Data-Driven Computer-Aided Molecular, Material, and Process Design for Efficient Separation Systems

Dissertation
zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

von M.Sc. Zihao Wang

geb. am 18.07.1995 in Hengyang, Hunan, China

genehmigt durch die Fakultät für Verfahrens- und Systemtechnik der Otto-von-Guericke-Universität Magdeburg

Gutachter:
Prof. Dr.-Ing. Kai Sundmacher
Prof. Dr.-Ing. Teng Zhou
Prof. Dr. Gürkan Sin

Promotionskolloquium am 28.03.2025

Abstract

Chemical separation systems are essential across various industries, but they are often highly energy intensive. Reducing the energy demand for separation is an important step to lower production costs, minimize environmental impacts, and promote the sustainable development of separation technologies. The energy efficiency of separation systems depends not only on the operating conditions but also on the selection of separating agents used to facilitate the separation, such as solvents or adsorbents. Computer-aided molecular and process design (CAMPD) can be used to identify the optimal separating agents and operating conditions. However, this design method is usually challenging because it often uses nonlinear mathematical models to describe the complex overall system. These models must then be integrated into an optimization problem, which requires advanced numerical methods to maximize the overall performance of the cross-scale system.

In this dissertation, several data-driven approaches are proposed to accelerate computer-aided molecular, material, and process design. These approaches cover various applications for efficient separation systems, including optimal molecular design, large-scale material screening, process optimization, and integrated molecular/material and process design.

To accelerate the identification of optimal separating agents, data-driven models are developed to predict separation performance based on the properties or structures of the separating agents. These models are further used to identify the optimal separating agents through surrogate-based optimization or large-scale screening to maximize separation performance in specific applications. For molecular discovery, molecular property targeting and molecular mapping methods are introduced and demonstrated to be effective for optimal solvent design. For materials discovery, two types of machine learning models are developed: an end-to-end model for accurate predictions and an interpretable model that provides insights into the predictions. Both models are very efficient and suitable for large-scale screening of metal-organic frameworks (MOFs) targeting energy-efficient gas separation.

Furthermore, a data-driven CAMPD approach is proposed to integrate the identification of optimal molecules and materials into process optimization. This method combines data-driven process models, optimization algorithms, and molecular property targeting for the simultaneous design of optimal molecules/materials and process parameters, improving the

overall process performance. Taking one step further, Bayesian optimization is integrated into the data-driven CAMPD approach to reduce the high data demand typically required for accurate surrogate modeling. The resulting BayesCAMPD approach offers a data-efficient and closed-loop solution to CAMPD tasks by iteratively performing data-driven modeling, surrogate-based optimization, and solution validation.

The effectiveness of the data-driven approaches proposed in this dissertation is demonstrated using two different separation processes, extractive distillation and pressure swing adsorption. These approaches are practical and computationally efficient in advancing the development of efficient separation systems with broad applications in chemical engineering.

Zusammenfassung

Chemische Trennprozesse sind in verschiedenen Industrien wichtig, aber oft sehr energieintensiv. Den Energieverbrauch für Stofftrennungen zu reduzieren, ist ein wesentlicher Schritt zur Senkung der Produktionskosten und zur Minimierung der Umweltauswirkungen. Daher muss eine nachhaltige Entwicklung von energiesparenden Trenntechnologien gefördert werden. Die Energieeffizienz von Trennsystemen hängt nicht nur von den Betriebsbedingungen ab, sondern auch von der Auswahl der Hilfsstoffe, die zur Erleichterung der Trennung eingesetzt werden, z. B. Lösungsmitteln und Adsorbenzien. Mit Hilfe des computergestützten Molekül- und Prozessdesigns (CAMPD) können die optimalen Hilfsstoffe und Betriebsbedingungen ermitteln werden. Diese Design-Methode ist jedoch oft schwierig, da hierbei häufig nichtlineare mathematische Modelle zur Beschreibung des komplexen Gesamtsystem verwendet werden und diese Modelle dann auch noch in ein Optimierungsproblem integriert werden müssen. Dieses Problem kann nur mit fortgeschrittenen numerischen Methoden gelöst werden, um die Gesamtleistung des Systems (Molekül/Material/Prozess) skalenübergreifend zu maximieren.

In der vorliegenden Dissertation werden verschiedene datengetriebene Ansätze vorgeschlagen, um das computergestützte Molekül-, Material- und Prozessdesign zu beschleunigen. Diese Ansätze decken verschiedene Anwendungen von effiziente Trennsysteme ab, darunter optimales Moleküldesign, großangelegtes Material-Screening, Prozessoptimierung und integriertes Molekül-/Material- und Prozessdesign.

Um das Auffinden von optimal geeigneten Molekülen und Materialien zu beschleunigen, werden datengetriebene Modelle entwickelt, mit denen man die Trennleistung auf der Grundlage der Eigenschaften oder Strukturen der Stofftrenn-Hilfsmittel abschätzen kann. Diese Modelle werden weiterhin verwendet, um die optimalen Hilfsstoffe durch Surrogat-basierte Optimierung oder großangelegtes Screening zu identifizieren, um die Trennleistung in spezifischen Anwendungen zu maximieren. Im Bereich der Moleküloptimierung werden die Methodiken „Molekül-Eigenschafts-Targeting“ und „Molekül-Mapping“ für ein effizientes optimales Lösungsmitteldesign eingeführt und deren Wirksamkeit anhand von Beispielen demonstriert. Im Bereich der Materialoptimierung werden zwei Arten von maschinellen Lernmodellen entwickelt: ein End-to-End-Modell für akkurate Vorhersagen und ein

interpretierbares Modell, welches Einblicke in die Vorhersagen bietet. Beide Modelle sind sehr gutgeeignet für das großangelegte Screening von metallorganischen Gerüstmaterialien (MOFs) zur energieeffizienten Gastrennung.

Weiterhin wird ein datengetriebener CAMPD-Ansatz vorgeschlagen, um die Ermittlung optimaler Moleküle und Materialien in die Prozessoptimierung zu integrieren. Diese Methode kombiniert datengestützte Prozessmodelle, Optimierungsalgorithmen und „Molekül-Eigenschafts-Targeting“, um gleichzeitig optimale Moleküle/Materialien und Prozessparameter zu entwerfen, wodurch die Gesamtleistung des Prozesses verbessert wird. In einem weiteren Schritt wird die Bayes’sche Optimierung in den datengestützten CAMPD-Ansatz integriert, um den oft hohen Datenbedarf für eine genaue Surrogat-Modellierung zu reduzieren. Der daraus resultierende BayesCAMPD-Ansatz bietet eine dateneffiziente und geschlossene Lösung von CAMPD-Aufgaben, indem er datengestützte Modellierung, Surrogat-basierte Optimierung und Lösungsvalidierung iterativ durchführt.

Die Wirksamkeit der in der Dissertation vorgeschlagenen datengestützten Ansätze werden anhand von zwei verschiedenen Trennprozessen, der extraktiven Destillation und der Druckwechseladsorption, demonstriert. Sie sind praktisch und rechnerisch effizient bei der Entwicklung effizienter Trennsysteme mit breiten Anwendungsmöglichkeiten in der chemischen Technik.

Table of Contents

1 Introduction

1.1 Motivation and objectives

Chemical processes are essential across numerous industries, including petrochemicals, pharmaceuticals, manufacturing, agriculture, and many others. They enable the transformation of energy and raw materials into products important for other industrial sectors and final consumers.[1] The chemical industry is the largest industrial energy consumer and accounts for around one-quarter of total industrial energy consumption. In 2020, it consumed 383 billion kilowatt hours, representing more than half of the electricity and heat used by all private households in Germany.[2] In the chemical industry, the components of large quantities of chemical mixtures are separated into purer forms. Such chemical separation processes are fundamental but energy intensive, contributing to approximately 10–15% of global energy consumption.[3] Therefore, reducing energy demand is crucial to lower production costs and minimize environmental impacts, promoting the sustainable development of chemical industries. This aligns with “Design for Separation” and “Maximize Efficiency” of the 12 Principles of Green Engineering.

Over the years, significant efforts have been dedicated to the chemical sector to improve energy efficiency, focusing on reducing fuel and power energy consumption. As a matter of fact, a 48% reduction in energy consumption has been achieved since 1990.[1] Optimizing chemical separation processes is a direct way to reduce energy demands and production costs while enhancing product quality and system efficiency, resulting in efficient and sustainable production. By performing process optimization, the optimal operating conditions such as pressure and temperature can be determined. Beyond process optimization, the selection of appropriate chemicals is essential for efficient chemical separation processes, as they directly affect energy consumption, economic feasibility, and environmental impact. For instance, selecting a suitable solvent can significantly reduce separation difficulty, energy demand, and solvent usage. This can be systematically achieved through computer-aided molecular design (CAMD) to identify the optimal chemicals for specific separation processes. Furthermore, integrating process optimization with the selection of chemicals usually leads to enhanced energy efficiency, which is achieved through computer-aided molecular and process design (CAMPD). CAMPD is an important method in the research and development stage across various engineering applications, focusing on the determination of suitable chemicals (such as

solvents and adsorbents) and optimal process operating conditions to achieve specific objectives such as minimized energy demands or production costs.

Common CAMPD methods often combine mathematical models and optimization algorithms. Mathematical models are used to describe all phenomena relevant to the system under investigation, such as thermodynamics, molecular properties, phase equilibrium, conservation laws, reaction kinetics, and unit operations, among others. Once the optimization problem is formulated with these models, optimization algorithms are used to solve it and identify the best solution for specified criteria. However, the complexity of these mathematical models can lead to difficulties in solving the optimization problem, requiring sophisticated numerical methods and solvers. This is particularly noticeable for large and complex systems, where many mathematical models with different levels of nonlinearity and nonconvexity are integrated, posing significant computational challenges for optimization.

Recent advances in data-driven approaches, particularly machine learning (ML), have offered promising solutions to various engineering tasks. Leveraging available data and advanced ML techniques, data-driven models can be developed to capture the behaviors of complex systems and guide their optimization. This has been proven practical in different science and engineering applications, such as materials discovery and designs[4,5]. In addition, these strategies are gaining attention in surrogate modeling for complex systems. Surrogate models with high computational efficiency are developed to replace complex mathematical models. Consequently, the optimization difficulty can be significantly reduced by using these efficient surrogate models.

In the early stage of process development, the focus is primarily on process-level performance indicators such as product quality, energy demand, and economic benefit. Equipment- and phase-level performance, such as temperature distribution and vapor-liquid equilibrium, can be temporarily ignored. Given these, it is often sufficient and more efficient to develop surrogate models for the entire system to estimate important process-level performance, rather than developing surrogate models for every mathematical model involved. Therefore, data-driven approaches can be considered effective and efficient for the CAMPD to discover better molecules and materials for efficient separation systems and identify optimal operating conditions that maximize overall system performance.

This dissertation explores the potential of advanced data-driven approaches for Process Systems Engineering (PSE) applications, with a particular focus on computer-aided molecular, material, and process design, to accelerate the optimal design of efficient separation systems. It aims to provide practical data-driven solutions for a broad range of applications in chemical and materials engineering, while also offering valuable insights for industrial practices and future research in the field. Accordingly, the research strategy will focus on data-driven solutions to achieve the following objectives.

- Optimal molecular design of solvents for extractive distillation.
- Identification of optimal metal-organic frameworks (MOFs) as adsorbents for gas separation.
- Integrated design of solvents and extractive distillation processes.
- Integrated design of MOFs and pressure swing adsorption processes.

1.2 Outline

Chapter 2 provides an overview of the research fundamentals related to molecular design, materials discovery, and process optimization in separation systems. It also introduces computational techniques such as data-driven modeling, mathematical optimization, and Bayesian optimization.

Part I (**Chapters 3** and **4**) focuses on molecular discovery. **Chapter 3** introduces molecular property targeting and molecular mapping techniques for optimal molecular design and demonstrates their effectiveness through the optimal design of solvents for extractive distillation. **Chapter 4** introduces a data-driven CAMPD approach that incorporates the molecular property targeting technique for efficient integrated molecular and process design. Additionally, Bayesian optimization is integrated to improve the CAMPD approach by reducing data demand for accurate surrogate modeling. Both data-driven CAMPD approaches are demonstrated by the integrated design of solvents and extractive distillation processes.

Part II (**Chapters 5** and **6**) focuses on materials discovery. **Chapter 5** develops two types of machine learning models for the identification of optimal MOFs. Both approaches are demonstrated efficient for large-scale adsorbent screening targeting energy-efficient gas separation. **Chapter 6** integrates MOF selection with process optimization. Both simulation-

based and data-driven optimization approaches are conducted for selecting suitable adsorbents and designing efficient pressure swing adsorption systems.

Chapter 7 concludes the dissertation by summarizing its contributions and suggesting directions for future research.

Figure 1-1 illustrates the sequence and interrelations among the four chapters that constitute the original contributions of this dissertation (**Chapters 3–6**).

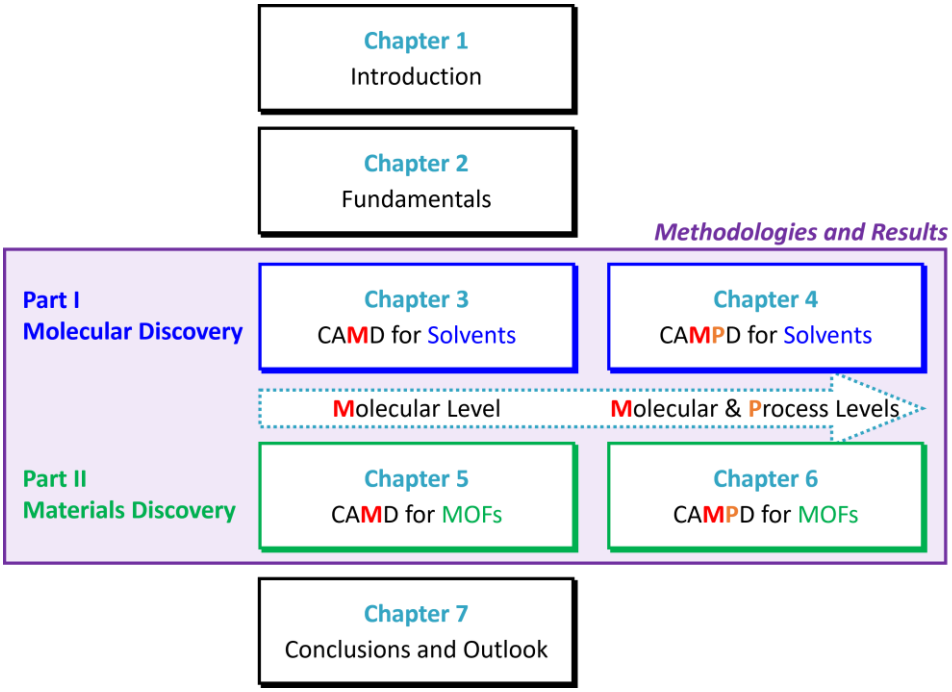


Figure 1-1. Schematic outline of the dissertation.

Some results and parts of this dissertation have been published[6-9] or are intended for future publication and, therefore, will not be explicitly cited within this dissertation. The main content of **Chapters 3–6** is primarily based on the following works.

- **Wang Z**, Zhou T, Sundmacher K. Data-driven integrated design of solvents and extractive distillation processes. *AIChE Journal*. 2023; 69(12): e18236
- **Wang Z**, Zhou Y, Zhou T, Sundmacher K. Identification of optimal metal–organic frameworks by machine learning: Structure decomposition, feature integration, and predictive modeling. *Computers & Chemical Engineering*. 2022; 160: 107739.

- **Wang Z**, Zhou T, Sundmacher K. Interpretable machine learning for accelerating the discovery of metal–organic frameworks for ethane/ethylene separation. *Chemical Engineering Journal*. 2022; 444: 136651.
- **Wang Z**, Zhou T, Sundmacher K. Molecular property targeting for optimal solvent design in extractive distillation processes. In: Kokossis AC, Georgiadis MC, Pistikopoulos E, eds. *Computer Aided Chemical Engineering*. Elsevier; 2023: 1247–1252.
- **Wang Z**, Zhou T, Sundmacher K. BayesCAMPD: Data-efficient and closed-loop integrated molecular and process design using Bayesian optimization. *To be submitted*.
- **Wang Z**, Zhou T, Sundmacher K. Integrated adsorbent selection and process design: Simulation-based and data-driven optimization approaches. *To be submitted*.

2 Fundamentals

2.1 Separation processes

In chemical production systems, one of the major tasks is separating large quantities of chemical mixtures into pure components. Chemical separation processes such as distillation, drying, and evaporation, are typically the most energy-requiring operations in chemical and petroleum refining industries, accounting for about half of US industrial energy use and 10–15% of the nation's total energy consumption.[3] Reducing the energy demand of separation processes is a pivotal step to lower production costs and minimize environmental impacts, promoting the sustainable development of chemical production systems.

2.1.1 Extractive distillation

Distillation is the process of separating components of a liquid mixture by successive evaporation and condensation according to their different boiling points. It is one of the most widely used separation techniques in various industrial applications, such as oil refineries, petrochemical plants, and natural gas processing facilities.[10] However, it is the major energy-intensive separation process as 49% of the energy consumed in separation processes is used for distillation.[3]

For components with high relative volatility, distillation is the preferred separation process due to the ease of achieving high purity. However, for mixtures with close boiling points (such as C4, C5, and C6 hydrocarbons) or mixtures that form azeotropes (such as ethanol/water and acetone/methanol), the separation by conventional distillation processes becomes challenging and energy intensive. In such cases, alternative techniques such as extractive distillation are considered effective in facilitating separation and reducing energy consumption.

In extractive distillation, a suitable solvent that interacts with a preferred affinity for one of the components is introduced to alter the relative volatility of mixtures being separated, allowing for efficient separation through regular distillation.[11,12] Extractive distillation has been widely applied in the petrochemical and pharmaceutical industry for difficult-to-separate mixtures such as acetone/methanol[13-16] and ethanol/water[17-20]. In general, for a binary mixture, the ED process consists of two columns, i.e., an extractive distillation column (EDC) and a solvent recovery column (SRC). In the EDC, solvent is fed to the upper part. One of the components is purified and obtained in the distillate, and the other is withdrawn with the solvent from the

bottom. The bottom product is then taken to the SRC, in which the other component is obtained in the distillate, and the solvent is recovered and recycled from the bottom (**Figure 2-1**).

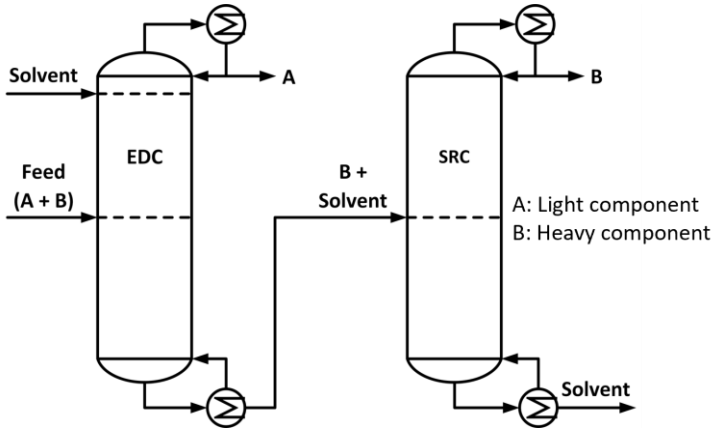


Figure 2-1. ED process for separating close-boiling or azeotropic mixtures.

2.1.2 Pressure swing adsorption

Separating and purifying gas mixtures are critical across various industries. Pressure swing adsorption (PSA) is a versatile technique that leverages the varying affinities of different gases for a specific adsorbent material.[21] By manipulating operating pressures, PSA efficiently separates the desired gas component through concessive adsorption and desorption.

PSA can efficiently produce high-purity gases and can be tailored for the separation of various gas mixtures depending on the chosen adsorbent. Selective adsorbent materials (e.g., activated carbon, zeolites, etc.) are used to preferentially adsorb the target or undesired gas species at high pressures. The PSA system then swings to a lower pressure to release the adsorbed gas. For example, PSA can be used to produce high-purity oxygen from air.[22,23] Specifically, air is fed into a vessel containing adsorbents that preferentially adsorb nitrogen over oxygen, allowing pure oxygen to be produced. Once the adsorbent reaches its adsorption capacity, it can be regenerated by decreasing the pressure, thus releasing the adsorbed nitrogen.

PSA has numerous applications beyond oxygen production, such as industrial production of high-purity nitrogen[24,25], removal of carbon dioxide in hydrogen manufactured by natural gas reforming[26-28], and separation of carbon dioxide in biogas upgrading[29-31]. Moreover, in the frame of carbon capture and storage, active research is underway to explore PSA for capturing CO_2 from power plants to mitigate greenhouse gas emissions.[32-34]

Compared to cryogenic distillation, PSA systems are energy-efficient because they operate at near-ambient temperatures. The industrial separation of olefins from paraffins for light

hydrocarbons typically relies on high-pressure cryogenic distillation at low temperatures, which requires significant energy input from refrigeration systems.[3,35] In such cases, PSA systems can be considered an energy-efficient solution.

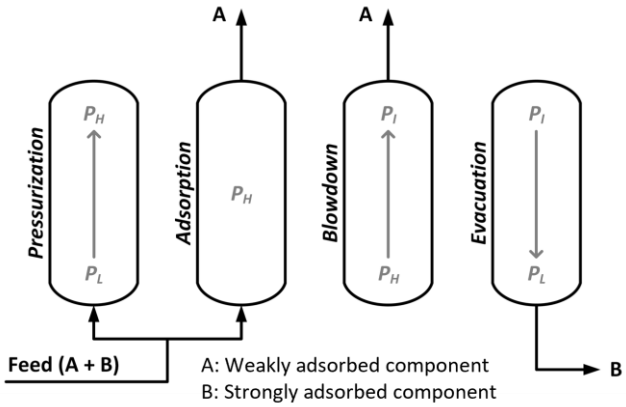


Figure 2-2. VPSA process for gas separation.

Vacuum pressure swing adsorption (VPSA) is a variation of PSA technology, where the adsorption step is performed at pressures higher than ambient and the desorption is achieved under vacuum. VPSA processes have superior regeneration effects and high product recovery rates. A PVSA cycle typically consists of four steps: pressurization, adsorption, blowdown, and evacuation/desorption (**Figure 2-2**).

Taking N_2/CO_2 separation with CO_2-selective adsorbents as an example, the cycle operates as follows:

- ***Pressurization***. The adsorption column begins at the low pressure (desorption pressure, P_L). Pressurized feed is used to raise the pressure of the column from P_L to high pressure (adsorption pressure, P_H).
- ***Adsorption***. Once the column is pressurized, the valve at the end of the column is opened and the pressurized feed flows through the column. CO_2 is adsorbed and N_2 exits from the end.
- ***Blowdown***. Once the column has become saturated with CO_2, the inlet valve is closed, and the column is depressurized by opening the valve at the end of the column.
- ***Evacuation***. After the column is depressurized, the valve at the front end is opened while the valve at the end of the column is closed. CO_2 is recovered by decreasing the pressure to vacuum. Once the CO_2 is removed, the adsorption column is ready for the next cycle.

2.2 Separating agents

In chemical separation processes, a mass separating agent is a chemical added to facilitate the separation of desired components, resulting in reduced energy consumption, improved product purity, or relaxed experimental conditions. The selection of a separating agent is critical for the separation process as it directly influences the separation efficiency, economic viability, and environmental impact.

2.2.1 Organic solvent

Organic solvents are versatile chemicals widely used across numerous scientific disciplines and industries due to their ability to dissolve a vast array of substances. Their applications include dissolving reactants to facilitate reactions, extracting valuable components from natural resources, and purifying solid compounds through recrystallization among others.[36-41] In extractive distillation, an organic solvent is commonly used to alter the relative volatility of close-boiling or azeotropic mixtures to be separated, allowing for improved separation efficiency and reduced energy consumption.

The selection of solvent is key for the viability of an extractive distillation process, and therefore, different aspects should be considered for the selection of a suitable solvent.[11,12] First, the solvent should be able to manipulate the relative volatility of components to be separated. For instance, the solvent should have a high selectivity, i.e., preferential interaction with one component over the other in terms of a binary mixture. Second, it is essential that the solvent can be recovered easily to recycle the solvent back to the extractive distillation column. That is to say, the solvent should have a high relative volatility with the preferentially interacting compound. Third, often but not exclusively, the introduced solvent should not form an azeotrope with the components to be separated. Additionally, other properties of the solvent also influence the separation performance. For instance, low heat capacity and enthalpy of vaporization can reduce energy demand. Environmental, health, and safety impacts can be considered to improve sustainability.

In addition to organic solvents, other types of separating agents such as ionic liquids, deep eutectic solvents, and mixtures of different solvents are increasingly being explored for efficient extractive distillation. Further comprehensive introductions of these advanced solvents can be found in recent works and reviews.[42-44]

2.2.2 Metal-organic framework

Metal-organic frameworks (MOFs) have emerged as an extensive class of crystalline materials, due to their large surface area, high porosity, and customizable functionality.[45] These characteristics make MOFs highly versatile for a wide range of potential applications in gas separation, gas storage, catalysis, and beyond.[46-52] Their modular building blocks (i.e., metal nodes and organic linkers) enable the tailoring of MOF structures with desirable properties for specific applications. This has been confirmed by the successful synthesis of tens of thousands of novel MOFs over the past decade.[53,54] For instance, a new MOF structure called CALF-20 is reported as a promising adsorbent for industrial-scale CO_2 capture, because of its high CO_2 adsorption capacity, high CO_2 selectivity over N_2, and stability during adsorption-desorption cycles.[55]

Compared to traditional adsorbents such as activated carbon and zeolites, MOFs offer several key advantages, including exceptional surface area, diverse structures, and tunable pore structure and functionality.[56] This translates to highly selective adsorption, allowing them to capture desired molecules while excluding undesired ones. Additionally, the enhanced adsorption capacity can significantly improve efficiency and productivity in large-scale separation processes. Despite these attractive features, some challenges remain. Stability has been considered an important factor limiting their applicability. MOFs can be susceptible to degradation under certain conditions, such as exposure to moisture, high temperatures, or specific chemicals.[57] Moreover, scalability and cost-effectiveness are major concerns, as large-scale production for industrial applications can be difficult and expensive. However, the potential benefits of MOFs make them a highly promising area of research with the potential to revolutionize various adsorption applications. Researchers are actively working to overcome these challenges to unlock the full potential of MOFs for large-scale industrial applications.

In addition to MOFs, other types of adsorbents such as covalent organic frameworks and composites of different porous materials are increasingly being explored for efficient adsorption. Further comprehensive introductions of these advanced adsorbents can be found in recent works and reviews.[58-60]

2.3 Computer-aided molecular and process design

In the past, most separating agents such as solvents and adsorbents used in the chemical industry were not systematically selected or designed, primarily depending on domain

knowledge and expert experience.[61] Thus, these separating agents may not be the optimal candidates to meet the separation requirements. In this context, identifying better alternatives through systematic screening or design strategies can improve energy efficiency of separation processes, and therefore, reduce emissions and pollution.

Computer-aided molecular design (CAMD) offers a promising approach for the systematic selection and design of solvents that fulfill a set of target molecular properties or process performance indicators.[62-64] By leveraging modern molecular property models and process models, CAMD methods have been widely used to design solvents for various applications, including gas absorption[65], liquid-liquid extraction[66,67], chemical reaction[68-71], and extractive distillation[72-74] among others[74-76].

While many CAMD studies endeavor to discover chemicals with the ultimate goal of being incorporated into industrial processes, few have explicitly considered the complex relationship between a particular molecule and the process.[77,78] This relationship unfortunately is essential since process performance is often highly sensitive to the chosen molecule. For separation processes such as extraction, crystallization, and adsorption, their feasibility and efficiency are highly dependent on not only the process operating conditions, but also the selection of separating agents.[63,79-81] Taking their interplay into account, the selection of separating agents and optimization of separation processes should be carried out simultaneously. Therefore, the process design needs to be integrated into the CAMD for efficient separation. This integrated approach is known as computer-aided molecular and process design (CAMPD), where molecules and processes are optimized simultaneously to improve the overall process performance. Different CAMPD approaches have been successfully applied to a wide range of processes, such as liquid-liquid extraction[82-84], gas absorption[85-87], pressure swing adsorption[88,89], extractive distillation[90], and chemical reaction[91-93].

2.4 Computational techniques

2.4.1 Data-driven modeling

Mathematical models are fundamental to the simulation, optimization, and control of chemical processes.[94,95] Accurate modeling and simulation of processes usually benefit from the increasing complexity of underlying models, which also leads to increased computational demands in applications such as process optimization.[96] To address this challenge, different

strategies such as surrogate modeling have been developed to replace complex models, thereby reducing the computational effort in function evaluation and model-based optimization.[97,98]

With the continuous advances in artificial intelligence and related subjects (e.g., machine learning, data science, and digitalization), data-driven techniques are transforming and revolutionizing both fundamental research and industrial practices across various disciplines.[99-103] In recent years, data-driven modeling has received substantial attention for its ability to effectively handle large datasets and approximate complex systems[96,102,104,105], providing practical solutions for a wide range of applications[106-114].

Data-driven modeling creates shortcut models for complex systems that are computationally expensive or unknown. In chemical and process engineering, data-driven models are usually built based on Kriging[98,107,113,115] and artificial neural networks (ANNs)[108,116-119], and they are subsequently used to replace the original physics-based models to alleviate computational burdens in process optimization. Some examples of applications in engineering fields include optimization of distillation columns[98,107], optimization of carbon fiber production plant[120], and control of pharmaceutical manufacturing system[121]. These studies demonstrate that data-driven models enable computationally efficient optimizations. Overall, data-driven modeling has been recognized as an emerging tool to build surrogate models (with high accuracy and low complexity) to capture the behavior of complex systems, enabling the efficient design and optimization of chemical processes.

2.4.2 Mathematical optimization

Mathematical optimization involves finding the best solution from all possible solutions to maximize or minimize a function. In terms of a function *f* defined on a domain *X*, the goal of optimization (in the case of minimization problems) is to systematically search the domain for a point $\boldsymbol{x}^* \in X$ such that $f(\boldsymbol{x}^*) \leq f(x)$ for all $\boldsymbol{x} \in X$. Typically, *X* is a subset of the Euclidean space $\mathbb{R}^n$, often constrained by a set of conditions that elements of *X* have to satisfy. The domain *X* is known as the search space, and the elements of *X* are called candidate solutions. The function *f* is called the objective function, and a feasible solution that minimizes the objective function is the optimal solution.

In general, three key components are integrated when formulating an optimization problem: objective function, decision variable, and constraint. Formulating an optimization problem involves translating a real-world problem into the mathematical equations and variables

comprising these three components. The objective function *f* is the function to be optimized (minimized or maximized). The decision variables, often denoted as the vector $\boldsymbol{x}$, are the unknown and controllable parameters that need to be adjusted to optimize the objective function. These variables can be discrete or continuous. Constraints define the feasible region within which the optimization algorithm must search for the optimal solution, limiting the possible values for the decision variables. Constraints can be divided into two categories: equality constraint *h* and inequality constraint *g*.

An optimization problem (for minimization) can be generally formulated as:

$$\begin{aligned} & \min_{\boldsymbol{x}} \ f(\boldsymbol{x}) \\ \textit{s.t.} \quad & h(\boldsymbol{x}) = 0 \\ & g(\boldsymbol{x}) \leq 0 \\ & \boldsymbol{x} \in \mathbb{R}^n \end{aligned}$$

The objective function is optimized with respect to decision variables in the presence of constraints on those variables. If an optimization problem involves more than one objective function to be optimized simultaneously, it is called multi-objective optimization. These objectives usually conflict with each other. For example, in terms of materials for Aerospace engineering, one might want to minimize both cost and weight while maximizing strength. Since improving one objective often degrades at least one of the other objectives, trade-offs can be identified, leading to a Pareto set of optimal solutions. All these Pareto optimal solutions are considered equally good. To solve multi-objective optimization problems, various algorithms are available, such as linear scalarization, epsilon-constraint method, and evolutionary algorithms.[122]

Optimization problems arise in various applications across different fields. In chemical production systems, process optimization is regularly applied to identify the operating conditions that maximize process efficiency and productivity under given product quality constraints. First, mathematical models of the process are developed and validated using experimental data. These models are then used as in silico representation of the process, allowing for the determination of optimal operating conditions using either derivative-based or derivative-free algorithms.[123]

2.4.3 Bayesian optimization

Bayesian optimization is a sequential design strategy used to optimize expensive-to-evaluate functions.[124,125] Unlike classical optimization techniques, Bayesian optimization routines rely on a statistical model that approximates the function of interest, guiding the algorithm to make the most informed decisions.[126] This model is computationally cheaper to evaluate than the actual objective function and can efficiently direct the search for the optimal solution. Bayesian optimization can deliver impressive performance even when optimizing complex functions under limited evaluation budgets.

One of the key advantages of Bayesian optimization is that it does not require the objective function to have a known mathematical form. Instead, it considers the objective function as a black box, requiring only measurements at selected points on demand.[126] In Bayesian optimization, the unknown objective function is treated as a random function, and a prior (usually a Gaussian process model) is placed over it. The optimization process follows three main steps. ***Modeling*****:** Using the available data (i.e., initial observations of the objective function), the prior is updated to form the posterior distribution (i.e., Gaussian process model trained with the initial dataset) over the objective function. ***Optimization*****:** The posterior distribution, in turn, is used to construct an acquisition function, which is used to determine the point to make the next observation (i.e., the optimal solution for the objective function) by using mathematical optimization techniques. ***Validation*****:** After the objective function at the suggested point is validated, the newly observed information is added to the dataset to update the posterior. This ***modeling-optimization-validation*** procedure iterates until the termination condition is reached (e.g., budget exhausted). In general, Bayesian optimization can efficiently explore the search space and identify the optimum for the objective function being studied.

In the context of data-driven modeling for complex systems, developing sufficiently accurate surrogate models to describe system behaviors often necessitates extensive data covering the entire region of interest (e.g., the operation window of a separation unit in a chemical process).[116] However, obtaining such data can be resource-intensive. Due to the limited data and high dimensionality encountered in engineering design tasks, constructing a globally valid approximation model remains difficult. For such situations, data-efficient approaches are of great importance in alleviating the burden of data collection, where Bayesian optimization can stand out.

As a data-efficient optimization method, Bayesian optimization is particularly advantageous for high-dimensional optimization problems where the objective function is difficult or expensive to evaluate. Evidenced by successful applications in reaction condition optimization[127-130], materials discovery[131-134], and reactor design[135,136], Bayesian optimization exhibits remarkable advantages in achieving state-of-the-art performance with minimal experimental or computational costs, making it a valuable tool for complex, high-dimensional optimization problems in engineering and scientific research.

PART I. MOLECULAR DISCOVERY

3 Optimal Solvent Design for Extractive Distillation Processes

This chapter introduces molecular property targeting and molecular mapping techniques for optimal molecular design. They are developed based on the continuous-molecular-targeting (CoMT) approach[137], which considers a continuous molecular structure space in terms of property-related parameters. Characterizing molecular structures by molecular properties, the molecular property targeting and molecular mapping approaches circumvent discrete molecular decisions in the optimal molecular design.

In this chapter, solvents are optimally designed for extractive distillation (ED) processes by directly targeting desirable molecular properties. First, data-driven process models are established to estimate key performance indicators of the ED process with the most important process-relevant properties of the solvent. Subsequently, solvent design is performed in two steps: molecular property targeting and molecular mapping. In the first step, optimal molecular properties are obtained from model-based optimization, and hypothetical target molecules featuring the desirable properties are thereby generated. In the subsequent step, real solvents that approximate the optimal property profiles of hypothetical molecules are identified from a real solvent database.

The proposed molecular property targeting approach is illustrated using an industrially relevant case: the separation of the close-boiling mixture 1,3-butadiene/1-butene (C_4H_6/C_4H_8), as introduced in **Appendix A**.

3.1 Data-driven process modeling

To efficiently evaluate the performance of solvents in separation systems, process models that reflect the impact of solvents on the process performance are required. Such models can also be used for the optimal design of solvents to discover better alternatives that present improved process performance while satisfying desired separation requirements.

Mechanistic models across different levels of separation systems (molecular interaction, thermodynamics, phase equilibrium, mass and energy transfer, etc.) are commonly used to provide insights for simulation and optimization purposes. In a data-driven manner, efficient

process surrogate models can be built to directly link solvents to their corresponding process performance indicators, achieving efficient process evaluation and optimization (**Figure 3-1**). Such models can direct the design of solvents by identifying the optimal values of molecular properties that maximize the process performance. For the C_4H_6/C_4H_8 separation, N-methyl-2-pyrrolidone (NMP) is recognized as a benchmark solvent since it is commercially used in butadiene extraction processes.[138-140] The optimal operating conditions are pre-determined by process optimization using NPM as the solvent. Subsequently, under these operating conditions, the optimal solvent design is carried out. Thus, it can be considered as the search for potentially better alternatives to the industrially used solvent NMP for butadiene extraction. For simplification, in this chapter, only the extractive distillation column (EDC) is considered to demonstrate the molecular property targeting method for the optimal solvent design. The entire ED process consisting of EDC and solvent recovery column (SRC) are further considered in the CAMPD in **Chapter 4**.

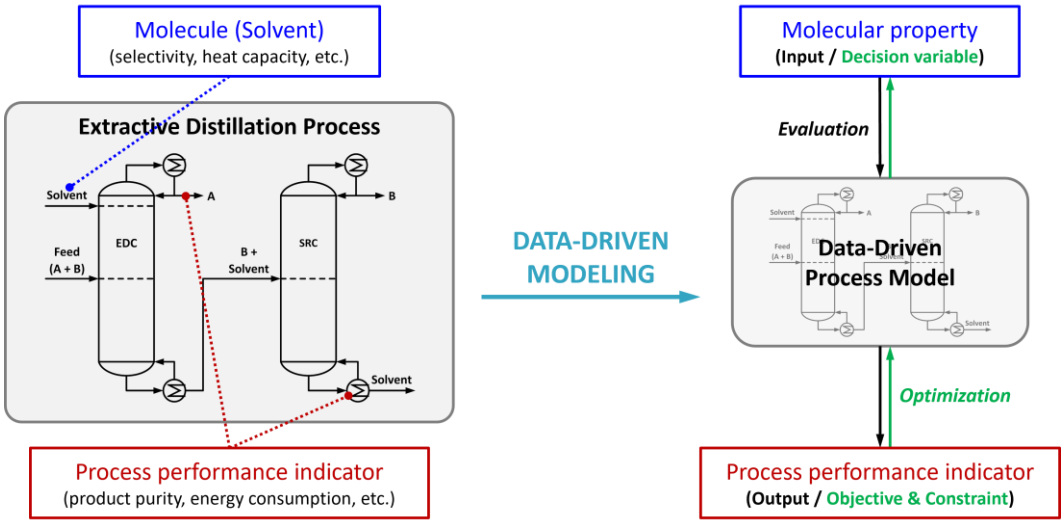


Figure 3-1. Data-driven modeling of the ED process for the optimal solvent design.

Taking advantage of the CoMT method[141], discrete molecular decision variables are circumvented by defining a hypothetical molecule that is represented by continuous parameters, i.e., process-relevant molecular properties. Five molecular properties are considered, including selectivity at infinite dilution, molar heat capacity, molecular weight, density, and viscosity.

Selectivity at infinite dilution (S^{∞}) is used to describe the capability of solvents in purifying C_4H_8 in the EDC. The selectivity of C_4H_6 over C_4H_8 at infinite dilution is expressed as,

$$S^{\infty}_{C_4H_6/C_4H_8} = \frac{\gamma^{\infty}_{C_4H_8}}{\gamma^{\infty}_{C_4H_6}}$$

where γ^{∞} is the infinite dilution activity coefficient. It is calculated at 25 °C in this case.

Molar heat capacity (C_p) reflects the energy demand in heating the solvent-contained mixtures. It is calculated at 25 °C and 1 bar in this case. Molecular weight (MW) is easy to obtain and can characterize the boiling point of the solvent to a large extent. Moreover, density (ρ) and viscosity (μ) are considered because they affect the transport of materials in the system. Both are calculated at 25 °C and 1 bar in this case.

The dataset for data-driven modeling contains input-output pairs for 126 organic solvents selected from the Aspen Plus component database[142]. The inputs are the aforementioned five process-relevant molecular properties of the solvent. The outputs are the key performance indicators of the ED process, including product purity (C_4H_8 purity of the EDC distillate, x_{C4H8}) and energy demand (described by EDC reboiler heat duty, Q_{EDC}). The dataset is generated by rigorous process simulation in Aspen Plus based on the UNIFAC thermodynamic model. It is randomly divided into two sets, i.e., a training set (80%) for model development and a test set (20%) for model evaluation. Feature scaling is applied to the input space using z-score normalization.

Table 3-1. Hyperparameters and corresponding options for hyperparameter optimization.

Hyperparameter	Options
The number of hidden layers (N_{HL})	1, 2
The number of neurons in each hidden layer (N_{HN})	[1, 4]
Activation function	ELU, Sigmoid, Softplus, Tanh

Feedforward neural network (FNN), the most straightforward type of artificial neural networks, is used to build the data-driven process model and implemented using PyTorch[143]. To constrain model complexity and reduce overfitting, the FNN has up to two hidden layers with a maximum of four neurons in each layer subject to hyperparameter optimization. Different types of non-linear activation functions are considered, including exponential linear unit (ELU), Sigmoid, Softplus, and hyperbolic tangent (Tanh).[144] To determine the optimal FNN architectures for data-driven process models, five-fold cross-validation is performed for the

optimization of hyperparameters. Hyperparameters and their corresponding options are provided in **Table 3-1**.

Table 3-2. Optimal hyperparameter settings.

Model	N_{HL}	N_{HN}	Activation function
x_{C4H8}	1	2	Sigmoid
Q_{EDC}	1	4	Softplus

With the optimal FNN hyperparameters identified by the five-fold cross-validation (**Table 3-2**), process models are developed using the training data and evaluated using the test data. The process performance of each solvent is predicted using five models derived from the five-fold cross-validation, and the point and the error bar in **Figure 3-2** show the average and standard deviation of these predictions, respectively. As they present satisfactory accuracy in the prediction of C_4H_8 purity and reboiler heat duty, the data-driven models are subsequently used for the optimal solvent design.

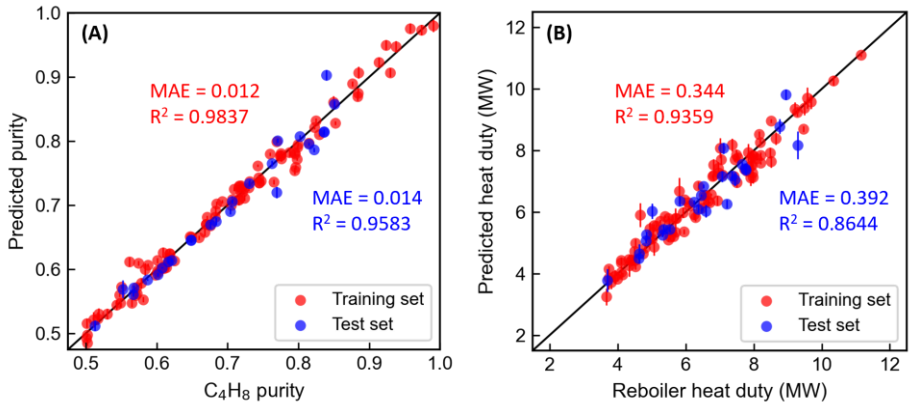


Figure 3-2. Performance of the data-driven models for (A) C_4H_8 purity and (B) reboiler heat duty.

3.2 Molecular property targeting

Based on the developed data-driven models, a multi-objective optimization problem is formulated to maximize the C_4H_8 purity while minimizing the heat duty, as follows:

$$\min_{\boldsymbol{y}} \; (1 - f_{\mathrm{C4H8}}(\boldsymbol{y}),\; f_Q(\boldsymbol{y}))$$

$$s.t. \quad \boldsymbol{y}_L \leq \boldsymbol{y} \leq \boldsymbol{y}_U$$

where f_{C4H8} and f_Q are the data-driven models for the estimation of C_4H_8 purity and reboiler heat duty, respectively, $\boldsymbol{y}$ is the molecular property space, and L and U denote the lower and upper bounds. Box constraints are set on each variable based on the corresponding minimum and maximum values from the dataset. The lower and upper bounds of each molecular decision variable are summarized in **Table 3-3**.

Table 3-3. Upper and lower bounds of molecular decision variables.

Variable	Symbol	Unit	Lower bound	Upper bound
Selectivity at infinite dilution	S^{∞}	–	0.844	1.642
Molar heat capacity	C_p	J/(mol·K)	114.7	395.5
Molecular weight	MW	g/mol	71.12	172.27
Density	ρ	kg/m^3	701	1618
Viscosity	μ	mPa·s	0.23	8.43

The multi-objective optimization problem is solved using the non-dominated sorting genetic algorithm (NSGA-II)[145] implemented in Pymoo[146]. With a population size of 100, the optimization converges in 100 generations, obtaining a set of Pareto-optimal solutions (i.e., hypothetical target solvent molecules). The hypothetical molecules are considered to be the optimal solutions in the design space. **Figure 3-3** depicts the objective function values for the hypothetical target molecules featuring optimal molecular properties. Z-score normalization is applied to the heat duty (i.e., objective function 2) so that the magnitudes of two objective functions are comparable.

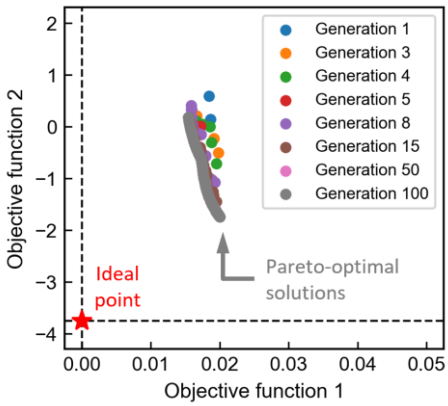


Figure 3-3. Multi-objective optimization results for the optimal solvent design.

3.3 Molecular mapping

With the industrially used solvent NMP as the benchmark, the target is to find better solvent alternatives that allow for a higher C_4H_8 purity and a lower reboiler heat duty under the given operating conditions. **Figure 3-4** shows the estimated process performance for hypothetical molecules (in gray) and benchmark solvent NMP (in green). The hypothetical molecules are closer to the ideal point than NMP, demonstrating that the molecular property targeting step successfully identifies better solutions.

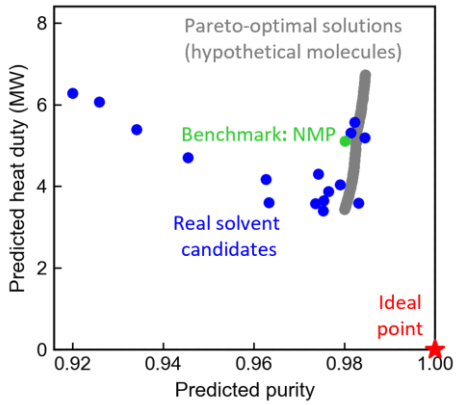


Figure 3-4. Process performance estimated by the data-driven models for hypothetical molecules, NMP, and real solvent candidates.

In the molecular mapping step, the hypothetical target molecules obtained from the molecular property targeting are mapped into real solvents. The molecular mapping is performed by searching a large database consisting of 1259 real solvents, which is derived from the Aspen Plus component database[142] with the exclusion of solvents used in the development of data-driven models. A preliminary criterion to find the real solvents closest to the hypothetical target molecules is based on the Euclidean distance in the molecular property space. Thereby, optimal real solvents that approximate the optimal property values are identified.

Nineteen solvent candidates are obtained from the molecular mapping. Their estimated process performance is also presented in **Figure 3-4**. It is observed that two solvent candidates are closer to the ideal point than the hypothetical molecules. Therefore, they could in principle show better process performance than the hypothetical molecules. A detailed simulation of the ED process on the 19 solvent candidates proves that 16 of them are technically viable to achieve the separation of C_4H_6/C_4H_8. Among them, nine solvents present decreased reboiler heat duty yet lower C_4H_8 purity compared to NMP under the identical operating conditions. Besides,

three solvents, methyl cyanoacetate, glutaronitrile, and 1,4-dicyano-2-butene (blue dots from bottom to top in the green area of **Figure 3-5**), are better alternatives to the benchmark solvent NMP, because they allow for a higher product purity and lower energy demand under the specified operating conditions. Vertical and horizontal lines in green represent the C_4H_8 purity and heat duty achieved by the benchmark solvent.

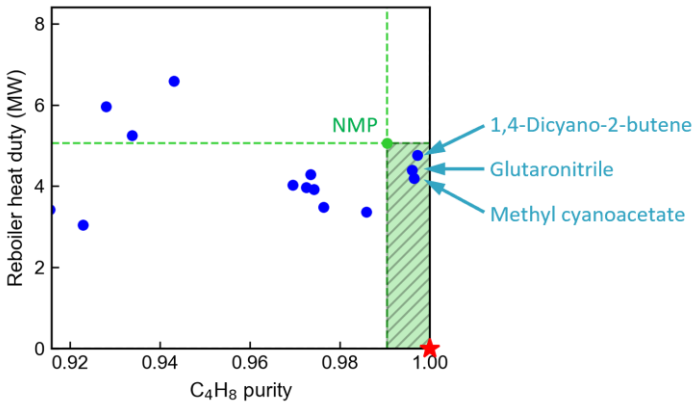


Figure 3-5. Process performance evaluated via rigorous process simulation.

Table 3-4. Molecular properties and the corresponding process performance of the solvents.

Solvent	NMP	Methyl cyanoacetate	Glutaronitrile	1,4-Dicyano-2-butene
CAS number	872-50-4	105-34-0	544-13-8	18715-38-3
Molecular formula	C_5H_9NO	$C_4H_5NO_2$	$C_5H_6N_2$	$C_6H_6N_2$
S^{∞}	1.642	1.753	1.666	1.689
MW (g/mol)	99.13	99.09	94.12	106.13
ρ (kg/m^3)	1027	1117	983	1002
C_p (J/(mol·K))	161.7	192.5	191.0	200.9
μ (mPa·s)	1.89	2.82	6.17	7.13
x_{C4H8}	0.9904	0.9965	0.9960	0.9972
Q_{EDC} (MW)	5.059	4.198	4.394	4.761

For the three solvent candidates, their molecular properties and corresponding process performance are summarized in **Table 3-4**. These solvent candidates have similar molecular properties (except for viscosity) to the benchmark solvent NMP. A correlation analysis performed on the dataset indicates that a solvent with a higher selectivity at infinite dilution

could lead to a higher C_4H_8 purity. This can also be inferred from **Table 3-4**. All three solvent candidates have higher selectivity and higher C_4H_8 purity than NMP. Therefore, the infinite dilution selectivity of the solvent can be considered a vital property in designing solvents for energy-efficient separation of C_4H_8 and C_4H_6.

4 Integrated Design of Solvents and Extractive Distillation Processes

Chapter 3 presents a data-driven CAMD approach focusing on the optimal solvent design under specified process operating conditions. However, process optimization is not considered to identify the optimal operating conditions that maximize process performance for the optimal solvent. Therefore, in this chapter, two data-driven CAMPD approaches are introduced to integrate molecular and process design, aiming to determine both the optimal molecules and the operating conditions simultaneously.

In **Section 4.1**, an efficient CAMPD approach is proposed for integrated molecular and process design using data-driven modeling. Data-driven process surrogates are developed to directly estimate key process performance indicators based on solvent properties and process operating parameters. Surrogate-based optimization is then employed to enhance process performance, through which optimal solvent properties and corresponding process parameters are obtained. Real solvents that approximate these optimal properties are subsequently identified from a large solvent database. Finally, the performance of the optimal solvent and corresponding process parameters is validated by rigorous process simulations.

To reduce data demand and improve the CAMPD efficiency further, BayesCAMPD approach is proposed in **Section 4.2** for the integrated molecular and process design using Bayesian optimization. This approach offers a data-efficient and closed-loop solution for data-driven CAMPD, enabled by an iterative process of data-driven modeling, model-based optimization, and solution validation. By inferring from observed data, BayesCAMPD continuously suggests and validates promising molecular and process settings until convergence.

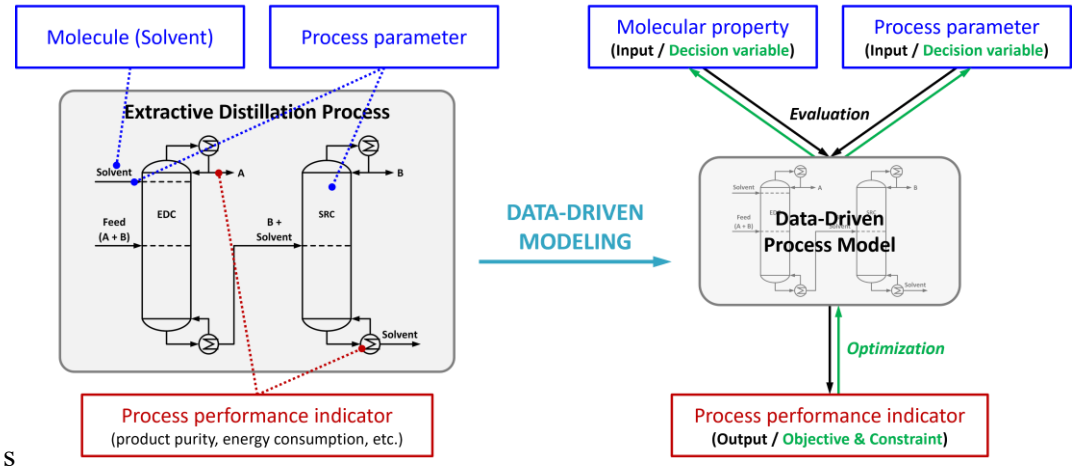
Both data-driven CAMPD approaches are illustrated by the separation of 1,3-butadiene/1-butene (C_4H_6/C_4H_8) using extractive distillation, which is the same case considered for the data-driven CAMD approach as discussed in **Chapter 3**.

4.1 Data-driven integrated molecular and process design

In CAMPD, solvent physical properties and process parameters are optimized simultaneously to maximize the overall process performance. Taking advantage of the CoMT method[137],

discrete molecular decision variables are circumvented by defining a hypothetical molecule that is represented by continuous parameters, i.e., molecular physical properties.

In a data-driven manner, process models are built to directly link both solvent properties and process parameters with their corresponding process performance (**Figure 4-1**). Using such models, the CAMPD problem can be efficiently solved to identify an ideal hypothetical solvent (represented by a set of optimal properties) and the corresponding optimal process operating conditions showing the highest process performance. In a subsequent step, the hypothetical target molecule is mapped onto real solvents, which is consistent with the molecular property targeting and molecular mapping methods introduced in **Chapter 3**. Data and code for implementing the data-driven CAMPD approach in this section are available in the GitHub repository[147].



s

Figure 4-1. Schematic diagram of data-driven integrated solvent and process design.

4.1.1 Data-driven process modeling

To enable CAMPD, data-driven process models are developed to estimate key process indicators based on solvent properties and process operating conditions. In addition to the five molecular properties introduced in **Section 3.1**, relative volatility at infinite dilution is considered to characterize the difficulty of recovering solvent in the solvent recovery column (SRC). Relative volatility between the solute (i.e., C_4H_6) and solvent at infinite dilution (α^{∞}) in the SRC reflects the separation efficiency in the solvent recovery step. The relative volatility of C_4H_6 over the solvent at infinite dilution is expressed as,[148]

$$\alpha^{\infty}_{C_4H_6/\text{solvent}} = \frac{P^0_{C_4H_6}}{P^0_{\text{solvent}}} \cdot \frac{\gamma^{\infty}_{C_4H_6}}{\gamma^{\infty}_{\text{solvent}}}$$

where P^0 is saturated vapor pressure. It is calculated at 25 °C in this case.

Therefore, a total of six molecular properties are considered in the CAMPD problem, including selectivity at infinite dilution, relative volatility at infinite dilution, molar heat capacity, molecular weight, density, and viscosity. In addition to the solvent properties, seven process parameters are considered in the CAMPD. In both EDC and SRC, the number of stages (N), reflux ratio (R), and operating pressure (P) are considered key process parameters for optimization, in addition to the solvent-to-feed ratio (S/F). The solvent-to-feed ratio is a global variable associated with the entire ED process, while other process parameters are local variables involved either in the EDC or SRC. Thus, in total 13 decision variables are considered. The number of stages is a discrete variable while other variables are continuous.

Data-driven process models are established for the EDC and SRC separately instead of the entire ED process, because the computational cost of data generation and modeling increases exponentially with the number of process parameters considered. The dataset for data-driven process modeling contains input-output pairs for 130 different solvents derived from the Aspen Plus component database[142]. For each solvent, the process performance is calculated under different process parameters by rigorous process simulations in Aspen Plus based on the UNIFAC thermodynamic model. By performing a full factorial design of computational experiments (DoCE) as shown in **Table 4-1**, the EDC and SRC datasets are obtained, containing ~560k and ~396k data points, respectively.

Table 4-1. Full factorial DoCE of process parameters for initial sampling.

Column	Variable	Considered levels	Unit
EDC	N_{EDC}	40, 45, 50, 55, 60, 65, 70, 75, 80	–
	R_{EDC}	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	–
	P_{EDC}	3.5, 4.0, 4.5, 5.0, 5.5, 6.0	bar
	S/F	1, 2, 3, 4, 5, 6, 7, 8	–
SRC	N_{SRC}	8, 10, 12, 14, 16, 18, 20	–
	R_{SRC}	0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0	–
	P_{SRC}	3.5, 4.0, 4.5, 5.0, 5.5, 6.0	bar
	S/F	1, 2, 3, 4, 5, 6, 7, 8	–

As process models are built for EDC and SRC separately, the inputs include six properties of the solvent and four process parameters (N, R, P, and S/F). The outputs are the key performance indicators of the ED process, including product purity (C_4H_8 purity in the EDC distillate and C_4H_6 purity in the SRC distillate, denoted by x_{C4H8} and x_{C4H6}) and energy demand (described by the reboiler heat duty of the EDC and SRC, Q_{EDC} and Q_{SRC}). Thus, four data-driven process models need to be trained to evaluate the process performance from the solvent properties and process parameters. The feedforward neural network (FNN) is used to build data-driven process models using PyTorch[143]. Each of the datasets is randomly split into two subsets, i.e., a training set (80%) for model development and a test set (20%) for evaluation.

To determine the optimal FNN architectures for data-driven process models, five-fold cross-validation is performed to optimize model-related hyperparameters using the training data. Hyperparameters considered for optimization and their corresponding options are provided in **Table 4-2**. To constrain model complexity and reduce overfitting, the FNN has up to two hidden layers with a maximum of 24 neurons in each layer. Different types of non-linear activation functions are considered, including ELU, Sigmoid, Softplus, and Tanh.[144]

Table 4-2. Hyperparameters and corresponding options for hyperparameter optimization.

Hyperparameter	Options
The number of hidden layers (N_{HL})	1, 2
The number of neurons in each hidden layer (N_{HN})	[8, 24]
Activation function	ELU, Sigmoid, Softplus, Tanh

Table 4-3. Optimal hyperparameter settings.

Model	N_{HL}	N_{HN}	Activation function
x_{C4H8}	2	14	Softplus
Q_{EDC}	1	13	ELU
x_{C4H6}	1	14	Sigmoid
Q_{SRC}	2	22	Softplus

By minimizing the average prediction errors on the validation data in the five-fold cross-validation, the optimal hyperparameter settings are determined (**Table 4-3**).

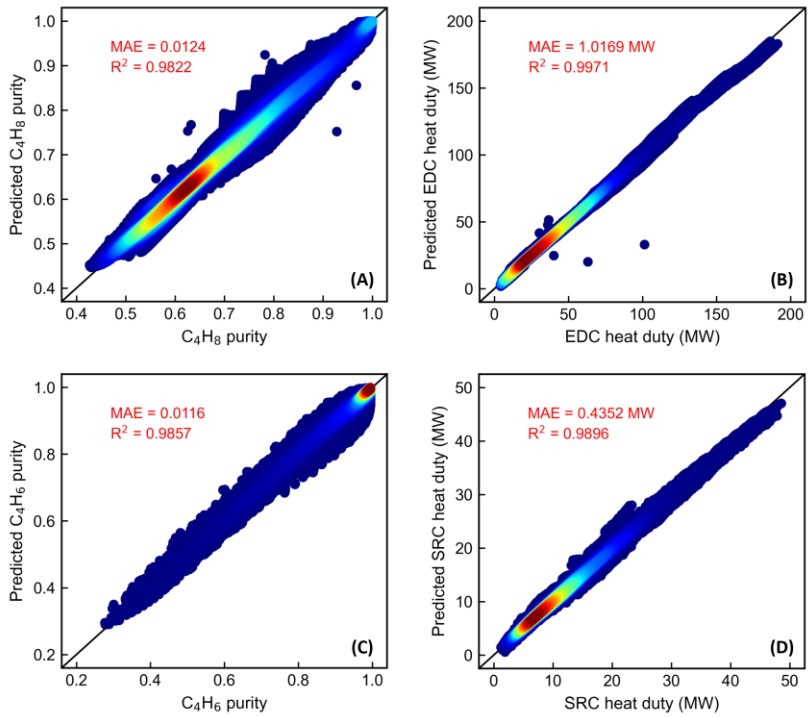


Figure 4-2. Performance of the data-driven models on the training data for (A) x_{C4H8}, (B) x_{C4H6}, (C) Q_{EDC}, and (D) Q_{SRC}.

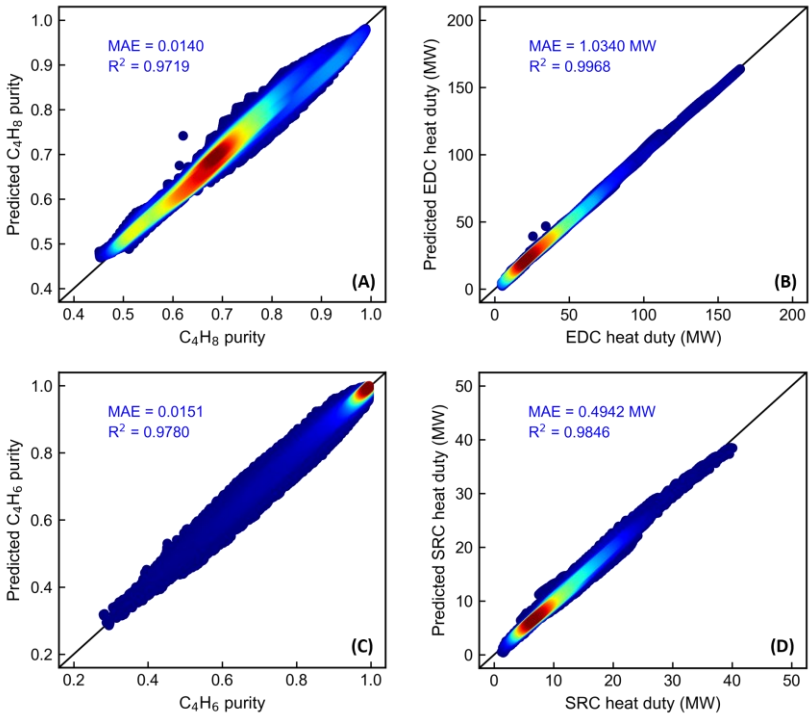


Figure 4-3. Performance of the data-driven models on the test data for (A) x_{C4H8}, (B) x_{C4H6}, (C) Q_{EDC}, and (D) Q_{SRC}.

Using the optimal FNN architectures, process models are developed with the training data and evaluated with the test data. It takes 30 minutes to complete the training of these process models for x_{C4H8}, x_{C4H6}, Q_{EDC}, and Q_{SRC}. All models present high accuracy on both training and test data (**Figure 4-2** and **Figure 4-3**), enabling accurate estimations of the process performance given molecular properties of the solvent and process parameters. The distribution of data in **Figure 4-2** and **Figure 4-3** is indicated by colors where red and blue represent high and low densities of data points, respectively.

4.1.2 Model-based optimization and solvent mapping

Based on the data-driven process models, a multi-objective optimization problem, that aims to minimize the total number of distillation stages and the total heat duty, is formulated to identify the optimal solvent properties and process parameters.

$$\begin{aligned} & \min_{\boldsymbol{y},\boldsymbol{z}} \left(N_{EDC} + N_{SRC},\ f_{Q_{\mathrm{EDC}}}(\boldsymbol{y},\boldsymbol{z}) + f_{Q_{\mathrm{SRC}}}(\boldsymbol{y},\boldsymbol{z})\right) \\ \textit{s.t.} \quad & f_{x_{\mathrm{C_4H_8}}}(\boldsymbol{y},\boldsymbol{z}) \geq 0.995 \\ & f_{x_{\mathrm{C_4H_6}}}(\boldsymbol{y},\boldsymbol{z}) \geq 0.995 \\ & \boldsymbol{y}_L \leq \boldsymbol{y} \leq \boldsymbol{y}_U \\ & \boldsymbol{z}_L \leq \boldsymbol{z} \leq \boldsymbol{z}_U \end{aligned}$$

where f is the developed data-driven model, $\boldsymbol{y}$ and $\boldsymbol{z}$ denote the solvent property and process parameter spaces, respectively, and L and U denote the lower and upper bounds. Box constraints are set on each variable based on their corresponding minimum and maximum values from the dataset. Among all the decision variables involved in the CAMPD, the number of stages (i.e., N_{EDC} and N_{SRC}) is a discrete variable while the others are continuous variables. The lower and upper bounds of the decision variables are summarized in **Table 4-4**.

The multi-objective optimization problem is solved using the NSGA-II[145] implemented in Pymoo[146], where a rounding operator is applied for discrete variables. With a population size of 1000, it terminates at the 100th iteration reaching the maximal number of evaluations. This generates a set of Pareto-optimal solutions consisting of optimal hypothetical molecules (represented by solvent properties) and corresponding process parameters. **Figure 4-4** depicts the objective function values for the optimal solutions. As two objective functions are defined based on the total number of distillation stages and the total reboiler heat duty, min-max normalization and z-score normalization are respectively applied to them so that their

magnitudes are comparable. Such an integrated design problem is solved within one minute, indicating that the data-driven modeling approach can substantially reduce the optimization complexity and enable efficient CAMPD.

Table 4-4. Upper and lower bounds of decision variables involved in the CAMPD.

Variable	Symbol	Unit	Lower bound	Upper bound
Selectivity at infinite dilution	S^{∞}	–	0.907	1.642
Relative volatility at infinite dilution	$\log_{10}(\alpha^{\infty})$	–	1.326	5.944
Molar heat capacity	C_p	J/(mol·K)	125.9	413.0
Molecular weight	MW	g/mol	70.09	200.32
Density	ρ	kg/m^3	712	1182
Viscosity	μ	mPa·s	0.34	2.47
Number of stages in the EDC	N_{EDC}	–	40	80
Reflux ratio of the EDC	R_{EDC}	–	1.00	10.00
Operating pressure of the EDC	P_{EDC}	bar	3.50	6.00
Number of stages in the SRC	N_{SRC}	–	8	20
Reflux ratio of the SRC	R_{SRC}	–	0.20	2.00
Operating pressure of the SRC	P_{SRC}	bar	3.50	6.00
Solvent-to-feed ratio	S/F	–	1.00	8.00

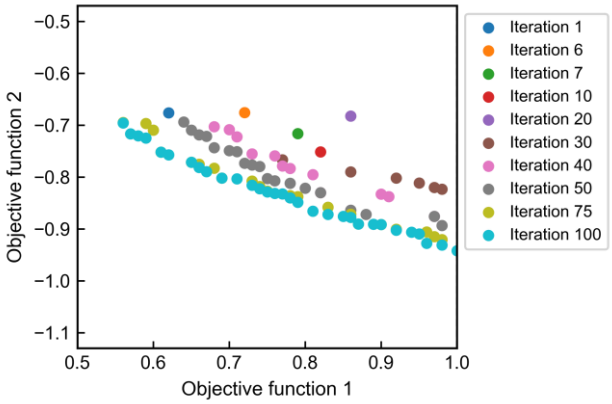


Figure 4-4. Multi-objective optimization results for the CAMPD.

In the follow-up step, the obtained hypothetical target molecules (Pareto solutions characterized by a set of optimal properties for each solution) are mapped onto real solvents. Such solvent mapping is performed by searching a large database consisting of 1248 solvents,

which is derived from the Aspen Plus component database[142] with the exclusion of solvents used in the development of data-driven models. The identification of real solvents closest to the hypothetical target molecules is based on the Euclidean distance in the molecular property space scaled by z-score normalization.

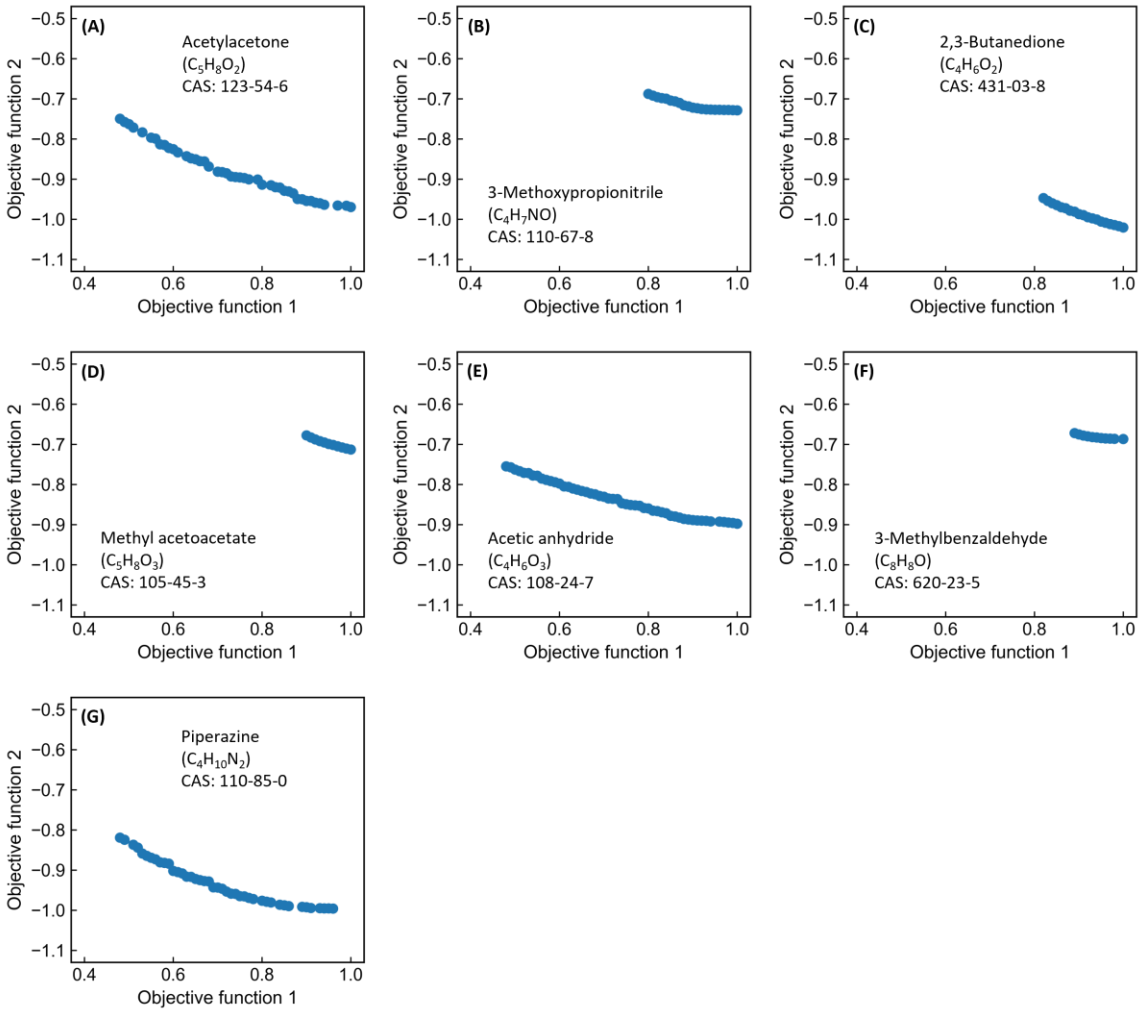


Figure 4-5. Pareto front obtained in the multi-objective process optimization for the solvents identified by CAMPD.

After searching for the optimal real solvents to match the hypothetical molecules, nine solvent candidates are obtained. To find their corresponding optimal process parameters, process optimization is performed for each solvent based on the data-driven process models. In the optimization, the molecular properties are fixed, and the process parameters are regarded as decision variables. For seven solvents, the process optimization successfully generates Pareto-optimal solutions that potentially satisfy the purity specifications (**Figure 4-5**). For the other two solvents, the process optimization fails to give any solution because of the insufficient

separation capability of the solvent, or the high energy demand required for the studied separation task.

Rigorous simulations of the ED process are performed for these seven solvents to evaluate their actual process performance. Compared to the process simulation, it is found that the developed models generally underestimate the product purity and slightly overestimate the heat duty (**Figure 4-6**). In terms of 2,3-butanedione, it cannot satisfy the purity specifications as the C_4H_8 purity of the EDC distillate is lower than 99.5% (**Figure 4-6C**). Although some solutions in the methyl acetoacetate case can approach the purity specifications, the required energy demand for the ED process is relatively high (**Figure 4-6D**). Acetylacetone and acetic anhydride (**Figure 4-6A** and **E**) are considered suitable candidates as high product purity and low energy demand can be simultaneously achieved.

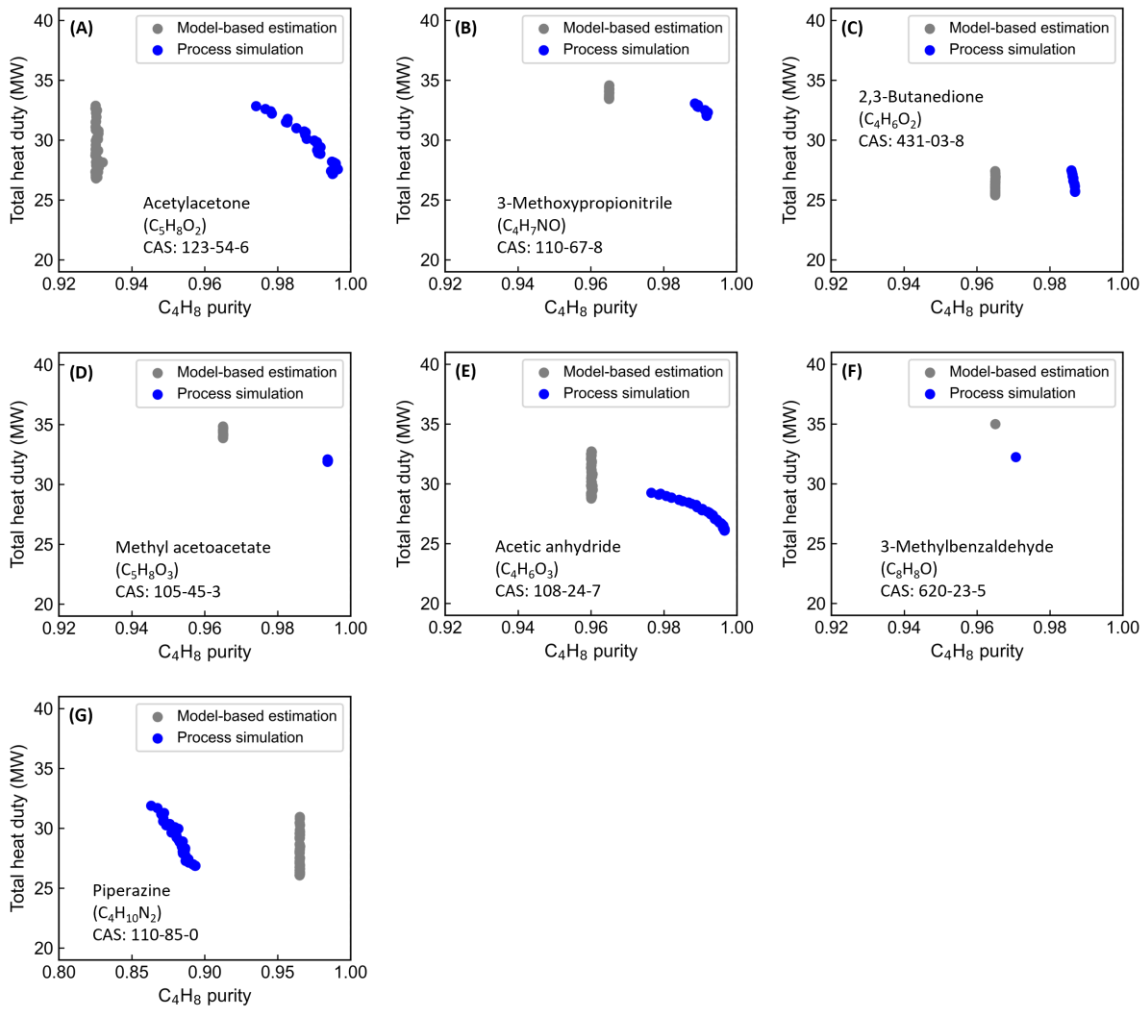


Figure 4-6. Process performance estimated by the data-driven models and evaluated via rigorous simulations for the solvents identified by CAMPD.

4.1.3 Comparison and analysis

To benchmark the performance of identified optimal solvents, N-methyl-2-pyrrolidone (NMP) is used as a reference solvent since it is commercially used in butadiene extraction processes.[138-140] The optimal process parameters for NMP are obtained by data-driven modeling and surrogate-based optimization, and therefore such an NMP-based ED process can be considered as a benchmark for the data-driven CAMPD. First, datasets for the EDC and SRC are generated through rigorous process simulation in Aspen Plus based on the UNIFAC thermodynamic model. The process performance is evaluated under different process parameters generated by the DoCE shown in **Table 4-1**. After removing simulation data with errors, the EDC and SRC datasets contain 4312 and 2036 data points, respectively. Considering the FNN hyperparameters listed in **Table 4-5**, the optimal settings (**Table 4-6**) are determined using five-fold cross-validation.

Table 4-5. Hyperparameters and corresponding options for hyperparameter optimization.

Hyperparameter	Options
The number of hidden layers (N_{HL})	1, 2
The number of neurons in each hidden layer (N_{HN})	[1, 16]
Activation function	ELU, Sigmoid, Softplus, Tanh

Table 4-6. Optimal hyperparameter settings.

Model	N_{HL}	N_{HN}	Activation function
x_{C4H8}	2	10	Tanh
Q_{EDC}	2	11	ELU
x_{C4H6}	2	9	Tanh
Q_{SRC}	2	15	ELU

Given the optimal hyperparameter settings, surrogates for the ED process that use NMP as the solvent are established with the training data. All models show sufficiently high accuracy on both training and test data (**Figure 4-7**), indicating their strong applicability.

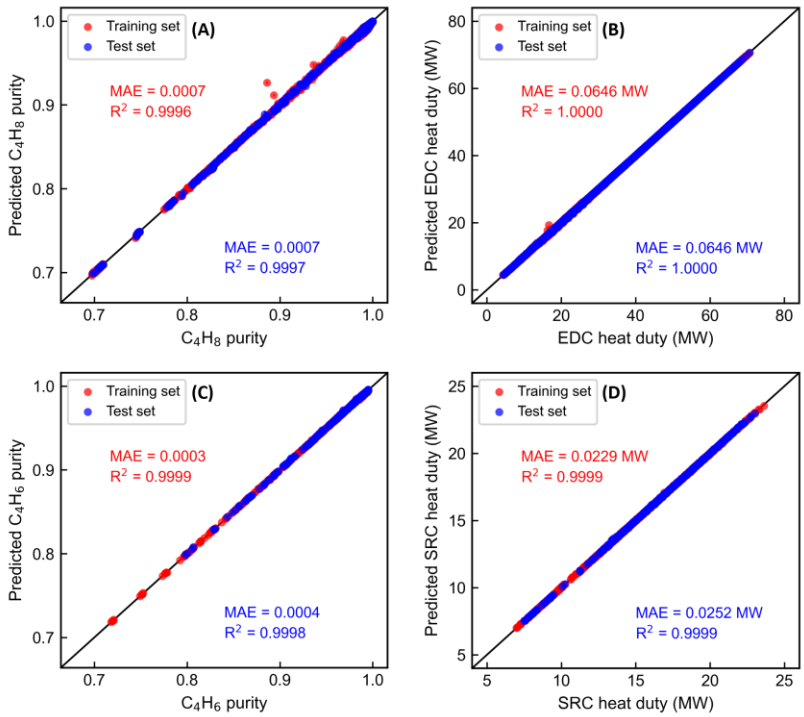


Figure 4-7. Performance of the data-driven models for the NMP-based ED process in predicting (A) x_{C4H8}, (B) x_{C4H6}, (C) Q_{EDC}, and (D) Q_{SRC}.

Based on the data-driven process models, a multi-objective optimization problem is formulated to identify the optimal process parameters for the NMP-based ED process.

$$\min_{\boldsymbol{z}} \left(N_{EDC} + N_{SRC},\ f_{Q_{\mathrm{EDC}}}(\boldsymbol{y},\boldsymbol{z}) + f_{Q_{\mathrm{SRC}}}(\boldsymbol{y},\boldsymbol{z})\right)$$

$$s.t. \quad f_{x_{\mathrm{C4H8}}}(\boldsymbol{z}) \geq 0.9945$$

$$f_{x_{\mathrm{C4H6}}}(\boldsymbol{z}) \geq 0.9945$$

$$\boldsymbol{z}_L \leq \boldsymbol{z} \leq \boldsymbol{z}_U$$

where f is the developed data-driven model, and $\boldsymbol{z}$ is the process parameter space. The constraints on the product purity are slightly relaxed because of the difficulties encountered in identifying feasible solutions.

The multi-objective optimization problem is solved using the NSGA-II[145] implemented in Pymoo[146], where a rounding operator is applied for discrete variables. It converges in 205 iterations with a population size of 100, obtaining a set of Pareto-optimal solutions (i.e., optimal process parameters). **Figure 4-8** depicts the objective function values for the NMP-based ED process under the optimal process parameters.

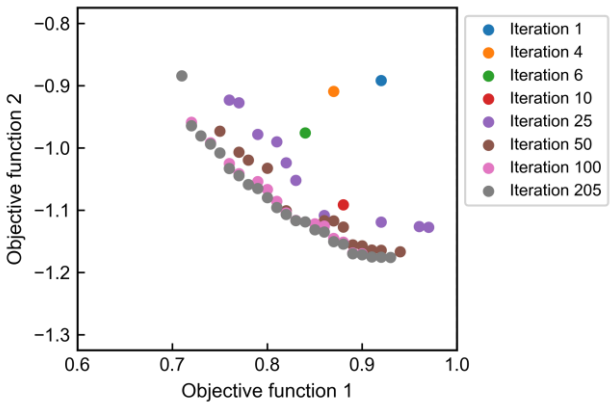


Figure 4-8. Multi-objective optimization results for the NMP-based ED process.

Subsequently, a detailed simulation of the NMP-based ED process is performed under the optimal process parameters obtained from the surrogate-based process optimization. Compared to the process simulation, the developed models accurately estimate the heat duty whereas slightly underestimate the product purity (**Figure 4-9**).

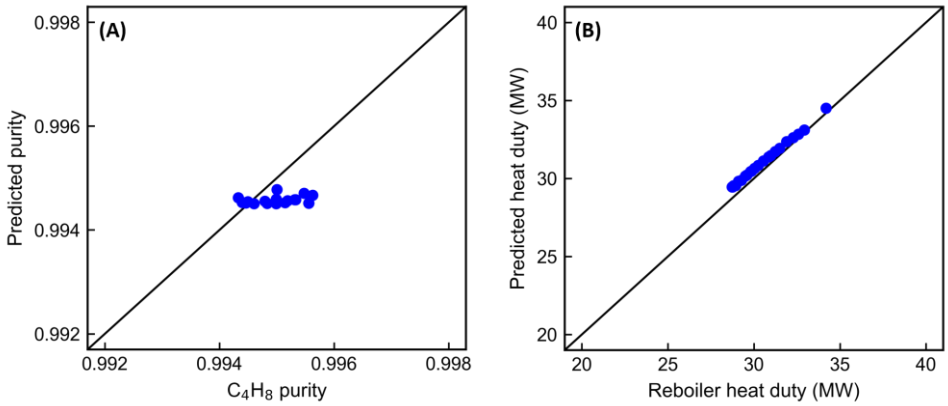


Figure 4-9. Process performance estimated by the data-driven models and evaluated via rigorous simulations for the NMP-based ED process.

The process parameters showing the best process performance in the dataset are regarded as the reference. Compared to the reference process, the optimal ED process obtained from the surrogate-based process optimization reduces the total heat duty by 2.66%. The energy-saving is not significant because the reference process is already approaching the global optimum. The optimal process parameters and corresponding process performance for the reference and optimal ED process are provided in **Table 4-7**. Such an optimal NMP-based ED process obtained from the surrogate-based process optimization is further considered a benchmark for the data-driven CAMPD.

Table 4-7. Optimal process parameters and corresponding process performance.

		Reference process	Optimal process
Process parameter	N_{EDC}	75	80
	R_{EDC}	4.00	5.99
	P_{EDC}	3.50	3.50
	N_{SRC}	12	12
	R_{SRC}	0.60	0.60
	P_{SRC}	3.50	3.50
	S/F	3.00	2.48
Process performance	x_{C4H8}	0.9955	0.9956
	Q_{EDC}	16.66	16.87
	x_{C4H6}	0.9954	0.9955
	Q_{SRC}	12.85	11.85
	Q_H	29.51	28.72

For each of the seven solvents identified by the data-driven CAMPD, its optimal solution is extracted for comparison, as shown in **Figure 4-10**. The region in green indicates that the purity constraint of 0.995 is satisfied. It is observed that two solvent candidates (acetylacetone and acetic anhydride) show lower heat duty than the benchmark solvent NMP while satisfying purity specifications. The other five solvents either show a higher heat duty or are unable to satisfy the purity requirements.

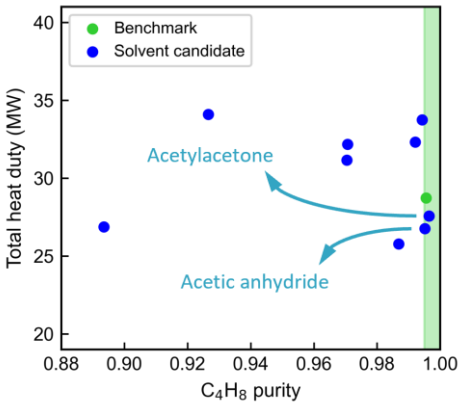


Figure 4-10. Process performance of the identified optimal solvent candidates.

The physical properties of the two solvent candidates, optimal process parameters, and the corresponding process performance are summarized in **Table 4-8**. For better reference, the values for NMP are included as well. The two solvent candidates have similar molecular properties with NMP (except relative volatility and viscosity). It can be found that for the solvent showing a high selectivity, a low solvent-to-feed ratio (*S*/*F*) and a high reflux ratio in the EDC (R_{EDC}) are generally required to satisfy the purity specifications. On average, the two solvents identified from CAMPD can reduce the overall heat duty (Q_H) of the ED process by 5.42% compared to the benchmark.

Table 4-8. Molecular properties of NMP and two candidate solvents, optimal process parameters, and the corresponding process performance.

		Benchmark	Candidate 1	Candidate 2
Solvent	Name	NMP	Acetylacetone	Acetic anhydride
	CAS number	872-50-4	123-54-6	108-24-7
	Molecular formula	C_5H_9NO	$C_5H_8O_2$	$C_4H_6O_3$
Solvent property	S^∞	1.642	1.473	1.701
	$\log_{10}(\alpha^\infty)$	3.664	2.012	2.344
	C_p	161.7	163.2	160.9
	MW	99.13	100.12	102.09
	ρ	1027	969	1074
	μ	1.89	0.76	0.84
Process parameter	N_{EDC}	80	79	69
	R_{EDC}	5.99	9.29	9.87
	P_{EDC}	3.50	3.50	3.50
	N_{SRC}	12	8	10
	R_{SRC}	0.60	1.38	0.95
	P_{SRC}	3.50	3.50	3.50
	S/*F*	2.48	2.13	1.99
Process performance	x_{C4H8}	0.9956	0.9965	0.9952
	Q_{EDC}	16.87	18.27	18.11
	x_{C4H6}	0.9955	0.9963	0.9952
	Q_{SRC}	11.85	9.30	8.65
	Q_H	28.72	27.57	26.76

4.2 Data-driven integrated molecular and process design using Bayesian optimization

Section 4.1 demonstrated that the ED process can be accurately approximated in a data-driven manner, and the data-driven CAMPD approach can efficiently identify the optimal solvents and process operating conditions to reduce energy demand. However, such a data-driven approach involves substantial computational costs for data generation, which is required to develop sufficiently accurate models for the system being investigated. This can be a notable drawback for its applications. To reduce data demand and improve the efficiency of the proposed data-driven CAMPD approach, a Bayesian optimization-based method, called BayesCAMPD, is introduced in this section. In BayesCAMPD, closed-loop optimization is accomplished by integrating data-driven modeling, model-based optimization, and solution validation. This iterative process can reduce the data demand for accurate surrogate modeling, leading to more efficient optimization. In this section, a key improvement in data-driven modeling is that surrogate models are built for the entire ED process, rather than for two separate columns. This is made possible by the data-efficient Bayesian optimization, which allows for accurate modeling with reduced data demand.

BayesCAMPD is performed to simultaneously identify optimal solvent and process parameters to maximize the performance of the ED process. The search space is defined by both molecular and process variables. Molecular variables are molecular properties including selectivity at infinite dilution (S^{∞}), molar heat capacity (C_p), and heat of vaporization (ΔH_{vap}). They are calculated at 25 °C and 1 bar and used to represent molecular variables in surrogate modeling. The property design space (upper and lower bounds) is defined by the corresponding maximum and minimum property values of 1563 solvents derived from the Aspen Plus component database[142]. Process-related variables include the number of stages (N), reflux ratio (R), and operating pressure (P) for both extractive distillation column (EDC) and solvent recovery column (SRC), along with the solvent-to-feed ratio (S/F) for the entire ED process. The design space is kept the same as in **Section 4.1**. In summary, three molecular variables (S^{∞}, C_p, and ΔH_{vap}) and seven process variables (N_{EDC}, R_{EDC}, P_{EDC}, N_{SRC}, R_{SRC}, P_{SRC}, and S/F) are considered. Among all the decision variables, the number of stages (i.e., N_{EDC} and N_{SRC}) is discrete while the others are continuous variables. **Table 4-9** lists the upper and lower bounds for all decision variables considered in the CAMPD.

Table 4-9. Upper and lower bounds of decision variables considered in the CAMPD.

Variable	Symbol	Unit	Lower bound	Upper bound
Selectivity at infinite dilution	S^{∞}	–	0.01	456.12
Molar heat capacity	C_p	J/(mol·K)	19.21	1853.12
Heat of vaporization	ΔH_{vap}	kJ/mol	19.95	258.08
Number of stages in the EDC	N_{EDC}	–	40	80
Reflux ratio of the EDC	R_{EDC}	–	1.00	10.00
Operating pressure of the EDC	P_{EDC}	bar	3.50	6.00
Number of stages in the SRC	N_{SRC}	–	8	20
Reflux ratio of the SRC	R_{SRC}	–	0.20	2.00
Operating pressure of the SRC	P_{SRC}	bar	3.50	6.00
Solvent-to-feed ratio	S/F	–	1.00	8.00

4.2.1 BayesCAMPD workflow

Figure 4-11 illustrates the BayesCAMPD workflow, which mainly involves four phases and is described as follows.

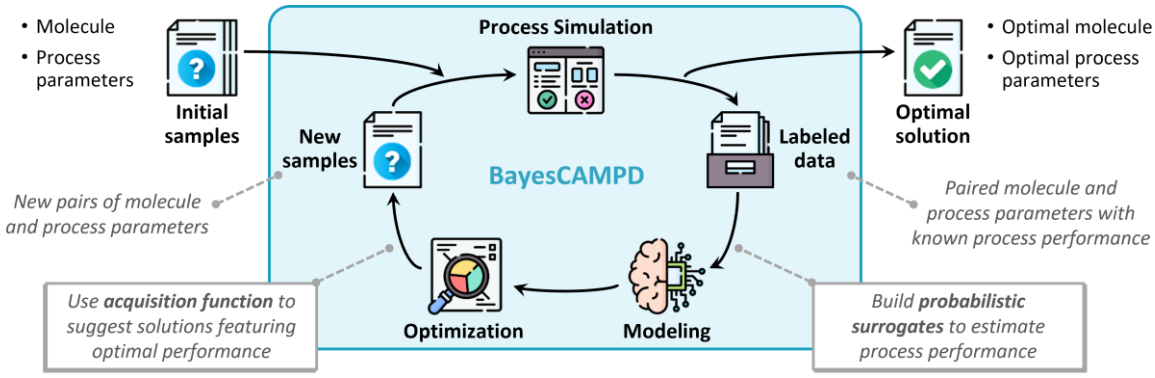


Figure 4-11. Schematic diagram of the BayesCAMPD workflow.

Phase 1: ***Initialization***. BayesCAMPD starts with an initial dataset of labeled data. Process parameters are generated using Latin hypercube sampling (LHS) and solvents are randomly selected from the solvent list, forming pairs of solvents and process parameters as the initial samples. Subsequently, process performance in terms of product purity and reboiler heat duty is obtained as sample labels by rigorous simulation in Aspen Plus using the UNIFAC thermodynamic model. It should be noted that process performance data may not be available for some samples due to convergence issues encountered in process simulation. Under this

situation, additional pairs of solvent and process parameters are generated for process simulation using random sampling, until enough initial labeled samples are collected.

Phase 2: ***Data-driven modeling***. Using the collected data, surrogate models are developed to quickly estimate process performance based on solvent properties and process parameters. Given the capability for uncertainty estimation and efficiency with small datasets, the Gaussian process is used for surrogate modeling. It is implemented with Scikit-learn[149] using a squared exponential kernel as the covariance function. Z-score normalization is applied to the output data. Additionally, prior to surrogate modeling, data transformation is performed to convert purities into real numbers using the logit function. Accordingly, inverse transformation is performed for purity prediction. This guarantees that the predicted product purity always falls into the range between 0 and 1.

$$\alpha = \mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

$$p = \mathrm{logit}^{-1}(\alpha) = \frac{1}{1+e^{-\alpha}}$$

Phase 3: ***Model-based optimization***. In Bayesian optimization, an acquisition function is used to guide the navigation over the search space to identify promising solutions. Expected improvement is a commonly used acquisition function that evaluates the expected amount of improvement in the objective function, which is calculated using mean and standard deviation values estimated by the surrogate model of the objective function. To find the optimal solution, optimization is performed to maximize the expected improvement. Consequently, the optimal solvent properties and process parameters that have the potential to minimize the objective function (e.g., energy demand) while satisfying constraints (e.g., product purity) are identified. Considering the complexity of such a mixed-integer nonlinear optimization problem, it is solved using the differential evolution algorithm implemented in SciPy[150]. As a result, an optimal hypothetical molecule described by solvent properties and the corresponding optimal process parameters are simultaneously determined. In a follow-up step, the obtained hypothetical molecule is mapped onto real solvents by minimizing its Euclidean distance to the hypothetical target molecule in the property space. The above stochastic optimization is performed five times and consequently, five pairs of solvent and process parameters are suggested as promising solutions for further validation. The above-described optimization problem is formulated as follows:

$$\begin{aligned} & \max_{\boldsymbol{y},\boldsymbol{z}} \ EI(\boldsymbol{y},\boldsymbol{z}) \\ s.t. \quad & g_L \leq g(\boldsymbol{y},\boldsymbol{z}) \\ & \boldsymbol{y}_L \leq \boldsymbol{y} \leq \boldsymbol{y}_U \\ & \boldsymbol{z}_L \leq \boldsymbol{z} \leq \boldsymbol{z}_U \end{aligned}$$

where *EI* is the acquisition function (i.e., expected improvement of the CAMPD objective function) and *g* is the surrogate model for purity constraint; $\boldsymbol{y}$ and $\boldsymbol{z}$ denote solvent property and process parameter spaces, respectively; *L* and *U* represent the lower and upper bounds, respectively.

Phase 4: ***Solution validation***. For the suggested candidate solutions, rigorous process simulation is performed to evaluate their performance. Promising solvent and process parameters are identified if the solution is validated to present improved objective function value while satisfying purity constraints. Lastly, the closed-loop BayesCAMPD workflow continues by incorporating these newly labeled samples and updating the model (**Phase 2**).

An early stopping criterion with a patience of 20 iterations is employed. Specifically, the entire workflow terminates if the objective function of suggested solutions shows no improvement in 20 consecutive iterations. It is worth noting that issues may arise over iterations in two circumstances: (i) the optimization in **Phase 3** fails to find a feasible solution satisfying the constraints, and (ii) the validation in **Phase 4** fails due to the convergence issues in process simulation. In both circumstances, BayesCAMPD cannot gain new knowledge about the process. Therefore, extra sampling and labeling are performed to obtain 10 additional labeled samples to augment the dataset. In short, BayesCAMPD continuously proposes and validates promising solvents and process parameters until the entire workflow terminates. Data and code for implementing the BayesCAMPD approach in this section are available in the GitHub repository[151].

Figure 4-12 illustrates the differences between the BayesCAMPD approach and the data-driven CAMPD approach introduced in **Section 4.1** (denoted as OneshotCAMPD). The OneshotCAMPD approach can be considered a conventional data-driven route where CAMPD is performed in a one-shot manner without Bayesian optimization. Surrogate models are usually constructed using a relatively large dataset, followed by model-based optimization to maximize process performance. Process simulations are conducted to validate process performance and consequently, identify the optimal solution. Although the BayesCAMPD

approach adopts the same modeling-optimization-validation strategy, it is conducted in a closed-loop manner. Starting with a relatively small dataset, it continuously enlarges the dataset and updates the model with new samples acquired from sequential optimization and validation. In addition, the surrogate model should be able to estimate prediction uncertainty, which is crucial for considering explore-exploit tradeoffs in searching for optimal solutions. Overall, the OneshotCAMPD has a higher data demand to accurately approximate the system, which poses a huge challenge to high dimensional and complex problems. In comparison, the BayesCAMPD has a significantly lower data demand, whereas increased computational resources are necessary for repeated modeling and optimization.

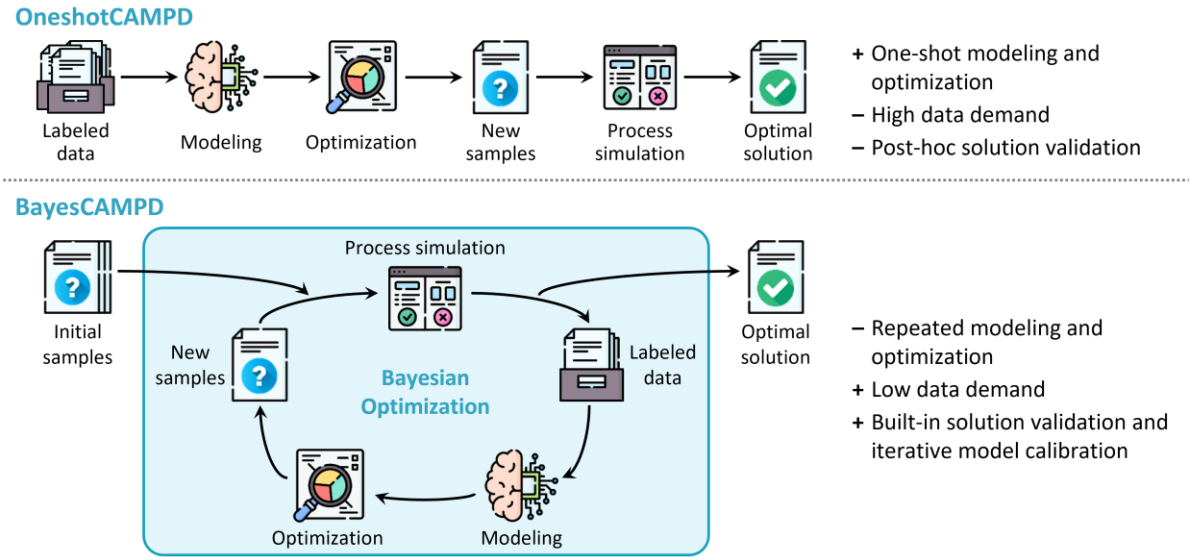


Figure 4-12. Comparison between the OneshotCAMPD and BayesCAMPD workflows.

4.2.2 BayesCAMPD performance

For simplicity, the reboiler heat duty (Q_H) of the entire ED process is used to describe the energy demand. Taking different sizes of initial samples into account, the BayesCAMPD is performed to minimize the objective function of Q_H. As different initial datasets are collected, BayesCAMPD has different starting points and consequently, the optimization can vary significantly. The optimal solvent and process parameters featuring minimized Q_H are identified when the BayesCAMPD workflow terminates. As shown in **Figure 4-13A**, six out of eight cases achieve improved performance. "Start" and "End" respectively represent the best observation in the initial dataset and the optimal solution obtained. For some cases, "End" points are absent, indicating that the BayesCAMPD fails to identify feasible solutions. Compared to their corresponding initial best observation, these six cases decrease Q_H by 10.11 MW on average.

Furthermore, the computational costs associated with labeling, modeling, and optimization for each case are displayed in **Figure 4-13B**. Initial labeling corresponds to the process simulation performed to collect the initial labeled dataset, while extra labeling refers to the process simulation performed for extra sampling and solution validation. Modeling cost is directly related to the model development using Gaussian processes, and optimization cost accounts for the identification of optimal solutions based on established surrogate models. The cost of initial labeling increases with the size of initial samples, while the cost of modeling and optimization depends on the number of iterations executed in the BayesCAMPD. This explains the less computational time for the cases with 256 and 896 initial samples, which only undergo 20 iterations. Due to the small dataset and high modeling efficiency, the labeling and modeling costs are insignificant compared to the optimization cost.

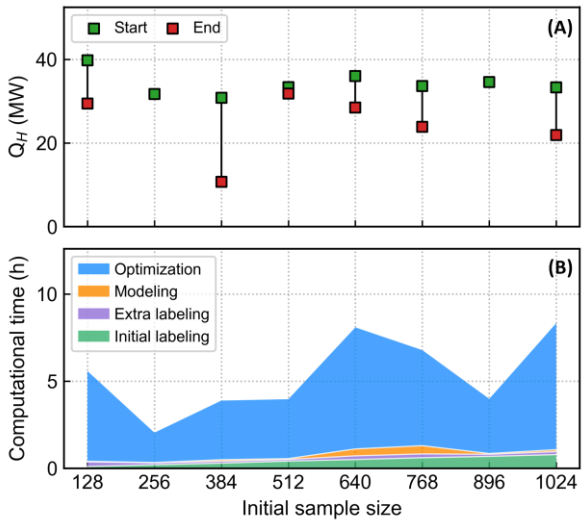


Figure 4-13. Performance of BayesCAMPD starting with different sizes of initial samples: (A) process performance represented by Q_H and (B) computational costs associated with labeling, modeling, and optimization.

The CAMPD problem with constraints on product purity has a high complexity, which poses a substantial challenge for BayesCAMPD to obtain practically feasible solutions. First, model-based optimization should be able to obtain feasible solutions satisfying the constraints. Different numbers of iterations executed in BayesCAMPD lead to a varying number of optimization attempts (1 iteration includes 5 optimization attempts, orange circles in **Figure 4-14A**). For the eight cases considered, an optimization success rate of 42.9% is achieved on average (blue bars in **Figure 4-14B**), which demonstrates the complexity of such a constrained optimization task. Second, the suggested solution should be able to be validated by process

simulation. The number of converged simulations is presented in **Figure 4-14A** with green squares, and a simulation success rate of 53.0% is achieved on average (green bars in **Figure 4-14B**). Overall, the above low success rates indicate difficulties in finding feasible solutions. Lastly, the validated feasible solution can be considered as an optimal solution provided that it presents a lower objective function value while satisfying product purity specifications. All three aspects demonstrate the challenge of BayesCAMPD, as failures can take place at every step. This is evident in the two cases starting with 256 and 896 initial samples, where BayesCAMPD fails to obtain promising solutions within 20 iterations (**Figure 4-13**).

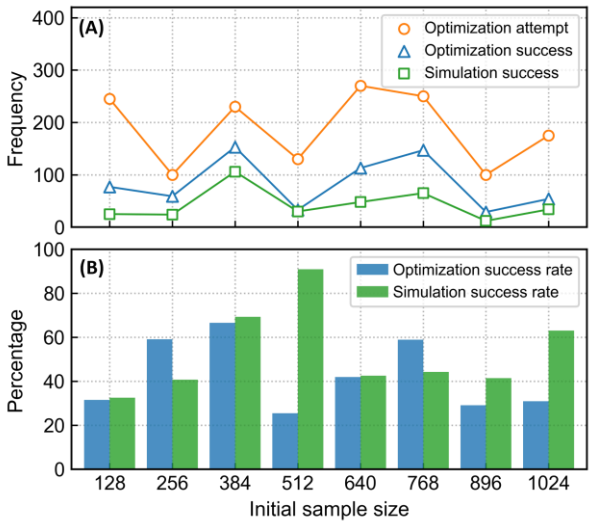


Figure 4-14. Analysis of optimizations and simulations within the BayesCAMPD starting with different sizes of initial samples: (A) number of total optimization attempts, optimization successes, and simulation successes, and (B) success rates of optimizations and simulations.

Therefore, BayesCAMPD is rerun for the two cases with a higher patience of stopping criterion, increased from 20 to 30 iterations. As it turns out, both cases successfully identify feasible solutions, leading to an average decrease of 5.34 MW in Q_H compared to their respective initial best observations (**Figure 4-15**). However, this improvement comes with an increase in computational costs. This supplementary investigation underscores that, by allowing for a higher patience in the stopping criterion, feasible solutions with improved performance can be obtained at the expense of increased computational costs.

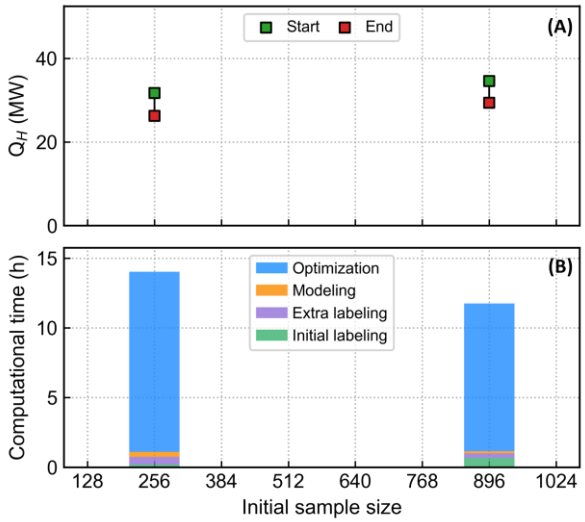


Figure 4-15. Performance of the BayesCAMPD with a patience of 30 in the stopping criterion: (A) process performance represented by Q_H and (B) computational costs.

The results of the six successful BayesCAMPD executions are listed in **Table 4-10** with their validated process performance. Chemical formula, CAS number, and molecular properties of each solvent are provided in **Table 4-11**.

Table 4-10. Optimal solvents and process parameters identified by BayesCAMPD.

Initial sample size	Optimal solvent	Optimal process parameter							Q_H (MW)
		N_{EDC}	R_{EDC}	P_{EDC} (bar)	N_{SRC}	R_{SRC}	P_{SRC} (bar)	S/F	
128	o-Nitroanisole	50	10.00	4.15	17	0.40	5.62	1.86	29.45
384	Ethylene glycol	73	1.02	4.65	11	0.53	3.53	1.00	10.74
512	Glycolaldehyde	79	9.08	6.00	8	1.79	4.34	3.41	31.86
640	Glycolaldehyde	73	7.17	6.00	16	0.20	3.50	3.90	28.51
768	Ethylene glycol	71	10.00	6.00	20	1.54	4.46	1.00	23.87
1024	Acetic anhydride	70	3.19	3.58	13	1.87	3.53	2.68	21.92

Among the six successful cases, the one starting with 384 initial samples obtains the best process performance, where ethylene glycol is identified as the optimal solvent. For this case, details of the BayesCAMPD performance are further illustrated. The BayesCAMPD executes 46 iterations, with several improvements in process performance observed at the 1st, 3rd, 4th, 5th, 15th, 18th, and 26th iterations. The best solution is found at the 26th iteration, decreasing Q_H down to 10.74 MW (**Figure 4-16A**). Feasible solutions refer to candidate solutions suggested

by optimization that are successfully validated via process simulation and satisfy purity constraints, while infeasible solutions are those violating the purity constraints. The green step line depicts the changes in Q_H of the best observations. The validated product purity of all the candidate solutions is presented in **Figure 4-16C-D**.

Table 4-11. Information on the optimal solvents identified by BayesCAMPD.

Solvent	Chemical formula	CAS number	S^{∞}	C_p (J/(mol·K))	ΔH_{vap} (kJ/mol)
o-Nitroanisole	$C_7H_7NO_3$	91-23-6	2.779	120.7	73.67
Ethylene glycol	$C_2H_6O_2$	107-21-1	6.896	148.4	67.22
Glycolaldehyde	$C_2H_4O_2$	141-46-8	1.511	159.3	62.85
Acetic anhydride	$C_4H_6O_3$	108-24-7	1.701	160.9	47.07

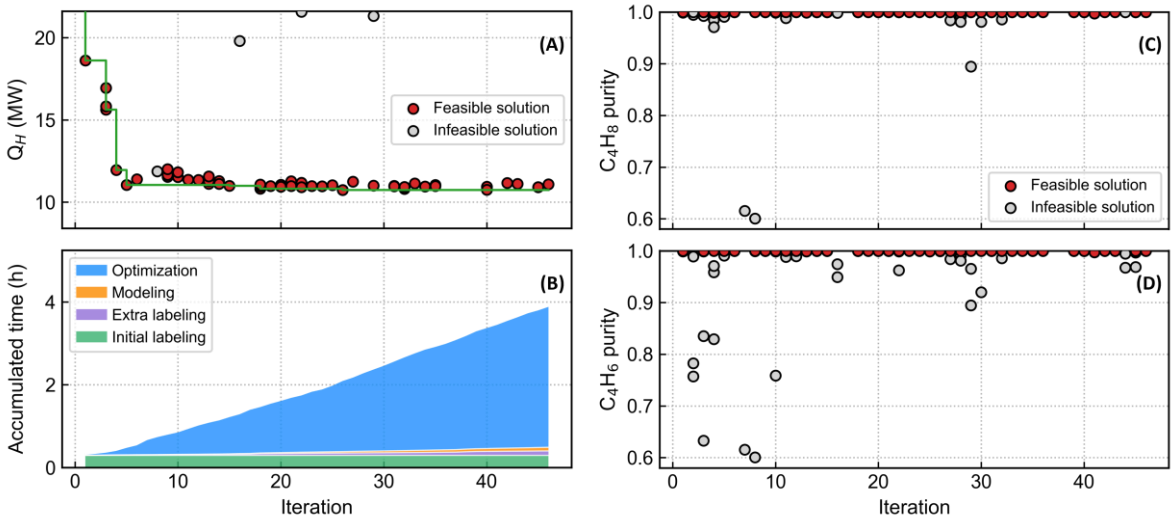


Figure 4-16. Performance of BayesCAMPD starting with 384 initial samples: (A) process performance of candidate solutions represented by Q_H, (B) accumulated computational costs, (C) C_4H_8 purities of solutions, and (D) C_4H_6 purities of solutions.

As shown in **Figure 4-16A**, although lots of feasible solutions are identified, most of them present slightly higher Q_H values than their current best observations. After executing for 20 iterations without any further improvement (from iteration 27 to iteration 46), the stopping criterion is satisfied and BayesCAMPD terminates at the 46th iteration. As shown in **Figure 4-16B**, the case starting with 384 initial samples takes approximately 4 hours, with the optimization cost accounting for the majority. The cost for data-driven modeling is negligible due to the high efficiency of Gaussian processes on small datasets.

As ethylene glycol has been identified as the optimal solvent, further process optimization is performed to refine its optimal process parameters. For this purpose, BayesCAMPD has a smaller search space as molecular decision variables are constant. The optimal process parameters are determined by the BayesCAMPD approach with the identical modeling-optimization-validation procedure.

Taking different sizes of initial samples into account, the BayesCAMPD is performed to minimize the objective function of Q_H. Excluding the case starting with 1024 initial samples, the remaining ones fail to identify feasible solutions presenting improved process performance while satisfying purity constraints (**Figure 4-17**).

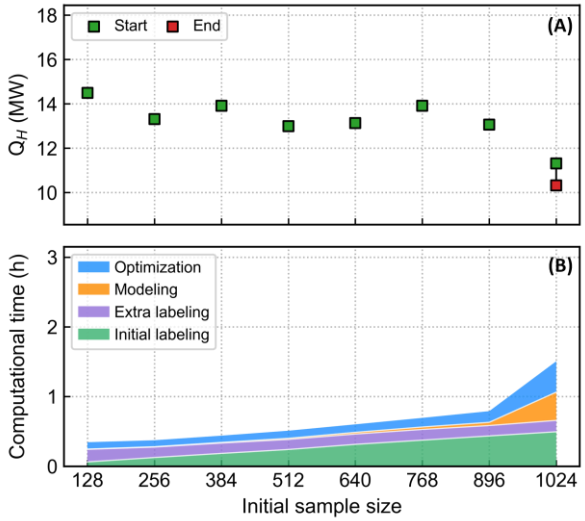


Figure 4-17. Performance of BayesCAMPD for process optimization starting with different sizes of initial samples: (A) process performance represented by Q_H and (B) computational costs.

For the case using 1024 initial samples, it obtains the optimal solution at the 4th iteration and terminates at the 24th iteration (**Figure 4-18**). Although feasible solutions satisfying purity constraints are obtained at the 5th and 17th iterations, they fail to achieve an improved process performance. As a result, BayesCAMPD successfully determines the optimal process parameters and reduces the heat duty further to 10.33 MW, representing a further decrease of 3.8% on the objective function. It runs for 24 iterations with approximately 1.5 hours. Process parameters and corresponding process performance of both solutions identified by CAMPD and process optimization are provided in **Table 4-12**.

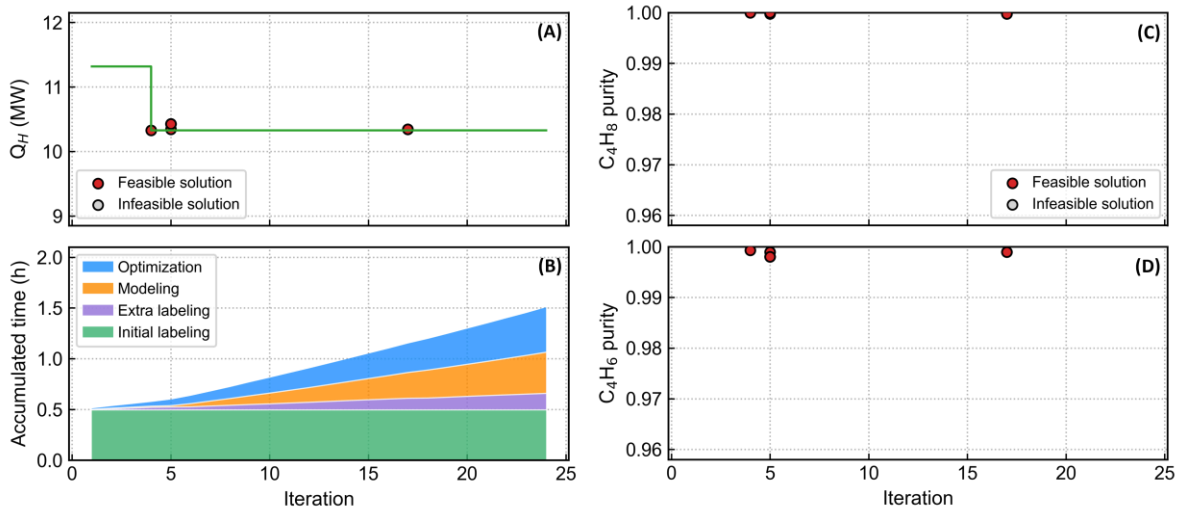


Figure 4-18. (A) Process performance of candidate solutions represented by Q_H, (B) accumulated computational costs of the BayesCAMPD, (C) C_4H_8 purity of solutions, and (D) C_4H_6 purity of solutions.

Table 4-12. Optimal process parameters determined by CAMPD and process optimization.

Method	Optimal process parameter							Process performance		
	N_{EDC}	R_{EDC}	P_{EDC}	N_{SRC}	R_{SRC}	P_{SRC}	S/F	Q_H (MW)	x_{C4H8}	x_{C4H6}
CAMPD	73	1.02	4.65	11	0.53	3.53	1.00	10.74	1.0000	0.9998
Process optimization	62	1.00	5.38	11	0.20	3.50	1.00	10.33	1.0000	0.9993

4.2.3 Comparison and analysis

For comparison, the CAMPD is performed using the OneshotCAMPD approach. To ensure consistency, the Gaussian process is used for data-driven modeling, as implemented in the BayesCAMPD approach. Based on the established surrogate models for the entire ED process, optimization is performed to minimize the objective function, and promising solutions are suggested for post-hoc validation.

Due to the one-shot characteristics of modeling and optimization, the OneshotCAMPD approach typically requires far more amount of initial data to accurately approximate the system across the full search space. In addition to the eight sample sizes investigated for BayesCAMPD (i.e., ranging from 128 to 1024), six larger sizes ranging from 1536 to 4096 are also considered. To increase the probability of identifying practically feasible solutions, 20 stochastic optimization attempts are performed in OneshotCAMPD. As a result, feasible

solutions are obtained in 8 out of the in total 14 cases, presenting a decreased Q_H value of 1.48 MW on average (**Figure 4-19A**). Therefore, the OneshotCAMPD approach can improve the process performance in an open-loop manner, which has also been demonstrated in **Section 4.1**. However, the improvement is very limited compared to that achieved by the BayesCAMPD. As OneshotCAMPD exploits, rather than explores, the search space, the optimization prefers a local search around the best observation to obtain better solutions. Therefore, a good starting point (i.e., a good sample in the initial dataset) is crucial to obtain promising solutions using the OneshotCAMPD approach.

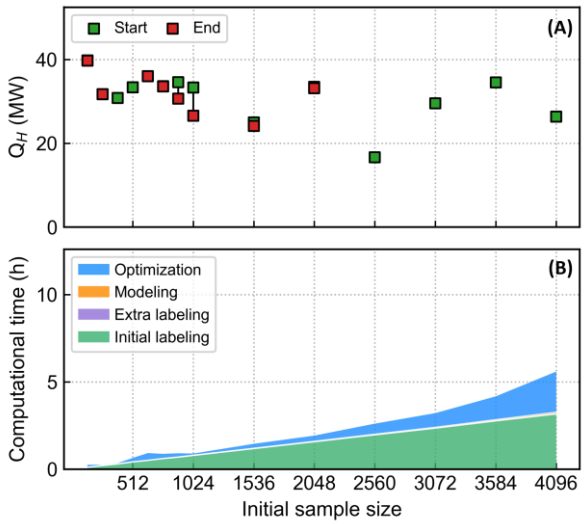


Figure 4-19. Performance of OneshotCAMPD starting with different sizes of initial samples: (A) process performance represented by Q_H and (B) computational costs associated with labeling, modeling, and optimization.

In general, the computational cost of OneshotCAMPD increases with the size of initial samples (**Figure 4-19B**), whereas it is lower than that of the BayesCAMPD due to the avoidance of repeated modeling and optimization. Nevertheless, the BayesCAMPD approach can achieve significantly better performance at an affordable computational cost, offering a computationally efficient solution for data-driven CAMPD tasks.

PART II. MATERIALS DISCOVERY

5 Accelerated Screening of Metal-Organic Frameworks for Pressure Swing Adsorption

In this chapter, data-driven approaches are introduced to expedite the screening of metal-organic frameworks (MOFs) for gas separation applications.

In **Section 5.1**, an end-to-end machine learning (ML) method integrating feature learning is proposed to predict adsorption capacity of MOFs. By combining feature embedding and molecular graph convolution, this approach learns both chemical and geometric features from MOF building blocks, which are subsequently used to correlate MOF adsorption capacity. Such ML models can accurately and efficiently estimate the adsorption capacity of MOFs from their structures, accelerating the discovery of MOFs with high selectivity for gas separation.

In **Section 5.2**, an interpretable ML method that combines feature engineering with straightforward tree-structure models is introduced to predict the adsorption preference of MOFs. Using different feature engineering methods, numerical descriptors or fingerprints that characterize MOF structures are calculated and subsequently used to correlate MOF's adsorption preference. Such ML models can provide interpretable and easy-to-understand insights into the model's decision-making, thereby facilitating the discovery of MOFs with specific adsorption preferences for gas separation.

Both approaches for MOF screening are illustrated by the separation of ethylene/ethane (C_2H_4/C_2H_6), as introduced in **Appendix B**.

5.1 MOF screening using end-to-end ML models

5.1.1 Computational details

MOF dataset

The hypothetical MOF (hMOF) database consists of 137,953 MOF structures, constructed from a library of 102 building blocks derived from known MOF structures.[152] To make this database machine-readable, several actions are undertaken to extract building blocks, identify topologies, and analyzing chemical information of the MOF structures from their crystallographic information files (CIFs). First, the MOF structures are translated into the

MOFid and MOFkey identifiers using the algorithm developed by Bucior et al.[153] Next, the building blocks and underlying topological networks of the MOF structures are extracted from these identifiers. Finally, data cleaning is performed to refine the MOF database. The data cleaning steps are as follows:

(1) Remove MOFs sharing duplicate MOFkey to ensure uniqueness, resulting in 45,254 remaining MOFs.
(2) Remove MOFs with incomplete MOFkey to ensure that both the chemical information and topology have been successfully identified, resulting in 39,676 remaining MOFs.
(3) Remove MOFs with invalid organic linker molecules, resulting in 33,480 MOFs.
(4) Retain only MOFs consisting of at most two types of organic linkers to reduce structural complexity, resulting in 9156 MOFs.

These 9156 MOFs, with identified chemical and topological information, are subsequently employed for molecular simulation and model development.

Molecular Simulation

Data is crucial to discover relationships between material structures and their process-relevant properties. While experimental data is always preferred for predictive modeling, it is often scattered and scarce. In this context, the grand canonical Monte Carlo (GCMC) simulation is recognized as a powerful tool for MOF discovery, due to its high efficiency in simulating the adsorption capacity of MOFs with satisfying accuracy.[154,155]

To evaluate the adsorption capacity of MOFs, GCMC simulations are performed using RASPA[156]. Each simulation includes 5000 equilibration cycles, followed by 20,000 production cycles. Interactions between non-bonded atoms are modeled by the Lennard-Jones (LJ) potential[157] with a cutoff distance of 12 Å. The LJ parameters for the relevant MOF atoms are taken from the DREIDING[158] and Universal[159] force fields. The LJ parameters between atoms of different types are calculated using the Lorentz-Berthelot mixing rule. The number of unit cells is adjusted so that each dimension of the simulation cell is at least twice the cutoff distance. Ethane and ethylene molecules are modeled using the united atom model of the Transferable Potentials for Phase Equilibria (TraPPE) force field[160], where the two-site LJ potential[161,162] describes both molecules. For several known MOFs, a good agreement between GCMC simulation results and experimental measurements from the literature[35,163-166] is observed (**Figure 5-1**), confirming the reliability of the configurations used in the GCMC simulations.

In addition, MOF geometric properties, such as void fraction and surface area, are computed using RASPA[156], while pore diameters are calculated with Zeo++[167].

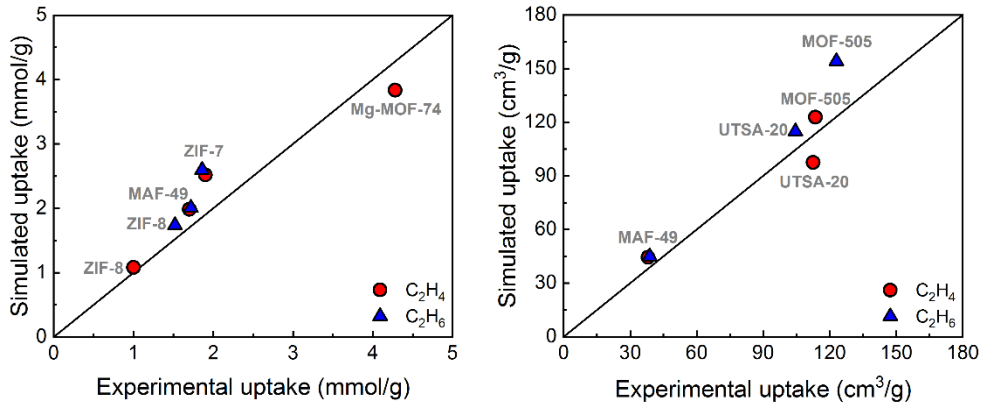


Figure 5-1. Comparison between experimental and GCMC simulated single-component C_2H_4 and C_2H_6 uptakes at 1 bar (296 K for MOF-505[163] and UTSA-20[163], 298 K for Mg-MOF-74[164] and ZIF-7[35], 303 K for ZIF-8[165], and 316 K for MAF-49[166]).

With the adsorption capacity obtained from GCMC simulations, performance metrics such as deliverable capacity and selectivity can be calculated. The deliverable capacity ΔN is defined as the difference between the uptakes at adsorption and desorption conditions, and the selectivity S is a key metric indicating the efficacy of an adsorbent for gas separation.

$$\Delta N_i = N_{i,ads} - N_{i,des}$$

$$S_{i/j} = \frac{N_i}{N_j} / \frac{y_i}{y_j}$$

where i and j are indexes of gas species, N_i is the uptake of the gas species i, and y_i is the mole fraction of gas i in the bulk phase.

Neural Network Architecture

An integrated neural network architecture is proposed to extract both chemical and geometric information of MOF structures and to estimate adsorption capacity, as illustrated in **Figure 5-2**. After decomposing MOF structures into metal nodes and organic linkers, chemical features are extracted by feature embedding and molecular graph convolution, respectively. Meanwhile, geometric features including embedded topology information and five key geometric properties (i.e., void fraction, pore limiting diameter, largest cavity diameter, and volumetric and gravimetric surface areas) are captured. Finally, these chemical and geometric features are combined to predict adsorption capacity.

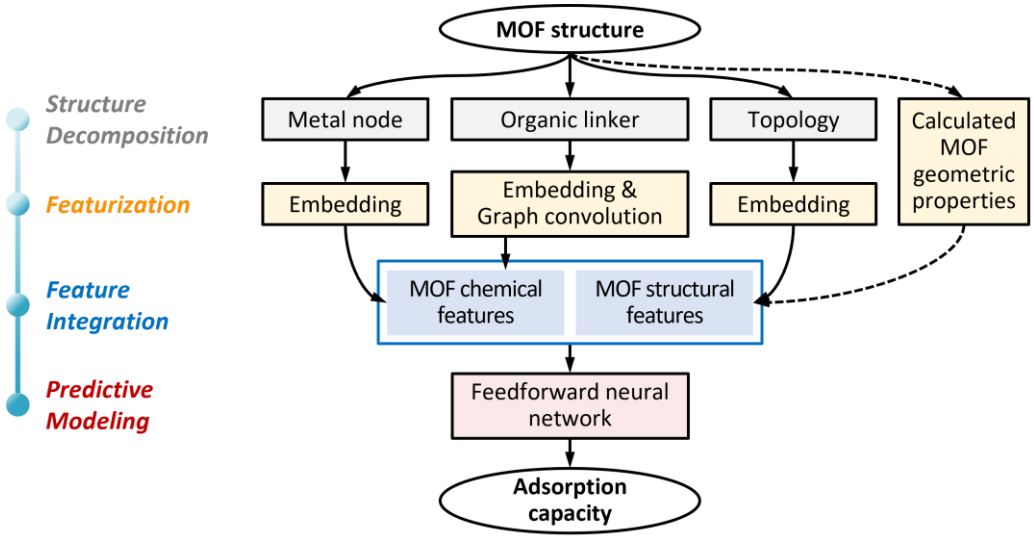


Figure 5-2. Schematic diagram of the proposed ML framework.

In the featurization stage, metal nodes and topologies are represented by chemical formulas and topology identifiers (e.g., "[Zn][Zn]" for the Zn paddlewheel metal node and "pcu" for the primitive cubic lattice topology) and encoded into real-valued vectors by word embedding.

Organic linkers exhibit significant structural diversity with the number of heavy atoms ranging from 4 to 102. Their chemical information is captured by representing each linker as a molecular graph where vertices and edges correspond to atoms and chemical bonds, respectively. For MOFs containing different organic linkers, the linkers in each MOF are represented as a graph composed of multiple unconnected subgraphs. The features of each atom in the organic linker are initially encoded using word embedding and then updated based on neighboring node features through graph convolution. After three layers of molecular graph convolution, the overall feature of the organic linkers is obtained from the features of all nodes using global pooling.

Finally, all chemical (metal and organic linker) and geometric (topology and geometric property) features are concatenated to form the input for a feedforward neural network (FNN) with three hidden layers, which are used to predict the adsorption capacity. Notably, all feature embeddings, graph convolutions, and the FNN are optimized as a whole to minimize the prediction error. **Figure 5-3** provides an example to illustrate the decomposition of a MOF structure and the subsequent prediction of adsorption capacity.

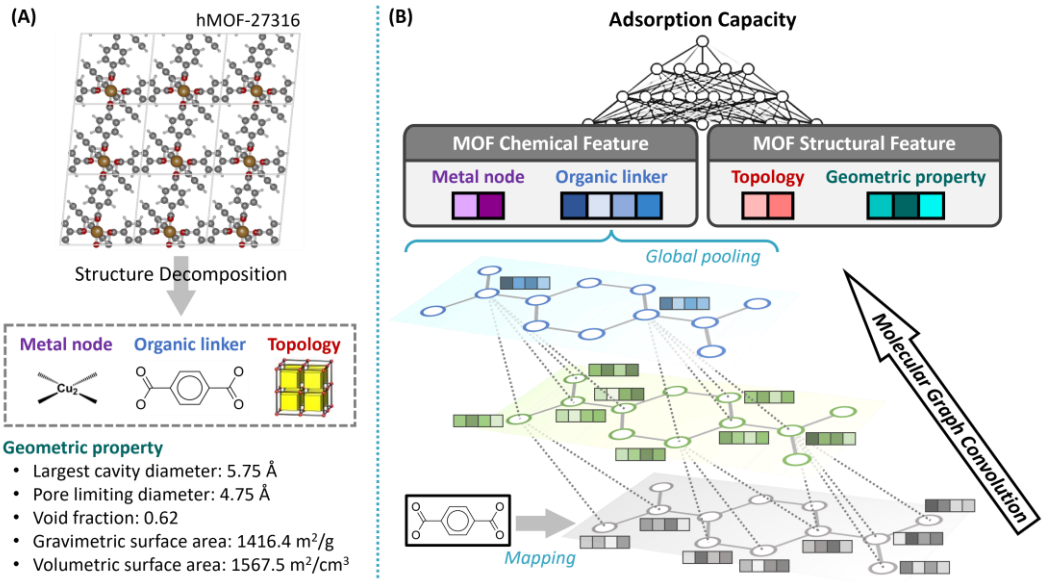


Figure 5-3. Predicting adsorption capacity from the MOF structure: (A) structure decomposition, and (B) featurization, feature integration, and prediction from the bottom to the top.

The proposed ML architecture is built using PyTorch[168] and PyG (PyTorch Geometric)[169]. The training, validation, and test sets account for 80%, 10%, and 10% of the employed dataset (corresponding to 7326, 915, and 915 MOFs). These three sets are used for model training, hyperparameter optimization, and model evaluation, respectively. Mean squared error (MSE) is used as the loss function. Hyperparameters subject to optimization are listed in **Table 5-1**. To prevent overfitting, an early stopping strategy with a patience of 10 epochs is employed. This means that if the model performance on the validation set does not improve for 10 consecutive epochs, the training process is terminated and the model with the lowest validation loss is the optimal model.

Table 5-1. Hyperparameters considered for the ML model.

Hyperparameter	Options
The number of neurons in each hidden layer	8, 16, 24, 32
Activation function	Tanh, ELU, ReLU, Sigmoid, Softplus
Batch size	64, 128, 256
Graph convolution method[169]	GINConv, GCNConv, AGNNConv, ClusterGCNConv, GATConv, GraphConv, LEConv, MFConv, SAGEConv

Data and code for implementing the end-to-end ML model in this section are available in the GitHub repository[170].

5.1.2 Analysis of structure-property relationships

The adsorptive separation of a typical cracked gas mixture (C_2H_4/C_2H_6, 15:1)[166,171] is studied by GCMC simulations at 1 bar and 298 K. **Figure 5-4** shows the relationships between the MOF geometric properties and C_2H_4 uptakes at 1 bar and 298 K. The maximal C_2H_4 uptakes occur at void fractions of 0.6–0.8, pore limiting diameters of 4–8 Å, volumetric surface areas of 1500–2500 m^2/cm^3, and gravimetric surface areas of 2000–3500 m^2/g. Similar trends are observed for C_2H_6 uptakes at the same adsorption conditions, as shown in **Figure 5-5**.

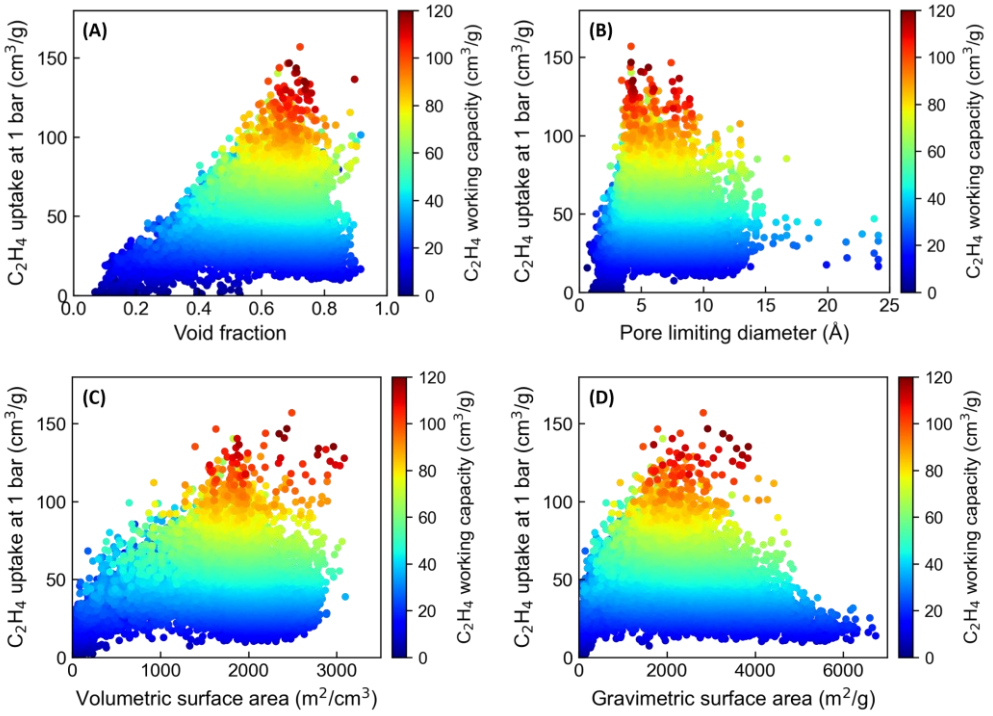


Figure 5-4. Relationships between geometric properties and C_2H_4 uptakes at 1 bar/298 K: (A) void fraction, (B) pore limiting diameter, C) volumetric surface area, and (D) gravimetric surface area.

To evaluate the capability of MOFs in separating C_2H_4 and C_2H_6, the C_2H_6/C_2H_4 selectivity is calculated considering both the uptake gap and the gas composition difference. The relationships between the C_2H_6/C_2H_4 selectivity and MOF geometric properties are visualized in **Figure 5-6**. The top five MOFs with the highest selectivity (larger than 3.5) have a void fraction of 0.41, a pore limiting diameter of 3.33 Å, a volumetric surface area of 639 m^2/cm^3, and a gravimetric surface area of 361 m^2/g on average. High separation selectivity can be achieved by MOFs with relatively low pore limiting diameters and surface areas. However, the opposite is not always true. Considering the implicit relationships between the separation capacity and geometric properties of MOFs, quantitative models are highly desirable to predict the adsorption uptakes and further calculate the selectivity from MOF structures.

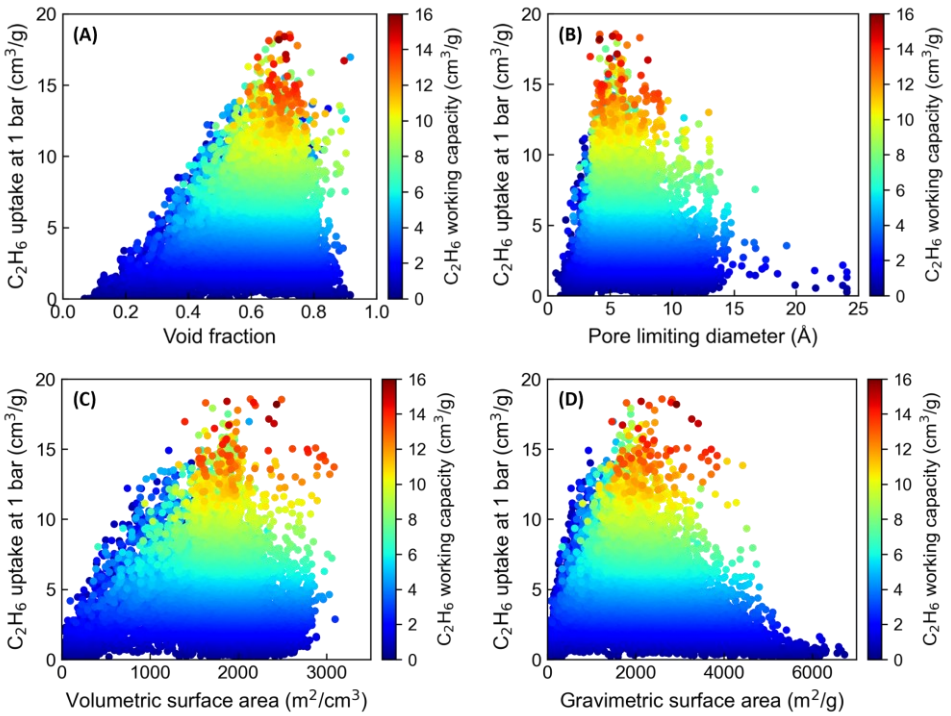


Figure 5-5. Relationships between geometric properties and C_2H_6 uptakes at 1 bar/298 K: (A) void fraction, (B) pore limiting diameter, (C) volumetric surface area, and (D) gravimetric surface area.

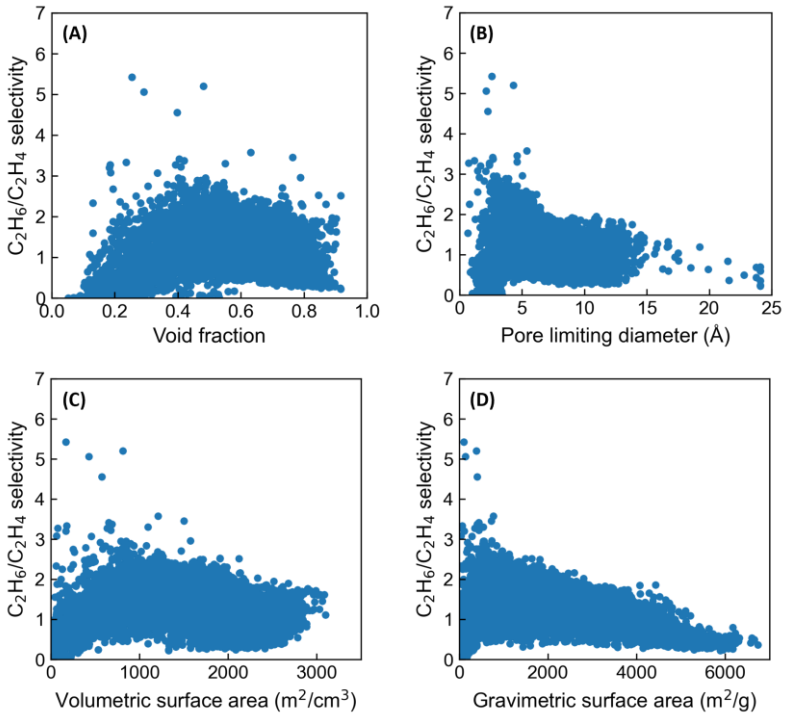


Figure 5-6. Relationships between MOF geometric properties and C_2H_6/C_2H_4 selectivity at 1 bar/298 K: (A) void fraction, (B) pore limiting diameter, (C) volumetric surface area, and (D) gravimetric surface area.

5.1.3 Model development

Employing the framework in **Figure 5-2**, two ML models are trained to predict the C_2H_6 and C_2H_4 equilibrium uptakes, from which the selectivity is determined. Considering all hyperparameter combinations, the optimal ML configurations (listed in **Table 5-2**) are determined by the grid search method. The model performance is evaluated with the mean absolute error (MAE) and coefficient of determination (R^2), as shown in **Table 5-3**. Adsorption uptakes predicted by the ML models are compared with simulation results, as visualized in **Figure 5-7**. In general, ML models achieve satisfying predictions, with an MAE of 5.79 cm^3/g and 0.77 cm^3/g on the test set for C_2H_4 and C_2H_6, respectively. Additionally, the two ML models show MAE values of 6.03 cm^3/g and 0.80 cm^3/g on the validation set.

Table 5-2. Optimal hyperparameter settings.

Model type	Target	Optimal hyperparameters
w/ chemical features	C_2H_4 uptake at 1 bar/298 K	16, Tanh, 256, GINConv
	C_2H_6 uptake at 1 bar/298 K	16, ELU, 256, GINConv
w/o chemical features	C_2H_4 uptake at 1 bar/298 K	32, Softplus, 64, -
	C_2H_6 uptake at 1 bar/298 K	24, Sigmoid, 128, -

Table 5-3. Model performance in predicting C_2H_4 and C_2H_6 uptakes.

Target	Model type	Dataset	MAE (cm^3/g)	R^2
C_2H_4 uptake at 1 bar/298 K	w/ chemical features	Training	5.00	0.9259
		Validation	6.03	0.8871
		Test	5.79	0.8955
	w/o chemical features	Training	8.63	0.7762
		Validation	9.33	0.7389
		Test	9.00	0.7423
C_2H_6 uptake at 1 bar/298 K	w/ chemical features	Training	0.62	0.9419
		Validation	0.80	0.8994
		Test	0.77	0.8965
	w/o chemical features	Training	1.12	0.7964
		Validation	1.21	0.7683
		Test	1.17	0.7721

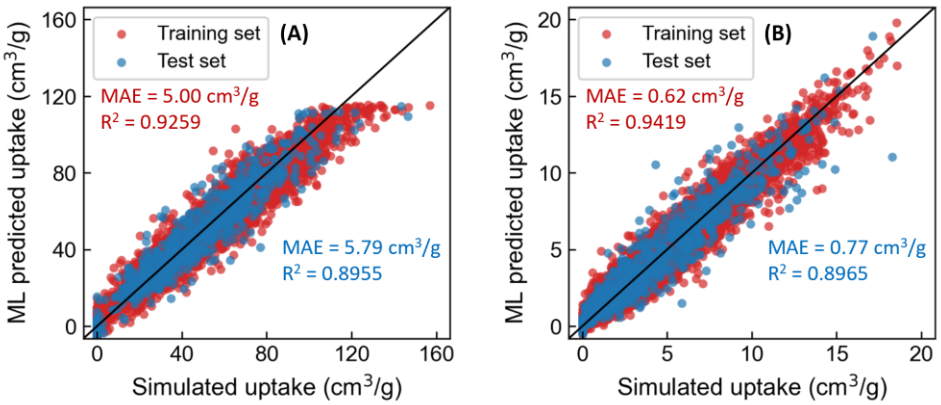


Figure 5-7. Performance of ML models in predicting (A) C_2H_4 and (B) C_2H_6 uptakes at 1 bar/298 K.

The proposed ML method considers both chemical and geometric information of MOFs. To demonstrate the importance of MOF chemical features in the prediction of C_2H_4 and C_2H_6 adsorption uptakes, two new ML models are trained using geometric features only. The corresponding optimal hyperparameters and model performance are presented in **Table 5-2** and **Table 5-3**, respectively. When using only the geometric features as inputs, the MAE of the model on the identical test set is 9.00 cm^3/g and 1.17 cm^3/g for C_2H_4 and C_2H_6, respectively. The removal of chemical features leads to a significant decrease in model performance, which can also be observed in **Figure 5-8**. This indicates that the incorporation of MOF chemical information significantly improves prediction accuracy, proving the significance of chemical features in the discovery of MOFs for C_2H_4/C_2H_6 separation.

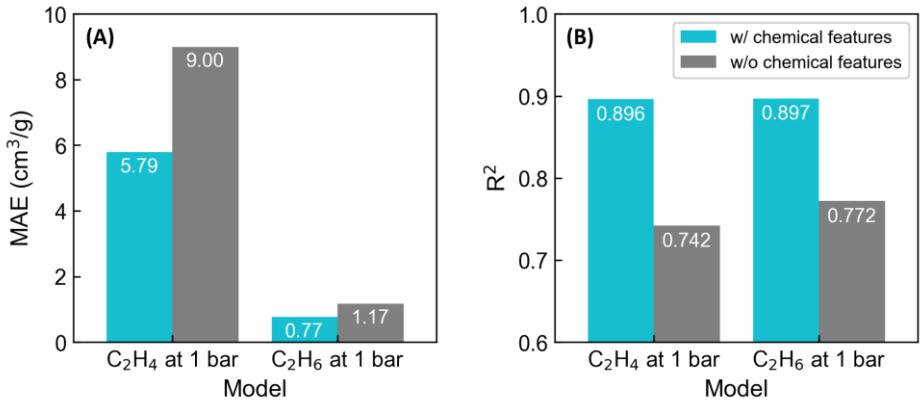


Figure 5-8. Performance of ML models on the test set in predicting C_2H_4 and C_2H_6 uptakes.

5.1.4 MOF screening

To demonstrate the application of ML models for identifying optimal MOFs, a large dataset comprising 21,384 new MOF structures is extracted from the hMOF database using the data

cleaning process described in **Subsection 5.1.1**. These MOF structures, which contain three distinct organic linkers, are more complex than the ones used for model development, which contain no more than two types of organic linkers. Subsequently, ML-assisted large-scale screening is conducted by employing the developed ML models to predict C_2H_6 and C_2H_4 uptakes. The top 100 MOFs with the highest C_2H_6/C_2H_4 selectivity are identified, and GCMC simulations are performed to validate their practical performance (**Figure 5-9**). Although the ML models tend to overestimate selectivity, a C_2H_6/C_2H_4 selectivity of 5.52 is confirmed by GCMC simulations, which is higher than the maximum selectivity of 5.06 in the training dataset. **Figure 5-10** summarizes the ID numbers, metal nodes, and organic linkers of the top three MOFs with the highest GCMC-derived selectivity ranging from 4.94 to 5.52.

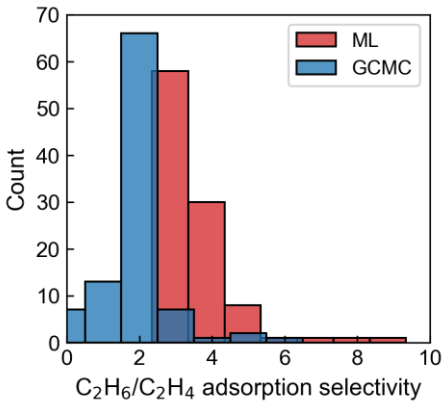


Figure 5-9. Comparison between ML predictions and GCMC simulations for the top 100 MOFs.

Importantly, the GCMC simulation for the top 100 MOFs requires over 140 hours, whereas the ML-assisted screening of the 21,384 MOFs is completed within 2 minutes, demonstrating the high efficiency of ML methods in accelerating MOF discovery. Overall, the integrated ML models are accurate and efficient for discovering highly selective MOFs for the separation of C_2H_6 and C_2H_4.

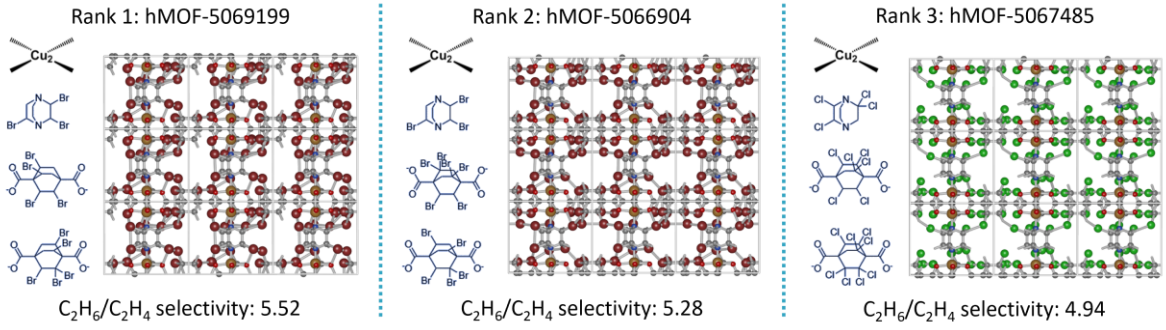


Figure 5-10. Top MOF candidates identified for the C_2H_4/C_2H_6 separation.

5.2 MOF screening using interpretable ML models

5.2.1 Computational details

Figure 5-11 presents the workflow of developing interpretable ML models for the discovery of MOFs for C_2H_4/C_2H_6 separation. Feature engineering is first conducted to calculate descriptors for MOFs structures. On this basis, ML models can be developed to classify MOFs into C_2H_4-selective or C_2H_6-selective adsorbents. With the insights obtained by interpreting these ML models, promising MOFs featuring desirable structural characteristics can be efficiently identified from large databases.

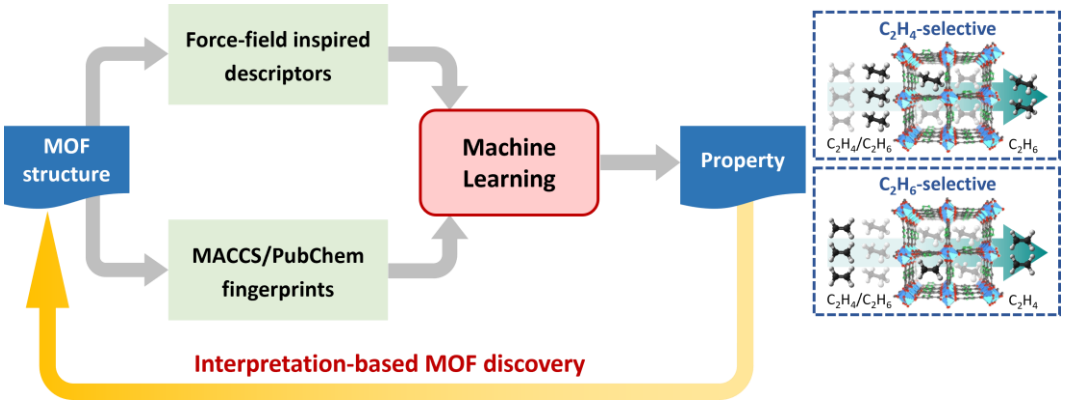


Figure 5-11. Schematic diagram of interpretable ML models for MOF discovery.

MOF Dataset

The dataset used is identical to that in **Section 5.1** and consists of 9156 MOFs with simulated C_2H_6 and C_2H_4 uptakes, as detailed in **Subsection 5.1.1**. Based on the simulated adsorption data, each MOF is classified as either C_2H_4-selective or C_2H_6-selective. C_2H_4-selective MOFs preferentially adsorb C_2H_4 over C_2H_6 and have a C_2H_6/C_2H_4 selectivity lower than 1, while C_2H_6-selective MOFs preferentially adsorb C_2H_6 over C_2H_4 and have a C_2H_6/C_2H_4 selectivity higher than 1. After excluding MOFs with zero C_2H_6 and C_2H_4 uptakes, a refined dataset of 8800 MOFs with their C_2H_6/C_2H_4 selectivity is obtained. **Figure 5-12** shows the distribution of C_2H_6/C_2H_4 selectivity for these 8800 MOFs, where 2617 are identified as C_2H_4-selective adsorbents while the others are C_2H_6-selective.

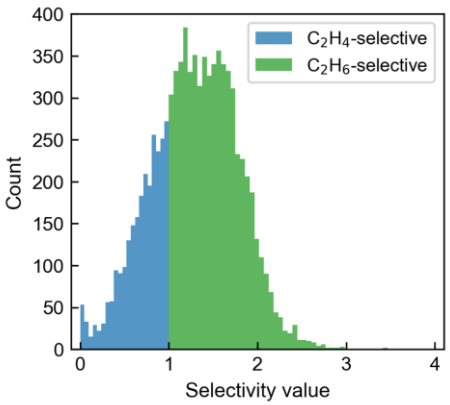


Figure 5-12. Distribution of the C_2H_6/C_2H_4 selectivity of 8800 MOFs.

Feature Engineering

To characterize MOF structures, two different approaches are used to obtain numerical features such as descriptors and molecular fingerprints. Material descriptors can appropriately represent the physical, chemical, or topological characteristics of materials in a numerical format. In contrast, molecular fingerprints are binary bit strings (i.e., sequences of 0s and 1s) encoded from molecular structures. Each bit corresponds to a predefined substructure or functional group, and its value indicates the presence or absence of that substructure or functional group. Both descriptors and fingerprints can serve as inputs for ML models to predict adsorption performance of materials.

Classic force-field inspired descriptors (CFID)[172] are a set of 1557 chemo-structural descriptors. It allows differentiating between material structures and provides a great advantage over many conventional methods as it is independent of using primitive, conventional, or supercell structures of a material. Based on the CIFs of MOF structures, CFID descriptors are calculated using Pymatgen[173] and Matminer[174].

In terms of molecular fingerprints, two commonly used ones are considered for the characterization of MOF structures: MACCS (Molecular ACCess System) keys and PubChem fingerprints. MACCS keys contain 166 types of substructures while the PubChem fingerprints encode molecular fragments with 881 binary digits. The definitions of substructures and fragments for both molecular fingerprints are available in the document[175,176]. Based on the CIFs of MOF structures, MACCS and PubChem fingerprints are calculated using Open Babel[177] and PaDEL[178].

Consequently, numerical features including CFID descriptors, MACCS fingerprints, and PubChem fingerprints are calculated for the 8800 MOFs and are subsequently used as inputs for ML models to predict adsorption performance of materials.

Interpretable Machine Learning Model

Random forest (RF), an ensemble of decision tree algorithms, is used as a classification model to predict the adsorption preferences of MOFs based on their features (i.e., CFID descriptors and two types of molecular fingerprints). Compared to other ML methods such as neural networks, the RF is less computationally expensive and more interpretable. For this binary classification task, MOFs are categorized as C_2H_6-selective (positive class) or C_2H_4-selective (negative class). The RF model outputs two probability values ranging from 0 to 1 that indicate the likelihood of a MOF being C_2H_6-selective or C_2H_4-selective. The summation of these two probabilities equals to 1. MOFs with a C_2H_6-selective probability larger than 0.5 are classified as C_2H_6-selective adsorbents, otherwise, they are classified as C_2H_4-selective adsorbents.

Data and code for implementing the interpretable ML model in this section are available in the GitHub repository[179].

5.2.2 Model development

The entire dataset is divided into three parts: training (80%), validation (10%), and test (10%) sets. To improve model accuracy, the parameters of the RF model are optimized using the training set, while the hyperparameters listed in **Table 5-4** are optimized based on performance evaluation on the validation set. With the optimal hyperparameters, the performance of the final RF model is evaluated using the test set. The model training, hyperparameter optimization, and final evaluation are performed using Scikit-learn[149]. Two statistical metrics, accuracy and F1 score, are used to quantify the performance of the ML classification models.

Table 5-4. Hyperparameters considered for the ML model.

Hyperparameter	Options
The number of trees (N_{tree})	20, 30, 40, 50, 60
The maximum depth of the tree (N_{max_depth})	16, 20, 24, 28, 32
The minimum number of samples required to be at a leaf node ($N_{min\_leaf\_sample}$)	2, 4, 6, 8, 10
The ratio between the minimum number of samples required to split an internal node and $N_{min\_leaf\_sample}$ ($N_{min\_node\_sample/min\_leaf\_sample}$)	2, 3, 4, 5, 6

Three RF models are developed using different feature sets: CFID, MACCS, and PubChem. Considering all hyperparameter combinations, the optimal hyperparameters are determined by the grid search method and listed in **Table 5-5**. The performance of three RF models is summarized in **Table 5-6**. Among the three models, the CFID-based one is the best with an overall accuracy of 0.86 on the test set and a higher F1 score than other models.

Table 5-5. Optimal hyperparameter combinations for the ML models.

Model	Optimal hyperparameters
CFID	30, 20, 6, 3
MACCS	60, 32, 2, 4
PubChem	60, 32, 2, 3

Table 5-6. Model performance in the prediction of adsorption preference.

Model	Accuracy			F1 score		
	Training	Validation	Test	Training	Validation	Test
CFID	0.95	0.90	0.86	0.97	0.93	0.90
MACCS	0.84	0.78	0.77	0.89	0.86	0.85
PubChem	0.88	0.82	0.78	0.92	0.88	0.85

The two RF models developed with molecular fingerprints (MACCS and PubChem) show similar predictive performance. The PubChem-based model is slightly better than the MACCS-based model across all sets. Therefore, the PubChem model is selected as the representative fingerprint-based model for subsequent analysis, along with the CFID-based model.

The CFID feature set includes 1557 descriptors (most of them are continuous) to capture MOF physical, chemical, and topological information, whereas the PubChem fingerprints only use 881 binary variables to indicate the presence of specific substructures. Therefore, the CFID descriptors provide a more comprehensive representation of MOF structures, resulting in a ML model with better performance.

5.2.3 Model interpretation

To selectively adsorb trace amounts of C_2H_6 from abundant C_2H_4, C_2H_6-selective MOFs should be selected. The CFID- and PubChem-based models have demonstrated their ability to

distinguish between C_2H_4- and C_2H_6-selective MOFs. To better understand these models, it is important to gain insights into how different features impact predictions.

As the RF is an ensemble tree model, the Tree SHAP algorithm[180] is used to interpret model predictions. It can quantify the impact of each feature on the model's output in terms of both magnitude (significant or insignificant) and direction (positive or negative) aspects. A positive SHAP value indicates that a specific feature has a positive impact on C_2H_6-selective probability, increasing the likelihood of a MOF being classified as C_2H_6-selective. Conversely, a negative SHAP value indicates a negative impact of the feature, leading to a low C_2H_6-selective probability (equivalent to a high C_2H_4-selective probability). The absolute value of the SHAP value represents the significance of impact.

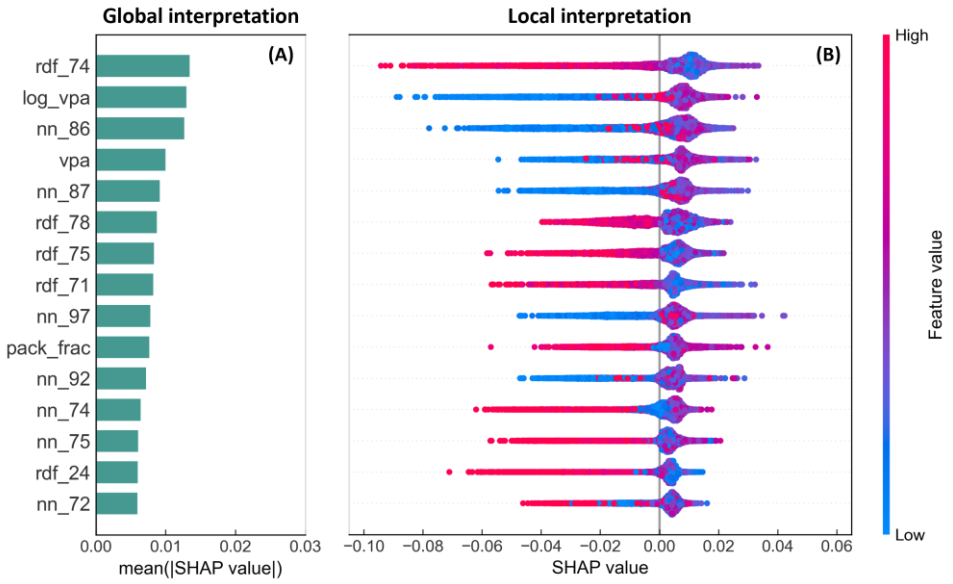


Figure 5-13. Global interpretation (average feature importance) and local interpretation (SHAP value distribution) of the CFID-based model.

For the CFID-based model, SHAP values are calculated for each feature and each MOF. **Figure 5-13A** shows the averaged importance of features and **Figure 5-13B** shows the distribution of SHAP values for each feature across all MOFs. The top 15 features are presented in descending order of importance, which is calculated by the average of the absolute SHAP values. Descriptions of these features can be found in the literature[172]. Notably, the feature "rdf_74", one of the descriptors derived from the radial distribution function, has the largest impact on the output of the CFID-based ML model.

The color bar in **Figure 5-13B** represents the scaled value of the features, allowing all features to be compared within the same range. From the distribution of the feature values, further insights into the feature impact can be obtained. For example, high "rdf_74" values are presented in red and generally exhibit negative SHAP values, while low "log_vpa" values are presented in blue and also show negative SHAP values. Both conditions result in a lower probability of a MOF being C_2H_6-selective. In other words, C_2H_6-selective MOFs are not supposed to have high "rdf_74" values or low "log_vpa" values.

Screening MOFs based on these insights is challenging because setting thresholds for continuous CFID descriptors is not straightforward. For example, while C_2H_6-selective MOFs should avoid high "rdf_74" values, it is generally difficult to determine the exact cutoff value. Moreover, these features are derived from the physical, chemical, and topological properties of MOFs, which cannot directly guide the structural synthesis and design of new MOFs. In contrast, the PubChem-based model relies solely on binary variables that indicate the presence or absence of specific substructures. This makes the PubChem-based model more practical and useful for MOF screening as well as functionalization and design.

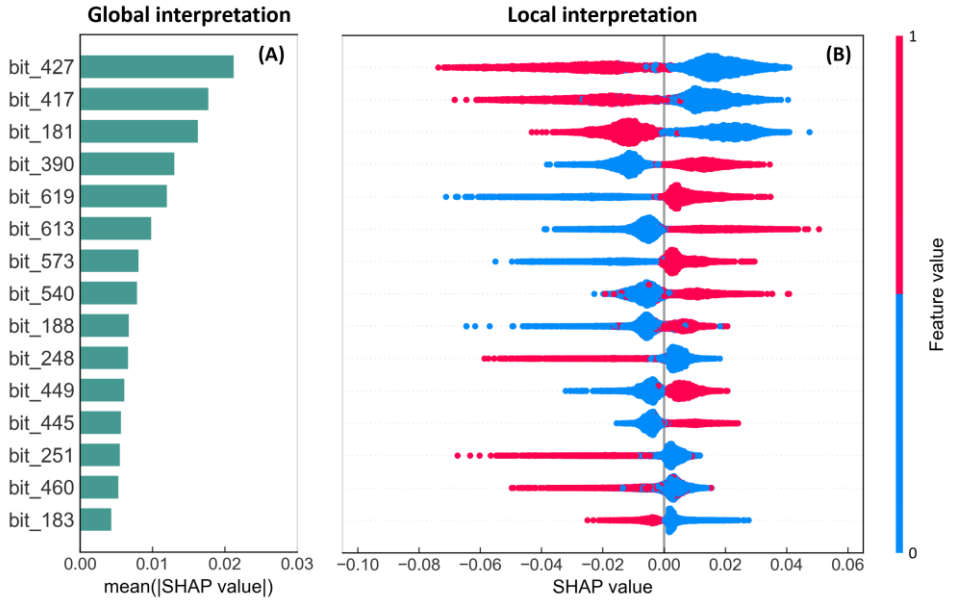


Figure 5-14. Global interpretation (average feature importance) and local interpretation (SHAP value distribution) of the PubChem-based model.

For the PubChem-based model, SHAP values are calculated for each feature and each MOF. The top 15 most important PubChem features are shown in **Figure 5-14A**. They are indicated by their indexes in the PubChem fingerprint list and their descriptions are available in the

document[176]. The feature "bit_427" presents the most significant impact on the predictions because of its highest average absolute SHAP value. It denotes the molecular substructure C#CC, a three-carbon chain with one triple bond.

When "bit_427" is 1 (presented in red), the substructure C#CC is present in the MOF. It generally exhibits negative SHAP values and decreases the probability of a MOF being C_2H_6-selective. This suggests that C_2H_6-selective MOFs usually do not have the substructure C#CC. A similar trend is observed for "bit_417" (substructure C#C) and "bit_181" (saturated or aromatic six-membered rings containing heteroatoms). Conversely, the presence of "bit_390" (C~N~C, carbon-nitrogen-carbon chains) is associated with positive SHAP values, which potentially increases the probability of a MOF being C_2H_6-selective. Thus, carbon-nitrogen-carbon chains are desirable for C_2H_6-selective MOFs. This also applies to other substructures, such as "bit_619" (CC=CCO), "bit_613" (CNCCC), and "bit_573" (C=CCO). With these insights gained from model interpretation, MOFs can be selected or tailor-made based on the desired presence of absence of specific substructures.

5.2.4 MOF screening

To leverage the insights gained by interpreting the PubChem-based model, the large MOF database described in **Subsection 5.1.4** is used for MOF screening. First, the PubChem fingerprints are calculated for these MOF structures. As these structures are more complex, fingerprints cannot be obtained for a small subset (220 MOFs), resulting in 21,164 MOFs being available for screening. Using these fingerprints, the presence and absence of specific substructures are analyzed to identify highly C_2H_6-selective MOFs. Finally, GCMC validation is conducted exclusively on these promising MOF candidates.

In MOF screening, considering too few feature specifications can result in a large number of MOF candidates, leading to a high simulation cost. Conversely, considering too many features may narrow down the pool excessively, potentially excluding some promising MOFs. To balance these trade-offs, a threshold of 5% is set to determine the number of features considered for screening, ensuring that no more than 5% of the 21,164 MOFs are selected for further validation.

Upon analyzing the top 20 important features, it is found that the requirements for the 3rd and 9th features (i.e., bit_181 and bit_188) are mutually exclusive. According to **Figure 5-14**, for a MOF to be C_2H_6-selective, "bit_181" (saturated or aromatic six-membered rings containing

heteroatoms) should be absent, while "bit_188" (at least two saturated or aromatic six-membered rings containing heteroatoms) should be present. The absence of "bit_181" implies that MOFs should not contain any saturated or aromatic six-membered rings containing heteroatoms, which directly contradicts the requirement for the presence of "bit_188". Therefore, these two features will not be considered in MOF screening.

Figure 5-15 illustrates the relationship between the number of feature specifications considered in screening and the number of MOFs retained. To meet the 5% threshold, 13 features need to be considered. These 13 features correspond to the top 15 most important features, excluding the 3rd and 9th features. The detailed feature specifications for screening are as follows: "bit_427", "bit_417", "bit_248", "bit_251", "bit_460", and "bit_183" should be 0 (indicating the absence of these corresponding substructures), while "bit_390", "bit_619", "bit_613", "bit_573", "bit_540", "bit_449", and "bit_445" should be 1 (indicating the presence of these corresponding substructures). After applying these 13 feature specifications, 583 MOFs are retained for subsequent GCMC validation.

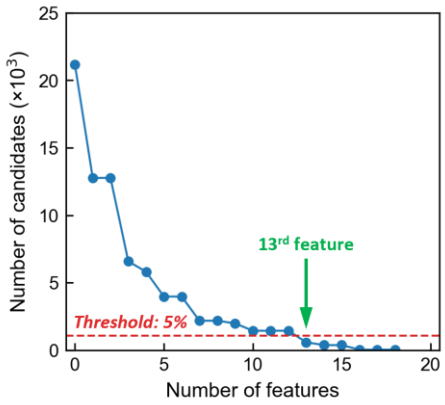


Figure 5-15. Relationship between the number of feature specifications considered and the number of MOFs preserved.

C_2H_6 and C_2H_4 uptakes for the 583 MOF candidates are calculated using GCMC simulations. After excluding 14 MOFs with zero uptakes, C_2H_6/C_2H_4 selectivity is calculated for the remaining 569 MOFs, as shown in **Figure 5-16**. Among them, 93.8% (534 MOFs) are successfully validated as C_2H_6-selective adsorbents, while the remaining 35 MOFs are C_2H_4-selective adsorbents. It demonstrates that the insights gained from interpreting the PubChem-based model enable a practical and efficient discovery of C_2H_6-selective MOFs from large MOF databases.

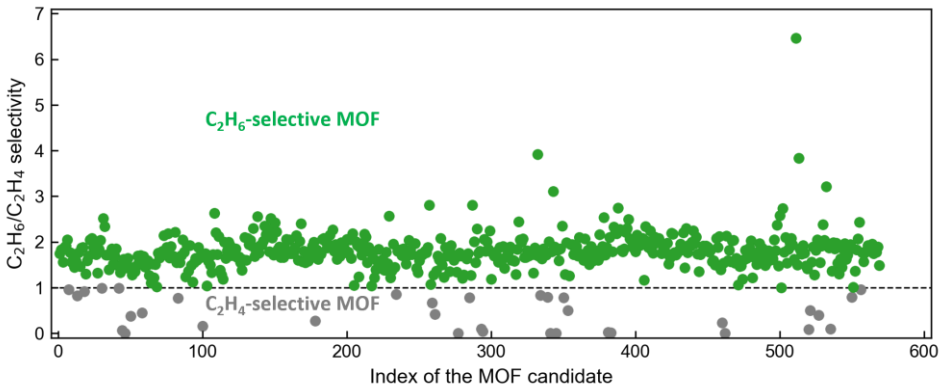


Figure 5-16. GCMC-derived selectivity of the MOF candidates.

Notably, among the 534 identified C_2H_6-selective MOFs, hMOF-5067000 shows the highest C_2H_6/C_2H_4 selectivity of 6.46. Its structure is visualized using iRASPA[181] and the density distribution of the adsorbed C_2H_4 and C_2H_6 molecules is calculated from GCMC adsorption data, as shown in **Figure 5-17A**. From the adsorption equilibrium state depicted in **Figure 5-18**, it is observed that C_2H_4 and C_2H_6 have similar adsorption sites (around the pore center). Despite this, the ratio of adsorbed C_2H_4 molecules to C_2H_6 molecules in the crystal is about 2. Given that the molar ratio of C_2H_4 to C_2H_6 in the bulk phase is 15, a high C_2H_6/C_2H_4 selectivity of around 7.5 is obtained. This promising MOF structure is not identified by the end-to-end ML model in **Section 5.1** because its C_2H_6/C_2H_4 selectivity is underestimated, yielding a value of 1.54.

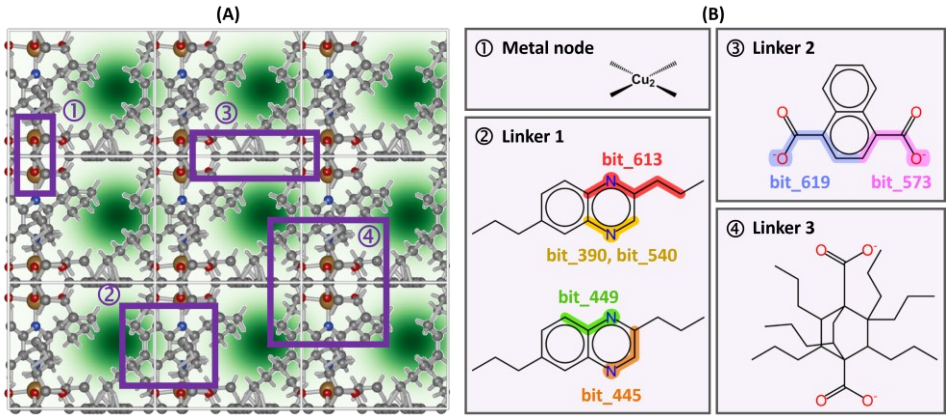


Figure 5-17. (A) Density distribution of adsorbed C_2H_4 and C_2H_6 (dark color indicates high density), and (B) metal node and organic linkers with key substructures highlighted for hMOF-5067000.

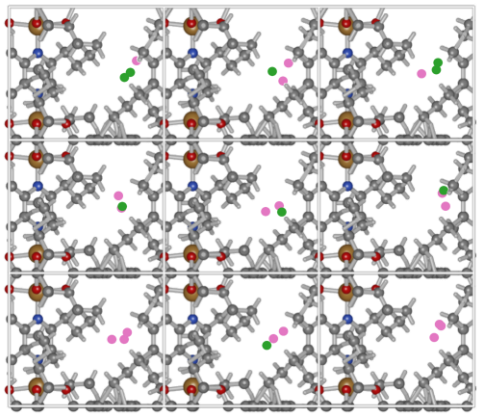


Figure 5-18. Locations of C_2H_4 (in pink) and C_2H_6 molecule centers (in green) in the crystal of hMOF-5067000 at equilibrium state.

Figure 5-17B lists the metal node and three organic linkers that constitute the MOF structure. The presence of favorable substructures analyzed from model interpretation is highlighted along with their feature indexes. This confirms that the substructures represented by "bit_427", "bit_417", "bit_248", "bit_251", "bit_460", and "bit_183" are absent in highly C_2H_6-selective MOFs, which fully aligns with the insights gained from the model interpretation.

The features related to metal nodes are not considered in the MOF screening due to their relatively low importance. Nevertheless, it is found that the presence of copper (Cu) generally shows positive SHAP values (**Figure 5-19**), indicating that copper is a favorable metal node in C_2H_6-selective MOFs. This finding is consistent with hMOF-5067000, which features copper as its metal node.

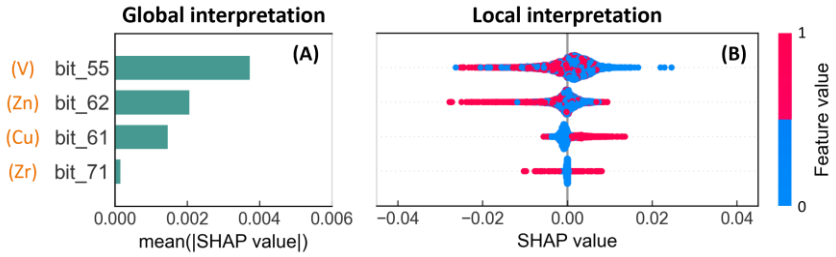


Figure 5-19. Global and local interpretations on metal features for the PubChem-based model.

Overall, the structural characteristics identified through model interpretation are valuable and useful for discovering MOFs that can selectively adsorb C_2H_6 over C_2H_4.

6 Integrated Metal-Organic Framework and Pressure Swing Adsorption Design

Chapter 5 presents two ML approaches to accelerate the identification of optimal MOFs, targeting desired process-relevant properties. However, adsorption process conditions are neglected, and the practical performance of MOFs in the adsorption process is not evaluated. In this chapter, CAMPD is conducted to identify both the optimal adsorbent and the corresponding PSA system simultaneously, where detailed process modeling and optimization are considered in the identification of optimal MOFs.

The integrated material and process design is illustrated by the separation of ethylene/ethane (C_2H_4/C_2H_6) using pressure swing adsorption, which is the same case studied for the data-driven MOF discovery discussed in **Chapter 5**.

6.1 Adsorption process modeling

To evaluate the practical performance of adsorbents in an adsorption process, adsorption isotherm models are essential because they provide mechanism information of the adsorption process. By incorporating mathematical models describing the adsorption process, they can be used to determine the optimal operating conditions, which are important for the development and optimization of adsorption systems.

6.1.1 MOF database and molecular simulation

CoRE MOF 2019 database is a collection of computation-ready, experimental metal-organic framework structures.[54] Compared to the hMOF database, it contains a wider variety of MOF structures that have been experimentally synthesized and show greater diversity in building blocks. This database is used to identify suitable adsorbents for C_2H_4/C_2H_6 separation using adsorption processes. Structures with a disorder are not considered, resulting in a candidate list of 10,143 MOFs as potential adsorbents for further evaluation.

To obtain the adsorption isotherm, the adsorption capacity of MOFs at different pressures is required. Therefore, GCMC simulations are performed using RASPA[156], for pure ethane and ethylene at 298 K and 10 different pressures (from 0.01 to 10 bar). Each simulation includes 5000 equilibration cycles, followed by 10,000 production cycles. The LJ parameters for the

relevant MOF atoms are taken from the Universal force field[159]. Other settings for the GCMC simulation are identical as introduced in **Subsection 5.1.1**.

6.1.2 Adsorption isotherm model fitting

Adsorption isotherm models that describe the variation of gas absorbed with pressure are important for the development and optimization of adsorption processes. Based on the adsorption data from GCMC simulations, MOFs that have low ethane and ethylene adsorption capacities (less than 0.1 mol/kg) are eliminated. As a result, a collection of 9549 MOFs with their C_2H_4 and C_2H_6 uptakes at 298 K and 10 different pressures is obtained to model adsorption isotherms.

Langmuir equation is used to model the single-component adsorption isotherms for pure ethane and ethylene.

$$q = \frac{q_{sat} K P}{1 + K P}$$

where q is the adsorption loading at pressure P, q_{sat} is the saturation adsorption loading, and K is the Langmuir adsorption constant. Based on 10 adsorption data points, the parameters are estimated using nonlinear least-squares implemented in SciPy[150].

6.1.3 Multi-component adsorption isotherm model

Multi-component adsorption isotherm model describes the competitive adsorption behavior of a mixture of multiple components on the adsorbent. Based on the single-component isotherm models, the equilibrium concentration of each component on the adsorbent at given conditions can be calculated using the following extended Langmuir equation.

$$q_i = \frac{q_{sat,i} K_i y_i P}{1 + \sum_{j=1}^{n} K_j y_j P}$$

where q_i is the adsorption loading of component i, K_i is the Langmuir adsorption constant of component i, y_i is the mole fraction of component i in the gas mixture, and n is the number of components in the gas mixture. As the gas mixture contains ethane and ethylene, n is 2.

6.1.4 Pressure swing adsorption process

Vacuum pressure swing adsorption (VPSA), a variation of PSA technology, is considered for the separation of C_2H_4 and C_2H_6 to produce polymer-grade C_2H_4 (>99.9%). Descriptions of the

VPSA process are introduced in **Subsection 2.1.2**. Mathematically, the VPSA process is described by partial differential algebraic equations (PDAEs) as detailed in **Appendix C**. A C2 product of ethylene and ethane (0.85/0.15)[182] is considered for separation via the VPSA process.

Seven key process parameters are considered for optimization: adsorption pressure (P_H), intermediate pressure (P_I), desorption pressure (P_L), adsorption time (t_{ads}), desorption time (t_{des}), length of adsorption column (L), and feed velocity (u_0). To reduce the number of decision variables, times for pressurization and depressurization (i.e., pressurization and blowdown steps in the VPSA cycle) are not considered. The pressurization and depressurization steps are allowed to run until the column is fully pressurized or depressurized, respectively.

6.2 Sequential MOF selection and PSA optimization

In this section, the MOF selection and PSA optimization are sequentially conducted. The optimal MOFs are first selected according to their process-relevant properties, and then the operating conditions of the PSA system are optimized for each MOF selected.

6.2.1 Adsorbent selection

Based on the adsorption isotherm data, single-component adsorption isotherm models are fitted for each MOF involved. On average, the fitted isotherm models of 9549 MOF structures present R^2 values of 0.9877 and 0.9780 for ethylene and ethane, respectively. This indicates that most of these fitted adsorption isotherm models can accurately describe the pressure-dependent adsorption behavior of ethylene and ethane on MOFs. Subsequently, equilibrium adsorption loadings for mixtures of ethylene and ethane at 1 bar and 298 K are calculated using the extended Langmuir equation, and therefore, the C_2H_6/C_2H_4 and C_2H_4/C_2H_6 selectivity is calculated.

For each MOF in the adsorbent candidate list, the VPSA process is simulated using 50 different process parameter settings generated by Sobol sampling. Both one-step and two-step purification processes are considered. In one-step purification, C_2H_6-selective adsorbents are preferred. C_2H_6 is adsorbed and C_2H_4 product is directly obtained from the adsorption step. Conversely, in two-step purification, C_2H_4-selective adsorbents are preferred. C_2H_4 is firstly adsorbed in the adsorption step and then produced during the evacuation step. For both strategies, C_2H_4 purity is separately visualized with C_2H_6/C_2H_4 and C_2H_4/C_2H_6 selectivity. In the one-step purification, the adsorbents cannot achieve polymer-grade ethylene production (**Figure 6-1A**). However, in the two-step purification, there is a clear trend that high C_2H_4/C_2H_6

selectivity is essential to achieve high C_2H_4 purity (**Figure 6-1B**). This indicates that two-step purification using C_2H_4-selective adsorbents is more effective for polymer-grade ethylene production, and that the C_2H_4/C_2H_6 selectivity is a crucial factor in the selection of adsorbents. Therefore, the following will consider C_2H_4-selective adsorbents and two-step ethylene purification processes.

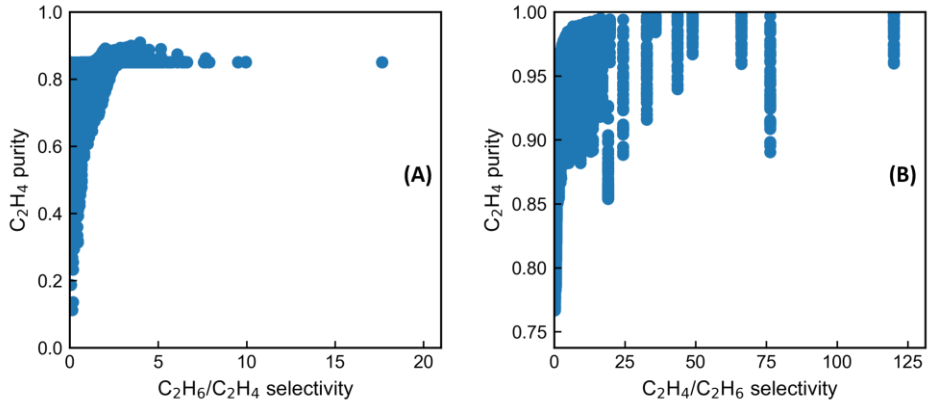


Figure 6-1. Relationships between selectivity of adsorbents and purity of C_2H_4 produced from the VPSA process via (A) one-step purification and (B) two-step purification.

Moreover, the relationship between C_2H_4 purity and recovery is presented. In the upper right of **Figure 6-2A**, there is a trade-off between product purity and recovery. High product purity usually leads to low product recovery. Focusing on the high-purity region (**Figure 6-2B**), it is not difficult to obtain high-purity products (>99.0%) at good recovery rates. However, identifying a suitable adsorbent that can produce polymer-grade ethylene (>99.9%) is challenging. This also leads to a recovery rate lower than 0.3.

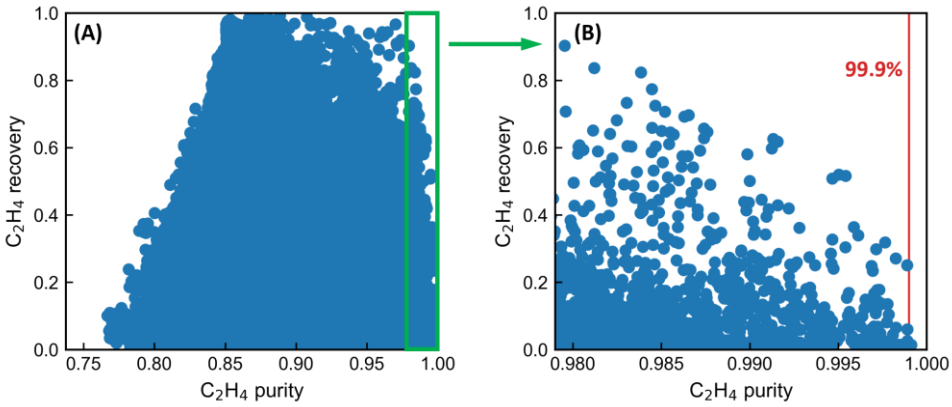


Figure 6-2. Relationship between purity and recovery of C_2H_4 produced from the VPSA process using different adsorbents: (A) the entire purity-recovery space and (B) the high-purity region.

Following the insights gained from the relationship between adsorbent property and separation performance (**Figure 6-1B**), the top 10 MOFs showing the highest C_2H_4/C_2H_6 selectivity are selected as promising adsorbents for polymer-grade ethylene production. Their unique identifiers in the Cambridge Structural Database[183] and calculated C_2H_4/C_2H_6 selectivity at 1 bar and 298 K are listed in **Table 6-1**. Using these unique identifiers, the crystal information and structures of MOFs can be obtained from the Cambridge Structural Database[183] or the CoRE MOF 2019 database[54].

Table 6-1. Adsorbent candidates selected for polymer-grade ethylene production.

Adsorbent	C_2H_4/C_2H_6 selectivity
WOWGEU02	120.08
TATFOL	76.34
YEYMEU	66.18
ASALIP	48.88
NEFTUP	43.56
XOPKIX	35.81
YUNJIB	32.74
HIDMEO	24.32
QIVBUT	19.59
CUKXEM	19.04

Table 6-2. Fitted adsorption isotherm parameters of selected adsorbent candidates.

Adsorbent	$q_{sat,ethane}$ (mol/kg)	$q_{sat,ethylene}$ (mol/kg)	$K_{sat,ethane}$ (Pa^{-1})	$K_{sat,ethylene}$ (Pa^{-1})
WOWGEU02	2.1891	0.9072	2.4147×10^{-7}	4.8526×10^{-9}
TATFOL	0.7597	0.0660	2.3091×10^{-7}	3.4828×10^{-8}
YEYMEU	17.8415	0.0454	3.7595×10^{-8}	2.2345×10^{-7}
ASALIP	179.7443	0.0165	4.9173×10^{-9}	1.0932×10^{-6}
NEFTUP	3.7343	1.8551	1.0666×10^{-7}	4.9291×10^{-9}
XOPKIX	2.1031	0.4854	1.4190×10^{-6}	1.7166×10^{-7}
YUNJIB	4.9532	1.5224	4.8386×10^{-8}	4.8088×10^{-9}
HIDMEO	1.3500	0.9700	8.4779×10^{-8}	4.8515×10^{-9}
QIVBUT	3.2555	139.6299	4.1195×10^{-6}	4.9020×10^{-9}

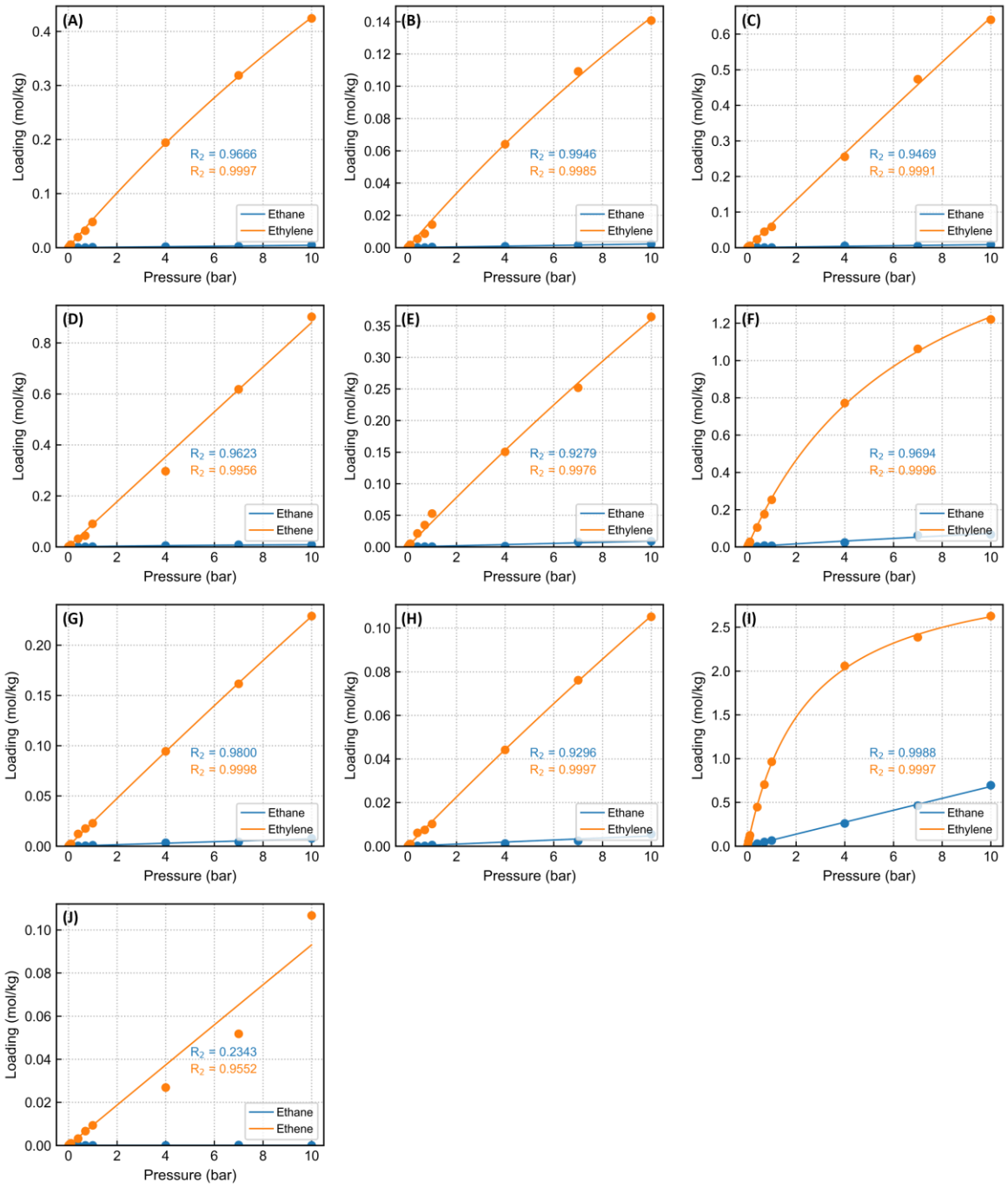


Figure 6-3. Adsorption data and fitted adsorption isotherm models for ethane and ethylene on different adsorbents: (A) WOWGEU02, (B) TATFOL, (C) YEYMEU, (D) ASALIP, (E) NEFTUP, (F) XOPKIX, (G) YUNJIB, (H) HIDMEO, (I) QIVBUT, and (J) CUKXEM.

For these adsorbent candidates, their adsorption isotherm data (in dots) and fitted isotherm models (in lines) are presented in **Figure 6-3**. For the first nine adsorbents, the R^2 values of ethylene isotherm models are higher than 0.9950, indicating the accuracy of these fitted isotherm models. However, the fitted isotherm models for ethane generally present a lower accuracy due to the relatively low magnitude of the ethane adsorption loadings. The last

adsorbent, "CUKXEM", is excluded for further PSA optimization because of the insufficient accuracy of the fitted isotherm models. Isotherm parameters of the first nine selected adsorbent candidates are shown in **Table 6-2**.

6.2.2 Process optimization

For each of the nine adsorbent candidates, PSA optimization is performed to determine the optimal operating conditions that maximize ethylene recovery, while satisfying the ethylene purity constraint of 0.999.

$$\begin{aligned} & \max_{\boldsymbol{z}} \ recovery(\boldsymbol{z}) \\ s.t. \quad & \text{PDAE model} \\ & purity \geq 0.999 \\ & \boldsymbol{z}_L \leq \boldsymbol{z} \leq \boldsymbol{z}_U \end{aligned}$$

where $\boldsymbol{z}$ denotes the process parameter space, and L and U represent the lower and upper bounds, respectively. All process decision variables are continuous, and their upper and lower bounds are identical as in the literature[184] (**Table 6-3**). The PSA optimization problem is solved using the genetic algorithm (GA) implemented in Pymoo[146]. A population size of 20 is used for the GA-based PSA optimization.

Table 6-3. Upper and lower bounds of process decision variables.

Variable	Symbol	Unit	Lower bound	Upper bound
Adsorption pressure	P_{H}	bar	1	10
Intermediate pressure	P_{I}	bar	0.11	3
Desorption pressure	P_{L}	bar	0.1	0.5
Adsorption time	t_{ads}	s	10	1000
Desorption time	t_{des}	s	10	1000
Length of adsorption column	L	m	1	7
Feed velocity	u_0	m/s	0.1	2

The optimization problem is also solved using the Bayesian optimization (BO) approach introduced in **Section 4.2**. The initial sample size for the BO approach ranges from 64 to 512, and the optimal results are presented for further comparison. In the BO-based PSA optimization, the mathematical formulation is modified by replacing the PDAEs that describe the VPSA

process with surrogate models that estimate purity and recovery rates based on process parameters.

$$\begin{aligned} & \max_{\boldsymbol{z}} \left(EI_{f_{recovery}}(\boldsymbol{z}),\ f_{purity}(\boldsymbol{z})\right) \\ s.t.\quad & f_{purity} \geq 0.999 \\ & \boldsymbol{z}_L \leq \boldsymbol{z} \leq \boldsymbol{z}_U \end{aligned}$$

where *EI* is the acquisition function (i.e., expected improvement of the recovery rate) and *f* represents the surrogate model. The surrogate models are developed using the Gaussian process implemented in BoTorch[185], and surrogate-based optimization is conducted using the NSGA-II[145] implemented in Pymoo[146] with a population size of 20 and a maximum of 100 generations.

In terms of GA-based process optimization, five out of nine selected MOFs are validated as being able to produce polymer-grade ethylene using the VPSA process (**Table 6-4**). These five MOFs are exactly the top five adsorbents showing the highest C_2H_4/C_2H_6 selectivity. It confirms that high C_2H_4/C_2H_6 selectivity is important to achieve high C_2H_4 purity. Among the five adsorbents, only "WOWGEU02" and "YEYMEU" can achieve acceptable recovery rates of 0.4830 and 0.2834, respectively. This can also be confirmed by the results of multi-objective PSA optimization as shown in **Figure 6-4**, where the trade-off between purity and recovery is identified for each adsorbent. The multi-objective optimization is conducted using the NSGA-II[145] implemented in Pymoo[146] with a population size of 100.

Table 6-4. GA-based PSA optimization results for polymer-grade ethylene production.

Adsorbent	Process parameters							Performance	
	P_H (bar)	P_I (bar)	P_L (bar)	t_{ads} (s)	t_{des} (s)	L (m)	u_0 (m/s)	Purity	Recovery
WOWGEU02	7.2030	0.1686	0.1538	10.0	968.4	6.97	1.60	0.9990	0.4821
TATFOL	9.3628	0.1100	0.1000	57.9	132.7	4.62	1.90	0.9990	0.0392
YEYMEU	9.9173	0.2114	0.1061	10.0	82.5	5.23	1.14	0.9990	0.2832
ASALIP	9.8909	0.3987	0.1989	10.5	12.9	1.51	1.74	0.9990	0.0398
NEFTUP	9.9535	0.1138	0.1070	10.2	10.1	3.57	1.30	0.9990	0.0354

In terms of BO-based process optimization, only two adsorbents are validated as being able to produce polymer-grade ethylene using the VPSA process (**Table 6-5**). These two adsorbents (i.e., "WOWGEU02" and "YEYMEU") are the same ones identified by the GA-based

optimization. Their recovery rates are 0.4715 and 0.2544, which are slightly lower than those achieved by the GA. For the other three adsorbents, “TATFOL”, “ASALIP”, and “NEFTUP”, the BO approach cannot find feasible solutions in the PSA optimization.

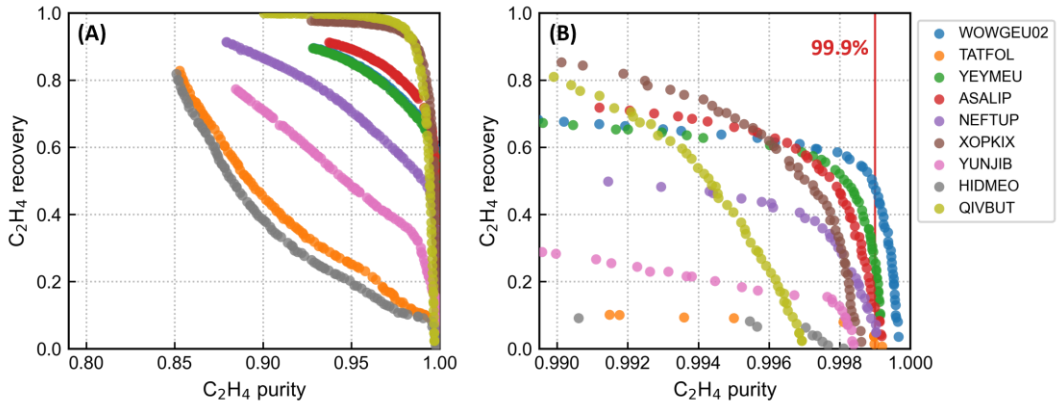


Figure 6-4. Product purity-recovery trade-off identified by multi-objective PSA optimization: (A) the entire purity-recovery space and (B) the high-purity region.

Table 6-5. BO-based PSA optimization results for polymer-grade ethylene production.

Adsorbent	Process parameter							Performance	
	P_H (bar)	P_I (bar)	P_L (bar)	t_{ads} (s)	t_{des} (s)	L (m)	u_0 (m/s)	Purity	Recovery
WOWGEU02	7.9029	0.2007	0.1958	10.27	994.19	6.99	1.60	0.9990	0.4711
YEYMEU	8.9063	0.1539	0.1005	10.81	101.88	4.51	0.21	0.9990	0.2593

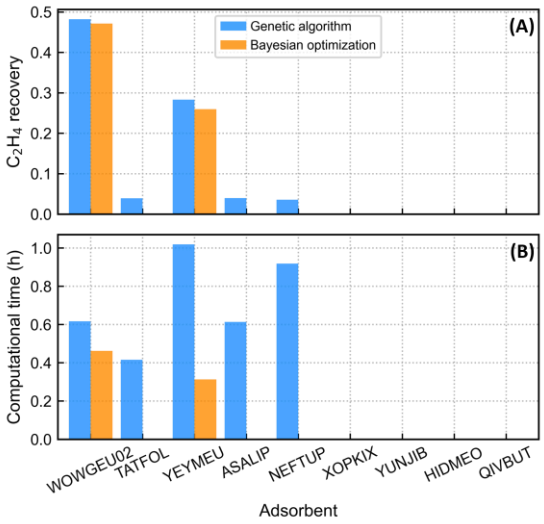


Figure 6-5. Comparison between GA- and BO-based PSA optimization in (A) the achieved product recovery and (B) the corresponding computational cost.

Figure 6-5 compares the performance and computational cost of GA and BO-based PSA optimization approaches for the nine MOF candidates. While GA outperforms in identifying the optimal process parameters for various adsorbents in polymer-grade ethylene production, it incurs higher computational costs compared to BO.

For the investigated PSA system, it is challenging to achieve a high recovery rate for polymer-grade ethylene production. In this context, two-stage PSA systems offer a viable alternative allowing for improved purity and recovery rates, which has been demonstrated on carbon capture[186-188]. To accomplish the C_2H_4/C_2H_6 separation using two-stage PSA systems, the C_2H_4 purity is initially increased to a certain level (for example 0.99) in the first stage and refined in the second stage to satisfy the desired purity (i.e., 0.999). Such a two-stage PSA system can maintain high recovery rates without compromising the purity of the final product. For simplicity, two stages of the VPSA process are optimized separately using the identical search space of process parameters. Both GA and BO approaches are studies. The purity of the product obtained in the first stage is set at 0.987.

Table 6-6. GA-based optimization results for the two-stage VPSA process.

Adsorbent	First stage		Second stage		Total recovery
	Purity	Recovery	Purity	Recovery	
WOWGEU02	0.9870	0.5908	0.9990	0.6474	0.3824
TATFOL	0.9870	0.1051	0.9990	0.1134	0.0119
YEYMEU	0.9870	0.5976	0.9990	0.6770	0.4045
ASALIP	0.9870	0.6688	0.9990	0.7595	0.5079
NEFTUP	0.9870	0.4920	0.9990	0.4664	0.2295
XOPKIX	0.9870	0.8724	0.9990	0.8357	0.7290
YUNJIB	0.9870	0.1858	0.9990	0.1763	0.0328
HIDMEO	0.9871	0.0965	0.9990	0.0977	0.0094
QIVBUT	0.9870	0.7854	0.9990	0.5680	0.4461

In terms of the GA-based optimization, all nine adsorbents are demonstrated to produce polymer-grade ethylene via the two-stage VPSA process (**Table 6-6**). Among them, six adsorbents achieve a recovery rate higher than 0.20. It is worth noting that some adsorbents (e.g., “XOPKIX” and “QIVBUT”) can achieve polymer-grade ethylene production at a high

recovery rate using the two-stage VPSA process, although they fail to produce polymer-grade ethylene using the one-stage VPSA process. For each adsorbent, the optimal process parameters of the two-stage VPSA, determined by the GA-based optimization, are provided in **Table D-1** and **Table D-2** of **Appendix D**, respectively.

Similarly, the BO-based optimization demonstrates that all nine adsorbents can produce polymer-grade ethylene via the two-stage VPSA process (**Table 6-7**). Among them, six adsorbents achieve an acceptable recovery rate higher than 0.20, which aligns with the results obtained from the GA-based optimization. For each of the nine adsorbents, the optimal process parameters of the two-stage VPSA, determined by the BO-based optimization, are provided in **Table D-3** and **Table D-4** of **Appendix D**, respectively.

Table 6-7. BO-based optimization results for the two-stage VPSA process.

Adsorbent	First stage		Second stage		Total recovery
	Purity	Recovery	Purity	Recovery	
WOWGEU02	0.9873	0.6953	0.9990	0.7062	0.4910
TATFOL	0.9870	0.1093	0.9990	0.1282	0.0140
YEYMEU	0.9873	0.6872	0.9990	0.7011	0.4818
ASALIP	0.9874	0.7486	0.9990	0.7711	0.5773
NEFTUP	0.9875	0.5206	0.9990	0.5432	0.2828
XOPKIX	0.9870	0.8943	0.9990	0.8596	0.7688
YUNJIB	0.9875	0.3066	0.9990	0.3736	0.1145
HIDMEO	0.9870	0.0974	0.9990	0.0997	0.0097
QIVBUT	0.9871	0.8330	0.9990	0.6913	0.5758

For both GA- and BO-based optimization, “XOPKIX” is identified as the optimal adsorbent for polymer-grade ethylene production using the two-stage VPSA process, as it achieves a recovery rate higher than 0.70. In comparison, the total recovery rates of the solutions identified by the BO-based optimization are generally higher than those identified by the GA-based optimization. In terms of computational cost averaged over the nine adsorbents, the BO-based optimization method is 47.3% less expensive than the GA-based method. Overall, the BO approach offers a more efficient solution for the optimization of the two-stage VPSA process in achieving high recovery rates.

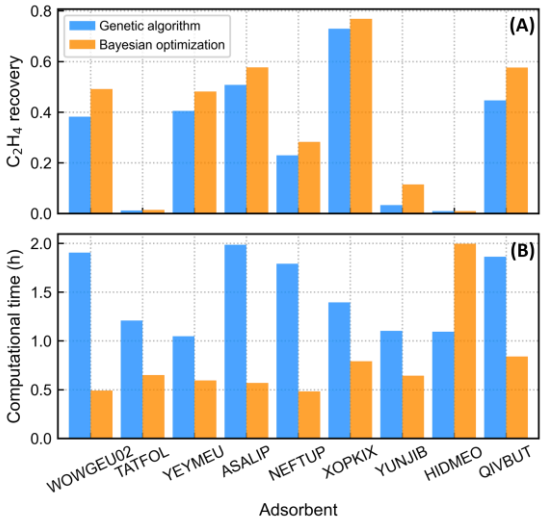


Figure 6-6. Comparison between GA- and BO-based two-stage PSA optimization: (A) the achieved product recovery and (B) the corresponding computational cost.

6.3 Integrated MOF and PSA design

To integrate MOF selection with PSA optimization, the C_2H_4/C_2H_6 selectivity of the MOF is considered as a decision variable together with the VPSA process parameters. This integrated MOF and PSA design task is a CAMPD problem. It is solved to identify the optimal adsorbent and operating conditions that maximize ethylene recovery, while satisfying the ethylene purity constraint of 0.999.

$$\begin{aligned} & \max_{\boldsymbol{y},\boldsymbol{z}} \ recovery(\boldsymbol{y},\boldsymbol{z}) \\ s.t. \quad & \text{PDAE model} \\ & purity \geq 0.999 \\ & \boldsymbol{y}_L \leq \boldsymbol{y} \leq \boldsymbol{y}_U \\ & \boldsymbol{z}_L \leq \boldsymbol{z} \leq \boldsymbol{z}_U \end{aligned}$$

where $\boldsymbol{y}$ and $\boldsymbol{z}$ denote the adsorbent property and process parameter spaces, and L and U represent the lower and upper bounds, respectively. All the material and process decision variables are continuous, and their upper and lower bounds are listed in **Table 6-8**. Similarly, the CAMPD problem is solved using the GA implemented in Pymoo[146], with a population size of 40.

Table 6-8. Upper and lower bounds of material and process decision variables.

Variable	Symbol	Unit	Lower bound	Upper bound
C_2H_4/C_2H_6 selectivity	$S_{C2H4/C2H6}$	–	0.0566	120.0786
Adsorption pressure	P_H	bar	1	10
Intermediate pressure	P_I	bar	0.11	3
Desorption pressure	P_L	bar	0.1	0.5
Adsorption time	t_{ads}	s	10	1000
Desorption time	t_{des}	s	10	1000
Length of adsorption column	L	m	1	7
Feed velocity	u_0	m/s	0.1	2

The CAMPD problem is also solved using the BayesCAMPD approach introduced in **Section 4.2**. The initial sample size for the BO-based CAMPD ranges from 64 to 512, and the optimal results are presented for further comparison. In the BO-based integrated MOF and PSA design, the mathematical formulation is modified by replacing the PDAEs that describe the VPSA process with surrogate models that estimate purity and recovery rates based on the material property and process parameters.

$$\begin{aligned} & \max_{\boldsymbol{y},\boldsymbol{z}} \left(EI_{f_{recovery}}(\boldsymbol{y},\boldsymbol{z}),\ f_{purity}(\boldsymbol{y},\boldsymbol{z})\right) \\ s.t.\quad & f_{purity} \geq 0.999 \\ & \boldsymbol{y}_L \leq \boldsymbol{y} \leq \boldsymbol{y}_U \\ & \boldsymbol{z}_L \leq \boldsymbol{z} \leq \boldsymbol{z}_U \end{aligned}$$

where EI is the acquisition function (i.e., expected improvement of the recovery rate) and f represents the surrogate model. The surrogate models are developed using the Gaussian process implemented in BoTorch[185], and surrogate-based optimization is conducted using the NSGA-II[145] implemented in Pymoo[146] with a population size of 40 and a maximum of 100 generations. For both GA- and BO-based approaches, the molecular property targeting and molecular mapping methods introduced in **Chapter 3** are adapted to identify real adsorbents from the CoRE MOF database during the optimization process.

First, the one-stage VPSA process is studied for the production of polymer-grade ethylene. The GA-based CAMPD approach identifies "WOWGEU02" as the optimal adsorbent that

maximizes the recovery rate while satisfying the C_2H_4 purity constraint of 0.999. The optimal process parameters for "WOWGEU02" are listed in **Table 6-9**. The recovery rate achieved by the GA-based CAMPD is 0.4809, closely matching the result obtained from the GA-based PSA optimization for the same adsorbent (**Table 6-4**). In comparison, the BO-based CAMPD approach also identifies "WOWGEU02" as the optimal adsorbent but achieves a lower recovery rate.

Table 6-9. Results of the GA-based integrated MOF and one-stage PSA design.

Adsorbent	Process parameter							Performance	
	P_H (bar)	P_I (bar)	P_L (bar)	t_{ads} (s)	t_{des} (s)	L (m)	u_0 (m/s)	Purity	Recovery
WOWGEU02	5.1667	0.1159	0.1064	10.0	981.1	6.40	0.85	0.9990	0.4809
WOWGEU02	6.2815	0.2084	0.1859	11.1	645.5	6.02	0.97	0.9991	0.4327

As demonstrated by the sequential MOF selection and PSA optimization in **Section 6.2**, the two-stage VPSA process can achieve polymer-grade ethylene production at higher recovery rates. Therefore, the two-stage VPSA process is also considered for the integrated MOF and PSA design using both GA- and BO-based methods. It should be noted that, for simplicity, the optimal adsorbent and process parameters are separately determined for two stages. The purity of the product obtained from the first stage is set at 0.987.

Table 6-10. Results of the GA- and BO-based integrated MOF and two-stage PSA design.

Method	First stage			Second stage			Total recovery
	Adsorbent	Purity	Recovery	Adsorbent	Purity	Recovery	
GA	XOPKIX	0.9870	0.8957	XOPKIX	0.9990	0.8485	0.7600
BO	XOPKIX	0.9871	0.8921	XOPKIX	0.9990	0.8464	0.7551

In terms of the GA-based CAMPD, "XOPKIX" is identified as the optimal adsorbent for both stages of the VPSA process (**Table 6-10**), resulting in a product recovery rate of 0.7600. Similarly, the BO-based CAMPD approach also identifies "XOPKIX" as the optimal adsorbent for both stages, achieving a slightly lower product recovery rate of 0.7551. Both approaches achieve a nearly identical product recovery rate to that obtained from the BO-based two-stage PSA optimization for the same adsorbent (**Table 6-7**). The optimal process parameters for both stages of the VPSA process, as determined by the GA- and BO-based CAMPD approaches, are

presented in **Table 6-11**. Notably, the BO-based CAMPD approach demonstrates a 35.8% reduction in computational cost compared to the GA-based approach.

Overall, the BO-based CAMPD approach is more efficient for the integrated MOF and two-stage PSA design. The adsorbent "XOPKIX" is an optimal adsorbent for polymer-grade ethylene production at a high recovery rate.

Table 6-11. Process parameters identified by the GA- and BO-based CAMPD approach for the two-stage VPSA process.

Method	Stage	Adsorbent	P_H (bar)	P_I (bar)	P_L (bar)	t_{ads} (s)	t_{des} (s)	L (m)	u_0 (m/s)
GA	First	XOPKIX	3.5094	0.6892	0.1000	10.0	764.1	6.73	0.16
	Second	XOPKIX	2.6927	0.9503	0.1000	29.0	286.3	7.00	0.10
BO	First	XOPKIX	3.5269	0.7199	0.1107	11.5	544.7	6.75	1.84
	Second	XOPKIX	5.6059	1.8853	0.1010	10.1	231.8	6.98	1.84

Among the MOFs investigated, only a few are capable of producing polymer-grade ethylene with a satisfactory recovery rate using the one-stage VPSA process. However, the two-stage VPSA process reduces the separation difficulty, enabling more MOFs to be identified as suitable adsorbents for producing polymer-grade ethylene at higher recovery rates.

In summary, four research strategies are explored for C_2H_4/C_2H_6 separation: (1) sequential MOF selection and one-stage PSA optimization, (2) sequential MOF selection and two-stage PSA optimization, (3) integrated MOF and one-stage PSA design, and (4) integrated MOF and two-stage PSA design. When applied to these different strategies, the GA- and BO-based approaches exhibit distinct advantages and limitations, particularly in terms of separation performance and computational costs. Both approaches are considered practical solutions for selecting suitable adsorbents and designing efficient adsorption systems for polymer-grade ethylene production.

Data and code for implementing the sequential and integrated material and process design approach in this chapter are available in the GitHub repository[189].

7 Conclusions and Outlook

7.1 Summary

In this dissertation, various data-driven approaches are proposed and demonstrated to accelerate computer-aided molecular, material, and process design. These approaches cover a wide range of applications, including optimal molecular design, large-scale material screening, chemical process optimization, and integrated molecular/material and process design.

Molecular property targeting and molecular mapping methods are first introduced to enable efficient optimal solvent design by characterizing solvents with their molecular properties. Data-driven models are developed to calculate key performance indicators of processes based on solvent properties, serving as surrogates for mechanistic process models. Optimal solvents are identified to maximize the separation performance of extractive distillation processes through surrogate-based optimization. The molecular property targeting approach is demonstrated to be effective in the optimal molecular design of solvents.

To integrate molecular design with process optimization, data-driven process models are extended to estimate process performance based on solvent properties and process parameters. By incorporating data-driven models, optimization algorithms, and the molecular property targeting approach, optimal solvents and process parameters that improve process performance are efficiently identified and validated. Such a data-driven CAMPD approach proves practical for the integrated design of solvents and extractive distillation processes. Taking one step further, Bayesian optimization is integrated into the data-driven CAMPD approach to reduce the high data demand for accurate surrogate modeling. The resulting BayesCAMPD approach offers a data-efficient and closed-loop solution for CAMPD tasks by iterating data-driven modeling, surrogate-based optimization, and solution validation. Through the same case study, its superiority in computational efficiency and result quality is demonstrated. The use of Bayesian optimization significantly reduces data demand, making it particularly beneficial for applications with limited data.

Compared to solvent molecules, MOFs exhibit greater structural complexity, as they consist of metal clusters that are connected by organic ligands. To accelerate materials discovery, two types of machine learning models are developed for efficient MOF screening: an end-to-end model for accurate predictions and an interpretable model that provides insights into the

predictions. The end-to-end model utilizes a neural network, that incorporates feature embedding and molecular graph convolution, to characterize MOF structures for correlating their adsorption capacities. It learns both chemical and geometric features from MOF building blocks, enabling accurate and efficient estimation of MOF's adsorption capacities. The interpretable model, on the other hand, combines feature engineering with straightforward tree-structure models to learn the MOF's adsorption preference for gas separation. It provides understandable insights into the model's decision-making process, identifying key structural characteristics that are crucial for designing high-performance materials. Both approaches are efficient for large-scale screening, accelerating the identification of optimal MOFs for energy-efficient gas separation.

Process optimization is further integrated with MOF selection to identify suitable adsorbents for gas separation from a process-level perspective, where the practical performance of MOFs is evaluated using PSA systems. Both simulation-based optimization and Bayesian optimization approaches are applied for the sequential MOF selection and PSA optimization, as well as for the integrated MOF and PSA design. The consideration of a two-stage PSA system allows more MOFs to be identified as suitable adsorbents, capable of achieving high product purity and recovery rates. Both approaches are demonstrated as efficient and practical for selecting suitable adsorbents and designing efficient adsorption systems, offering distinct advantages in solution quality and computational costs across different tasks.

Overall, these data-driven approaches have proven to be effective and computationally efficient in advancing the development of efficient separation systems, with broad applications in process systems engineering, materials engineering, and chemical engineering.

7.2 Limitations and future work

In terms of the end-to-end ML model, complex MOF structures are decomposed into their fundamental building blocks, such as metal nodes and organic linkers. The features of these building blocks are learned separately and then aggregated as the MOF features to characterize the MOF structures for property prediction. However, this approach does not consider the interrelations between different building blocks. The MOF structure can be represented as a large molecular graph where metal nodes and organic linkers are vertices and edges, whereas the organic linker can be represented as a small molecular graph. Therefore, a hierarchical graph convolutional network can be proposed to capture both the local structural information

of the individual building blocks and the global structural information of the entire MOF structure. This hierarchical approach can learn a more comprehensive representation of MOFs and thereby improve predictive accuracy by considering the connectivity and interactions between the MOF building blocks.

For integrated molecular/material and process design, the BayesCAMPD approach has been demonstrated practical and efficient through the cases of ED and VPSA processes. Currently, all surrogate models in the BayesCAMPD are developed using the Gaussian processes (GP) because they are well-suited for small datasets and do not require complex hyperparameter optimization. Large datasets are beneficial for BayesCAMPD in guiding optimization, particularly for complex systems with numerous decision variables. However, the computational cost associated with model development and prediction using GP increases exponentially with the size of the datasets. This poses a significant challenge for the BayesCAMPD approach in balancing dataset size and computational efficiency. In this context, Bayesian neural networks (BNNs) can be considered an alternative solution for surrogate modeling due to their ability to handle large datasets and estimate prediction uncertainty. Integrating BNNs can improve the scalability of BayesCAMPD for applications involving large datasets. However, it is important to limit the hyperparameter search space to ensure efficient model training with BNNs.

For the production of polymer-grade ethylene, identifying MOFs with high recovery rates is challenging because of the similar physical and chemical properties of ethylene and ethane. The two-stage VPSA process is demonstrated better than the one-stage VPSA, allowing for more MOFs to be used for the production of polymer-grade ethylene at high recovery rates. Since the two stages are optimized separately, co-optimization of both stages can further improve the recovery rate. Additionally, alternative PSA cycles can be explored to accomplish the desired separation in a single stage, thereby reducing both capital investments and operational costs for polymer-grade ethylene production.

Appendix A: Separation of 1,3-Butadiene and 1-Butene

C4 hydrocarbons, co-produced in ethylene production by steam cracking of feedstocks such as naphtha, contain significant quantities of valuable unsaturated compounds such as butene, isobutylene, and butadiene.[190,191] For example, butadiene is an industrially important precursor to synthetic rubber, and therefore it needs to be recovered from the C4 mixture. However, similar physicochemical properties (especially the boiling point) of these components make them very difficult to separate by conventional methods. Adding a suitable solvent into such a mixture to alter the relative volatility, and ED can be readily conducted to achieve efficient separation. Therefore, for this separation task, it is required to identify a suitable solvent and determine the optimal process parameters to extract butadiene from the mixture efficiently.

For the separation of C4 hydrocarbons (butadiene extraction), organic solvents such as N-methyl-2-pyrrolidone (NMP), dimethyl formamide (DMF), and acetonitrile (ACN) are frequently used in different industrial applications. Among them, the NMP shows distinct advantages in energy consumption and environmental impact over other solvents, and it is commercially used in BASF's butadiene extraction technology[140].

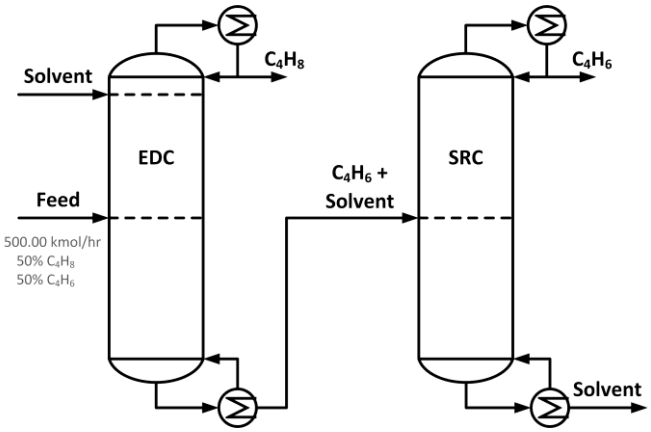


Figure A-1. ED process for the separation of 1,3-butadiene and 1-butene.

Herein, the separation of a simplified C4 mixture, 1,3-butadiene (C_4H_6) and 1-butene (C_4H_8), is taken as an example to demonstrate the proposed approaches in **Chapters 3–4** for the optimal solvent design as well as the integrated solvent and ED design. Solvents that selectively interact with C_4H_6 are considered, and therefore, the C_4H_8 and C_4H_6 are produced from the EDC and SRC, respectively (**Figure A-1**). For this purpose, a simplified setup is studied to allow for a

clear interpretation of the results. The feed consists of an equimolar mixture of C_4H_6 and C_4H_8 with a total flow rate of 500 kmol/hr. Other process parameters are specified below.

- In the EDC, pure solvent and feed are introduced to the third stage and the middle of the column, respectively.
- The feed for the SRC is introduced in the middle.
- The pressure drop at each distillation tray is 1 kPa.
- The distillate rate is 250 kmol/hr in both EDC and SRC.

With such an ED process, it is expected to obtain C_4H_8 and C_4H_6 products with a purity higher than 99.5% in the EDC and SRC distillates, respectively.

Appendix B: Separation of Ethylene and Ethane

Ethylene, primarily produced by steam cracking of naphtha, is one of the major industrial chemicals used in the production of polymers and industrial chemicals.[191,192] In ethylene production, the separation of ethylene (C_2H_4) and ethane (C_2H_6) is crucial, but challenging and energy intensive due to their similar properties. Traditionally, cryogenic distillation is used for the industrial separation of C_2H_4/C_2H_6 mixtures under high pressures and low temperatures (typically 7–28 bar and 183–258 K) with high distillation towers (more than 100 trays), which results in high energy demand and capital investment.[182,193] In contrast, porous materials-based adsorptive separation is a promising alternative due to its high energy efficiency and operational simplicity.

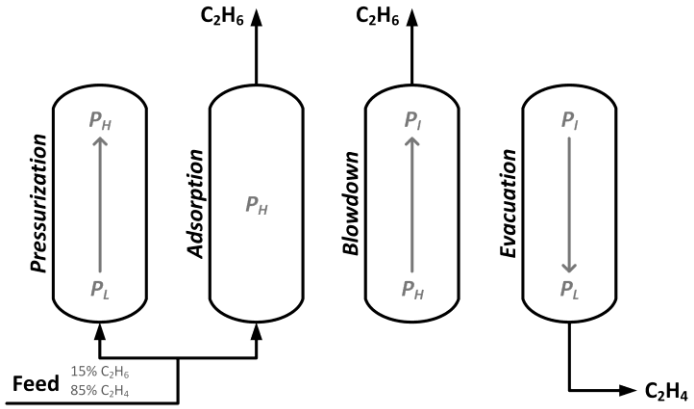


Figure B-1. VPSA process for the separation of ethylene and ethane.

Herein, the separation of C_2H_4/C_2H_6 is taken as an example to demonstrate the proposed approaches in **Chapters 5–6** for the accelerated MOF screening as well as the integrated MOF and PSA design. Adsorbents that selectively adsorb C_2H_4 are considered, and therefore, the C_2H_4 is produced through two-step rather than one-step purification (**Figure B-1**). For this purpose, a simplified setup is studied to allow for a clear interpretation of the results (details are provided in **Appendix C**).

With such a VPSA process, it is expected to obtain ethylene product in polymer grade (>99.9% purity) to produce polyethylene.

Appendix C: Mathematical Models for Pressure Swing Adsorption

Mathematical models describing the VPSA process involve a system of coupled partial differential equations (PDEs) and nonlinear algebraic equations taken from Leperi et al.[188] and Yancy-Caballero et al.[184]

The following assumptions are made[188]:

- The ideal gas law accurately describes the gas phase.
- The viscosity of the gas is independent of pressure.
- There are no radial effects in the concentration, pressure, or temperature in either the gas or solid phase.
- The Ergun equation is used to represent the axial pressure drop.
- The particle size and void fraction are constant throughout the bed.
- The linear driving force (LDF) model is used to describe gas diffusion into the adsorbent.
- The adsorption process is operated at isothermally at 298 K.

All the equations used to describe the VPSA process are put into dimensionless forms. The dimensionless variables are as follows.

$$\bar{P} = \frac{P}{P_0}, \qquad x_i = \frac{q_i}{q_s}, \qquad \bar{u}_z = \frac{u_z}{u_0}, \qquad \tau = \frac{tu_0}{L}, \qquad Z = \frac{z}{L}$$

The following component mass balance is used to calculate the mole fraction of ethylene (y_i) in the gas phase.

$$\frac{\partial y_i}{\partial \tau} = \frac{1}{Pe}\left(\frac{\partial^2 y_i}{\partial Z^2} + \frac{1}{\bar{P}}\frac{\partial \bar{P}}{\partial Z}\frac{\partial y_i}{\partial Z}\right) - \bar{u}_z\frac{\partial y_i}{\partial Z} + \frac{\psi}{\bar{P}}\left((y_i - 1)\frac{\partial x_i}{\partial \tau} + y_i\sum_{j,j\neq i}\frac{\partial x_j}{\partial \tau}\right)$$

where $Pe = \frac{u_0 L}{D_\mathrm{L}}$ and $\psi = \frac{(1-\varepsilon)}{\varepsilon}\frac{RT_0 q_s \rho_s}{P_0}$. The axial dispersion coefficient, D_L, is given by the following equation:

$$D_\mathrm{L} = 0.7D_\mathrm{m} + r_\mathrm{p}u_0$$

The mole fraction of ethane is calculated through the equation:

$$y_{C_2H_4} + y_{C_2H_6} = 1$$

The overall mass balance results in the following equation for calculating the total pressure:

$$\frac{\partial \bar{P}}{\partial \tau} = \left(-\bar{P}\frac{\partial \bar{u}_z}{\partial Z} - \bar{u}_z\frac{\partial \bar{P}}{\partial Z}\right) - \psi \sum_i \frac{\partial x_i}{\partial \tau}$$

The LDF model is used to calculate the mass transfer between the gas phase and the solid phase.

$$\frac{\partial x_i}{\partial \tau} = \frac{k_i L}{u_0}(x_i^* - x_i)$$

$$x_i^* = \frac{q_i^*}{q_s}$$

The pressure drop throughout the column is calculated using the Ergun equation.

$$-\frac{\partial \bar{P}}{\partial Z} = \frac{150\mu(1-\varepsilon)^2 L u_0}{4{r_p}^2\varepsilon^3 P_0}\bar{u}_z + \frac{1.75(1-\varepsilon)L{u_0}^2}{2r_p\varepsilon^3 P_0}\left(\sum_i MW_i y_i C_g\right)\bar{u}_z|\bar{u}_z|$$

$$C_g = \frac{\bar{P}P_0}{RT_0}$$

For all the equations above, the parameters used for the PSA simulation are provided in **Table C-1**, and other variables are summarized in **Table C-3**.

Table C-1. Parameters used in the VPSA cycle.

Parameter	Symbol	Unit	Value	Reference
Bed void fraction	ε	–	0.37	Leperi et al.[188]
Radius of adsorbent pellet	r_p	m	1×10^{-3}	Leperi et al.[188]
Molar loading scaling factor	q_s	mol/kg	5.84	Leperi et al.[188]
Ethylene mole fraction	y_0	–	0.85	Leo et al.[182]
Gas viscosity	μ	Pa·s	1.01815×10^{-5}	Kestin et al.[194]
Diffusion coefficient	D_m	m^2/s	1.14×10^{-5}	Mueller and Cahill[195]
Ethylene molecular weight	MW_{C2H4}	kg/mol	0.0280532	NIST[196]
Ethane molecular weight	MW_{C2H6}	kg/mol	0.0300690	NIST[196]
Ethylene mass transfer coefficient	k_{C2H4}	s^{-1}	0.0125	Bachman et al.[197]
Ethane mass transfer coefficient	k_{C2H6}	s^{-1}	0.0037	Bachman et al.[197]

To solve these equations, the initial and boundary conditions of the column need to be known. Since the first step in the VPSA cycle is the pressurization step, the pressure in the bed is

initially at the evacuation pressure. After startup, the initial conditions of each step are assumed to be the same as the bed profile at the end of the previous step.

The boundary conditions for each step are provided in **Table C-2**.

Table C-2. Boundary conditions of different steps in the VPSA cycle.

Step	Column end	Pressure	Mole fraction
Pressurization	Bottom	$\bar{P} = \bar{P}_{\mathrm{L}} \rightarrow 1$	$y_i = y_0$
	Top	$\frac{\partial \bar{P}}{\partial Z} = 0$	$\frac{\partial y_i}{\partial Z} = 0$
Adsorption	Bottom	$\bar{P} = 1.02$	$y_i = y_0$
	Top	$\bar{P} = 1$	$\frac{\partial y_i}{\partial Z} = 0$
Evacuation	Bottom	$\frac{\partial \bar{P}}{\partial Z} = 0$	$\frac{\partial y_i}{\partial Z} = 0$
	Top	$\bar{P} = 1 \rightarrow \bar{P}_{\mathrm{I}}$	$\frac{\partial y_i}{\partial Z} = 0$
Blowdown	Bottom	$\bar{P} = \bar{P}_{\mathrm{I}} \rightarrow \bar{P}_{\mathrm{L}}$	$\frac{\partial y_i}{\partial Z} = 0$
	Top	$\frac{\partial \bar{P}}{\partial Z} = 0$	$\frac{\partial y_i}{\partial Z} = 0$

Using the method of lines, the spatial derivatives are discretized into 10 volume elements by the finite volume method (FVM)[198] with a weighted essentially nonoscillatory (WENO) scheme[199]. In this way, the system of PDEs is converted into a set of ordinary differential equations (ODEs) and solved using the backward differentiation formula (BDF) method[200] implemented in SciPy[150]. The integration of ODEs is speeded up by leveraging the Numba compiler[201].

A single bed is used to simulate the VPSA cycle. Different steps are conducted in sequence until a cyclic steady state is achieved, which happens when the changes in the state variables are less than 1×10^{-3} between the final conditions in the last step and the initial conditions in the first step in the dimensionless variable. In addition, to ensure no accumulation in the column, the ratio of the gas entering the column to the gas exiting the column over the entire cycle needs to match within a tolerance of 5×10^{-3}. Once the bed has reached the cyclic steady state, the process performances such as product purity and recovery rate are evaluated.

Table C-3. Variables involved in mathematical models of the VPSA process.

Symbol	Variable	Unit
C_g	Molar concentration	mol/m^3
D_L	Axial dispersion coefficient	m^2/s
L	Column length	m
P	Pressure	bar
P_0	Adsorption pressure	bar
$\overline{P}$	Dimensionless pressure	–
q	Molar loading	mol/kg
q^*	Equilibrium molar loading	mol/kg
R	Universal gas constant	J/(mol·K)
t	Time	s
T_0	Adsorption temperature	K
u_0	Inlet gas velocity	m/s
u_z	Superficial gas velocity	m/s
$\bar{u}_z$	Dimensionless gas velocity	m/s
x	Dimensionless molar loading	–
x^*	Dimensionless equilibrium molar loading	–
y	Gas mole fraction	–
z	Bed length coordinate	m
Z	Dimensionless bed length coordinate	–
μ	Gas viscosity	Pa·s
ρ_s	Adsorbent density	kg/m^3
τ	Dimensionless time	–

Appendix D: Optimal Process Parameters of Two-Stage VPSA Processes

Table D-1. Process parameters of the first VPSA stage identified by the GA-based optimization.

Adsorbent	P_H (bar)	P_I (bar)	P_L (bar)	t_{ads} (s)	t_{des} (s)	L (m)	u_0 (m/s)
WOWGEU02	7.7731	1.4283	0.1007	10.6	999.1	7.00	0.50
TATFOL	5.7421	0.1888	0.1000	21.5	762.8	7.00	0.28
YEYMEU	8.6273	1.5504	0.1000	10.0	1000.0	7.00	1.45
ASALIP	6.7450	1.3393	0.1000	10.0	996.4	7.00	0.10
NEFTUP	2.3785	0.3359	0.1000	10.0	1000.0	7.00	0.75
XOPKIX	6.8911	1.3868	0.1756	10.0	963.3	7.00	0.10
YUNJIB	9.2296	0.5443	0.1002	21.8	866.8	7.00	1.18
HIDMEO	9.3026	0.3214	0.1038	28.4	968.4	7.00	1.88
QIVBUT	4.2522	1.5880	0.1000	94.8	214.3	3.73	0.10

Table D-2. Process parameters of the second VPSA stage identified by the GA- based optimization.

Adsorbent	P_H (bar)	P_I (bar)	P_L (bar)	t_{ads} (s)	t_{des} (s)	L (m)	u_0 (m/s)
WOWGEU02	7.6876	2.1271	0.1000	10.0	952.8	7.00	0.62
TATFOL	3.0282	0.1385	0.1000	34.5	999.9	7.00	1.35
YEYMEU	8.4867	2.5504	0.1001	10.0	935.2	7.00	0.43
ASALIP	7.8049	2.8819	0.1702	10.0	989.9	7.00	2.00
NEFTUP	5.6661	0.8768	0.1000	10.0	664.3	7.00	1.83
XOPKIX	3.5617	1.0008	0.2246	17.1	442.0	7.00	0.10
YUNJIB	7.9136	0.7657	0.1000	47.2	1000.0	7.00	1.99
HIDMEO	8.9263	0.3056	0.1000	23.8	847.2	7.00	1.81
QIVBUT	2.6061	1.3257	0.1760	376.0	147.3	6.82	0.11

Table D-3. Process parameters of the first VPSA stage identified by the BO-based optimization.

Adsorbent	P_H (bar)	P_I (bar)	P_L (bar)	t_{ads} (s)	t_{des} (s)	L (m)	u_0 (m/s)
WOWGEU02	1.0998	0.2651	0.1001	10.0	667.3	7.00	2.00
TATFOL	2.4215	0.1148	0.1003	46.5	912.6	7.00	1.88
YEYMEU	1.1617	0.2655	0.1031	10.1	678.2	7.00	1.99
ASALIP	1.1040	0.2717	0.1022	10.1	999.5	7.00	0.11
NEFTUP	1.0267	0.1452	0.1021	10.0	530.8	7.00	0.48
XOPKIX	3.7642	0.7311	0.1006	10.0	492.7	7.00	0.15
YUNJIB	1.4589	0.1101	0.1004	14.1	829.2	7.00	2.00
HIDMEO	4.0408	0.1415	0.1001	33.2	675.1	7.00	0.84
QIVBUT	1.7078	0.9931	0.1617	164.7	199.9	3.47	0.75

Table D-4. Process parameters of the second VPSA stage identified by the BO-based optimization.

Adsorbent	P_H (bar)	P_I (bar)	P_L (bar)	t_{ads} (s)	t_{des} (s)	L (m)	u_0 (m/s)
WOWGEU02	1.0151	0.3150	0.1007	10.1	596.5	7.00	2.00
TATFOL	1.6484	0.1100	0.1002	62.3	714.2	6.99	1.86
YEYMEU	1.0040	0.2983	0.1005	10.1	844.8	7.00	0.15
ASALIP	1.3761	0.4937	0.1008	10.0	903.1	6.84	0.10
NEFTUP	1.5875	0.3062	0.1002	10.0	842.6	7.00	0.86
XOPKIX	1.8288	0.6479	0.1418	10.2	179.0	6.22	2.00
YUNJIB	1.0956	0.1101	0.1001	10.5	581.7	7.00	2.00
HIDMEO	3.6156	0.1266	0.1003	29.6	503.3	6.99	1.64
QIVBUT	9.9985	1.4201	0.4921	10.0	161.0	2.68	0.69

Bibliography

1. *2023 Facts and Figures of the European Chemical Industry*. European Chemical Industry Council. Accessed 21 August 2024. https://cefic.org/a-pillar-of-the-european-economy/facts-and-figures-of-the-european-chemical-industry/
2. *Study Summary "Blackbox Chemical Industry"*. Bund für Umwelt und Naturschutz Deutschland; 2023. Accessed 21 August 2024. https://www.bund.net/service/publikationen/detail/publication/factsheet-studie-blackbox-chemieindustrie-zusammenfassung/
3. Sholl DS, Lively RP. Seven chemical separations to change the world. *Nature*. 2016;532(7600):435-437.
4. Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. *Nature*. 2023;624(7990):80-85.
5. Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*. 2018;361(6400):360-365.
6. Wang Z, Zhou T, Sundmacher K. Data-driven integrated design of solvents and extractive distillation processes. *AIChE Journal*. 2023;69(12):e18236.
7. Wang Z, Zhou T, Sundmacher K. Molecular property targeting for optimal solvent design in extractive distillation processes. In: Kokossis AC, Georgiadis MC, Pistikopoulos E, eds. *Computer Aided Chemical Engineering*. Elsevier; 2023:1247-1252.
8. Wang Z, Zhou Y, Zhou T, Sundmacher K. Identification of optimal metal–organic frameworks by machine learning: Structure decomposition, feature integration, and predictive modeling. *Computers & Chemical Engineering*. 2022;160:107739.
9. Wang Z, Zhou T, Sundmacher K. Interpretable machine learning for accelerating the discovery of metal–organic frameworks for ethane/ethylene separation. *Chemical Engineering Journal*. 2022;444:136651.
10. Sun S, Lü L, Yang A, Shen W. Extractive distillation: Advances in conceptual design, solvent selection, and separation strategies. *Chinese Journal of Chemical Engineering*. 2019;27(6):1247-1256.
11. Gerbaud V, Rodriguez-Donis I, Hegely L, Lang P, Denes F, You X. Review of extractive distillation. Process design, operation, optimization and control. *Chemical Engineering Research and Design*. 2019;141:229-271.

12. Sprakel LM, Kamphuis P, Nikolova AL, Keijsper DJ, Schuur B. Solvent selection for extractive distillation processes to separate close-boiling polar systems. *Chemical Engineering Research and Design*. 2019;144:123-134.
13. Kossack S, Kraemer K, Gani R, Marquardt W. A systematic synthesis framework for extractive distillation processes. *Chemical Engineering Research and Design*. 2008;86(7):781-792.
14. You X, Rodriguez-Donis I, Gerbaud V. Improved design and efficiency of the extractive distillation process for acetone–methanol with water. *Industrial & Engineering Chemistry Research*. 2015;54(1):491-501.
15. Luyben WL. Comparison of extractive distillation and pressure-swing distillation for acetone–methanol separation. *Industrial & Engineering Chemistry Research*. 2008;47(8):2696-2707.
16. Shang X, Ma S, Pan Q, et al. Process analysis of extractive distillation for the separation of ethanol–water using deep eutectic solvent as entrainer. *Chemical Engineering Research and Design*. 2019;148:298-311.
17. Li G, Bai P. New operation strategy for separation of ethanol–water by extractive distillation. *Industrial & Engineering Chemistry Research*. 2012;51(6):2723-2729.
18. Gil ID, Gómez JM, Rodríguez G. Control of an extractive distillation process to dehydrate ethanol using glycerol as entrainer. *Computers & Chemical Engineering*. 2012;39:129-142.
19. Zhu Z, Ri Y, Li M, Jia H, Wang Y, Wang Y. Extractive distillation for ethanol dehydration using imidazolium-based ionic liquids as solvents. *Chemical Engineering and Processing-Process Intensification*. 2016;109:190-198.
20. Kiss AA, Ignat RM. Optimal economic design of an extractive distillation process for bioethanol dehydration. *Energy Technology*. 2013;1(2-3):166-170.
21. Sircar S. Pressure swing adsorption. *Industrial & Engineering Chemistry Research*. 2002;41(6):1389-1392.
22. Santos J, Cruz P, Regala T, Magalhaes F, Mendes A. High-purity oxygen production by pressure swing adsorption. *Industrial & Engineering Chemistry Research*. 2007;46(2):591-599.
23. Chin C, Kamin Z, Bahrun MHV, Bono A. The production of industrial-grade oxygen from air by pressure swing adsorption. *International Journal of Chemical Engineering*. 2023;2023(1):2308227.

24. Bulfin B, Buttsworth L, Lidor A, Steinfeld A. High-purity nitrogen production from air by pressure swing adsorption combined with $SrFeO_3$ redox chemical looping. *Chemical Engineering Journal*. 2021;421:127734.
25. Shirley AI, Lemcoff NO. High-purity nitrogen by pressure-swing adsorption. *AIChE Journal*. 1997;43(2):419-424.
26. Sircar S, Golden T. Purification of hydrogen by pressure swing adsorption. *Separation Science and Technology*. 2000;35(5):667-687.
27. Delgado JA, Águeda V, Uguina M, Sotelo J, Brea P, Grande CA. Adsorption and diffusion of H_2, CO, CH_4, and CO_2 in BPL activated carbon and 13X zeolite: evaluation of performance in pressure swing adsorption hydrogen purification by simulation. *Industrial & Engineering Chemistry Research*. 2014;53(40):15414-15426.
28. Luberti M, Ahn H. Review of Polybed pressure swing adsorption for hydrogen purification. *International Journal of Hydrogen Energy*. 2022;47(20):10911-10933.
29. Bahrun MHV, Bono A, Othman N, Zaini MAA. Carbon dioxide removal from biogas through pressure swing adsorption–A review. *Chemical Engineering Research and Design*. 2022;183:285-306.
30. Augelletti R, Conti M, Annesini MC. Pressure swing adsorption for biogas upgrading. A new process configuration for the separation of biomethane and carbon dioxide. *Journal of Cleaner Production*. 2017;140:1390-1398.
31. Santos MPS, Grande CA, Rodrigues AE. Pressure swing adsorption for biogas upgrading. Effect of recycling streams in pressure swing adsorption design. *Industrial & Engineering Chemistry Research*. 2011;50(2):974-985.
32. Riboldi L, Bolland O. Overview on pressure swing adsorption (PSA) as CO_2 capture technology: state-of-the-art, limits and potentials. *Energy Procedia*. 2017;114:2390-2400.
33. Chao C, Deng Y, Dewil R, Baeyens J, Fan X. Post-combustion carbon capture. *Renewable and Sustainable Energy Reviews*. 2021;138:110490.
34. Pires J, Martins F, Alvim-Ferraz M, Simões M. Recent developments on carbon capture and storage: An overview. *Chemical Engineering Research and Design*. 2011;89(9):1446-1460.
35. Gucuyener C, Van Den Bergh J, Gascon J, Kapteijn F. Ethane/ethene separation turned on its head: selective ethane adsorption on the metal–organic framework ZIF-7 through a gate-opening mechanism. *Journal of the American Chemical Society*. 2010;132(50):17704-17706.

36. Shuai L, Luterbacher J. Organic solvent effects in biomass conversion reactions. *ChemSusChem*. 2016;9(2):133-155.
37. Chemat F, Abert Vian M, Ravi HK, et al. Review of alternative solvents for green extraction of food and natural products: Panorama, principles, applications and prospects. *Molecules*. 2019;24(16):3007.
38. Choi YH, Verpoorte R. Green solvents for the extraction of bioactive compounds from natural products using ionic liquids and deep eutectic solvents. *Current Opinion in Food Science*. 2019;26:87-93.
39. Gani R, Jiménez-González C, Constable DJ. Method for selection of solvents for promotion of organic reactions. *Computers & Chemical Engineering*. 2005;29(7):1661-1676.
40. Wissel K, Riegger LM, Schneider C, et al. Dissolution and recrystallization behavior of Li_3PS_4 in different organic solvents with a focus on N-methylformamide. *ACS Applied Energy Materials*. 2023;6(15):7790-7802.
41. Fickelscherer RJ, Ferger CM, Morrissey SA. Effective solvent system selection in the recrystallization purification of pharmaceutical products. *AIChE Journal*. 2021;67(5):e17169.
42. Pereiro A, Araújo J, Esperança J, Marrucho I, Rebelo L. Ionic liquids in separations of azeotropic systems–A review. *The Journal of Chemical Thermodynamics*. 2012;46:2-28.
43. Neubauer M, Wallek T, Lux S. Deep eutectic solvents as entrainers in extractive distillation–A review. *Chemical Engineering Research and Design*. 2022;184:402-418.
44. Dai C, Lei Z, Xi X, Zhu J, Chen B. Extractive distillation with a mixture of organic solvent and ionic liquid as entrainer. *Industrial & Engineering Chemistry Research*. 2014;53(40):15786-15791.
45. Furukawa H, Cordova KE, O'Keeffe M, Yaghi OM. The chemistry and applications of metal–organic frameworks. *Science*. 2013;341(6149):1230444.
46. Bao Z, Wang J, Zhang Z, et al. Molecular sieving of ethane from ethylene through the molecular cross-section size differentiation in gallate-based metal–organic frameworks. *Angewandte Chemie*. 2018;130(49):16252-16257.
47. Lin R-B, Xiang S, Xing H, Zhou W, Chen B. Exploration of porous metal–organic frameworks for gas separation and purification. *Coordination Chemistry Reviews*. 2019;378:87-103.
48. Li J-R, Kuppler RJ, Zhou H-C. Selective gas adsorption and separation in metal–organic frameworks. *Chemical Society Reviews*. 2009;38(5):1477-1504.

49. Chen Y-Z, Zhang R, Jiao L, Jiang H-L. Metal–organic framework-derived porous materials for catalysis. *Coordination Chemistry Reviews*. 2018;362:1-23.
50. Qiu T, Liang Z, Guo W, Tabassum H, Gao S, Zou R. Metal–organic framework-based materials for energy conversion and storage. *ACS Energy Letters*. 2020;5(2):520-532.
51. Xue Z, Li Y, Zhang Y, et al. Modulating electronic structure of metal–organic framework for efficient electrocatalytic oxygen evolution. *Advanced Energy Materials*. 2018;8(29):1801564.
52. Murray LJ, Dincă M, Long JR. Hydrogen storage in metal–organic frameworks. *Chemical Society Reviews*. 2009;38(5):1294-1314.
53. Witman M, Ling S, Anderson S, et al. In silico design and screening of hypothetical MOF-74 analogs and their experimental synthesis. *Chemical Science*. 2016;7(9):6263-6272.
54. Chung YG, Haldoupis E, Bucior BJ, et al. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. *Journal of Chemical & Engineering Data*. 2019;64(12):5985-5998.
55. Lin J-B, Nguyen TT, Vaidhyanathan R, et al. A scalable metal–organic framework as a durable physisorbent for carbon dioxide capture. *Science*. 2021;374(6574):1464-1469.
56. Ghanbari T, Abnisa F, Daud WMAW. A review on production of metal organic frameworks (MOF) for CO_2 adsorption. *Science of The Total Environment*. 2020;707:135090.
57. Vikrant K, Kumar V, Kim K-H, Kukkar D. Metal–organic frameworks (MOFs): potential and challenges for capture and abatement of ammonia. *Journal of Materials Chemistry A*. 2017;5(44):22877-22896.
58. Waller PJ, Gándara F, Yaghi OM. Chemistry of covalent organic frameworks. *Accounts of Chemical Research*. 2015;48(12):3053-3063.
59. Ahmed I, Jhung SH. Composites of metal–organic frameworks: preparation and application in adsorption. *Materials Today*. 2014;17(3):136-146.
60. Muñoz-Senmache JC, Kim S, Arrieta-Pérez RR, Park CM, Yoon Y, Hernández-Maldonado AJ. Activated carbon–metal organic framework composite for the adsorption of contaminants of emerging concern from water. *ACS Applied Nano Materials*. 2020;3(3):2928-2940.
61. Chemmangattuvalappil NG. Development of solvent design methodologies using computer-aided molecular design tools. *Current Opinion in Chemical Engineering*. 2020;27:51-59.

62. Gani R. Computer-aided methods and tools for chemical product design. *Chemical Engineering Research and Design*. 2004;82(11):1494-1504.
63. Chai S, Song Z, Zhou T, Zhang L, Qi Z. Computer-aided molecular design of solvents for chemical separation processes. *Current Opinion in Chemical Engineering*. 2022;35:100732.
64. Alshehri AS, Gani R, You F. Deep learning and knowledge-based methods for computer-aided molecular design—toward a unified approach: State-of-the-art and future directions. *Computers & Chemical Engineering*. 2020;141:107005.
65. Burger J, Papaioannou V, Gopinath S, Jackson G, Galindo A, Adjiman CS. A hierarchical method to integrated solvent and process design of physical CO_2 absorption using the SAFT-γ Mie approach. *AIChE Journal*. 2015;61(10):3249-3269.
66. Song Z, Zhang C, Qi Z, Zhou T, Sundmacher K. Computer-aided design of ionic liquids as solvents for extractive desulfurization. *AIChE Journal*. 2018;64(3):1013-1025.
67. Ten JY, Liew ZH, Oh XY, Hassim MH, Chemmangattuvalappil N. Computer-aided molecular design of optimal sustainable solvent for liquid-liquid extraction. *Process Integration and Optimization for Sustainability*. 2021;5(2):269-284.
68. Struebing H, Ganase Z, Karamertzanis PG, et al. Computer-aided molecular design of solvents for accelerated reaction kinetics. *Nature Chemistry*. 2013;5(11):952-957.
69. Zhou T, Wang J, McBride K, Sundmacher K. Optimal design of solvents for extractive reaction processes. *AIChE Journal*. 2016;62(9):3238-3249.
70. Keßler T, Kunde C, Linke S, Sundmacher K, Kienle A. Integrated computer-aided molecular and process design: Green solvents for the hydroformylation of long-chain olefines. *Chemical Engineering Science*. 2022;249:117243.
71. Wang Y, Achenie LEK. Computer aided solvent design for extractive fermentation. *Fluid Phase Equilibria*. 2002;201(1):1-18.
72. Cignitti S, Rodriguez-Donis I, Abildskov J, You X, Shcherbakova N, Gerbaud V. CAMD for entrainer screening of extractive distillation process based on new thermodynamic criteria. *Chemical Engineering Research and Design*. 2019;147:721-733.
73. Zhou T, Song Z, Zhang X, Gani R, Sundmacher K. Optimal solvent design for extractive distillation processes: A multiobjective optimization-based hierarchical framework. *Industrial & Engineering Chemistry Research*. 2019;58(15):5777-5786.
74. Pavurala N, Achenie LEK. Identifying polymer structures for oral drug delivery–A molecular design approach. *Computers & Chemical Engineering*. 2014;71:734-744.

75. Chai S, Li E, Zhang L, Du J, Meng Q. Crystallization solvent design based on a new quantitative prediction model of crystal morphology. *AIChE Journal*. 2022;68(1):e17499.
76. Karunanithi AT, Acquah C, Achenie LE, Sithambaram S, Suib SL. Solvent design for crystallization of carboxylic acids. *Computers & Chemical Engineering*. 2009;33(5):1014-1021.
77. Austin ND, Sahinidis NV, Trahan DW. Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques. *Chemical Engineering Research and Design*. 2016;116:2-26.
78. Alshehri AS, You F. Machine learning for multiscale modeling in computational molecular design. *Current Opinion in Chemical Engineering*. 2022;36:100752.
79. Adjiman CS, Sahinidis NV, Vlachos DG, Bakshi B, Maravelias CT, Georgakis C. Process systems engineering perspective on the design of materials and molecules. *Industrial & Engineering Chemistry Research*. 2021;60(14):5194-5206.
80. Zhou T, McBride K, Linke S, Song Z, Sundmacher K. Computer-aided solvent selection and design for efficient chemical processes. *Current Opinion in Chemical Engineering*. 2020;27:35-44.
81. Skiborowski M. Synthesis and design methods for energy-efficient distillation processes. *Current Opinion in Chemical Engineering*. 2023;42:100985.
82. Papadopoulos AI, Linke P. Integrated solvent and process selection for separation and reactive separation systems. *Chemical Engineering and Processing: Process Intensification*. 2009;48(5):1047-1060.
83. Scheffczyk J, Schäfer P, Fleitmann L, et al. COSMO-CAMPD: a framework for integrated design of molecules and processes based on COSMO-RS. *Molecular Systems Design & Engineering*. 2018;3(4):645-657.
84. Fleitmann L, Kleinekorte J, Leonhard K, Bardow A. COSMO-susCAMPD: Sustainable solvents from combining computer-aided molecular and process design with predictive life cycle assessment. *Chemical Engineering Science*. 2021;245:116863.
85. Zhang X, Ding X, Song Z, Zhou T, Sundmacher K. Integrated ionic liquid and rate-based absorption process design for gas separation: Global optimization using hybrid models. *AIChE Journal*. 2021;67(10):e17340.
86. Lee YS, Galindo A, Jackson G, Adjiman CS. Enabling the direct solution of challenging computer-aided molecular and process design problems: Chemical absorption of carbon dioxide. *Computers & Chemical Engineering*. 2023;174:108204.

87. Gopinath S, Jackson G, Galindo A, Adjiman CS. Outer approximation algorithm with physical domain reduction for computer-aided molecular and separation process design. *AIChE Journal*. 2016;62(9):3484-3504.
88. Zhang X, Zhou T, Sundmacher K. Integrated metal–organic framework and pressure/vacuum swing adsorption process design: Descriptor optimization. *AIChE Journal*. 2022;68(2):e17524.
89. Zhang X, Zhou T, Sundmacher K. Integrated metal–organic framework (MOF) and pressure/vacuum swing adsorption process design: MOF matching. *AIChE Journal*. 2022;68(9):e17788.
90. Chen Y, Gani R, Kontogeorgis GM, Woodley JM. Integrated ionic liquid and process design involving azeotropic separation processes. *Chemical Engineering Science*. 2019;203:402-414.
91. Zhou T, McBride K, Zhang X, Qi Z, Sundmacher K. Integrated solvent and process design exemplified for a Diels–Alder reaction. *AIChE Journal*. 2015;61(1):147-158.
92. Gertig C, Fleitmann L, Schilling J, Leonhard K, Bardow A. Rx-COSMO-CAMPD: Enhancing reactions by integrated computer-aided design of solvents and processes based on quantum chemistry. *Chemie Ingenieur Technik*. 2020;92(10):1489-1500.
93. Zhang L, Pang J, Zhuang Y, Liu L, Du J, Yuan Z. Integrated solvent-process design methodology based on COSMO-SAC and quantum mechanics for TMQ (2,2,4-trimethyl-1,2-H-dihydroquinoline) production. *Chemical Engineering Science*. 2020;226:115894.
94. Agi DT, Jones KD, Watson MJ, et al. Computational toolkits for model-based design and optimization. *Current Opinion in Chemical Engineering*. 2024;43:100994.
95. Kalakul S, Zhang L, Fang Z, et al. Computer aided chemical product design–ProCAPD and tailor-made blended products. *Computers & Chemical Engineering*. 2018;116:37-55.
96. McBride K, Sundmacher K. Overview of surrogate modeling in chemical process engineering. *Chemie Ingenieur Technik*. 2019;91(3):228-239.
97. Biegler LT, Lang Y, Lin W. Multi-scale optimization for process systems engineering. *Computers & Chemical Engineering*. 2014;60:17-30.
98. Quirante N, Javaloyes J, Caballero JA. Rigorous design of distillation columns using surrogate models based on Kriging interpolation. *AIChE Journal*. 2015;61(7):2169-2187.
99. Grossmann IE, Harjunkoski I. Process systems engineering: Academic and industrial perspectives. *Computers & Chemical Engineering*. 2019;126:474-484.

100. Venkatasubramanian V. The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE Journal*. 2019;65(2):466-478.
101. Zavala VM. Outlook: How I learned to love machine learning (a personal perspective on machine learning in process systems engineering). *Industrial & Engineering Chemistry Research*. 2023;62(23):8995-9005.
102. Zhou T, Song Z, Sundmacher K. Big data creates new opportunities for materials research: A review on methods and applications of machine learning for materials design. *Engineering*. 2019;5(6):1017-1026.
103. Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV, Van Geem KM. Machine learning in chemical engineering: Strengths, weaknesses, opportunities, and threats. *Engineering*. 2021;7(9):1201-1211.
104. Zhou T, Gani R, Sundmacher K. Hybrid data-driven and mechanistic modeling approaches for multiscale material and process design. *Engineering*. 2021;7(9):1231-1238.
105. Pistikopoulos EN, Barbosa-Povoa A, Lee JH, et al. Process systems engineering – The generation next? *Computers & Chemical Engineering*. 2021;147:107252.
106. Winz J, Nentwich C, Engell S. Surrogate modeling of thermodynamic equilibria: Applications, sampling and optimization. *Chemie Ingenieur Technik*. 2021;93(12):1898-1906.
107. Keßler T, Kunde C, McBride K, et al. Global optimization of distillation columns using explicit and implicit surrogate models. *Chemical Engineering Science*. 2019;197:235-245.
108. Schweidtmann AM, Mitsos A. Deterministic global optimization with artificial neural networks embedded. *Journal of Optimization Theory and Applications*. 2019;180(3):925-948.
109. Streb A, Mazzotti M. Performance limits of neural networks for optimizing an adsorption process for hydrogen purification and CO_2 capture. *Computers & Chemical Engineering*. 2022;166:107974.
110. Leperi KT, Yancy-Caballero D, Snurr RQ, You F. 110th Anniversary: Surrogate models based on artificial neural networks to simulate and optimize pressure swing adsorption cycles for CO_2 capture. *Industrial & Engineering Chemistry Research*. 2019;58(39):18241-18252.

111. Hwangbo S, Al R, Sin G. An integrated framework for plant data-driven process modeling using deep-learning with Monte-Carlo simulations. *Computers & Chemical Engineering*. 2020;143:107071.
112. Yang L, Liu S, Chang C, Yang S, Shen W. An efficient and invertible machine learning-driven multi-objective optimization architecture for light olefins separation system. *Chemical Engineering Science*. 2024;285:119553.
113. Hasan MF, Baliban RC, Elia JA, Floudas CA. Modeling, simulation, and optimization of postcombustion CO_2 capture for variable feed concentration and flow rate. 2. Pressure swing adsorption and vacuum swing adsorption processes. *Industrial & Engineering Chemistry Research*. 2012;51(48):15665-15682.
114. Li J, Suvarna M, Pan L, Zhao Y, Wang X. A hybrid data-driven and mechanistic modelling approach for hydrothermal gasification. *Applied energy*. 2021;304:117674.
115. McBride K, Kaiser NM, Sundmacher K. Integrated reaction–extraction process for the hydroformylation of long-chain alkenes with a homogeneous catalyst. *Computers & Chemical Engineering*. 2017;105:212-223.
116. Nentwich C, Engell S. Application of surrogate models for the optimization and design of chemical processes. In: *2016 International Joint Conference on Neural Networks*. IEEE; 2016:1291-1296.
117. Lu J, Wang Q, Zhang Z, et al. Surrogate modeling-based multi-objective optimization for the integrated distillation processes. *Chemical Engineering and Processing: Process Intensification*. 2021;159:108224.
118. Ma K, Sahinidis NV, Bindlish R, Bury SJ, Haghpanah R, Rajagopalan S. Data-driven strategies for extractive distillation unit optimization. *Computers & Chemical Engineering*. 2022;167:107970.
119. Song Z, Shi H, Zhang X, Zhou T. Prediction of CO_2 solubility in ionic liquids using machine learning methods. *Chemical Engineering Science*. 2020;223:115752.
120. Golkarnarenji G, Naebe M, Badii K, et al. Multi-objective optimization of manufacturing process in carbon fiber industry using artificial intelligence techniques. *IEEE Access*. 2019;7:67576-67588.
121. Içten E, Nagy ZK, Reklaitis GV. Process control of a dropwise additive manufacturing system for pharmaceuticals using polynomial chaos expansion based surrogate model. *Computers & Chemical Engineering*. 2015;83:221-231.
122. Emmerich MT, Deutz AH. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. *Natural Computing*. 2018;17:585-609.

123. Destro F, Inguva PK, Srisuma P, Braatz RD. Advanced methodologies for model-based optimization and control of pharmaceutical processes. *Current Opinion in Chemical Engineering*. 2024;45:101035.
124. Ureel Y, Dobbelaere MR, Ouyang Y, et al. Active machine learning for chemical engineers: A bright future lies ahead! *Engineering*. 2023;
125. Wang K, Dowling AW. Bayesian optimization for chemical products and functional materials. *Current Opinion in Chemical Engineering*. 2022;36:100728.
126. Garnett R. *Bayesian optimization*. Cambridge University Press; 2023.
127. Shields BJ, Stevens J, Li J, et al. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*. 2021;590(7844):89-96.
128. Schweidtmann AM, Clayton AD, Holmes N, Bradford E, Bourne RA, Lapkin AA. Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives. *Chemical Engineering Journal*. 2018;352:277-282.
129. Clayton AD, Schweidtmann AM, Clemens G, et al. Automated self-optimisation of multi-step reaction and separation processes using machine learning. *Chemical Engineering Journal*. 2020;384:123340.
130. Hickman RJ, Aldeghi M, Häse F, Aspuru-Guzik A. Bayesian optimization with known experimental and design constraints for chemistry applications. *Digital Discovery*. 2022;1(5):732-744.
131. Kusne AG, Yu H, Wu C, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nature Communications*. 2020;11(1):5966.
132. Gantzler N, Deshwal A, Doppa JR, Simon CM. Multi-fidelity Bayesian optimization of covalent organic frameworks for xenon/krypton separations. *Digital Discovery*. 2023;2(6):1937-1956.
133. Häse F, Aldeghi M, Hickman RJ, Roch LM, Aspuru-Guzik A. Gryffin: An algorithm for Bayesian optimization of categorical variables informed by expert knowledge. *Applied Physics Reviews*. 2021;8(3):031406.
134. Li Z, Achenie LE, Xin H. An adaptive machine learning strategy for accelerating discovery of perovskite electrocatalysts. *ACS Catalysis*. 2020;10(7):4377-4384.
135. Savage T, Basha N, McDonough J, Matar OK, del Rio Chanona EA. Multi-fidelity data-driven design and analysis of reactor and tube simulations. *Computers & Chemical Engineering*. 2023;179:108410.

136. Begall MJ, Schweidtmann AM, Mhamdi A, Mitsos A. Geometry optimization of a continuous millireactor via CFD and Bayesian optimization. *Computers & Chemical Engineering*. 2023;171:108140.
137. Bardow A, Steur K, Gross J. Continuous-molecular targeting for integrated solvent and process design. *Industrial & Engineering Chemistry Research*. 2010;49(6):2834-2840.
138. Mantingh J, Kiss AA. Enhanced process for energy efficient extraction of 1,3-butadiene from a crude C4 cut. *Separation and Purification Technology*. 2021;267:118656.
139. Kim Y, Kim S, Lee B. Simulation of 1,3-butadiene extractive distillation process using N-methyl-2-pyrrolidone solvent. *Korean Journal of Chemical Engineering*. 2012;29:1493-1499.
140. BASF Butadiene Extraction Technology. Accessed June 24, 2024, https://www.basf.com/cn/en/products/chemicals/Intermediates/solutions/butadiene-extraction.html
141. Stavrou M, Lampe M, Bardow A, Gross J. Continuous molecular targeting-computer-aided molecular design (CoMT-CAMD) for simultaneous process and solvent design for CO_2 capture. *Industrial & Engineering Chemistry Research*. 2014;53(46):18029-18041.
142. Satola B. Searchable list of Aspen Plus components. Accessed June 02, 2022, https://chejunkie.com/knowledge-base/aspen-plus-components-list/
143. Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*. 2019:32.
144. Apicella A, Donnarumma F, Isgrò F, Prevete R. A survey on modern trainable activation functions. *Neural Networks*. 2021;138:14-32.
145. Deb K, Jain H. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation*. 2014;18(4):577-601.
146. Blank J, Deb K. Pymoo: Multi-objective optimization in python. *IEEE Access*. 2020;8:89497-89509.
147. Wang Z, Zhou T, Sundmacher K. *GitHub rsepository for* Data-driven integrated molecular and process design. https://github.com/zwang1995/data-driven-CAMPD
148. Brouwer T, Kersten SR, Bargeman G, Schuur B. Solvent pre-selection for extractive distillation using infinite dilution activity coefficients and the three-component Margules equation. *Separation and Purification Technology*. 2021;276:119230.
149. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*. 2011;12:2825-2830.

150. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*. 2020;17(3):261-272.
151. Wang Z, Zhou T, Sundmacher K. *GitHub repository for* Data-driven integrated molecular and process design using Bayesian optimization. https://github.com/zwang1995/BayesCAMPD
152. Wilmer CE, Leaf M, Lee CY, et al. Large-scale screening of hypothetical metal–organic frameworks. *Nature Chemistry*. 2012;4(2):83-89.
153. Bucior BJ, Rosen AS, Haranczyk M, et al. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Crystal Growth & Design*. 2019;19(11):6682-6697.
154. Moghadam PZ, Islamoglu T, Goswami S, et al. Computer-aided discovery of a metal–organic framework with superior oxygen uptake. *Nature Communications*. 2018;9(1):1378.
155. Chung YG, Gómez-Gualdrón DA, Li P, et al. In silico discovery of metal–organic frameworks for precombustion CO_2 capture using a genetic algorithm. *Science Advances*. 2016;2(10):e1600909.
156. Dubbeldam D, Calero S, Ellis DE, Snurr RQ. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Molecular Simulation*. 2016;42(2):81-101.
157. Tee LS, Gotoh S, Stewart WE. Molecular parameters for normal fluids. Lennard-Jones 12-6 Potential. *Industrial & Engineering Chemistry Fundamentals*. 1966;5(3):356-363.
158. Mayo SL, Olafson BD, Goddard WA. DREIDING: a generic force field for molecular simulations. *Journal of Physical Chemistry*. 1990;94(26):8897-8909.
159. Rappé AK, Casewit CJ, Colwell K, Goddard III WA, Skiff WM. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*. 1992;114(25):10024-10035.
160. Eggimann BL, Sunnarborg AJ, Stern HD, Bliss AP, Siepmann JI. An online parameter and property database for the TraPPE force field. *Molecular Simulation*. 2014;40(1-3):101-105.
161. Martin MG, Siepmann JI. Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes. *The Journal of Physical Chemistry B*. 1998;102(14):2569-2577.
162. Wick CD, Martin MG, Siepmann JI. Transferable potentials for phase equilibria. 4. United-atom description of linear and branched alkenes and alkylbenzenes. *The Journal of Physical Chemistry B*. 2000;104(33):8008-8016.

163. He Y, Krishna R, Chen B. Metal–organic frameworks with potential for energy-efficient adsorptive separation of light hydrocarbons. *Energy & Environmental Science*. 2012;5(10):9107-9120.
164. Liao Y, Zhang L, Weston MH, Morris W, Hupp JT, Farha OK. Tuning ethylene gas adsorption via metal node modulation: Cu-MOF-74 for a high ethylene deliverable capacity. *Chemical Communications*. 2017;53(67):9376-9379.
165. Mondal SS, Hovestadt M, Dey S, et al. Synthesis of a partially fluorinated ZIF-8 analog for ethane/ethene separation. *CrystEngComm*. 2017;19(39):5882-5891.
166. Liao P-Q, Zhang W-X, Zhang J-P, Chen X-M. Efficient purification of ethene by an ethane-trapping metal–organic framework. *Nature Communications*. 2015;6(1):8697.
167. Willems TF, Rycroft CH, Kazi M, Meza JC, Haranczyk M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*. 2012;149(1):134-141.
168. PyTorch. https://pytorch.org/
169. PyTorch Geometric. https://pytorch-geometric.readthedocs.io/
170. Wang Z, Zhou Y, Zhou T, Sundmacher K. *GitHub repository for* Machine learning for the discovery of metal–rganic frameworks. https://github.com/zwang1995/ML-MOF
171. Chen Y, Qiao Z, Wu H, et al. An ethane-trapping MOF PCN-250 for highly selective adsorption of ethane over ethylene. *Chemical Engineering Science*. 2018;175:110-117.
172. Choudhary K, DeCost B, Tavazza F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Physical Review Materials*. 2018;2(8):083801.
173. Ward L, Dunn A, Faghaninia A, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*. 2018;152:60-69.
174. Ong SP, Richards WD, Jain A, et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*. 2013;68:314-319.
175. MACCS keys. https://github.com/rdkit/rdkit/blob/master/rdkit/Chem/MACCSkeys.py
176. PubChem substructure fingerprint. https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt
177. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*. 2011;3:1-14.
178. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*. 2011;32(7):1466-1474.

179. Wang Z, Zhou T, Sundmacher K. *GitHub repository for* Interpretable machine learning for the discovery of metal–organic frameworks. https://github.com/zwang1995/IML-MOF
180. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*. 2020;2(1):56-67.
181. Dubbeldam D, Calero S, Vlugt TJ. iRASPA: GPU-accelerated visualization software for materials scientists. *Molecular Simulation*. 2018;44(8):653-676.
182. Leo MB, Dutta A, Farooq S. Process synthesis and optimization of heat pump assisted distillation for ethylene-ethane separation. *Industrial & Engineering Chemistry Research*. 2018;57(34):11747-11756.
183. Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge Structural Database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*. 2016;72(2):171-179.
184. Yancy-Caballero D, Leperi KT, Bucior BJ, et al. Process-level modelling and optimization to evaluate metal–organic frameworks for post-combustion capture of CO_2. *Molecular Systems Design & Engineering*. 2020;5(7):1205-1218.
185. Balandat M, Karrer B, Jiang D, et al. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in neural information processing systems*. 2020;33:21524-21538.
186. Park J-H, Beum H-T, Kim J-N, Cho S-H. Numerical analysis on the power consumption of the PSA process for recovering CO_2 from flue gas. *Industrial & Engineering Chemistry Research*. 2002;41(16):4122-4131.
187. Shen C, Liu Z, Li P, Yu J. Two-stage VPSA process for CO_2 capture from flue gas using activated carbon beads. *Industrial & Engineering Chemistry Research*. 2012;51(13):5011-5021.
188. Leperi KT, Snurr RQ, You F. Optimization of two-stage pressure/vacuum swing adsorption with variable dehydration level for postcombustion carbon capture. *Industrial & Engineering Chemistry Research*. 2016;55(12):3338-3350.
189. Wang Z, Zhou T, Sundmacher K. *GitHub repository for* CAMaterPD: Computer-aided material and process design. https://github.com/zwang1995/CAMaterPD
190. Streich T, Kömpel H, Geng J, Renger M. Secure the best benefit from C4 hydrocarbon processing-Part 1: Separation sequences. *Hydrocarbon Processing*. 2016:73-78.
191. Ren T, Patel M, Blok K. Olefins from conventional and heavy feedstocks: Energy use in steam cracking and alternative processes. *Energy*. 2006;31(4):425-451.

192. Leonzio G, Chachuat B, Shah N. Towards ethylene production from carbon dioxide: Economic and global warming potential assessment. *Sustainable Production and Consumption*. 2023;43:124-139.
193. Wang Y, Peh SB, Zhao D. Alternatives to cryogenic distillation: advanced porous materials in adsorptive light olefin/paraffin separations. *Small*. 2019;15(25):1900058.
194. Kestin J, Khalifa H, Wakeham W. The viscosity of five gaseous hydrocarbons. *The Journal of Chemical Physics*. 1977;66(3):1132-1134.
195. Mueller C, Cahill R. Mass spectrometric measurement of diffusion coefficients. *The Journal of Chemical Physics*. 1964;40(3):651-654.
196. Linstrom PJ, Mallard WG. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. National Institute of Standards and Technology.
197. Bachman JE, Reed DA, Kapelewski MT, et al. Enabling alternative ethylene production through its selective adsorption in the metal–organic framework Mn_2(m-dobdc). *Energy & Environmental Science*. 2018;11(9):2423-2431.
198. Webley PA, He J. Fast solution-adaptive finite volume method for PSA/VSA cycle simulation; 1 single step simulation. *Computers & Chemical Engineering*. 2000;23(11-12):1701-1712.
199. Jiang G-S, Shu C-W. Efficient implementation of weighted ENO schemes. *Journal of Computational Physics*. 1996;126(1):202-228.
200. Shampine LF, Reichelt MW. The matlab ode suite. *SIAM Journal on Scientific Computing*. 1997;18(1):1-22.
201. Lam SK, Pitrou A, Seibert S. Numba: A llvm-based python jit compiler. 2015:1-6.

List of Figures

List of Tables

List of Symbols

Greek Symbols

ρ	Density
μ	Viscosity
γ^{∞}	Infinite dilution activity coefficient
α^{∞}	Relative volatility at infinite dilution

Latin Symbols

C_p	Specific heat capacity
K	Langmuir adsorption constant
L	Length of adsorption column
MW	Molecular weight
N	Number of distillation stages
P	Pressure
P^0	Saturation vapor pressure
P_{H}	Adsorption pressure
P_{I}	Intermediate pressure
P_{L}	Desorption pressure
q	Adsorption loading
q_{sat}	Saturation adsorption loading
Q	Heat duty
R	Reflux ratio
S^{∞}	Selectivity at infinite dilution
S/F	Solvent-to-feed ratio
t_{ads}	Adsorption time
t_{des}	Desorption time
u_0	Feed velocity
$\boldsymbol{y}$	Property decision variables
$\boldsymbol{z}$	Process decision variables
ΔH_{vap}	Heat of vaporization

Declaration of Honor

I hereby declare that I produced this thesis without prohibited external assistance and that none other than the listed references and tools have been used.

In the case of co-authorship, especially in the context of a cumulative dissertation, the own contribution is correctly and completely stated. I did not make use of any commercial consultant concerning graduation. A third party did not receive any nonmonetary perquisites neither directly nor indirectly for activities which are connected with the contents of the presented thesis. All sources of information are clearly marked, including my own publications.

In particular I have not consciously:

- Fabricated data or rejected undesired results,
- Misused statistical methods with the aim of drawing other conclusions than those warranted by the available data,
- Plagiarized data or publications,
- Presented the results of other researchers in a distorted way.

I do know that violations of copyright may lead to injunction and damage claims of the author and also to prosecution by the law enforcement authorities.

I hereby agree that the thesis may need to be reviewed with an electronic data processing for plagiarism.

This work has not yet been submitted as a doctoral thesis in the same or a similar form in Germany or in any other country. It has not yet been published as a whole.

Magdeburg, 16.12.2024

Zihao Wang

Publication List

1. **Wang Z**, Zhou T, Sundmacher K. Data-driven integrated design of solvents and extractive distillation processes. *AIChE Journal*. 2023; 69(12): e18236.
2. **Wang Z**, Zhou T, Sundmacher K. Interpretable machine learning for accelerating the discovery of metal–organic frameworks for ethane/ethylene separation. *Chemical Engineering Journal*. 2022; 444: 136651.
3. **Wang Z**, Zhou Y, Zhou T, Sundmacher K. Identification of optimal metal–organic frameworks by machine learning: Structure decomposition, feature integration, and predictive modeling. *Computers & Chemical Engineering*. 2022; 160: 107739.
4. **Wang Z**, Zhou T, Sundmacher K. Molecular property targeting for optimal solvent design in extractive distillation processes. In: Kokossis AC, Georgiadis MC, Pistikopoulos E, eds. *Computer Aided Chemical Engineering*. Elsevier; 2023: 1247–1252.
5. **Wang Z**, Zhou T, Sundmacher K. A novel machine learning-based optimization approach for the molecular design of solvents. In: Montastruc L, Negny S, eds. *Computer Aided Chemical Engineering*. Elsevier; 2022: 1477–1482
6. Zhou T, **Wang Z**, Sundmacher K. A new machine learning framework for efficient MOF discovery: Application to hydrogen storage. In: Yamashita Y, Kano M, eds. *Computer Aided Chemical Engineering*. Elsevier; 2022: 1807–1812.
7. **Wang Z**, Zhou T, Sundmacher K. BayesCAMPD: Data-efficient and closed-loop integrated molecular and process design using Bayesian optimization. *To be submitted*.
8. **Wang Z**, Zhou T, Sundmacher K. Integrated adsorbent selection and process design: Simulation-based and data-driven optimization approaches. *To be submitted*.