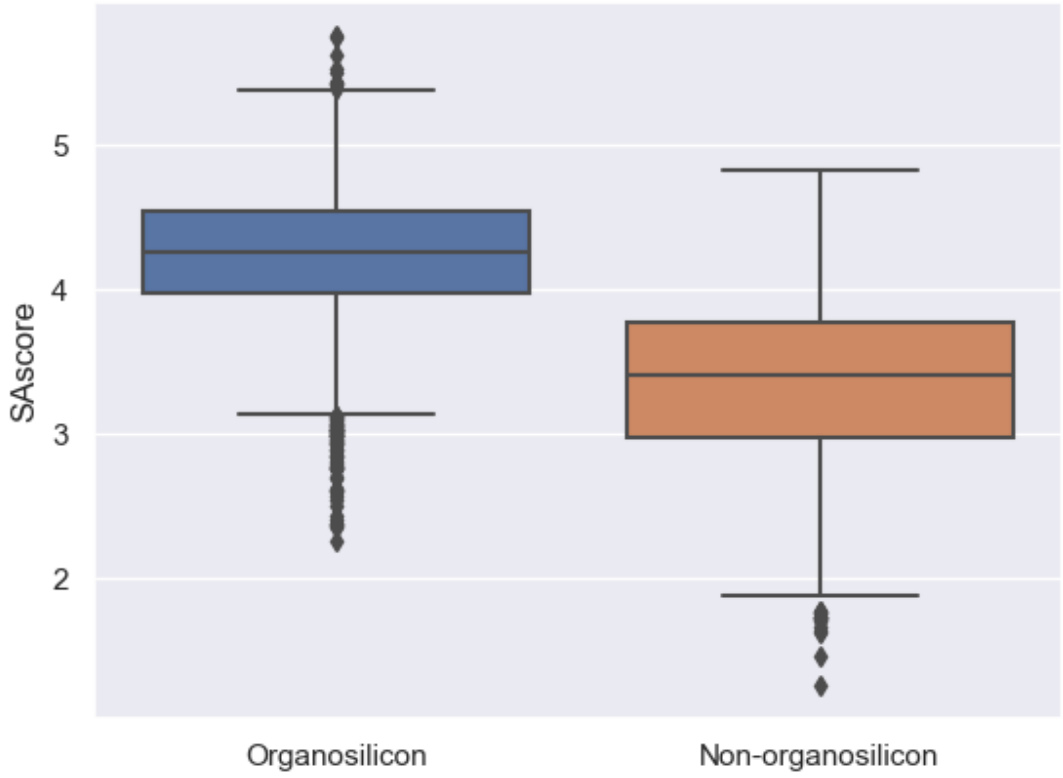


Figure 5.4: SAscore distribution of organosilicon coolants and non-organosilicon coolants

ranging from 2.77 to 5.79 with a median of 4.25, while the SAscore of non-organosilicons ranges from 1.26 to 4.83 with a median of 3.40. This comparison suggests that the addition of a silicon group can potentially decrease the synthesis barrier, leading to more accessible designs. Compounds with SAscore greater than 5 are more likely to be synthesized from a set of commercially available compounds, thus worth experimental validation.

Finally, we rank the performance of all candidate compounds taking into account their manufacturability. Figure 5.5 presents the heat transfer efficiency versus the SAScore for both organosilicons and non-organosilicon compounds. Candidate coolants in the upper right corner of this figure

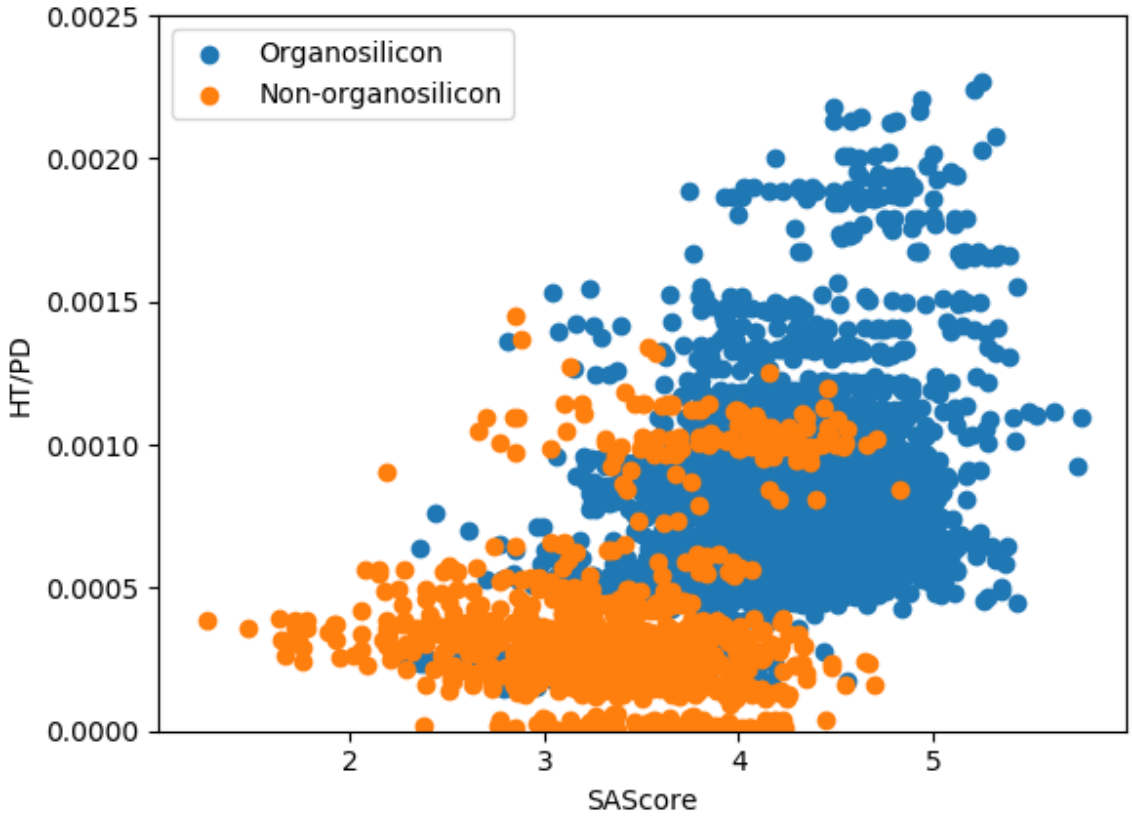


Figure 5.5: Compared to non-organosilicon compounds, organosilicon candidates exhibit higher heat transfer efficiency and higher synthetic accessibility.

exhibit satisfactory cooling performance and are predicted to be easier to synthesize. We maximize $\frac{HT}{PD} \times$ SAscore to identify the best performing coolants with the highest SAscore. Table 5.1 presents the top-10 coolants following this ranking metric. All of the top-10 best performing coolants are organosilicon compounds. This observation further supports organosilicons as ideal replacements of current commercial cooling fluids. Their properties should be subject to thorough experimental validation.

Canonical SMILES[75]	HT/PD	SAscore	Flash point (K)	Toxicity category
CCO[Si]C(C)CC	2.17E-3	4.93	304.39	Category III
COC(C)OC(C)[Si]C	1.95E-3	5.09	325.56	Category IV
COC(C)[Si]C(C)OC	1.94E-3	5.12	323.95	Category IV
CCC[Si]C(C)OC	2.02E-3	5.00	316.93	Category IV
COCCC(C)[Si]C	2.13E-3	4.49	308.39	Category IV
COCC(OC)[Si]C	1.90E-3	4.90	313.46	Category IV
COC[Si]C(C)OC	1.78E-3	5.11	322.02	Category IV
COC(C)OC[Si]C	1.77E-3	5.01	323.90	Category IV
CO[Si]C(C)OC	1.79E-3	5.17	302.93	Category IV
COC(C)O[Si]C	1.80E-3	5.11	304.87	Category IV

Table 5.1: Top ten molecules that are heat transfer efficient and relatively easy to synthesize

Chapter 6

Conclusions and future directions

6.1 Summary of contributions

Chapter 2 presents a framework for single-component design problems where the design targets are constant. We introduced the problem of electronic cooling fluid design and described a computer-aided molecular design approach. This approach allows us to discover molecules with desirable property values that can be used as drop-in replacements of industrial cooling fluids. We focused on a two-phase cooling system—microchannel heat sinks—and developed a metric to rank the performance of the candidate molecules. We also included metrics on environment and safety to further analyze these molecules. We conducted a kinetic stability test to guarantee that the candidate molecules are not subject to potential chemical transformation. Following this methodology, we are able to identify a solution pool of candidate molecules that have better cooling performance than the industrial cooling fluid HFE 7200. Among all candidate molecules, we identified four organic families with previous industrial applications. These families of compounds suggest promising directions for developing safe, biodegradable, low ozone depleting, and low-global-warming-potential coolants. Additionally, certain molecules posses superior heat transfer

properties and can potentially serve as an additive to boost thermal performance. The predicted performance of many candidate molecules motivates future analysis of their mixture properties.

In Chapter 3, we expanded the search area for superior electronic cooling fluids to include organosilicon compounds. Despite their wide encounter in a broad range of commercial applications, silicon containing structures are usually excluded from computer-aided molecular design applications due to the lack of reliable property prediction models. We aimed to develop reliable property estimation models to include organosilicon compounds in the design space. To do so, we proposed a new group selection method in the development of group contribution models. This method provides an effective way to automatically decompose molecular structures into subgroups consistently. After the functional group selection procedure, our methodology relies on ALAMO, a model selection framework, to build property models for eight target properties. The identified property models enable property estimation of a variety of organic compounds, including organosilicon structures. We performed cross-validation to examine the reliability of the proposed model. Both the training and test results demonstrate that our proposed method has good predictive power. The small model size indicates that this method can select property models that balance judiciously between overfitting and goodness of fit. We then applied the GC models to a coolant design problem to discover electronics cooling fluids containing silicon substructures. We are able to identify a solution pool of organosilicon compounds with better cooling performance than the commercial coolant HFE7200. A sensitivity analysis of the performance metric attests to the strength of the candidate molecules. Comparison to our previously identified non-organosilicon compounds suggested significant potential advantages of silicon containing cooling fluids.

In Chapter 4, we looked into chemical product design problems with time-varying design targets. Our proposed methodology integrates derivative-free optimization models into computer-aided chemical design frameworks. We presented an overview of derivative-free optimization algorithms

and discussed various implementations. This methodology demonstrates successful application to a chemical product design problem that involves the use of a first-principles simulator to predict rheological behavior of polymer blends. Our results indicate that a portfolio of DFO solvers is capable of identifying a diverse set of good solutions despite the complexity of the underlying chemical design space.

The final step to completely solve a chemical product design problem is figuring out whether the candidate compounds can be manufactured using readily available materials. This analysis is referred to as the retrosynthesis problem. In Chapter 5, we provided a review on computer-aided retrosynthetic strategies to understand the most recent development in the field and answer the question of manufacturability of the products. We reviewed how machine learning (ML) algorithms contribute to retrosynthetic strategies that rely on reaction libraries and explained ML based strategies without the use of reaction libraries. We provided discussion on potential challenges facing this field and offered our outlook. We then performed a manufacturability test on the previously identified candidate coolants to assess their ease of synthesis. This work touches upon multiple aspects of the chemical product design problems, aiming to offer solution strategies to complement existing tools.

In summary, contributions made in this work have expanded the search space of the CAMD framework and have improved the solution quality and fidelity of product design problems. These contributions can be used as a stepping stone for future development in this field. In the next section, we propose several research directions worth exploring.

6.2 Future research directions

6.2.1 Automated framework for QSPR modeling

As more chemical compounds become available, the amount of chemical information added to the online data base is growing. The chemical data, however, are not evenly distributed. Some compounds might have multiple measurements for one property while the same property may be missing entirely for other compounds. This unevenly distributed data sparsity can hinder the performance of machine learning-based property prediction models, presenting a challenge to researchers who may lack extensive experience in machine learning. Therefore, an automated platform with advanced modeling and bias reducing techniques can be a beneficial tool for the QSPR community to avoid the time-consuming model building and hyper-parameter tuning process. The automated framework should include procedures for data cleaning and pre-processing, feature selection, variable selection and model validation. The workflow should be streamlined so that existing QSPR models can be updated and retrained with the presence of new data.

In Chapter 3, we present a QSPR modeling framework that automates the process from data pre-processing to model building and validation. Our proposed functional group selection method is able to decompose any molecular structure into a subset of sub-molecular descriptors that are important to the given property data. This framework can serve as the foundation for a broader QSPR modeling scheme that enables a wide range of physicochemical and environmentally-related property estimations. Once such a framework becomes available, it can further improve the solution quality and general applicability of the CAMD approach.

6.2.2 Product design in process intensification

The design trends of miniaturized devices have significantly decreased system size and increased power density. In the chemical industry, the miniaturization trend is known as process intensification, where reduction in system size results in higher process efficiency, lower capital cost and higher product quality. Process intensification requires a holistic view on the process. From a macroscopic point of view, process intensification results from the use of smaller equipment or integrated systems, which reduce the number of steps and enhance the performance of the chemical process. On the microscopic end, an integrated system with multiple processes running simultaneously requires molecules to similarly exhibit multiple features at the same time, such as heat transfer properties, transport properties and liquid-liquid extraction properties. Hence, identifying suitable molecular structures is crucial to develop a fully intensified process. An interesting application would be to design compounds (reactants and reaction solvents) for a microreactor.

6.2.3 Closed-loop chemical product design

In the past decade, excitement has been growing around the idea of using machine learning algorithms to create an automated chemical design framework. These works, such as in [212–214, 224, 225], rely on variational autoencoders (VAEs), a type of deep generative models, to describe the molecular structures in a latent space. A VAE consists of an encoder that converts the SMILES representation of a molecule into a real-valued continuous representation in the latent space, and a decoder that maps a point from the continuous latent space back to SMILES string representation. After training, the latent space is organized by molecular properties, which allows the use of gradient-based optimization and new structures to be automatically generated from latent space operation.

Recently, the interest in forming a closed-loop material discovery system has been growing. Researchers attempt to integrate automated chemical product design frameworks and retrosynthesis

strategies with automated experimental instruments. Closing the loop requires incorporating inverse design where the inputs are desired material functionalities and outputs are ideal molecular structures or mixture formulations. Initial candidates are screened based on certain targets including toxicity, stability, and the likelihood of synthesis. Promising candidates are then subject to high-throughput experimentation and virtual screening for further characterization and optimization. Coley et al. offer detailed analysis on the trends of autonomous chemical discovery and insightful outlook in [226, 227]. This community is beginning to show some success [228–230]. An interesting problem would be to investigate the application to this area of the optimization methodologies developed in this thesis.

Bibliography

(1) Budde, F.; Ezekoye, O.; Hundertmark, T.; Prieto, M.; Simons, T. J. Chemicals 2025: Will the industry be dancing to a very different tune?, https://www.mckinsey.com/industries/chemicals/our-insights/chemicals-2025-will-the-industry-be-dancing-to-a-very-different-tune, Accessed: 02/28/2021, 2017.

(2) American Chemical Society CAS Registry, Available at www.cas.org/content/chemical-substances.

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

(4) *Chemical Product Design: Toward a perspective through case studies*; Ng, K. M., Gani, R., Dam-Johansen, K., Eds.; Computer Aided Chemical Engineering, Vol. 23; Elsevier: 2007.

(5) Duvedi, A. P.; Achenie, L. E. K. *Chemical Engineering Science* **1996**, *51*, 3727–3739.

(6) Duvedi, A. P.; Achenie, L. E. K. *Computers & Chemical Engineering* **1997**, *21*, 915–923.

(7) Sahinidis, N. V.; Tawarmalani, M.; Yu, M. *AIChE Journal* **2003**, *49*, 1761–1775.

(8) Warrier, P.; Sathyanarayana, A.; Patil, D. V.; France, S.; Joshi, Y.; Teja, A. S. *International Journal of Heat and Mass Transfer* **2012**, *55*, 3379–3385.

(9) Samudra, A.; Sahinidis, N. V. *Industrial & Engineering Chemistry Research* **2013**, *52*, 8518–8526.

(10) Sun, Y.; Samudra, A.; Sahinidis, N. V. *Industrial & Engineering Chemistry Research* **2019**, *58*, 4925–4935.

(11) Macchietto, S.; Odele, O.; Omatsone, O. *Chemical Engineering Research and Design* **1990**, *68*, 429–433.

(12) Karunanithi, A. T.; Achenie, L. E. K.; Gani, R. *Chemical Engineering Science* **2006**, *61*, 1247–1260.

(13) Karunanithi, A. T.; Acquah, C.; Achenie, L. E. K.; Sithambaram, S.; Suib, S. L. *Computers & Chemical Engineering* **2009**, *33*, 1014–1021.

(14) Satyanarayana, K. C.; Abildskov, J.; Gani, R. *Computers & Chemical Engineering* **2009**, *33*, 1004–1013.

(15) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. *Computers & Chemical Engineering* **1994**, *18*, 833–844.

(16) Vaidyanathan, R.; El-Halwagi, M. *Industrial & Engineering Chemistry Research* **1996**, *35*, 627–634.

(17) Gani, R.; Fredenslund, A. *Fluid Phase Equilibria* **1993**, *82*, 39–46.

(18) Klein, J. A.; Wu, D. T.; Gani, R. *Computers & Chemical Engineering* **1992**, *16*, S229–S236.

(19) Austin, N. D.; Samudra, A. P.; Sahinidis, N. V.; Trahan, D. W. *AIChE Journal* **2016**, *62*, 1514–1530.

(20) Karunanithi, A. T.; Achenie, L. E. K.; Gani, R. *Industrial & Engineering Chemistry Research* **2005**, *44*, 4785–4797.

(21) Churi, N.; Achenie, L. E. K. *Computers & Chemical Engineering* **1997**, *21*, S349–S354.

(22) Churi, N.; Achenie, L. E. K. *International Transactions of Operational Research* **1997**, *4*, 45–54.

(23) Austin, N. D.; Sahinidis, N. V.; Trahan, D. W. *Chemical Engineering Science* **2017**, *159*, 93–105.

(24) Folić, M.; Adjiman, C. S.; Pistikopoulos, E. N. *AIChE Journal* **2007**, *53*, 1240–1256.

(25) Strübing, H.; Konstantinidis, S.; Karamertzanis, P. G.; Pistikopoulos, E. N.; Galindo, A.; Adjiman, C. S. In *Process Systems Engineering*; Wiley-VCH Verlag GmbH & Co. KGaA: 2011, pp 267–305.

(26) Strübing, H.; Ganase, Z.; Karamertzanis, P. G.; Siougkrou, E.; Haycock, P.; Piccione, P. M.; Armstrong, A.; Galindo, A.; Adjiman, C. S. *Nature Chemistry* **2013**, *5*, 952–957.

(27) Austin, N. D.; Sahinidis, N. V.; Konstantinov, I. A.; Trahan, D. W. *AIChE Journal* **2018**, *64*, 104–122.

(28) Karunanithi, A.; Mehrkesh, A. *AIChE Journal* **2013**, *59*, 4627–4640.

(29) Song, Z.; Zhang, C.; Qi, Z.; Zhou, T.; Sundmacher, K. *AIChE Journal* **2018**, *64*, 1013–1025.

(30) Zhang, J.; Peng, D.; Song, Z.; Zhou, T.; Cheng, H.; Chen, L.; Qi, Z. *Chemical Engineering Science* **2017**, *162*, 355–363.

(31) McLeese, S. E.; Eslick, J. C.; Hoffmann, N. J.; Scurto, A. M.; Camarda, K. V. *Computers & Chemical Engineering* **2010**, *34*, 1476–1480.

(32) Siddhaye, S.; Camarda, K.; Southard, M.; Topp, E. *Computers & Chemical Engineering* **2004**, *28*, 425–434.

(33) Siddhaye, S.; Camarda, K. V.; Topp, E.; Southard, M. *Computers & Chemical Engineering* **2000**, *24*, 701–704.

(34) Gani, R. *Chemical Engineering Research and Design* **2004**, *82*, 1494–1504.

(35) Ng, L. Y.; Chong, F. K.; Chemmangattuvalappil, N. G. *Computers & Chemical Engineering* **2015**, *81*, 115–129.

(36) Austin, N. D.; Sahinidis, N. V.; Trahan, D. W. *Chemical Engineering Research and Design* **2016**, *116*, 2–26.

(37) Joback, K. G.; Reid, R. C. *Chemical Engineering Communications* **1987**, *57*, 233–243.

(38) Constantinou, L.; Gani, R. *AIChE Journal* **1994**, *40*, 1697–1710.

(39) Marrero, J.; Gani, R. *Fluid Phase Equilibria* **2001**, *183–184*, 183–208.

(40) Marrero, J.; Gani, R. *Industrial & Engineering Chemistry Research* **2002**, *41*, 6623–6633.

(41) Gani, R.; Harper, P. M.; Hostrup, M. *Industrial & Engineering Chemistry Research* **2005**, *44*, 7262–7269.

(42) Conte, R.; Martinho, A.; Matos, H. A.; Gani, R. *Industrial & Engineering Chemistry Research* **2008**, *47*, 7940–7954.

(43) Klincewicz, K. M.; Reid, R. C. *AIChE Journal* **1984**, *30*, 137–142.

(44) Nannoolal, Y.; Rarey, J.; Ramjugernath, D.; Cordes, W. *Fluid Phase Equilibria* **2004**, *226*, 45–63.

(45) Nannoolal, Y.; Rarey, J.; Ramjugernath, D. *Fluid Phase Equilibria* **2008**, *269*, 117–133.

(46) Nannoolal, Y.; Rarey, J.; Ramjugernath, D. *Fluid Phase Equilibria* **2009**, *281*, 97–119.

(47) Martin, T. M.; Young, D. M. *Chemical Research in Toxicology* **2001**, *14*, 1378–1385.

(48) Sun, Y.; Sahinidis, N. V.; Sundaram, A.; Cheon, M.-S. *Current Opinion in Chemical Engineering* **2020**, *27*, 98–106.

(49) Samudra, A.; Sahinidis, N. V. *AIChE Journal* **2013**, *59*, 3686–3701.

(50) Ertl, P.; Schuffenhauer, A. *Journal of Cheminformatics* **2009**, *1*.

(51) Hoefflinger, B., *Chips 2020. The Frontiers Collection*; Springer: Berlin, Heidelberg; Chapter ITRS: The International Technology Roadmap for Semiconductors.

(52) Thermal management technologies market analysis and segment forecasts to 2024, Grand View Research, 2016.

(53) Ebadian, M. A.; Lin, C. X. *ASME. J. Heat Transfer* **2011**, *133*, 110801–110801–11.

(54) Murshed, S. M. S.; de Castro, C. A. N. *Renewable and Sustainable Energy* **2017**, *78*, 821–833.

(55) Montreal Protocol on Substances that Deplete the Ozone Layer, Montreal, 16 September, 1987.

(56) Simons, R. E. Direct liquid immersion cooling for high power density microelectronics, Electronics Cooling, 1 May, 1996.

(57) The Kigali Amendment to the Montreal Protocal: Another Global Commitment to stop climate change, Kigali, December, 2016.

(58) Sathyanarayana, A.; Joshi, Y.; Im, Y. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2010 12th IEEE Intersociety Conference on*, 2010, pp 1–6.

(59) Joback, K. G. Designing molecules possessing desired physical property values, Ph.D. Thesis, Cambridge, MA: Department of Chemical Engineering, Massachusetts Institute of Technology, 1989.

(60) Harper, P. M.; Gani, R.; Kolar, P.; Ishikawa, T. *Fluid Phase Equilibria* **1999**, *158–160*, 337–347.

(61) Gani, R.; Nielsen, B.; Fredenslund, A. *AIChE Journal* **1991**, *37*, 1318–1332.

(62) Sahinidis, N. V.; Tawarmalani, M. *Computers & Chemical Engineering* **2000**, *24*, 2157–2169.

(63) Apostolakou, A.; Adjiman, C. In L. E. K. Achenie, R. Gani and V. Venkatasubramanian (eds.)*, Computer Aided Molecular Design: Theory and Practice,* Elsevier, Computer Aided Chemical Engineering, Volume 12 **2002**, 289–301.

(64) Churi, N.; Achenie, L. E. K. *Industrial & Engineering Chemistry Research* **1996**, *35*, 3788–3794.

(65) Warrier, P.; Sathyanarayana, A.; Bazdar, S.; Joshi, Y.; Teja, A. S. *Industrial & Engineering Chemistry Research* **2012**, *51*, 10517–10523.

(66) Odele, O.; Macchietto, S. *Fluid Phase Equilibria* **1993**, *82*, 47–54.

(67) Brignole, E. A.; Bottini, S.; Gani, R. *Fluid Phase Equilibria* **1986**, *29*, 125–132.

(68) Fredenslund, A.; Gmehling, J.; Michelson, M. L.; Ramussen, P.; Prausnitz, J. M. *Industrial & Engineering Chemistry Process Design and Development* **1977**, *16*, 450–462.

(69) Gopinath, S.; Jackson, G.; Galindo, A.; Adjiman, C. S. *AIChE Journal* **2016**.

(70) Giovanoglou, A.; Barlatier, J.; Adjiman, C. S.; Pistikopoulos, E. N.; Cordiner, J. L. *AIChE Journal* **2003**, *49*, 3095–3109.

(71) *Computer Aided Molecular Design Theory and Practice*; Achenie, L. E. K., Gani, R., Venkatasubramanian, V., Eds., New York, NY, 2002; Vol. 12.

(72) Bommareddy, S.; Chemmangattuvalappil, N. G.; Solvason, C. C.; Eden, M. R. *Computers & Chemical Engineering* **2010**, *34*, 1481–1486.

(73) Joback, K. G.; Stephanopoulos, G. *Advances in Chemical Engineering* **1995**, *21*, 257–311.

(74) Ceriani, R.; Gani, R.; Meirelles, A. J. A. *Fluid Phase Equilibria* **2009**, *283*, 49–55.

(75) Weininger, D. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(76) Catoire, L.; Naudet, V. *Journal of Physical and Chemical Reference Data* **2004**, *33*, 1083–1111.

(77) Poling, B. E.; Prausnitz, J. M.; O'Connell, J. P., *The properties of gases and liquids*, 5th; McGraw-Hill: New York, 2001.

(78) Reid, R. C.; Prausnitz, J. M.; Poling, B. E., *The properties of gases and liquids*, 4th; McGraw-Hill: New York, 1987.

(79) Gonzalez, M. H.; Eakin, B. E.; Lee, A. L., *Viscosity of Natural Gases: Monograph on API Research Project 65*; American Petrolium Institute: 1979.

(80) Estimations Programs Interface (EPI Suite), EPA, Available at http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm.

(81) Kawahara, A.; Chung, P. M. .-Y.; Kawaji, M. *International Journal of Multiphase Flow* **2002**, *28*, 1411–1435.

(82) Lockhart, R. W.; Martinelli, R. C. *Chemical Engineering Progress* **1949**, *45*, 39–48.

(83) Lee, J.; Mudawar, I. *International Journal of Heat and Mass Transfer* **2005**, *48*, 928–940.

(84) Lee, J.; Mudawar, I. *International Journal of Heat and Mass Transfer* **2005**, *48*, 941–955.

(85) Eide-Haugmo, I.; Brakstad, O. G.; Hoff, K. A.; SÃ¸rheim, K. R.; da Silva, E. F.; Svendsen, H. F. *Energy Procedia* **2009**, *1*, Greenhouse Gas Control Technologies 9, 1297–1304.

(86) Ku, H. H. *Journal of Research of the National Bureau of Standards–C. Engineering and Instrumentation* **1966**, *70C*, 263–273.

(87) Chemical Abstracts Service: Columbus, OH.

(88) Kirsch, P., *Modern Fluoroorganic Chemistry: Synthesis, Reactivity, Applications*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2013.

(89) Blowers, P.; Moline, D. M.; Tetrault, K. F.; Wheeler, R. R.; Tuchawena, S. L. *Environmental Science and Technology* **2008**, *42*, 1301–1307.

(90) Sciance, F. The transition from HFC-134a to a low-GWP refrigerant in mobile air conditioners, The Mobile Source Technical Review Subcommittee meeting, Wachington, D.C.: U.S. Environmental Protecting Agency (EPA), 2013.

(91) Bravo, I.; Diaz-de-Mera, Y.; Aranda, A.; Smith, K.; Shine, K. P.; Marston, G. *Physical Chemistry Chemical Physics* **2010**, *12*, 5115–5125.

(92) Vitcak, D. R.; Flynn, R. M. Process for the production of hydrofluoroethers, US Patent 5,750,797, 1998.

(93) Flynn, R. M.; Costello, M. G. Hydrofluoroether compounds and processes for their preparation and use, US Patent 7,691,282 B2, 2010.

(94) Costello, M. G.; Flynn, R. M.; Behr, F. E. Hydrofluoroether as a heat-transfer fluid, US Patent 7,651,627 B2, 2010.

(95) Toxicity Estimation Software Tool (TEST) version 5.1, Available at https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test#pubs.

(96) Wong, M. W. *Chemical Physics Letters* **1996**, *256*, 391–399.

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

(98) Bochevarov, A.; Harder, E.; Hughes, T.; Greenwood, J.; Braden, D.; Philipp, D.; Rinaldo, D.; Halls, M.; Zhang, J.; Friesner, R. *International Journal of Quantum Chemistry* **2013**, *113*, 2110–2142.

(99) Costello, M. G.; Flynn, R. M.; Behr, F. E. Hydrofluoroether as a heat-transfer fluid, U.S. Patent 7651627B2, Jan. 26th 2010.

(100) Mohapatra, S. C. *Electronics Cooling* **2006**, *12*, 1–6.

(101) Joback, K. G.; Stephanopoulos, G. *Proceedings of the 1989 Foundations of Computer-Aided Process Design Conference, Snowmass, CO,* Elsevier, Amsterdam **1990**, 195–230.

(102) Hukkerikar, A. S.; Sarup, B.; Kate, A. T.; Abildskov, J.; Sin, G.; Gani, R. *Fluid Phase Equilibria* **2012**, *321*, 25–43.

(103) *Design, Institute for Physical Property Data/AIChE* **2019**.

(104) Dortmund Data Bank Software and Separation Technology GmbH Parameters of the original UNIFAC model, http://www.ddbst.com/published-parameters-unifac.html.

(105) *Vapor-liquid Equilibria Using Unifac A Group-Contribution Method*; Elsevier: 1977.

(106) Hansen, H. K.; Rasmussen, P.; Fredenslund, A.; Schiller, M.; Gmehling, J. *Industrial & Engineering Chemistry Research* **1991**, *30*, 2352–2355.

(107) Wittig, R.; Lohmann, J.; Gmehling, J. *Industrial & Engineering Chemistry Research* **2003**, *42*, 183–188.

(108) Constantinou, L.; Gani, R.; O'Connell, J. P. *Fluid Phase Equilibria* **1995**, *103*, 11–22.

(109) Matsuda, H.; Yamamoto, H.; Kurihara, K.; Tochigi, K. *Fluid Phase Equilibria* **2007**, *261*, 434–443.

(110) Kolská, Z.; Kukal, J.; Zábranský, M.; Ružička, V. *Industrial & Engineering Chemistry Research* **2008**, *47*, 2075–2085.

(111) Cozad, A.; Sahinidis, N. V.; Miller, D. C. *AIChE Journal* **2014**, *60*, 2211–2227.

(112) IBM IBM ILOG CPLEX 12.9 User's Manual, (IBM ILOG CPLEX Division, Incline Village, NV), 2017.

(113) Material Property Search in Knovel, Knovel, https://app.knovel.com/web/data-search.v, Accessed: 02/09/2021.

(114) Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M. *Crystal Growth & Design* **2001**, *1*, 261–265.

(115) Kirkwood, J. G. *The Journal of Chemical Physics* **1939**, *7*, 911–919.

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

(117) 3M Heat transfer applications using 3M™ Novec™ Engineered Fluids, https://multimedia.3m.com/mws/media/1091997O/3m-novec-engineered-fluids-for-heat-transfer-line-card.pdf, Accessed: 02/23/2021, 2016.

(118) Austin, N. D.; Sahinidis, N. V.; Trahan, D. W. *Chemical Engineering Research and Design* **2016**, *116*, 2–26.

(119) Sundaram, A.; Ghosh, P.; Caruthers, J. M.; Venkatasubramanian, V. *AIChE Journal* **2001**, *47*, 1387–1406.

(120) Zhou, N.; Zhou, Y.; Sundmacher, K. *Chemical Engineering Science* **2017**, *159*, 207–216.

(121) Song, Z.; Zhang, C.; Qi, Z.; Zhou, T.; Sundmacher, K. *AIChE Journal* **2018**, *64*, 1013–1025.

(122) Zhang, N.; Qin, L.; Peng, D.; Zhou, T.; Cheng, H.; Chen, L.; Qi, Z. *Chemical Engineering Science* **2017**, *162*, 364–374.

(123) Zhou, T.; Wang, J.; McBride, K.; Sundmacher, K. *AIChE Journal* **2016**, *62*, 3238–3249.

(124) Marcoulaki, E. C.; Kokossis, A. C. *Chemical Engineering Science* **2000**, *55*, 2529–2546.

(125) Marcoulaki, E. C.; Kokossis, A. C. *Chemical Engineering Science* **2000**, *55*, 2547–2561.

(126) Gebreslassie, B. H.; Diwekar, U. M. *Computers & Chemical Engineering* **2015**, *78*, 1–9.

(127) Benavides, P. T.; Gebreslassie, B. H.; Diwekar, U. M. *Chemical Engineering Science* **2015**, *137*, 977–985.

(128) Rios, N. M.; Sahinidis, N. V. *Journal of Global Optimization* **2013**, *56*,
This paper presents what is currently the most comprehensive comparison of derivative-free optimization (DFO) algorithms and software. A total of 23 deterministic and stochastic DFO codes were compared on over 500 problems., 1247–1293.

(129) Austin, N. D.; Samudra, A. P.; Sahinidis, N. V.; Trahan, D. W. *AIChE Journal* **2016**, *62*, 1514–1530.

(130) Binder, K. *Reports on Progress in Physics* **1997**, *60*, 487–559.

(131) Xia, B.; Sun, D. *Computers and Electronics in Agriculture* **2002**, *34*, 5–24.

(132) Laird, B. B.; Ross, R. B.; Ziegler, T., *Chemical applications of density-functional theory*; American Chemical Society: Washigton, D.C., 1996.

(133) Amaran, S.; Sahinidis, N. V.; Sharda, B.; Bury, S. J. *Annals of Operations Research* **2016**, *240*, 351–380.

(134) Forrester, A. I. J.; Sobester, A.; Keane, A. J., *Engineering design via surrogate modelling - A practical guide*; John Wiley & Sons: 2008.

(135) Torczon, V. J. *SIAM Journal on Optimization* **1991**, *1*, 123–145.

(136) Torczon, V. J. *SIAM Journal on Optimization* **1997**, *7*, 1–25.

(137) Kolda, T. G.; Lewis, R. M.; Torczon, V. J. *SIAM Review* **2003**, *45*, 385–482.

(138) Audet, C.; Dennis Jr., J. E. *SIAM Journal on Optimization* **2006**, *17*, 188–217.

(139) Nelder, J. A.; Mead, R. *Computer Journal* **1965**, *7*, 308–313.

(140) Conn, A. R.; Scheinberg, K.; Vicente, L. N. *SIAM Journal on Optimization* **2009**, *20*, 387–415.

(141) Gilmore, P.; Kelley, C. T. *SIAM Journal on Optimization* **1995**, *5*, 269–285.

(142) Jones, D. R.; Perttunen, C. D.; Stuckman, B. E. *Journal of Optimization Theory and Application* **1993**, *79*, 157–181.

(143) Huyer, W.; Neumaier, A. *Journal of Global Optimization* **1999**, *14*, 331–355.

(144) Abramson, M. A.; Audet, C. *SIAM Journal on Optimization* **2006**, *17*, 606–609.

(145) Abramson, M. A.; Audet, C.; Dennis Jr., J. E.; Le Digabel, S. *SIAM Journal on Optimization* **2009**, *20*, 948–966.

(146) Cocchi, G.; Liuzzi, G.; Papini, A.; Sciandrone, M. *Computational Optimization and Applications* **2018**, *69*, 267–296.

(147) Huyer, W.; Neumaier, A. *ACM Transactions on Mathematical Software* **2008**, *35*, 1–25.

(148) Hansen, N. In *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, Lozano, J. A., Larrañaga, P., Inza, I., Bengoetxea, E., Eds.; Springer: 2006, pp 75–102.

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

(150) Ploskas, N.; Laughman, C.; Raghunathan, A. U.; Sahinidis, N. V. *Chemical Engineering Research and Design* **2018**, *131*, 16–28.

(151) Liu, J.; Ploskas, N.; Sahinidis, N. V. *Journal of Global Optimization* **2019**, *74*, 611–637.

(152) Read, D. J.; Auhl, D.; Das, C.; den Doelder, J.; Kapnistos, M.; Vittorias, I.; McLeish, T. C. B. *Science* **2011**, *333*, 1871–1874.

(153) Das, C.; Inkson, N. J.; Read, D. J.; Kelmanson, M. A.; McLeish, T. C. B. *Journal of Rheology* **2006**, *50*, 207–234.

(154) Holmström, K.; Göran, A. O.; Edvall, M. M. User's Guide for TOMLAB 7, http://tomopt.com; Tomlab Optimization, 2010.

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

(156) Sandia National Laboratories HOPSPACK Home Page, http://www.sandia.gov/hopspack/index.shtml.

(157) Vaz, A. I. F.; Vicente, L. N. *Optimization Online Digest* **2008**, 1–22.

(158) Custódio, A. L.; Vicente, L. N. SID-PSM: A pattern search method guided by simplex derivatives for use in derivative-free optimization, http://www.mat.uc.pt/sid-psm/; Departamento de Matemática, Universidade de Coimbra, Coimbra, Portugal, 2008.

(159) Le Digabel, S. *ACM Transactions on Mathematical Software* **2011**, *37*, 1–15.

(160) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. *Angewandte Chemie International Edition* **2016**, *55*, 5904–5937.

(161) Corey, E. J. *Pure and Applied Chemistry* **1967**, *14*, 19–38.

(162) Johnson, P. Y.; Burnstein, I.; Crary, J.; Evans, M.; Wang, T. *Expert System Applications in Chemistry* **1989**, 102–123.

(163) Gelernter, H. L.; Sanders, A. F.; Larsen, D. L.; Agarwal, K. K.; Boivie, R. H.; Spritzer, G. A.; Searleman, J. E. *Science* **1977**, *197*, 1041–1049.

(164) Krebsbach, D.; Gelernter, H.; Sieburth, S. M. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 595–604.

(165) Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. *Journal of the American Chemical Society* **1972**, *94*, 421–430.

(166) Corey, E. J.; Long, A. K.; Rubenstein, S. D. *Science* **1985**, *228*, 408–418.

(167) Ihlenfeldt, W.; Gasteiger, J. *Angewandte Chemie International Edition in English* **1996**, *34*, 2613–2633.

(168) Todd, M. H. *Chemical Society Reviews* **2005**, *34*, 247–266.

(169) Cook, A.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A. *WIREs Computational Molecular Science* **2012**, *2*, 79–107.

(170) Warr, W. A. *Molecular Informatics* **2014**, *33*, 469–476.

(171) Engkvist, O.; Norrby, P.; Selmi, N.; Lam, Y.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. *Drug Discovery Today* **2018**, *23*, 1203–1218.

(172) He, K.; Zhang, X.; Ren, S.; Sun, J. *CoRR* **2015**.

(173) Amodei, D. et al. *CoRR* **2015**.

(174) Silver, D. et al. *Nature* **2017**, *550*.

(175) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. *Journal of Chemical Information and Modeling* **2009**, *49*, 593–602.

(176) Segler, M. H. S.; Waller, M. P. *Chemistry A European Journal* **2017**, *23*, 5966–5971.

(177) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. *ACS Central Science* **2017**, *3*, 1237–1245.

(178) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate, 2016.

(179) Sutskever, I.; Vinyals, O.; Le, Q. V. Sequence to Sequence Learning with Neural Networks, 2014.

(180) Weininger, D. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(181) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Nguyen, Q. L.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. *ACS Central Science* **2017**, *3*, 1103–1113.

(182) Coley, C. W.; Green, W. H.; Jensen, K. F. *Accounts of Chemical Research* **2018**, *51*, 1281–1289.

(183) De Almeida, A. F.; Moreira, R.; Rodrigues, T. *Nature Reviews Chemistry* **2019**, *3*, 589–604.

(184) Klucznik, T. et al. *Chem* **2018**, *4*, 522–532.

(185) Gajewska, E. P.; Szymkuć, S.; Dittwald, P.; MichałStartek; Popik, O.; Mlynarski, J.; Grzybowski, B. A. *Chem* **2020**, *6*, 280–293.

(186) Llanos, E. J.; Leal, W.; Luu, D. H.; Jost, J.; Stadler, P. F.; Restrepo, G. *Proceedings of the National Academy of Sciences* **2019**, *116*, 12660–12665.

(187) Dai, H.; Li, C.; Coley, C. W.; Dai, B.; Song, L. *CoRR* **2020**.

(188) Plehiers, P. P.; Marin, G. B.; Stevens, C. V.; Geem, K. M. V. *Journal of Cheminformatics* **2018**, *10*.

(189) Coley, C. W., https://github.com/connorcoley/retrosim/blob/master/retrosim/utils/generate_retro_templates.py#L768 , Accessed: 02/01/2021, 2017.

(190) Coley, C. W.; Green, W. H.; Jensen, K. F. *Journal of Chemical Information and Modeling* **2019**, *59*, 2529–2537.

(191) Landrum, G. RDKit: Open-source cheminformatics, http://www.rdkit.org, Accessed:02/10/2021, 2021.

(192) Daylight Chemical Information Systems Inc., 4. SMARTSâ€"a Language for Describing Molecular Patterns, http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, Accessed:02/10/2021, 2019.

(193) Marcou, G.; Aires de Sousa, J.; Latino, D. A. R. S.; de Luca, A.; Horvath, D.; Rietsch, V.; Varnek, A. *Journal of Chemical Information and Modeling* **2015**, *55*, 239–250.

(194) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. *ACS Central Science* **2018**, *4*, 1465–1476.

(195) Walker, E.; Kammeraad, J.; Goetz, J.; Robo, M. T.; Tewari, A.; Zimmerman, P. M. *Journal of Chemical Information and Modeling* **2019**, *59*, 3645–3654.

(196) Maser, M. R.; Cui, A. Y.; Ryou, S.; DeLano, T. J.; Yue, Y.; Reisman, S. E. *Journal of Chemical Information and Modeling* **2021**, *61*, 156–166.

(197) Barnard, J. M. *Journal of Chemical Information and Computer Sciences* **1993**, *33*, 532–538.

(198) Raymond, J. W.; Willett, P. *Journal of Computer-Aided Molecular Design* **2002**, *16*, 521–533.

(199) Willett, P. *Journal of Medicinal Chemistry* **2005**, *48*, 4183–4199.

(200) Fukunishi, Y.; Kurosawa, T.; Mikami, Y.; Nakamura, H. *Journal of Chemical Information and Modeling* **2014**, *54*, 3259–3267.

(201) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. *Journal of Chemical Information and Modeling* **2018**, *58*, 252–261.

(202) Podolyan, Y.; Walters, M. A.; Karypis, G. *Journal of Chemical Information and Modeling* **2010**, *50*, 979–991.

(203) Li, J.; Eastgate, M. D. *Organic & Biomolecular Chemistry* **2015**, *13*, 7164–7176.

(204) Rogers, D.; Hahn, M. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

(205) Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604–610.

(206) Willett, P.; Barnard, J. M.; Downs, G. M. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 983–996.

(207) Bender, A.; Glen, R. C. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

(208) Baylon, J. L.; Cilfone, N. A.; Gulcher, J. R.; Chittenden, T. W. *Journal of Chemical Information and Modeling* **2019**, *59*, 673–688.

(209) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. *Journal of Cheminformatics* **2015**, *7*.

(210) Cadeddu, A.; Wylie, E. K.; Jurczak, J.; Wampler-Doty, M.; Grzybowski, B. A. *Angewandte Chemie International Edition* **2014**, *53*, 8108–8112.

(211) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. *ACS Central Science* **2018**, *4*, 120–131.

(212) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. *Molecular Informatics* **2018**, *37*, 1700111.

(213) Popova, M.; Isayev, O.; Tropsha, A. *Science Advances* **2018**, *4*.

(214) Schwalbe-Koda, D.; Gómez-Bombarelli, R. *CoRR* **2019**.

(215) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. *ACS Central Science* **2017**, *3*, 434–443.

(216) Nam, J.; Kim, J. *CoRR* **2016**.

(217) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. *ACS Central Science* **2019**, *5*, 1572–1583.

(218) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. *Chemical Science* **2018**, *9*, 6091–6098.

(219) Lin, K.; Xu, Y.; Pei, J.; Lai, L. *Chem. Sci.* **2020**, *11*, 3355–3364.

(220) Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. *Journal of Chemical Information and Modeling* **2020**, *60*, 47–55.

(221) Houben, C.; Lapkin, A. A. *Current Opinion in Chemical Engineering* **2015**, *9*, Energy and environmental engineering · Reaction engineering and catalysis, 1–7.

(222) Gromski, P. S.; Granda, J. M.; Cronin, L. *Trends in Chemistry* **2020**, *2*, 4–12.

(223) Peplow, M. *Nature* **2014**, *512*, 20–22.

(224) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. *ACS Central Science* **2018**, *4*, 268–276.

(225) Griffiths, R.-R.; Hernández-Lobato, J. M. *Chemical Science* **2020**, *11*, 577–586.

(226) Coley, C. W.; Eyke, N. S.; Jensen, K. F. *Angewandte Chemie International Edition* **2020**, *59*, 22858–22893.

(227) Coley, C. W.; Eyke, N. S.; Jensen, K. F. *Angewandte Chemie International Edition* **2020**, *59*, 23414–23436.

(228) Nikolaev, P.; Hooper, D.; Webber, F.; Rao, R.; Decker, K.; Krein, M.; Poleski, J.; Barto, R.; Maruyama, B. *npj Computational Materials* **2016**, *2*.

(229) Schneider, G. *Nature Reviews Drug Discovery* **2018**, *17*, 97–113.

(230) Button, A.; Merk, D.; Hiss, J. A.; Schneider, G. *Nature Machine Intelligence* **2019**, *1*, 307â€“315.

Appendix A

Group contribution model statistical analysis

A.1 Fitting correlation plots and error residual histograms

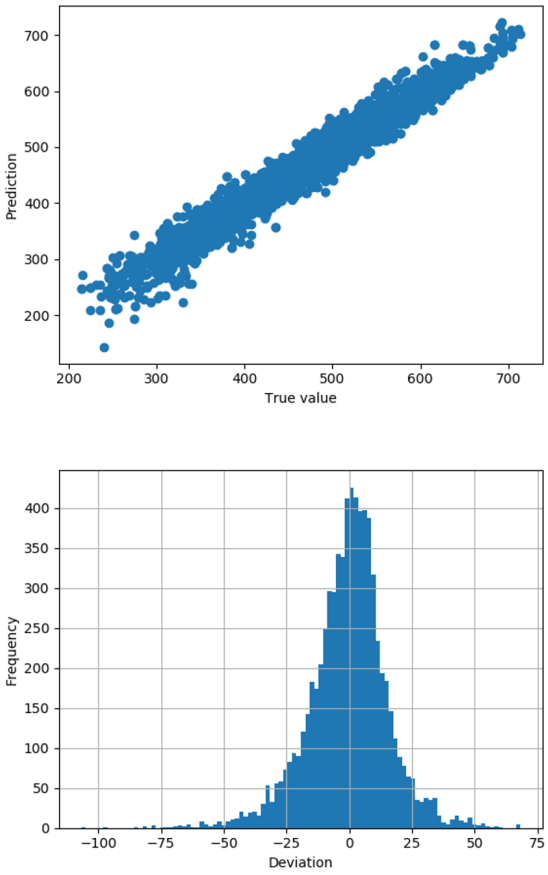


Figure A.1.1: Normal boiling point (K) parity plot ($R^2 = 0.96$) and error residual plot. 75.27% of the compounds deviate by no more than one training RMSE.

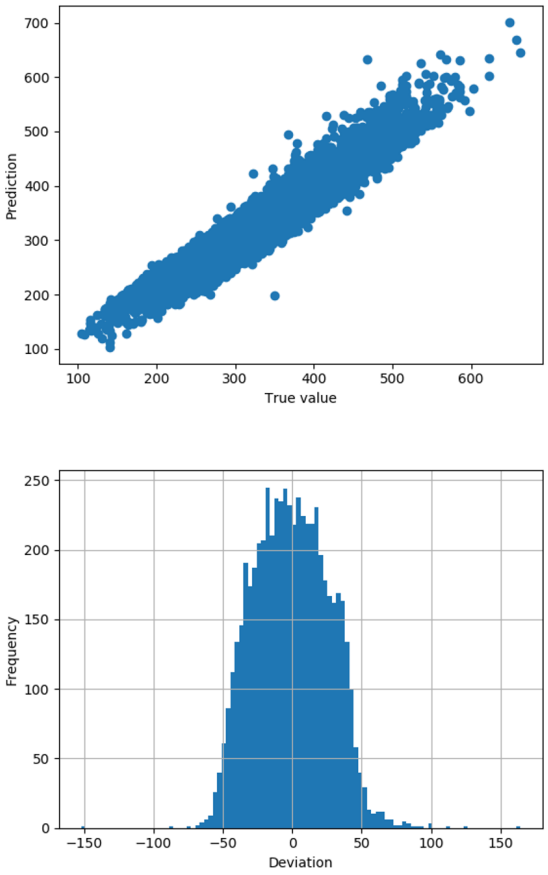


Figure A.1.2: Melting point (K) parity plot ($R^2 = 0.90$) and error residual plot. 64.04% of the compounds deviate by no more than one training RMSE.

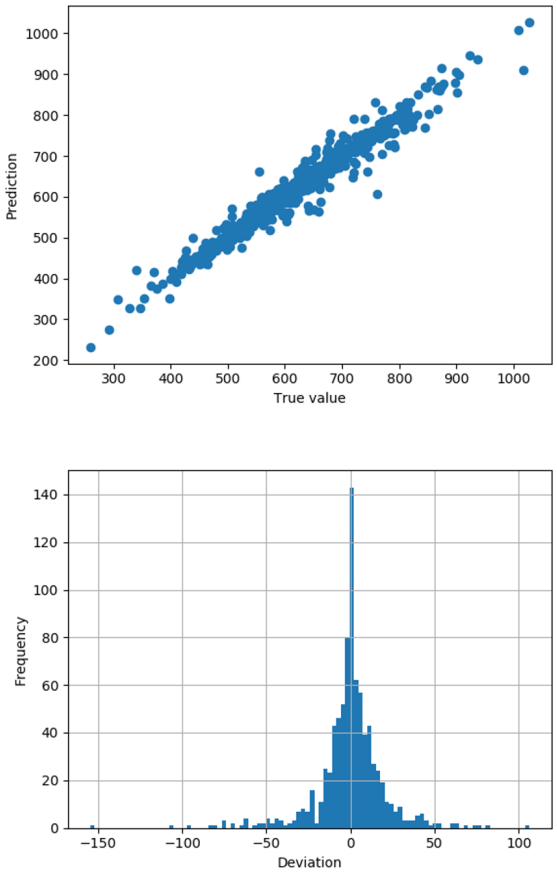


Figure A.1.3: Critical point (K) parity plot ($R^2 = 0.96$) and error residual plot. 81.94% of the compounds deviate by no more than one training RMSE.

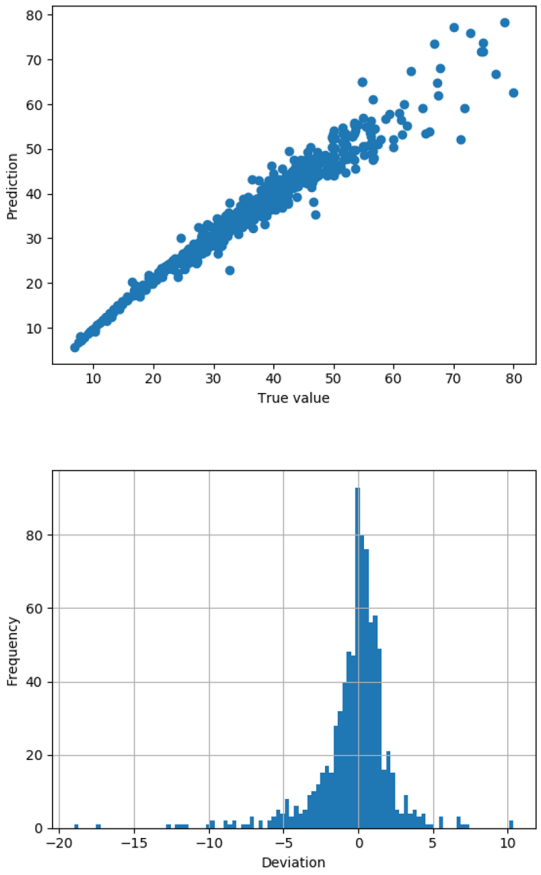


Figure A.1.4: Critical pressure (bar) parity plot ($R^2 = 0.95$) and error residual plot. 83.17% of the compounds deviate by no more than one training RMSE.

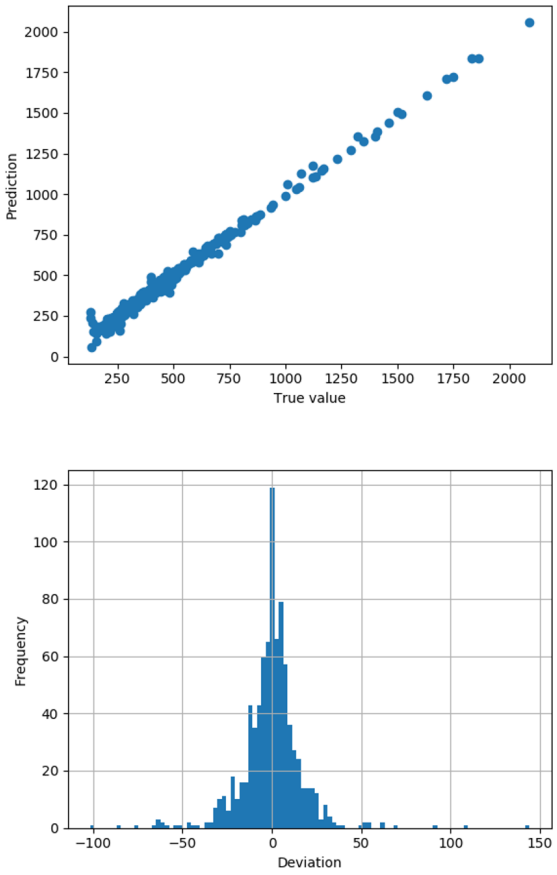


Figure A.1.5: Critical volume (cc/mol) parity plot ($R^2 = 0.99$) and error residual plot. 80.75% of the compounds deviate by no more than one training RMSE.

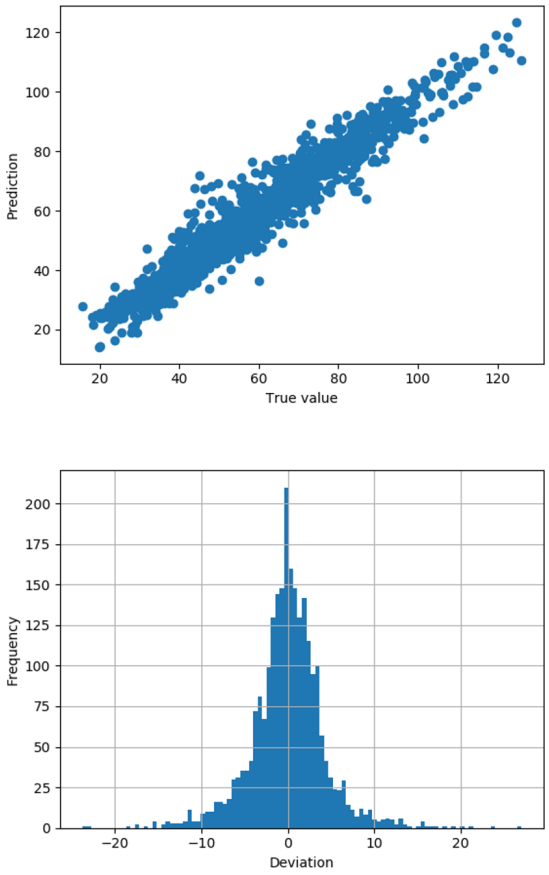


Figure A.1.6: Enthalpy of vaporization at 298 K (kJ/mol) parity plot ($R^2 = 0.95$) and error residual plot. 78.33% of the compounds deviate by no more than one training RMSE.

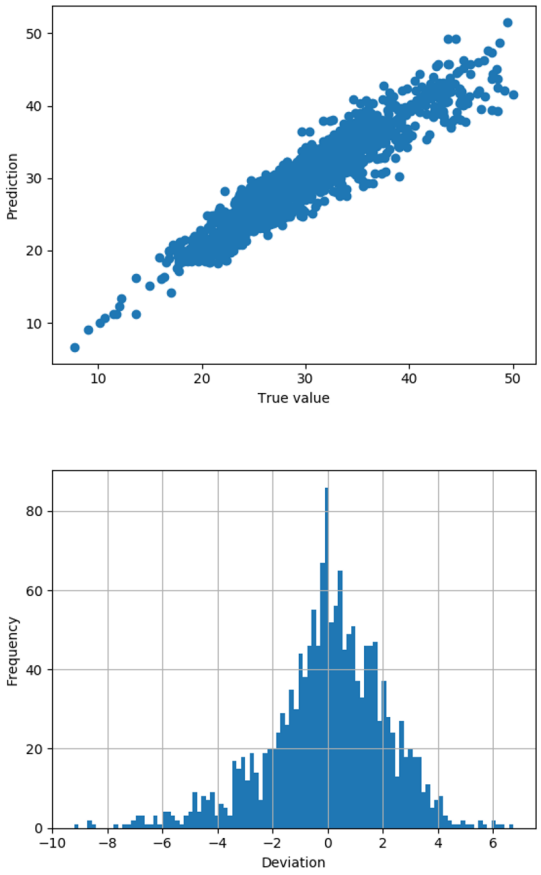


Figure A.1.7: Surface tension (N/m) parity plot ($R^2 = 0.89$) and error residual plot. 72.38% of the compounds deviate by no more than one training RMSE.

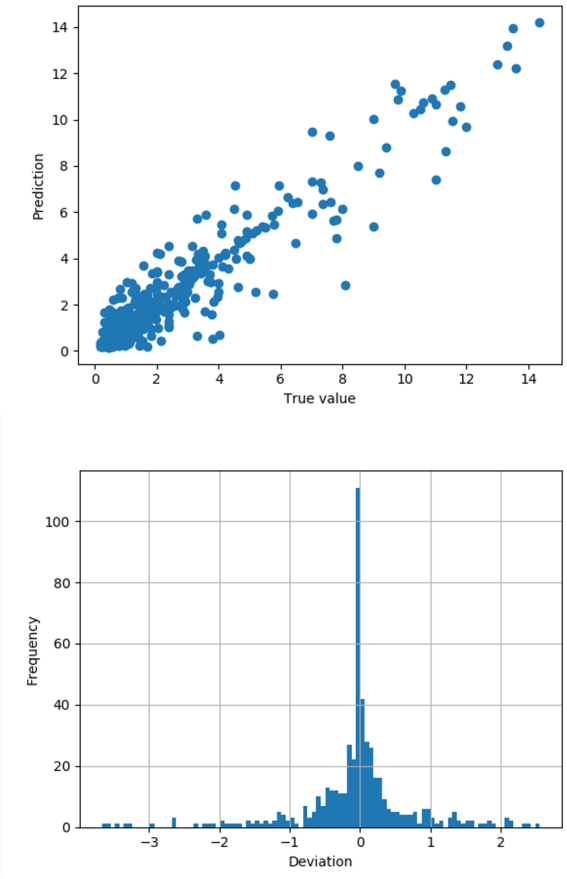


Figure A.1.8: Dynamic viscosity (mPa s) parity plot ($R^2 = 0.91$) and error residual plot. 81.87% of the compounds deviate by no more than one training RMSE.

A.2 GC coefficient

First-order GC coefficients are presented in Table A.1. Higher-order GC coefficients are shown in Table A.2 and A.3. All functional groups are represented using SMARTS strings [192]. For more information about SMARTS representation, please refer to https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

Table A.1: First-order group contribution coefficients for boiling point (T_b), melting point (T_m),critical temperature (T_c), critical pressure (P_c), critical volume (V_c), enthalpy of vaporization (H_{vap}),surface tension (σ), and dynamic viscosity (μ)

Groups	T_b (K)	T_m (K)	T_c (K)	P_c (bar)	V_c (cc/mol)	H_{vap} at 298K (kJ/mol)	σ at 298K (N/m)	μ at 298K (mPa s)
[CH3x0D1]	1.0209	-0.0531	1.7646	0.0300	43.0805	2.7705	-0.0514	1.4088
[CH2x0D2]	0.3220	0.0121	0.5490	0.0132	56.3625	4.3693	0.2864	0.1549
[CH1x0D3]	-0.5899	-0.0096	-0.7440	-0.0024	63.3063	3.0149	-0.4449	-1.2247
[CH0x0D4]	-1.3759	0.1463	-1.9683	-0.0217	68.0951	1.8031	1.2498	-2.0937
[CH2x0D1]=[CH1x0D2]	1.3068	-0.1403	2.3223	0.0360	86.3883	4.8931	0.4382	2.2522
[CH1x0D2]=[CH1x0D2]	0.6430	0.0295	1.6912	0.0221	100.1672	7.0035	0.0000	2.9716
[CH2x0D1]=[CH0x0D3]	0.5110	0.0351	1.6019	0.0180	91.3308	5.0483	-3.2278	2.2681
[CH1x0D2]=[CH0x0D3]	-0.1037	0.2469	1.0577	0.0055	111.3959	8.4481	-0.0624	1.6678
[CH0x0D3]=[CH0x0D3]	-1.0856	0.3411	0.6118	-0.0066	142.3349	5.9021	-1.7767	1.5300
[CH2x0D1]=[CH0x0D2]=[CH1x0D2]	0.0000	-0.4923	2.9679	0.0428	119.3953	0.0000	0.0000	0.0000
[CH2x0D1]=[CH0x0D2]=[CH0x0D3]	0.0000	0.0000	1.7204	0.0271	147.4960	0.0000	0.0000	0.0000
[CH1x0D1]#[CH0x0D2]	1.3459	0.0000	2.3435	0.0213	60.5633	5.6108	4.1718	0.7414
[CH0x0D2]#[CH0x0D2]	0.7216	0.0000	1.6982	0.0062	76.2532	10.4739	2.1019	0.0000
[cH1x2D2]	0.6162	0.0092	1.1269	0.0108	33.3729	3.9910	2.0491	0.6168
[cH0x3D3]	0.3233	0.0690	2.0756	-0.0005	31.1623	5.2344	1.0293	-0.2063
[cH0x2D3]	-0.1018	0.0687	0.8095	-0.0005	51.2809	5.1585	-1.8295	-0.5147
[nH0x2D3]	0.4322	0.0170	1.6070	-0.0134	49.0850	14.6145	0.0000	0.0000
[cH0x2D3][CH3x0D1]	0.8830	0.0503	2.5201	0.0264	89.8239	8.1802	0.4577	0.3996
[cH0x2D3][CH2x0D2]	0.1218	0.0635	1.0761	0.0130	105.2433	9.3052	0.3467	-0.7308
[cH0x2D3][CH1x0D3]	-0.8251	0.1087	-0.2898	-0.0047	117.6477	9.3730	-2.5452	-0.4773
[cH0x2D3][CH0x0D4]	-1.3738	0.2232	-2.1785	-0.0180	129.9231	6.5669	-0.0500	0.0000
[cH0x2D3][CH1x0D2]=[CH2x0D1]	1.2427	-0.0986	3.4328	0.0354	128.3551	12.1534	14.5742	1.1469
[cH0x2D3][CH1x0D2]=[CH1x0D2]	0.8129	0.1407	2.7994	0.0218	140.2713	18.3320	2.0496	1.9704
[cH0x2D3][CH0x0D3]=[CH2x0D1]	0.4571	0.0000	2.1902	0.0207	131.7121	13.2426	0.0000	0.0000
[cH0x2D3][CH0x0D2]#[CH1x0D1]	1.0185	0.0000	3.3867	0.0188	107.7926	0.0000	0.0000	0.0000
[cH0x2D3][CH0x0D2]#[CH0x0D2]	0.7133	0.0000	0.0000	0.0000	0.0000	0.0000	1.5112	0.0000
[OH1x0D1]	1.8490	0.0671	3.4563	0.0039	3.9188	22.3440	7.9289	3.6372
[cH0x2D3][OH1x0D1]	1.5537	0.1861	5.0092	-0.0135	47.4184	24.5300	5.3290	2.6425
[CH0x0D3](=[OH0x0D1])[OH1x0D1]	2.8718	0.2245	5.5594	0.0196	65.4029	37.5219	7.0766	2.7751
[cH0x2D3][CH0x0D3](=[OH0x0D1])[OH1x0D1]	2.8446	0.3660	6.6180	0.0197	116.7487	47.0824	0.0000	0.0000
[CH3x0D1][CH0x0D3]=[OH0x0D1]	1.8562	0.0574	4.5509	0.0351	106.8901	15.4852	5.8323	2.2336
[CH2x0D2][CH0x0D3]=[OH0x0D1]	1.0794	0.1239	2.5647	0.0225	119.7991	15.4571	4.3210	1.6621
[CH1x0D3][CH0x0D3]=[OH0x0D1]	0.1171	0.0873	1.2836	-0.0037	123.4155	12.0399	0.5823	-0.1661
[CH0x0D4][CH0x0D3]=[OH0x0D1]	-0.8519	0.2186	-0.1974	-0.0136	138.3349	13.2951	0.0000	-2.4479
[cH0x2D3][CH0x0D3]=[OH0x0D1]	0.7053	0.1379	2.3778	0.0127	110.4739	18.0010	2.2696	0.2189
[CH1x0D2]=[OH0x0D1]	1.7904	0.0577	3.3615	0.0137	38.5980	11.6608	4.8514	1.6201
[cH0x2D3][CH1x0D2]=[OH0x0D1]	1.6111	0.1376	4.9294	0.0198	112.7896	16.6865	10.5416	1.6422
[CH3x0D1][CH0x0D3](=[OH0x0D1])[OH0x0D2]	1.9155	0.0815	3.2748	0.0448	118.3483	16.6749	6.5017	1.7795
[CH2x0D2][CH0x0D3](=[OH0x0D1])[OH0x0D2]	1.2460	0.0000	2.3853	0.0238	137.3421	16.2438	5.8395	0.6173
[CH1x0D3][CH0x0D3](=[OH0x0D1])[OH0x0D2]	0.3353	0.0000	1.0221	0.0136	154.0530	14.5605	4.9780	-0.4352
[CH0x0D4][CH0x0D3](=[OH0x0D1])[OH0x0D2]	-0.2597	0.0000	0.0000	0.0000	0.0000	12.6563	5.0868	0.0000
[CH1x0D2](=[OH0x0D1])[OH0x0D2]	1.7406	-0.1239	3.2577	0.0236	69.3584	15.8112	6.4604	1.8932

Table A.1 continued from previous page

Groups	T_b (K)	T_m (K)	T_c (K)	P_c (bar)	V_c (cc/mol)	H_{vap} at 298K (kJ/mol)	σ at 298K (N/m)	μ at 298K (mPa s)
[cH0x2D3][CH0x0D3](=[OH0x0D1])[OH0x0D2]	0.6774	0.1267	2.3274	0.0174	119.6389	20.4784	0.0862	0.5828
[cH0x2D3][OH0x0D2][CH0x0D3]=[OH0x0D1]	0.6385	0.1487	0.0000	0.0000	0.0000	25.7500	0.0000	0.0000
[CH0x0D3](=[OH0x0D1])[OH0x0D2]	0.9080	0.0522	1.7800	0.0231	88.3501	13.6344	4.3380	0.9659
[CH3x0D1][OH0x0D2]	1.3669	-0.0162	2.3453	0.0341	63.2124	7.4327	3.8469	1.6988
[CH2x0D2][OH0x0D2]	0.5205	0.0000	0.8471	0.0162	75.4021	6.8645	1.7706	0.1482
[CH1x0D3][OH0x0D2]	-0.4427	0.0000	-0.7761	0.0084	67.7376	5.5508	-0.2370	-1.2992
[CH0x0D4][OH0x0D2]	-1.2008	0.1019	-1.6229	-0.0148	95.9725	4.3248	0.9989	-1.9527
[cH0x2D3][OH0x0D2]	0.1072	0.0982	2.1721	-0.0015	69.6025	9.3174	-0.2736	-0.7458
[CH2x0D2][NH2x0D1]	1.7887	0.1091	3.2029	0.0112	105.6172	14.2048	0.9339	2.6669
[CH1x0D3][NH2x0D1]	0.8810	0.1753	1.5988	0.0000	77.4960	12.9276	0.0425	0.5966
[CH0x0D4][NH2x0D1]	-0.0115	0.2929	0.0596	-0.0033	106.4155	11.5199	0.3706	-1.7896
[CH3x0D1][NH1x0D2]	1.6855	0.1307	4.3813	0.0137	86.3797	17.2697	8.1869	3.1359
[CH2x0D2][NH1x0D2]	0.8656	0.1797	2.4235	0.0000	107.4312	16.6199	7.9887	0.0000
[CH1x0D3][NH1x0D2]	-0.1852	0.1213	0.0000	0.0000	0.0000	16.4021	12.3792	0.0000
[CH3x0D1][NH0x0D3]	0.7574	-0.0335	-0.2248	0.0000	77.3695	0.0000	5.9798	0.0000
[CH2x0D2][NH0x0D3]	0.0000	0.1757	0.0000	0.0000	0.0000	9.3121	2.1737	0.0000
[cH0x2D3][NH2x0D1]	1.8517	0.1931	5.4797	0.0051	76.5962	20.5134	10.7479	1.6341
[cH0x2D3][NH1x0D2]	0.7568	0.1779	0.0000	0.0000	0.0000	29.5825	6.8585	0.0000
[cH0x2D3][NH0x0D3]	-0.4030	0.1946	1.0660	-0.0198	154.6316	18.0083	2.1002	-1.3680
[NH2x0D1]	1.6494	0.1245	3.2291	0.0107	39.4253	11.2031	4.5947	2.1962
[CH1x0D2]=[NH0x0D2]	0.9931	0.1422	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[CH2x0D2][CH0x0D2]#[NH0x0D1]	2.6740	0.0000	5.4234	0.0461	124.1950	21.4027	10.1512	1.9785
[CH1x0D3][CH0x0D2]#[NH0x0D1]	1.5603	0.0000	3.4111	0.0246	136.5768	20.0725	10.5487	0.0000
[CH0x0D4][CH0x0D2]#[NH0x0D1]	0.5344	0.2972	1.3497	0.0099	148.5771	15.5945	0.0000	0.0000
[cH0x2D3][CH0x0D2]#[NH0x0D1]	1.7846	0.1980	5.0264	0.0207	114.7926	19.8247	9.3491	1.4100
[CH0x0D2]#[NH0x0D1]	2.1336	0.0543	4.7341	0.0364	65.2455	15.9550	8.3508	2.1870
[NH0x0D2]=[CH0x0D2]=[OH0x0D1]	1.7001	0.0000	2.9126	0.0295	83.1158	8.4493	0.0000	0.0000
[CH2x0D2][N+1H0x0D3](=[OH0x0D1])[O-1H0x0D1]	2.6463	0.0000	5.6742	0.0208	135.5765	0.0000	11.1608	2.4800
[CH1x0D3][N+1H0x0D3](=[OH0x0D1])[O-1H0x0D1]	1.6575	0.0000	4.1088	0.0000	144.4960	0.0000	9.4163	1.3713
[CH0x0D4][N+1H0x0D3](=[OH0x0D1])[O-1H0x0D1]	0.0714	0.3312	0.0000	0.0000	0.0000	0.0000	9.8192	0.0000
[cH0x2D3][N+1H0x0D3](=[OH0x0D1])[O-1H0x0D1]	2.0374	0.1703	6.0600	0.0219	127.1729	0.0000	13.6828	0.0000
[OH0x0D1]=[N+1H0x0D3][O-1H0x0D1]	1.9771	0.0818	4.8845	0.0146	74.8516	0.0000	9.4245	2.4063
[OH0x0D2][NH0x0D2]=[OH0x0D1]	1.3156	0.0000	0.0000	0.0000	0.0000	0.0000	4.3481	0.0000
[OH0x0D2][N+1H0x0D3](=[OH0x0D1])[O-1H0x0D1]	1.8984	0.1331	0.0000	0.0000	0.0000	0.0000	8.8948	0.0000
[OH0x0D1]=[CH1x0D2][NH0x0D3]([CH2x0D2])[CH2x0D2]	2.5533	0.0000	0.0000	0.0000	0.0000	31.2651	0.0000	1.6876
[OH0x0D1]=[CH1x0D2][NH1x0D2][CH2x0D2]	3.9043	0.0000	0.0000	0.0000	0.0000	42.1356	0.0000	0.0000
[OH0x0D1]=[CH0x0D3][NH2x0D1]	3.4570	0.3397	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[OH0x0D1]=[CH0x0D3][NH1x0D2][CH3x0D1]	2.0070	0.2598	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[OH0x0D1]=[CH0x0D3][NH1x0D2][CH2x0D2]	2.5055	0.2603	0.0000	0.0000	0.0000	45.8651	0.0000	0.0000
[OH0x0D1]=[CH0x0D3][NH0x0D3]([CH3x0D1])[CH3x0D1]	3.3948	0.1723	0.0000	0.0000	0.0000	33.4896	0.0000	0.0000
[OH0x0D1]=[CH0x0D3][NH0x0D3]([CH2x0D2])[CH2x0D2]	1.4793	0.2373	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[OH0x0D1]=[CH0x0D3][NH1x0D2][CH0x0D3]=[OH0x0D1]	0.0000	0.4597	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[cH0x2D3][CH0x0D3](=[OH0x0D1])[NH2x0D1]	2.9006	0.4545	0.0000	0.0000	0.0000	51.6468	0.0000	0.0000
[cH0x2D3][NH1x0D2][CH1x0D2]=[OH0x0D1]	3.3128	0.0000	8.7075	0.0239	157.7926	0.0000	0.0000	0.0000
[cH0x2D3][NH0x0D3][CH1x0D2]=[OH0x0D1]	1.6335	0.0000	0.0000	0.0000	0.0000	34.0806	0.0000	0.0000

Table A.1 continued from previous page

Groups	T_b (K)	T_m (K)	T_c (K)	P_c (bar)	V_c (cc/mol)	H_{vap} at 298K (kJ/mol)	σ at 298K (N/m)	μ at 298K (mPa s)
[cH0x2D3][CH0x0D3](=[OH0x0D1])[NH1x0D2]	0.0000	0.2491	0.0000	0.0000	0.0000	40.0991	0.0000	0.0000
[cH0x2D3][NH1x0D2][CH0x0D3]=[OH0x0D1]	1.7945	0.3061	0.0000	0.0000	0.0000	44.2768	0.0000	0.0000
[cH0x2D3][NH0x0D3][CH0x0D3]=[OH0x0D1]	0.7350	0.2646	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[NH1x0D2][CH0x0D3](=[OH0x0D1])[NH1x0D2]	2.6432	0.2957	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[NH2x0D1][CH0x0D3](=[OH0x0D1])[NH1x0D2]	0.0000	0.5009	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[NH2x0D1][CH0x0D3](=[OH0x0D1])[NH0x0D3]	0.0000	0.4513	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[NH0x0D3][CH0x0D3](=[OH0x0D1])[NH0x0D3]	0.0000	0.0000	0.0000	0.0000	0.0000	32.0240	0.0000	-0.9508
[cH0x2D3][NH1x0D2][CH0x0D3](=[OH0x0D1])[NH2x0D1]	2.3212	0.5448	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[cH0x2D3][NH1x0D2][CH0x0D3](=[OH0x0D1])[NH1x0D2]	0.0000	0.2960	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[NH1x0D2][CH0x0D3]=[OH0x0D1]	2.3010	0.1968	0.0000	0.0000	0.0000	42.2703	0.0000	0.0000
[CH2x0D2][ClH0x0D1]	1.7743	0.0000	3.4505	0.0276	85.5062	10.7916	6.0308	1.6977
[CH1x0D3][ClH0x0D1]	0.8095	-0.0600	1.8140	0.0128	103.1307	8.3682	4.3574	1.5587
[CH0x0D4][ClH0x0D1]	-0.3541	0.0000	0.4778	-0.0053	113.4155	8.7238	4.2576	0.0000
[ClH0x0D1][CH1x0D3][ClH0x0D1]	2.0013	0.0000	4.3110	0.0358	135.3222	14.6126	7.0230	2.2385
[ClH0x0D1][CH0x0D4][ClH0x0D1]	1.0687	0.1668	0.0000	0.0000	0.0000	12.9804	-8.6486	0.4664
[ClH0x0D1][CH0x0D4]([ClH0x0D1])[ClH0x0D1]	2.1453	0.0949	4.5589	0.0423	180.5765	16.0026	5.6597	1.2492
[CH2x0D2][FH0x0D1]	1.3507	0.0000	1.7964	0.0237	63.5765	0.0000	0.0000	1.6788
[CH1x0D3][FH0x0D1]	0.3420	0.0000	0.0000	0.0000	0.0000	0.0000	4.4380	0.5803
[CH0x0D4][FH0x0D1]	-1.3156	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.2197
[FH0x0D1][CH1x0D3][FH0x0D1]	1.1761	-0.1476	1.9331	0.0362	78.5765	9.8427	2.2640	1.8549
[FH0x0D1][CH0x0D4][FH0x0D1]	0.1966	0.0000	0.0000	0.0000	0.0000	0.0000	-2.4649	0.3831
[FH0x0D1][CH0x0D4]([FH0x0D1])[FH0x0D1]	0.8531	-0.1720	1.2714	0.0633	89.3562	1.8400	-5.1415	1.0006
[ClH0x0D1][CH0x0D4]([ClH0x0D1])[FH0x0D1]	1.7090	0.0000	3.1885	0.0471	138.6694	10.2524	1.0874	2.2294
[ClH0x0D1][CH1x0D3][FH0x0D1]	1.5052	0.0000	0.0000	0.0000	0.0000	11.4829	0.0000	0.0000
[ClH0x0D1][CH0x0D4]([FH0x0D1])[FH0x0D1]	1.1702	-0.1481	2.0125	0.0557	128.9877	4.3690	-3.8506	1.6887
[cH0x2D3][ClH0x0D1]	1.1225	0.0944	3.4612	0.0182	86.8873	8.4223	3.6598	0.3107
[cH0x2D3][FH0x0D1]	0.5095	0.0211	1.9835	0.0202	21.7911	4.3153	0.0000	0.2009
[cH0x2D3][IH0x0D1]	1.9620	0.0975	5.8426	0.0122	126.7926	18.6338	9.6053	0.0000
[cH0x2D3][BrH0x0D1]	1.4555	0.1184	4.2858	0.0064	99.0671	12.9514	6.0283	0.8702
[IH0x0D1]	1.9433	0.0000	5.0375	0.0064	79.6921	13.8968	10.2554	0.0000
[BrH0x0D1]	1.6492	0.0308	2.8826	-0.0001	-5.0652	7.0771	7.4055	2.0677
[FH0x0D1]	0.7906	-0.0806	1.2201	0.0237	9.3412	0.3053	-1.1616	1.1491
[ClH0x0D1]	1.2673	-0.0210	2.2944	0.0232	27.2863	5.3904	3.3327	1.3964
[CH1x0D2]=[NH0x0D2][OH1x0D1]	2.5120	0.3450	0.0000	0.0000	0.0000	32.7656	8.2107	0.0000
[CH0x0D3]=[NH0x0D2][OH1x0D1]	1.8329	0.3513	0.0000	0.0000	0.0000	34.2957	8.1246	0.0000
[cH0x2D3][CH1x0D2]=[NH0x0D2][OH1x0D1]	1.8778	0.3021	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[OH0x0D2][CH2x0D2][CH2x0D2][OH1x0D1]	2.8056	0.0000	4.9788	0.0241	141.5765	29.0429	13.3068	3.5879
[OH0x0D2][CH1x0D3][CH2x0D2][OH1x0D1]	1.7513	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[OH0x0D2][CH2x0D2][CH1x0D3][OH1x0D1]	1.8949	0.0000	3.7146	0.0000	150.4960	27.1651	9.9982	1.5135
[OH0x0D2][OH1x0D1]	2.4449	0.0000	3.3607	0.0029	35.3204	24.3914	0.0000	0.0000
[CH2x0D2][SH1x0D1]	2.0084	0.0000	3.8059	0.0228	87.6914	14.4928	3.6165	0.0000
[CH1x0D3][SH1x0D1]	1.0049	-0.5248	2.3966	0.0000	110.4960	10.5624	0.0000	0.0000
[CH0x0D4][SH1x0D1]	0.0000	0.0000	0.9278	-0.0105	120.4155	10.0532	0.0000	-1.6073
[cH0x2D3][SH1x0D1]	1.4491	0.1262	4.6595	0.0065	90.7926	0.0000	0.0000	0.0000
[SH1x0D1]	1.7202	-0.0432	3.5086	0.0205	24.4881	10.1020	4.1915	1.9104

Table A.1 continued from previous page

Groups	T_b (K)	T_m (K)	T_c (K)	P_c (bar)	V_c (cc/mol)	H_{vap} at 298K (kJ/mol)	σ at 298K (N/m)	μ at 298K (mPa s)
[CH3x0D1][SH0x0D2]	1.8767	0.0000	0.0000	0.0000	0.0000	12.3916	7.1833	0.0000
[CH2x0D2][SH0x0D2]	1.0728	0.0000	2.5976	0.0130	107.8976	13.0917	5.2000	0.0000
[CH1x0D3][SH0x0D2]	0.2310	0.0000	0.0000	0.0000	0.0000	12.1891	4.1810	0.0000
[CH0x0D4][SH0x0D2]	-0.4915	0.0000	0.0344	-0.0224	130.3349	9.6240	5.3277	0.0000
[cH0x2D3][SH0x0D2]	0.5180	0.0628	0.0000	0.0000	0.0000	18.0806	6.1763	0.0000
[SH0x0D3]=[OH0x0D1]	2.4588	0.0000	8.2570	0.0000	83.4960	38.3692	23.0797	1.7514
[OH0x0D1]=[SH0x0D4]=[OH0x0D1]	1.8723	0.1458	7.0692	0.0000	90.6837	45.3086	13.7066	1.5811
[OH0x0D1]=[SH0x0D3]([OH0x0D2])[OH0x0D2]	1.6399	0.0000	3.6344	0.0000	116.7711	21.0188	0.0000	0.0000
[OH0x0D1]=[SH0x0D4](=[OH0x0D1])[OH0x0D2]	1.6178	0.2300	0.0000	0.0000	0.0000	0.0000	14.0974	0.0000
[OH0x0D1]=[SH0x0D4](=[OH0x0D1])([OH0x0D2])[OH0x0D2]	2.3733	0.0000	7.8557	-0.0506	145.6335	29.8071	20.2310	1.7826
[cH0x2D3][SH0x0D3]=[OH0x0D1]	2.0455	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[cH0x2D3][SH0x0D4](=[OH0x0D1])=[OH0x0D1]	1.5360	0.2130	0.0000	0.0000	157.8738	0.0000	8.5923	0.0000
[PH0x0D3]	-0.4738	0.0000	10.9652	-0.1525	-157.7788	0.0000	-1.2484	0.0000
[OH0x0D1]=[PH1x0D3]([OH0x0D2])[OH0x0D2]	2.4457	0.0000	0.0000	0.0000	0.0000	0.0000	6.7448	0.0000
[OH0x0D1]=[PH0x0D4]([OH0x0D2])[OH0x0D2]	1.1549	0.0000	0.0000	0.0000	0.0000	25.0003	5.6806	0.0000
[OH0x0D1]=[PH0x0D4]([OH1x0D1])([OH0x0D2])[OH0x0D2]	2.4992	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[OH0x0D1]=[PH0x0D4]([OH0x0D2])([OH0x0D2])[OH0x0D2]	1.0496	0.0000	5.9965	-0.0574	111.8718	22.2173	7.4245	0.0093
[OH0x0D1]=[CH0x0D3]([OH0x0D2])[OH0x0D2]	1.1501	0.1945	2.6609	0.0120	103.6335	19.5655	5.5359	0.8858
[CH1x2D3R1][OH0x2D2][CH2x2D2R1]	2.0276	0.0459	2.9373	0.0069	109.5080	18.3150	12.8550	-0.4357
[CH1x2D3R1][OH0x2D2][CH1x2D3R1]	1.0158	0.0000	0.0000	0.0000	0.0000	12.7562	0.0000	0.0000
[CH2x2D2]	0.5408	0.0136	1.0981	0.0133	42.3168	3.5749	0.7817	0.4763
[CH1x2D3]	-0.0205	-0.0152	-1.6243	0.0004	57.8462	4.0151	2.6883	-2.9364
[CH0x2D4]	-1.1086	0.1468	0.8868	-0.0247	168.8401	0.6566	0.7796	-1.6425
[CH1x2D2]=[CH1x2D2]	1.0572	0.0404	2.4749	0.0112	66.1367	6.7128	1.2535	4.4515
[CH1x2D2]=[CH0x2D3]	0.2592	0.0539	0.2709	0.0000	23.7975	7.6098	0.7953	2.0831
[CH0x2D3]=[CH0x2D3]	-0.5034	0.1041	0.0000	0.0000	0.0000	9.2569	2.6177	0.0000
[CH2x0D1]=[CH0x2D3]	1.0050	0.0919	1.2588	0.0255	49.0672	5.4750	1.4584	3.3981
[NH1x2D2]	1.0420	0.1231	2.3650	-0.0071	37.7420	13.8603	7.4338	2.5918
[NH0x2D3]	0.0817	0.0056	0.4631	-0.0101	34.7550	5.8098	9.3485	-3.6908
[CH1x2D2]=[NH0x2D2]	2.0547	0.1602	0.0000	0.0000	0.0000	37.5213	0.0000	0.0000
[CH0x2D3]=[NH0x2D2]	0.5159	0.1256	0.0000	0.0000	0.0000	11.4548	0.0000	0.0000
[OH0x2D2]	0.7327	0.0192	1.8532	-0.0033	8.1796	4.5964	5.6654	0.6149
[CH0x2D3]=[OH0x0D1]	1.5300	0.1240	5.8435	0.0016	43.4254	17.6058	9.0861	2.8376
[SH0x2D2]	1.4141	0.0841	4.0946	-0.0025	0.0000	9.4319	12.2133	2.3442
[OH0x0D1]=[SH0x2D4]=[OH0x0D1]	4.1787	0.1648	10.6357	0.0007	78.8867	41.3067	12.4015	4.7048
[NH1x0D2]	0.6257	0.0800	0.9570	0.0000	45.3795	7.3125	2.3178	0.3232
[OH0x0D2]	0.2038	-0.0249	0.3172	0.0054	22.5253	2.5643	1.2621	-0.0388
[SH0x0D2]	0.8073	0.0176	2.2179	-0.0070	50.8490	9.6747	4.8577	0.7137
[CH0x0D3]=[OH0x0D1]	0.7605	0.0819	2.2074	0.0106	57.3424	11.5446	3.6517	1.6495
[CH1x0D3][NH0x0D3]	-1.3427	0.0000	0.0000	0.0000	0.0000	3.5554	0.0000	0.0000
[SiH0x0D4][OH1x0D1]	0.0000	0.2780	0.0000	0.0000	0.0000	15.6866	0.0000	1.6040
[SiH3x0D1]	1.1997	-0.3646	0.0000	0.0000	0.0000	9.8171	0.0000	0.0000
[SiH2x0D2]	0.0000	-0.4124	0.0000	0.0000	119.2903	0.0000	11.4041	-0.1056
[SiH1x0D3]	-0.6618	-0.3523	0.0000	0.0000	0.0000	0.0000	-2.2303	0.0000
[SiH0x0D4]	-1.5034	0.1025	-2.6899	-0.0039	137.6096	-0.6034	-1.1885	-2.4286

Table A.1 continued from previous page

Groups	T_b (K)	T_m (K)	T_c (K)	P_c (bar)	V_c (cc/mol)	H_{vap} at 298K (kJ/mol)	σ at 298K (N/m)	μ at 298K (mPa s)
[NH0x0D2]=[NH0x0D2]	1.1493	0.1011	4.6075	0.0000	91.0332	0.0000	0.0000	0.0000
[CH0x2D3]=[NH0x0D2]	0.6432	0.1699	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[CH0x2D3]=[CH1x0D2]	0.3105	0.1256	0.0000	0.0000	0.0000	7.4385	0.0000	0.0000
[NH0x0D2]=[OH0x0D1]	1.3517	0.1415	0.0000	0.0000	0.0000	0.0000	10.1700	0.0000
[CH0x2D3]=[CH0x0D3]	-0.5142	0.0000	1.1396	-0.0132	119.8389	8.4568	0.0000	0.0000
[CH0x0D3]=[NH0x0D2]	0.0000	0.1672	0.0000	0.0000	0.0000	0.0000	2.3643	0.0000
[CH0x0D3]=[NH1x0D1]	0.0000	0.1086	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[CH0x0D3]=[SH0x0D1]	1.0637	0.1869	0.0000	0.0000	0.0000	0.0000	5.2608	0.0000
[cH0x2D3][CH0x0D3](=[OH0x0D1])[NH0x0D3]	0.9696	0.3087	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[CH1x0D2](=[OH0x0D1])[NH1x0D2]	3.4871	0.2805	0.0000	0.0000	0.0000	39.9356	18.0396	3.5119
[CH1x0D2](=[OH0x0D1])[CH1x0D3]	1.3656	0.0000	1.0815	0.0250	120.7711	13.2651	0.0000	0.4540
[FH0x0D1][CH0x2D4][FH0x0D1]	0.4790	0.0000	0.0000	0.0000	0.0000	3.5097	-1.8639	0.0000
[SiH0x2D4]	-1.7108	0.1131	-2.2644	0.0131	117.2117	-0.7072	-5.8710	0.0000
[SiH1x2D3]	-0.7603	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[CH1x3D3]	0.4017	0.0433	0.3972	0.0057	36.4348	3.8776	1.3926	0.1920
[CH0x3D4]	-0.5820	0.1171	0.0000	0.0000	0.0000	3.3827	6.1501	-0.4373
[NH0x0D3]	-0.5051	0.1016	-0.7838	-0.0066	53.6841	3.2430	0.7996	-0.5209
[SiH0x0D3]	-0.5498	-0.3321	-1.1186	0.0172	133.6895	-5.9859	-13.1741	-1.5009
[SiH2x0D1]	1.1191	-0.2814	0.0000	0.0000	84.3370	0.0000	6.6314	0.0000
[oH0x2D2]	0.5655	0.0000	0.9384	-0.0071	0.0000	2.5427	4.4860	0.9354
[sH0x2D2]	1.1252	0.0241	2.2677	0.0013	35.3428	7.1896	5.8715	1.3255
[nH0x2D2]	0.7885	0.0643	2.4502	-0.0061	54.2162	8.3970	5.6065	1.0338
[nH1x2D2]	1.8876	0.2034	4.1252	0.0035	36.0951	22.0347	8.8919	2.0187
[SiH0x0D4][OH0x0D2]	0.0000	0.0000	-2.5348	0.0022	171.3893	0.0000	0.0000	0.0000

Table A.2: Higher-order group contribution coefficients for boiling point (T_b), melting point (T_m),critical temperature (T_c), critical pressure (P_c), and critical volume (V_c)

Groups	T_b (K)	T_m (K)	T_c (K)	P_c (bar)	V_c (cc/mol)
[CH1D3]([CH3D1])([CH3D1])	0.0000	0.0000	-0.1740	0.0000	0.0000
[CH0D4]([CH3D1])([CH3D1])([CH3D1])	0.0000	0.0000	-0.2111	0.0000	0.0000
[CH1D3]([CH3D1])[CH1D3]([CH3D1])	0.0964	-0.0367	0.1016	0.0000	0.0000
[CH1D3]([CH3D1])[CH0D4]([CH3D1])([CH3D1])	0.0758	0.0000	0.1141	0.0000	0.0000
[CH0D4]([CH3D1])([CH3D1])[CH0D4]([CH3D1])([CH3D1])	0.1112	0.0539	0.0886	0.0000	0.0000
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][CH0D3,CH1D2]=[CH0D3,CH1D2,CH2D1]	0.0553	0.0231	-0.0942	0.0000	0.0000
[CH3D1][CH0D3,CH1D2]=[CH0D3,CH1D2,CH2D1]	0.0000	-0.1137	-0.4451	0.0000	0.0000
[CH2D2][CH0D3,CH1D2]=[CH0D3,CH1D2,CH2D1]	-0.0307	-0.0508	-0.1598	0.0041	0.0000
[CH1D3,CH0D4][CH0D3,CH1D2]=[CH0D3,CH1D2,CH2D1]	0.0000	-0.0389	-0.2313	0.0000	0.0000
[CH1D3,CH0D4][CH1D2]=[OH0D1]	-0.3149	0.0000	1.0805	0.0000	0.0000
[CH3D1][CH0D3](=[OH0D1])[CH2D2]	0.2158	0.0000	-0.1691	0.0000	0.0000
[CH3D1][CH0D3](=[OH0D1])[CH1D3,CH0D4]	0.1192	0.0000	0.0935	0.0000	0.0000
[CH1D3,CH0D4][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0634	-0.9457	0.0133	0.0000
[CH0D3](=[OH0D1])[OH0D2][CH0D3](=[OH0D1])	0.0000	0.0000	-1.5156	0.0000	0.0000
[CH1D3][OH1D1]	-0.1589	0.0423	-0.7467	0.0000	0.0000
[CH0D4][OH1D1]	-0.3729	-0.0564	-1.1097	0.0000	0.0000
[CH3D1][CH0D3](=[OH0D1])[CH2D2,CH1D3,CH0D4][OH1D1]	0.0000	0.0000	-0.8724	0.0000	0.0000
[NH0D1]#[CH0D2][CH1D3,CH0D4][OH1D1]	0.0000	0.0000	0.9249	0.0000	0.0000
[OH1D1][CH0D4,CH1D3,CH2D2][CH0D3](=[OH0D1])[OH0D2]	-0.2763	0.0000	0.0000	0.0000	0.0000
[CH0D4,CH1D3,CH2D2]([OH1D1])[CH0D4,CH1D3,CH2D2]([OH1D1])	0.3877	0.0000	1.2843	-0.0107	0.0000
[CH0D4,CH1D3,CH2D2]([OH1D1])[CH0D4,CH1D3,CH2D2]([NH0D3,NH1D2,NH2D1])	0.0000	0.0000	0.4670	0.0085	0.0000
[CH0D4,CH1D3,CH2D2]([NH2D1])[CH0D4,CH1D3,CH2D2]([NH2D1])	0.0000	0.0000	1.0329	0.0000	0.0000
[CH1D3,CH2D2]([NH1D2])[CH1D3,CH2D2]([NH2D1])	0.0000	-0.2012	0.1200	0.0000	-32.5114
[CH2D2,CH1D3,CH0D4]([NH2D1,NH1D2,NH0D3])[CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0528	23.6566	0.0000	0.0000
[CH0D3](=[OH0D1])([OH1D1])[CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	-1.5111	0.0000	3.5751	0.0000	0.0000
[CH0D3](=[OH0D1])([OH1D1])[CH1D3,CH2D2][CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	3.0777	0.0000	0.0000
[CH1D3,CH2D2]([NH2D1])[CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.2478	0.0000	0.0000	0.0000
[CH3D1][OH0D2][CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	1.9103	0.0000	0.0000
[SH1D1][CH1D3][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.5676	0.0000	0.0000	0.0000
[CH1D3,CH2D2]([SH1D1])[CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	1.9593	0.0000	0.0000
[CH1D3,CH2D2]([CH0D2]#[NH0D1])[CH1D3,CH2D2]([CH0D2]#[NH0D1])	0.8288	0.0000	2.6936	0.0000	0.0000
[CH1D3,CH2D2]([OH1D1])[CH1D3,CH2D2]([CH0D2]#[NH0D1])	0.0000	0.0000	1.0406	0.0000	0.0000
[CH1D3,CH2D2]([SH1D1])[CH1D3,CH2D2]([SH1D1])	0.0000	0.0000	1.8153	0.0000	0.0000
[NH0D1]#[CH0D2][CH1D3,CH2D2][CH0D3](=[OH0D1])[OH0D2]	0.0000	0.0000	0.6872	0.0000	0.0000
[CH1D3,CH2D2]([CH0D3]=[OH0D1])[CH0D3](=[OH0D1])[OH0D2]	-0.0665	0.0000	-0.4451	0.0000	0.0000
[CH0D4,CH1D3,CH2D2,CH3D1][OH0D2][CH0D3,CH1D2]=[CH0D3,CH1D2,CH2D1]	0.0000	0.0000	0.0681	0.0000	0.0000
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][FH0D1]	0.0000	-0.1734	-0.5103	0.0000	0.0000
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][BrH0D1]	-0.1370	-0.0857	-0.2488	0.0000	61.3340
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][ClH0D1]	-0.0565	-0.0982	-0.0815	0.0000	0.0000
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][CH0D2]#[NH0D1]	-0.1677	0.0000	-0.8673	0.0000	0.0000
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][CH0D3](=[OH0D1])[OH0D2][CH0D4,CH1D3,CH2D3,CH3D1]	0.2660	0.1260	0.2924	0.0000	0.0000
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][CH1D2](=[OH0D1])	0.0000	0.0000	0.0000	0.0000	0.0000

Table A.2 continued from previous page

Groups	T_b (K)	T_m (K)	T_c (K)	P_c (bar)	V_c cc/mol
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.1217	0.4662	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][FH0D1,ClH0D1,BrH0D1,IH0D1]	0.0000	0.0536	0.2955	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][NH0D3,NH1D2,NH2D1]	0.0000	0.0000	0.1645	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][OH0D2]	0.0000	0.0000	0.0666	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][OH1D1]	0.0000	0.0314	-0.0042	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][CH0D2]#[NH0D1]	0.0000	0.0000	0.4617	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][SH1D1]	0.0000	0.0000	0.8773	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	-0.5072	0.0000	0.0000	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][CH0D3]=[OH0D1]	0.1765	0.0000	0.0000	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][OH0D2][CH1D2](=[OH0D1])	0.0000	0.0000	0.5775	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][OH0D2][CH0D3](=[OH0D1])	0.0000	0.0000	0.5963	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][CH0D3](=[OH0D1])[OH0D2]	0.0000	0.0000	0.0000	0.0000	0.0000
[cH0x2D3][CH1D3]([CH3D1])([CH3D1])	0.0000	0.0000	0.1554	0.0000	0.0000
[cH0x2D3][CH0D4]([CH3D1])([CH3D1])[CH3D1]	-0.2861	0.0000	0.6944	0.0000	0.0000
[cH0x2D3][CH0D4]([FH0D1])([FH0D1])[FH0D1]	-0.2823	0.1242	0.0000	0.0000	0.0000
[CH0x2D3,CH1x2D2,CH2x2D1]=[CH0x2D3][CH0D3](=[OH0D1])	0.0000	0.0559	0.0000	0.0000	0.0000
[CH0x2D3,CH1x2D2,CH2x2D1]=[CH0x2D3][CH3D1]	0.0000	0.1075	0.2727	0.0000	0.0000
[CH0x2D3,CH1x2D2,CH2x2D1]=[CH0x2D3][CH2D2]	0.2510	-0.0327	1.4219	0.0000	51.8285
[CH0x2D3,CH1x2D2,CH2x2D1]=[CH0x2D3][ClH0D1]	0.0000	0.1118	0.0000	0.0000	0.0000
[CH1x2D3][CH3D1]	-0.2543	-0.0096	0.2864	0.0000	0.0000
[CH1x2D3][CH2D2]	0.0465	-0.0088	0.5813	0.0023	0.0000
[CH1x2D3][CH1D3]	0.0000	0.0000	1.1389	0.0000	0.0000
[CH1x2D3][CH0D4]	0.0000	0.0495	0.0000	0.0000	0.0000
[CH1x2D3][CH1D2]=[CH1D2,CH2D1]	0.0000	0.0000	0.6410	0.0000	0.0000
[CH1x2D3][CH0D3]=[CH1D2,CH2D1]	0.0000	0.0000	0.0000	0.0000	0.0000
[CH1x2D3][ClH0D1]	0.0000	0.1217	0.0000	0.0000	0.0000
[CH1x2D3][OH1D1]	-0.1663	0.0000	1.2833	0.0000	0.0000
[CH1x2D3][NH2D1]	-0.4026	0.0000	-0.2439	0.0000	0.0000
[CH1x2D3][NH1D2][CH0D4,CH1D3,CH2D2,CH1D3]	0.0000	0.1513	0.0000	0.0000	0.0000
[CH1x2D3][NH0D3][CH0D4,CH1D3,CH2D2,CH1D3]	0.0000	0.0000	0.0000	0.0000	0.0000
[CH1x2D3][SH1D1]	0.0000	0.0000	0.9219	0.0000	0.0000
[CH1x2D3][CH0D2]#[NH0D1]	0.5770	0.0000	0.0000	0.0000	0.0000
[CH1x2D3][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	0.0000	0.0000	0.0000
[CH1x2D3][CH0D3](=[OH0D1])	0.0000	0.0000	0.0000	0.0000	0.0000
[CH1x2D3][OH0D2]	-0.0752	0.0000	-0.2239	0.0292	0.0000
[CH1x2D3][OH0D2][CH0D3](=[OH0D1])	0.0000	0.0237	1.4211	-0.0274	0.0000
[CH0x2D4][CH3D1]	0.0000	0.0000	-0.5940	0.0101	-16.3626
[CH0x2D4][CH2D2]	0.0560	-0.0407	-0.5812	0.0000	-28.7420
[CH0x2D4][OH1D1]	0.0000	0.0000	-0.7723	0.0000	-36.9198
[NH0x2D3][CH3D1]	0.0000	0.1071	0.4628	0.0000	0.0000
[NH0x2D3][CH2D2]	0.0000	0.0000	-0.3257	0.0000	0.0000
c1(A)c(A)cccc1	-0.1357	-0.0230	-1.5654	-0.0023	-7.7812
c1(A)cc(A)ccc1	0.0409	-0.0020	-1.0978	0.0000	3.3629
c1(A)ccc(A)cc1	0.0706	0.0300	-0.9591	0.0000	0.0000
c1(A)c(A)c(A)ccc1	0.1209	0.0000	1.8460	0.0000	0.0000

Table A.2 continued from previous page

Groups	T_b (K)	T_m (K)	T_c (K)	P_c (bar)	V_c cc/mol
c1(A)c(A)cc(A)cc1	0.0000	-0.0332	1.5505	0.0000	0.0000
c1(A)cc(A)cc(A)c1	-0.1541	0.0000	0.9679	0.0000	0.0000
c1(A)c(A)c(A)c(A)cc1	0.0000	-0.0004	-0.8420	0.0000	29.5423
c1(A)c(A)c(A)cc(A)c1	0.0000	0.0000	-0.9143	0.0000	0.0000
c1(A)c(A)cc(A)c(A)c1	0.0000	0.1265	-1.4690	0.0000	0.0000
n1c(A)cccc1	0.0000	0.0000	-1.2999	0.0000	0.0000
n1cc(A)ccc1	0.0000	-0.0396	-0.6080	0.0000	0.0000
n1ccc(A)cc1	0.0000	0.0361	-0.5734	0.0000	0.0000
n1c(A)cc(A)cc1	0.0000	0.0000	0.0000	0.0000	0.0000
n1c(A)ccc(A)c1	0.0000	0.0000	0.0000	0.0000	0.0000
n1c(A)cccc1(A)	0.0000	0.0000	-0.0169	0.0000	-75.3257
n1cc(A)c(A)cc1	0.0000	0.0000	0.0000	0.0000	0.0000
c1(A)c(A)c(A)c(A)c(A)c1	0.0000	0.0000	-0.6776	0.0000	0.0000
[cH0x2D3][SH0x0D4](=[OH0x0D1])(=[OH0x0D1])[NH2D1]	0.0000	-0.0940	0.0000	0.0000	0.0000
[OH1D1][CH0D3](=[OH0D1])[R0;D2][R0;D2][R0;D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	2.5806	0.0000	0.0000
[NH2D1][R0;D2][R0;D2][R0;D2][OH1D1]	0.0000	0.0000	0.6585	0.0000	0.0000
[OH1D1][R0;D2][R0;D2][R0;D2][OH1D1]	0.0000	0.0000	0.0000	0.0000	0.0000
[CH0D2](#[NH0D1])[R0;D2][R0;D2][R0;D2][CH0D2]#[NH0D1]	0.6892	0.0000	2.9863	0.0000	0.0000
[cH0x3D3][CH0x2D3,CH1x2D2]=[CH0x2D3,CH1x2D2]	0.0000	0.0000	-0.4013	0.0139	14.9223
[CH1x3D3]	0.0000	0.0000	0.3988	0.0000	0.0000
[CH0x3D4,CH0x4D4]	0.0000	0.0000	0.0000	0.0000	0.0000
c12ccccc1cccc2	0.2436	0.0000	-0.3165	0.0066	17.0875
c12ccccc1c(A)ccc2	0.0000	0.0000	-0.8537	0.0030	24.1623
c12ccccc1cc(A)cc2	0.0000	0.0000	-0.9097	-0.0060	0.0000
c1cc2ccc3cccc4ccc(c1)c2c34	0.0000	0.0000	1.6062	0.0046	0.0000
c1ccc2ncccc2c1	0.0000	0.0000	0.0261	0.0000	0.0000
c1ccc2cnccc2c1	0.0000	0.0000	0.6580	0.0000	0.0000

Table A.3: Higher-order group contribution coefficients for enthalpy of vaporization (H_{vap}),surface tension (σ), and dynamic viscosity (μ)

Groups	H_{vap} (kJ/mol)	σ (N/m)	μ (mPa s)
[CH1D3]([CH3D1])([CH3D1])	0.0000	0.0000	-0.1713
[CH0D4]([CH3D1])([CH3D1])([CH3D1])	0.0000	-1.7519	1.1490
[CH1D3]([CH3D1])[CH1D3]([CH3D1])	0.0000	0.0000	0.1534
[CH1D3]([CH3D1])[CH0D4]([CH3D1])([CH3D1])	0.0000	0.0000	0.0000
[CH0D4]([CH3D1])([CH3D1])[CH0D4]([CH3D1])([CH3D1])	1.1134	0.0000	0.0000
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][CH0D3,CH1D2]=[CH0D3,CH1D2,CH2D1]	0.0000	0.0000	-3.3332
[CH3D1][CH0D3,CH1D2]=[CH0D3,CH1D2,CH2D1]	0.0000	2.0596	-1.8568
[CH2D2][CH0D3,CH1D2]=[CH0D3,CH1D2,CH2D1]	0.0000	0.0000	-1.5292
[CH1D3,CH0D4][CH0D3,CH1D2]=[CH0D3,CH1D2,CH2D1]	0.0000	0.0000	-0.9851
[CH1D3,CH0D4][CH1D2]=[OH0D1]	0.0000	0.0000	0.4540
[CH3D1][CH0D3](=[OH0D1])[CH2D2]	0.0000	0.0000	-0.9664
[CH3D1][CH0D3](=[OH0D1])[CH1D3,CH0D4]	0.0000	2.7685	-0.3998
[CH1D3,CH0D4][CH0D3](=[OH0D1])[OH1D1]	0.0000	-1.9207	1.0712
[CH0D3](=[OH0D1])[OH0D2][CH0D3](=[OH0D1])	0.0000	0.0000	-0.3239
[CH1D3][OH1D1]	0.0000	-2.4881	-0.1853
[CH0D4][OH1D1]	0.0000	-5.0447	-1.2027
[CH3D1][CH0D3](=[OH0D1])[CH2D2,CH1D3,CH0D4][OH1D1]	0.0000	0.0000	0.0000
[NH0D1]#[CH0D2][CH1D3,CH0D4][OH1D1]	0.0000	0.0000	0.0000
[OH1D1][CH0D4,CH1D3,CH2D2][CH0D3](=[OH0D1])[OH0D2]	0.0000	0.0000	-0.6981
[CH0D4,CH1D3,CH2D2]([OH1D1])[CH0D4,CH1D3,CH2D2]([OH1D1])	0.0000	7.2661	-1.3162
[CH0D4,CH1D3,CH2D2]([OH1D1])[CH0D4,CH1D3,CH2D2]([NH0D3,NH1D2,NH2D1])	0.0000	0.0000	-0.2830
[CH0D4,CH1D3,CH2D2]([NH2D1])[CH0D4,CH1D3,CH2D2]([NH2D1])	0.0000	19.4354	-0.6231
[CH1D3,CH2D2]([NH1D2])[CH1D3,CH2D2]([NH2D1])	0.0000	0.0000	0.1859
[CH2D2,CH1D3,CH0D4]([NH2D1,NH1D2,NH0D3])[CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	0.0000
[CH0D3](=[OH0D1])([OH1D1])[CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	0.0000
[CH0D3](=[OH0D1])([OH1D1])[CH1D3,CH2D2][CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	0.0000
[CH1D3,CH2D2]([NH2D1])[CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	0.0000
[CH3D1][OH0D2][CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	0.0000
[SH1D1][CH1D3][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	0.0000
[CH1D3,CH2D2]([SH1D1])[CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	0.0000
[CH1D3,CH2D2]([CH0D2]#[NH0D1])[CH1D3,CH2D2]([CH0D2]#[NH0D1])	0.0000	0.0000	0.0000
[CH1D3,CH2D2]([OH1D1])[CH1D3,CH2D2]([CH0D2]#[NH0D1])	0.0000	0.0000	0.0000
[CH1D3,CH2D2]([SH1D1])[CH1D3,CH2D2]([SH1D1])	0.0000	0.0000	0.0000
[NH0D1]#[CH0D2][CH1D3,CH2D2][CH0D3](=[OH0D1])[OH0D2]	0.0000	-1.5926	0.0000
[CH1D3,CH2D2]([CH0D3]=[OH0D1])[CH0D3](=[OH0D1])[OH0D2]	0.0000	0.0000	-0.2432
[CH0D4,CH1D3,CH2D2,CH3D1][OH0D2][CH0D3,CH1D2]=[CH0D3,CH1D2,CH2D1]	0.0000	0.0000	-0.8034
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][FH0D1]	0.0000	0.0000	0.0000
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][BrH0D1]	0.0000	0.0000	0.0000
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][ClH0D1]	0.0000	0.0000	-0.7528
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][CH0D2]#[NH0D1]	0.0000	0.0000	-1.0321
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][CH0D3](=[OH0D1])[OH0D2][CH0D4,CH1D3,CH2D3,CH3D1]	0.0000	3.1669	-0.8823
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][CH1D2](=[OH0D1])	0.0000	0.0000	-0.7092
[CH0D3,CH1D2,CH2D1]=[CH0D3,CH1D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	-0.3517

Table A.3 continued from previous page

Groups	H_{vap} (kJ/mol)	σ (N/m)	μ (mPa s)
[cH0x2D3][CH1D3,CH2D2][FH0D1,ClH0D1,BrH0D1,IH0D1]	0.0000	0.0000	0.0386
[cH0x2D3][CH1D3,CH2D2][NH0D3,NH1D2,NH2D1]	0.0000	0.0000	0.3123
[cH0x2D3][CH1D3,CH2D2][OH0D2]	0.0000	0.0000	0.5741
[cH0x2D3][CH1D3,CH2D2][OH1D1]	0.0000	0.0000	0.1491
[cH0x2D3][CH1D3,CH2D2][CH0D2]#[NH0D1]	0.0000	0.0000	0.4116
[cH0x2D3][CH1D3,CH2D2][SH1D1]	0.0000	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][CH0D3]=[OH0D1]	0.0000	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][OH0D2][CH1D2](=[OH0D1])	0.0000	0.0000	0.0000
[cH0x2D3][CH1D3,CH2D2][OH0D2][CH0D3](=[OH0D1])	0.0000	0.0000	0.0735
[cH0x2D3][CH1D3,CH2D2][CH0D3](=[OH0D1])[OH0D2]	0.0000	-1.9111	0.0000
[cH0x2D3][CH1D3]([CH3D1])([CH3D1])	0.0000	0.0000	-1.2098
[cH0x2D3][CH0D4]([CH3D1])([CH3D1])[CH3D1]	0.0000	0.0000	0.0000
[cH0x2D3][CH0D4]([FH0D1])([FH0D1])[FH0D1]	0.0000	0.0000	-0.9025
[CH0x2D3,CH1x2D2,CH2x2D1]=[CH0x2D3][CH0D3](=[OH0D1])	0.0000	0.0000	0.0000
[CH0x2D3,CH1x2D2,CH2x2D1]=[CH0x2D3][CH3D1]	0.0000	0.0000	2.0830
[CH0x2D3,CH1x2D2,CH2x2D1]=[CH0x2D3][CH2D2]	0.0000	0.0000	-2.4023
[CH0x2D3,CH1x2D2,CH2x2D1]=[CH0x2D3][ClH0D1]	-6.8676	0.0000	0.0000
[CH1x2D3][CH3D1]	-1.3451	-2.5700	0.6570
[CH1x2D3][CH2D2]	0.0000	-0.3227	1.0140
[CH1x2D3][CH1D3]	0.0000	0.0000	1.9696
[CH1x2D3][CH0D4]	0.0000	-0.0956	0.0000
[CH1x2D3][CH1D2]=[CH1D2,CH2D1]	0.0000	0.0000	-0.0572
[CH1x2D3][CH0D3]=[CH1D2,CH2D1]	0.0000	0.0000	4.5859
[CH1x2D3][ClH0D1]	0.0000	0.0000	0.0000
[CH1x2D3][OH1D1]	0.0000	0.0000	0.0000
[CH1x2D3][NH2D1]	0.0000	0.0000	0.0000
[CH1x2D3][NH1D2][CH0D4,CH1D3,CH2D2,CH1D3]	0.0000	0.0000	2.8657
[CH1x2D3][NH0D3][CH0D4,CH1D3,CH2D2,CH1D3]	0.0000	0.0000	2.9762
[CH1x2D3][SH1D1]	0.0000	0.0000	0.0000
[CH1x2D3][CH0D2]#[NH0D1]	0.0000	0.0000	0.0000
[CH1x2D3][CH0D3](=[OH0D1])[OH1D1]	0.0000	4.2964	0.0000
[CH1x2D3][CH0D3](=[OH0D1])	0.0000	0.1115	0.0000
[CH1x2D3][OH0D2]	0.0000	-1.6513	1.2453
[CH1x2D3][OH0D2][CH0D3](=[OH0D1])	3.2105	0.0000	-0.9318
[CH0x2D4][CH3D1]	0.0000	-0.3422	0.2214
[CH0x2D4][CH2D2]	0.0000	-0.2736	0.0260
[CH0x2D4][OH1D1]	0.0000	0.0000	0.0000
[NH0x2D3][CH3D1]	0.0000	0.0000	1.1991
[NH0x2D3][CH2D2]	0.0000	0.0000	1.6961
c1(A)c(A)cccc1	-1.7629	0.0000	0.6324
c1(A)cc(A)ccc1	0.0000	0.0000	1.1548
c1(A)ccc(A)cc1	0.0000	0.0000	0.9462
c1(A)c(A)c(A)ccc1	1.1586	0.0000	-2.7555
c1(A)c(A)cc(A)cc1	0.0000	0.0000	-0.2551

Table A.3 continued from previous page

Groups	H_{vap} (kJ/mol)	σ (N/m)	μ (mPa s)
c1(A)cc(A)cc(A)c1	0.0000	0.0000	-2.1758
c1(A)c(A)c(A)c(A)cc1	0.0000	0.0000	2.6003
c1(A)c(A)c(A)cc(A)c1	0.0000	0.0000	-0.1700
c1(A)c(A)cc(A)c(A)c1	0.0000	0.0000	-0.0850
n1c(A)cccc1	-2.5737	0.0000	0.1680
n1cc(A)ccc1	0.0000	0.0000	0.2390
n1ccc(A)cc1	0.0000	0.0000	0.2698
n1c(A)cc(A)cc1	0.0000	0.0000	-0.0274
n1c(A)ccc(A)c1	-10.9799	0.0000	0.0000
n1c(A)cccc1(A)	0.0000	0.0000	0.0862
n1cc(A)c(A)cc1	0.0000	7.1924	0.0000
c1(A)c(A)c(A)c(A)c(A)c1	0.0000	0.0000	-0.1699
[cH0x2D3][SH0x0D4](=[OH0x0D1])(=[OH0x0D1])[NH2D1]	0.0000	0.0000	0.0000
[OH1D1][CH0D3](=[OH0D1])[R0;D2][R0;D2][R0;D2][CH0D3](=[OH0D1])[OH1D1]	0.0000	0.0000	0.0000
[NH2D1][R0;D2][R0;D2][R0;D2][OH1D1]	0.0000	11.3703	0.0000
[OH1D1][R0;D2][R0;D2][R0;D2][OH1D1]	0.0000	8.4836	0.0000
[CH0D2](#[NH0D1])[R0;D2][R0;D2][R0;D2][CH0D2]#[NH0D1]	0.0000	10.0574	0.0000
[cH0x3D3][CH0x2D3,CH1x2D2]=[CH0x2D3,CH1x2D2]	0.0000	0.0000	0.0000
[CH1x3D3]	0.0000	0.0000	0.1915
[CH0x3D4,CH0x4D4]	0.0000	0.0000	-0.3811
c12ccccc1cccc2	2.3790	0.0000	0.5562
c12ccccc1c(A)ccc2	0.0000	0.0000	0.5559
c12ccccc1cc(A)cc2	0.0000	0.0000	0.0000
c1cc2ccc3cccc4ccc(c1)c2c34	0.0000	0.0000	0.0000
c1ccc2ncccc2c1	0.0000	0.0000	0.4874
c1ccc2cnccc2c1	0.0000	0.0000	0.0000