Data Driven Synthesis Planning Applied to Zeolite Materials

by

Zach Jensen

Submitted to the Department of Materials Science and Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Materials Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2022

Author .....
Department of Materials Science and Engineering
December 6, 2021

Certified by .....
Elsa Olivetti
Esther and Harold E. Edgerton Associate Professor
Thesis Supervisor

Accepted by .....
Frances M. Ross
Chairman, Department Committee on Graduate Students

Data Driven Synthesis Planning Applied to Zeolite Materials

by

Zach Jensen

Submitted to the Department of Materials Science and Engineering
on December 6, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Materials Science and Engineering

Abstract

Materials discovery is critical for dealing with societal problems, but is a tedious process requiring substantial time and energy to accumulate knowledge. Computational techniques have accelerated understanding of material structure and properties, answering the question "What" materials to make for a specific application. These techniques have shifted the bottleneck in materials design to the synthesis and processing of materials, posing the question "How" to make a specified material. Zeolites are microporous, crystalline aluminosilicates described by this paradigm. Their relevance for chemical and "green" applications has led to sustained interest for many decades with substantial progress made in predicting hypothetical zeolites with databases of thousands of energetically favorable structures. However, only 255 of these structures have been synthesized and far fewer, approximately 20, are commercially viable pointing to synthesis as the major bottleneck in zeolite discovery and design. This thesis aims to improve the understanding of synthesis-structure relationships in zeolite materials through the use of data driven synthesis tools. It is guided by three questions: 1) How can zeolite synthesis data be automatically extracted on a large scale? 2) How can coupling of data-driven, first principles, and experimental approaches accelerate understanding of structure and processing relationships in zeolite materials? 3) In what ways can this data and discovered relationships be used to engineer improved zeolite materials?

Data driven synthesis planning requires large amounts of data to develop hypotheses about underlying trends and train machine learning (ML) models. The zeolite literature provides thousands of records of synthesis routes and the resulting zeolite structure but requires advanced information extraction techniques to obtain. This thesis utilizes and builds upon a natural language processing (NLP) pipeline to extract and format this data on realistic timescales. Algorithmic improvements for this pipeline along with additional components targeted specifically to unique linguistic components of zeolite literature are developed along with a researcher-computer interaction framework designed to optimize both extraction accuracy and efficiency by fixing mistakes made by the extraction algorithm. This extraction algorithm results in five, highly curated datasets related to zeolite synthesis representing the

largest collection of zeolite synthesis routes to the author's knowledge.

These datasets are used to study zeolite synthesis starting with organic structure directing agent (OSDA) design. Determining which OSDA molecule templates which zeolite structure is a difficult problem. The author extracts a dataset of known OSDA-zeolite pairs from the literature to study these relationships. Using an advanced featurization schemes for the OSDA, relationships between OSDAs and certain zeolite structures can be established. These relationships help answer thesis question two. A generative model is trained on the extracted data and validated through simulation to suggest potential OSDAs for a given zeolite structure providing tools to accelerate OSDA design addressing thesis question 3.

OSDAs are very important in zeolite formation but the rest of the hydrothermal variables also play a large role. This thesis utilizes failed experiment data to study the probability of zeolite crystallization and interprets the model results through Shapley values to determine impacts of specific hydrothermal synthesis variables. Using multi-fidelity data and Bayesian inference, zeolite crystallization curves are studied to determine nucleation and crystal growth behavior. Both of these tasks are done in pursuit of thesis question two. An additional generative model that predicts hydrothermal synthesis conditions given an OSDA-zeolite pair is developed presenting another tool to guide zeolite development looking to answer thesis question 3.

Finally, the thesis suggests high potential areas for future research and further exploration using the extracted data. It concludes with a brief commentary on the publication process and the necessity of data extraction.

Thesis Supervisor: Elsa Olivetti
Title: Esther and Harold E. Edgerton Associate Professor

Acknowledgments

I would first like to thank my thesis committee and sponsors for making this thesis possible.

In addition, many people have supported me through the thesis process. I would like to thank the following in no particular order:

- my advisor, Elsa, for her exceptional insights, guidance, flexibility, and commitment.
- my teachers through graduate school and before, for providing the necessary foundations and skills.
- my synthesis project team, Eddie, Alex, Chris, Elton, Rubayyat, Vineeth, and Kevin, for pushing the technical boundaries of what I thought was possible.
- my zeolite collaborators, Yuriy, Manolo, Soon, Daniel, and Sujay, for providing expertise and exposure to a fascinating field.
- my group members for stimulating insights, challenging questions, and fun outings.
- my friends, too many to name, for providing fun during much needed breaks.
- my parents, Kurt and Kathy, for their love and support as well as providing their kitchen for my "Wisconsin office."
- my brothers and soon to be sister, Nick, Chris, Mitch and Bekah, for their encouragement and necessary distractions.
- my new family, Dan, Darci, Michael, and Jenna, for support, fun distractions, and fascination with paint.
- and my wife, Lauren, for being my support, teammate, and motivation since way before this thesis and being by my side every step of the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Zeolites and Data Driven Synthesis

This chapter serves as an introduction to the thesis by providing the necessary background, motivation, and scientific gaps. First, background on zeolite structure, chemistry, and synthesis is provided followed by a summary of current applications of computational tools and data science in zeolite and inorganic synthesis as a whole. The chapter ends with an explicit explanation of the knowledge gaps this thesis aims to fill and an outline of proposed research activities to answer the thesis questions posed.

1.1 Zeolites

Catalysts are enabling materials; they are used in 95% of industrial chemical reactions[1] making them vital for progress in all chemical-related industries including food, energy, transportation, environmental conservation, healthcare and new material design.[2] This thesis examines zeolite materials, an important industrial heterogeneous catalyst. Zeolites are microporous, crystalline aluminosilicates with a wide range of applications in the chemical and petroleum industries even beyond

heterogeneous catalysis such as adsorption, separation, and ion exchange.[3,4] Beyond the chemical industries, zeolites have several important environmental and renewable energy applications including biomass conversion, CO_2 capture and conversion, NO_x abatement, and water purification.[5] The topological features of the zeolite such as pore structure, framework type, and heteroatom composition determine its performance in the target application.[6,7] As such, it is desirable to control the synthesis of a zeolite morphology specifically towards a target application.

1.1.1 A Brief History

Form the Greek words 'zein' for boil and 'lithos' for stone, zeolites were first discovered naturally by Cronstedt in 1756[8] and first synthesized by Sainte-Claire Deville in 1862.[9] However, most consider the founding fathers of the zeolite field to be Richard Barrer, who synthesized zeolite P and Q in 1948 by converting mineral phases in strong salt solutions at high temperatures,[10–12] and Robert Milton, who synthesized zeolites A, B, C, and X in the early 1950s using more reactive aluminosilicate gels.[13,14]

From there, the field expanded rapidly in the following decades. In 1961, two groups (Barrer/Denny and Kerr/Kokotailo) discovered the effect of using quaternary ammonium cations to template zeolite structures.[15,16] In 1967, researchers discovered these organic cations, specifically tetraethylammonium, could be used to make high-silica zeolites in the form of zeolite β.[17] The 1970s and 80s saw the invention of common industrial zeolite materials including ZSM-5[18] and silicate[19] along with the discovery of the fluoride ion as an alternative mineralizer,[20] advances in the understanding of zeolite crystallization,[21–23] breakthroughs in the use of polymeric templates,[24,25] and advances in the characterization of zeolite structures including NMR[26] and Raman spectroscopy.[27] The 1990s and 2000s ushered in computational modeling to the zeolite field[28–31] and heteroatom substitution for Si and Al including Ge,[32] Ga,[33] Mn,[34] Zn,[35] and Ti.[36] The most recent decade has seen many advances in synthesis strategies including predesigned organic templates,[37,38] tar-

geted heteroatom substitution,[39,40] topotactic transformations,[41,42] and inter-zeolite transitions.[43] To date, 255 unique zeolite topologies have been confirmed by the International Zeolite Association.[44]

1.1.2 Zeolite Chemistry and Structure

The fundamental unit of zeolites is the TO_4 tetrahedron, called the primary building unit, where T is a framework atom, typically Si or Al. Each pair of framework atoms is linked by an oxygen atom, building up a three-dimensional structure through corner sharing of the tetrahedron. While the placement of Si and Al on the framework sites is typically disordered, zeolites obey Löwenstein's Rule[45] which states that Al-O-Al linkages will not occur, allowing the Si/Al ratio in a zeolite to vary from 1 to infinity. To balance the negative charge associated with the AlO_4^- tetrahedron, alkali cations are incorporated into the structure along with absorbed water molecules giving an empirical formula $A_{x/n}[Si_{1-x}Al_xO_2]*mH_2O$ where x can vary from 0 to 0.5 and n is the charge of the cation.[46]

The linkage of primary tetrahedron results in secondary structures called secondary building units (SBU).[47] There are 23 known SBUs[48] containing a maximum of 16 T atoms that occur in zeolite frameworks. The zeolite unit cell always contains an integer number of SBUs.[49] These SBUs can be combined into larger structures called composite building units (CBU). These units are defined by the number of T atoms in each face of the CBU.[50]

Through combinations of CBUs, the macroscopic framework structure and pore geometry of the zeolite is defined. Channels extend infinitely through the zeolite structure and are defined to have at least one face large enough for guest species to pass through. Zeolites also have cavities,[51] which are similar to channels in the ability for guest molecules to penetrate the structure, but are not infinitely extended. Channels and cavities are described by the dimensionality of the pore and size of the rings comprising the openings. Typical zeolite channel opening range from 3-15

Å, allowing a wide selection of molecules to pass through.[44]

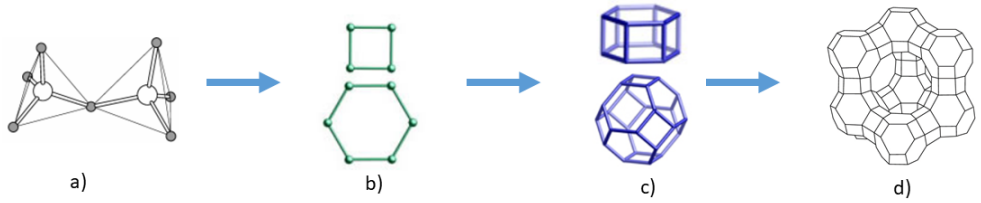


Figure 1-1. Various size scales of building units for the FAU zeolite structure. a) Primary building units, TO_4 tetrahedron linked by a corner sharing oxygen atom. b) Secondary building units (SBU), the two shown are 4 and 6. c) Composite building units (CBU), *d6r* and *sod*. d) Macroscale FAU structure (referred to as a framework)

Each unique zeolite framework is given a three letter code by the International Zeolite Association (IZA).[44] The code only refers to the connectivity of atoms in the zeolite framework. It does not define the composition, T atom distribution, cell dimensions, or symmetry. This leads to materials with the same three letter code but different compositions and framework density (FD), defined as the number of T atoms per 1000 $Å^3$. Besides providing an easy way to discriminate zeolite structures from dense aluminosilicates, FD is related to the pore volume and the volume within the zeolite accessible to outside species. Materials with different compositions and FDs can have different properties even if they share the same zeolite framework.[48] Within this thesis the terminology zeolite structure, zeolite phase, zeolite topology, and zeolite morphology all refer to the unique zeolite framework coded by the IZA.

Conventional zeolite frameworks use Si and Al as the framework elements. However, the zeolite field often considers materials with zeolite topologies and different chemistries as part of zeolite research. Technically, these structures are referred to a zeotypes but are often also referred to as zeolites. Common zeotype chemistries include aluminophosphates, borosilicates, germanosilicates, and titanosilicates. This thesis considers all types of zeolites and zeotypes and does not usually distinguish between the two.

1.1.3 Hydrothermal Synthesis of Zeolites

The most common way to synthesize a zeolite is with a hydrothermal approach. Hydrothermal synthesis crystallizes a material from solution at high temperature and pressure using water as the solvent. Zeolites are typically synthesized in a basic environment to mineralize the Si and Al source materials. Other reactants include alkali cations for charge neutralization and organic structure directing agent (OSDA) molecules.[52] A typical zeolite synthesis route is visualized in Figure 1-2. First, element sources such as Al and Si are mixed in a basic solution until they form an aluminosilicate gel. This gel is aged then placed in reactors and crystallized in a furnace. A convenient way to describe the zeolite synthesis space is breaking it down into two components: an organic piece and an inorganic piece.

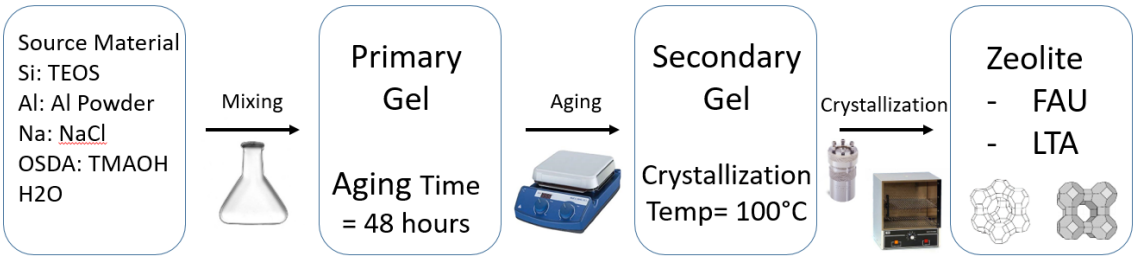


Figure 1-2. Example of a typical hydrothermal zeolite synthesis. The values and source materials come from ref.[53]

The organic piece is concerned with OSDA selection. The OSDA molecule acts as template for the structure of the zeolite. During synthesis, silicon and aluminum tetrahedra form around the OSDA constructing a zeolite structure with pores corresponding to the OSDA.[54] The relationship between the OSDA and the synthesized zeolite structure is complex and hard to predict, depending on the size, shape, flexibility, functional groups, and charge density of the OSDA.[55] Often these OSDA molecules are custom made in a pre-hydrothermal synthesis step, although it is also common to use commercial organic compounds for some zeolite structures.

The inorganic piece of zeolite synthesis encompasses all other synthesis variables outside of OSDA selection. There are many variables that play an important role in

determining if a zeolite will crystallize and which structure it will have, although the effects of all these parameters are not well understood. All the compositional variables including the amounts of Si, Al, alkali cations, OSDA, F, additional framework elements such as Ge, Ti, or B, and water play a role. Additionally, all the synthesis conditions, aging time and temperature, crystallization time and temperature, agitation of the reactor, reactor size, and pH are important factors to the nature of the product. Finally, the choice of source materials matters as well. Beyond choice of the OSDA, the choice of Si and Al source materials also affect the product formed as properties including precursor surface area and impurity concentration affects the nucleation and growth of zeolite phases.[56] There are many studies examine these synthesis variables' affect on the zeolite product, shown in Table 1.1, but most of the studies are limited to examining the effect of single variables on a constrained zeolite system.

Table 1.1. Selected examples of relationships between a synthesis parameter and a zeolite structure or property found in the literature.

Synthesis Parameter	Zeolite	Relationship	Ref.
Na Conc.	FAU/LTA	High [Na] favors LTA	[57]
Aging Time	FAU/LTA	Long aging times favor FAU	[57,58]
Si Source	FAU	Impurity level impacts nucleation and crystal size	[59]
Presence of Ge	IWW	Ge stabilizes small SBUs leading to large pores	[60]
OSDA	Global	Size, shape, charge distribution responsible for pore shape	[55,61]
Si/Al ratio	MFI	Particle size increase with Si/Al	[62]
Crystal Temp	*BEA	Crystal size increase with temp	[63]

1.1.4 Thermodynamics and Kinetics of Zeolite Synthesis

When considering the thermodynamics, typically only pure-silica systems are considered. Most zeolites are 6-14 kJ mol^{-1} less enthalpically stable than quartz, the most stable, ground-state phase.[64] Typical silica sources, including amorphous silicas and silica glasses, have very similar enthalpy values to zeolites. These sources will have higher entropy values making it unclear whether free energy will favor

the formation of zeolites from the precursors. Since the thermal energy at a typical synthesis temperature (100-200 °C) is 3.1 kJ mol^{-1} which is quite close to the energetic differences between silica precursors, zeolites, and quartz, thermodynamic arguments alone can rarely be used to predict the outcomes of a zeolite synthesis.[65]

Due to these small energetic differences in the reaction, kinetics plays a key role in zeolite formation. There are a few generally accepted phenomena that are consistent across all hydrothermal zeolite syntheses. Zeolite crystallization, shown in Figure 1-3, is described as a three step process: order evolution, nucleation, and crystal growth.[66]

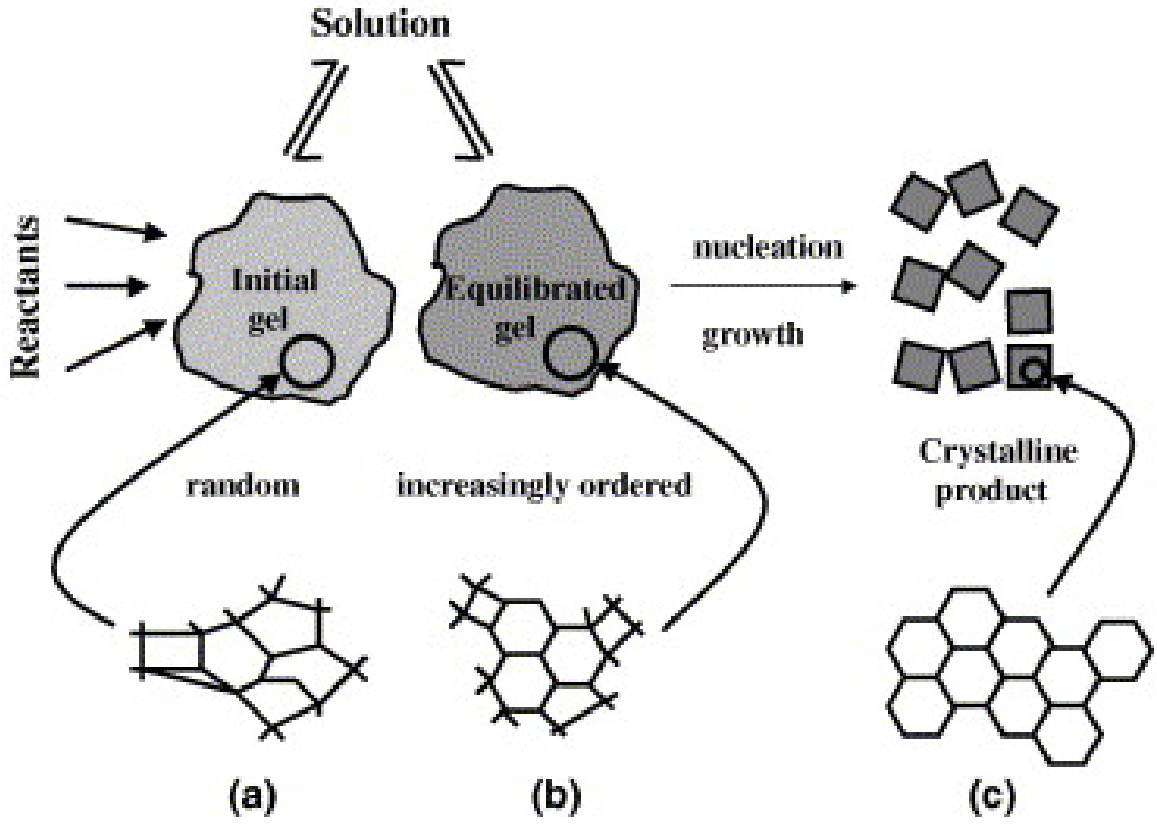


Figure 1-3. General kinetic assembly of zeolite crystals. From ref.[66]

When reactants are first mixed, a gel is formed called the primary amorphous phase. This phase is colloidal, non-equilibrated, and highly disordered. After a period of time, a steady-state intermediate phase forms called the secondary amorphous

phase. This phase exhibits local ordering, and equilibrium distributions of aluminosilicate anions are established.[67] Formation of the secondary amorphous phase is very important as the local ordering leads to clusters of critical size as described in classical nucleation theory. However, thus far in published research, the formation of this phase has proven difficult to measure and predict. While it is possible to measure when the secondary phase has formed with a combination of X-ray diffraction and solid state NMR,[68] it is time-consuming and not typically performed for intermediate materials. Another complication is that these kinetic processes can overlap making the boundaries between order evolution and nucleation unclear.[66]

Nucleation occurs when locally ordered clusters reach a critical size, consistent with nucleation theory. In contrast to dense materials, zeolites have much larger surface area, leading to the belief that other energetic terms play a role.[69] While there is some disagreement,[70] zeolite nucleation is typically considered to be heterogeneous, occurring on amorphous particles in the gel.[71] These amorphous particles are hard to characterize limiting understanding of the nucleation process. In addition, the forming and breaking of T-O-T bonds, responsible for the nucleus growth, are affected by cations in the system through complex interactions that are hard to predict.[66]

After nucleation, zeolite crystals grow into macroscopic sizes observable with visual inspection. Zeolites usually grow linearly in time with the rate dependent on temperature, concentration, cation type, and composition.[72] Zeolites have been observed to grow at a much slower rate than salts and simple molecular compounds.[73] A layer-by-layer adsorption model, limited by the nucleation of a new layer, agrees well with experimental observations made for several types of zeolites.[74]

The small energetic differences, role of kinetics, and large parameter space make predictions of zeolite products from synthesis variables difficult. Synthesis parameters will interact and correlate affecting the kinetics and products formed. Due to this complexity, researchers are typically limited to applying domain heuristics and

trial-and–error to synthesis which thus far has limited global, fundamental understanding.

1.2 Computational Advances in Zeolite Design

Computational tools have been widely used to study zeolite materials in the past decade. However, each of these approaches has left a gap in the research of zeolites that needs to be addressed.

1.2.1 Theoretical Zeolite Structures

By enumerating the possible combination of tetrahedral building units,[75] researchers are able to explore the zeolite structural space very efficiently, generating almost 3 million potential zeolite structures.[76] These structures have their pure silicon version's energy minimized using General Utility Lattice Program (GULP)[77] calculations with the Sanders-Leslie-Catlow[78] and van Beest-Kramer-van Santen[79] interatomic potentials. Of these structures, 314k (15%) have energy within +30 kJ/mol Si relative to quartz the same region as most of the known zeolites.[76,80] All of these structures are available in public databases,[76,80] and can be queried to find hypothetical zeolites suitable for selected applications.[81] These databases have existed for 10+ years, but only 255 zeolite structures have ever been synthesized[44] and far fewer are commercially available.[82] This highlights the synthesis bottleneck in zeolite development and deployment.

1.2.2 Simulation-based OSDA Design

Another recent methodological improvement in accelerating zeolite discovery is the computational design of OSDAs with simulation. In simulations, the OSDA is placed within the porosity of the zeolite.[38,83] Then molecular dynamic simulations calculate the energy of the system using either density functional theory or atomic potentials to determine forces.[84–87] Lowering this energy (often called "binding en-

ergy") indicates better templating ability of that OSDA with that specific zeolite structure and can be experimentally confirmed.[88,89] An additional improvement is selecting an OSDA to mimic the transition states of industrially relevant catalytic reactions.[90,91] This approach also uses DFT to calculate the energy of the transition state inside of the zeolite.

Although simulations are far faster than experimentally searching OSDA space, they are still too time consuming to be fully implemented into design pipelines. Replacing the DFT force calculation with machine learning can increase efficiency but the search space remains incredibly large.[92,93] To truly evaluate the search space, all suitable organic molecules need to be compared with all zeolite structures as binding energy alone cannot predict template suitability due to non-selective OSDAs. This is an incredibly vast space that cannot be comprehensively searched with simulation alone. Another compounding factor is the chemistry of zeolites. Almost all simulation approaches only consider zeolites in their pure silica forms, but chemistry can play a large role on the stability of the formed zeolite. Incorporating additional elements exponentially increases the search space again. To advance OSDA design and accelerate zeolite development, additional, more efficient and comprehensive techniques are needed to supplement current simulation efforts.

1.2.3 Early Machine Learning Studies

A few studies have used machine learning (ML) to study zeolite materials. Some of these early studies used ML to predict the zeolite structure from crystallographic data[94,95] and model mechanical properties.[96] A handful of studies have looked at zeolite synthesis with ML but have been limited to very small regions of the zeolite synthesis space[92,97,98] or OSDA-free synthesis[99] which ignores a very complex aspect of zeolite synthesis necessary for complete understanding of important systems. Other types of porous materials have successfully incorporated ML into research[100–103] indicating that ML has high potential to improve zeolite synthesis, but it will require large amounts of data and more complicated modeling that can ac-

count for the complexity of zeolite synthesis.

1.3 Data Driven Synthesis

While the computation prediction of materials structures and properties for inorganic materials has been largely successful resulting in many large structure-property databases,[104–106] the prediction of materials synthesis has lagged behind. While structure and property design is often driven by first principle approaches and mathematical models,[81,107,108] synthesis design is often done heuristically with domain knowledge accumulated over years in each material domain making the synthesis of novel materials difficult. This results in synthesis as the primary bottleneck for materials design in most systems.[109]

1.3.1 Organic Synthesis Planning

Attempts to generate synthesis routes for organic materials has proven effective. Building upon the concept of retrosynthesis, it is possible to treat organic chemistry as a graph with connections between organic molecule that can be transformed into each other.[110] Organic materials with specific properties can be generated,[111–113] and the organic chemistry graph searched until readily available organic precursors are found.[114,115] ML can be used to assist in accelerating organic synthesis through prediction of reaction outcomes,[116] generation of suitable reaction conditions,[117] and incorporation into high throughput experimental systems.[118] Models have also been created to gauge the synthesizability of molecules[119] which helps researchers screen for practical molecules for an application.

1.3.2 Inorganic Synthesis Planning

In contrast to organic molecules, inorganic materials rarely have well-defined intermediate materials, preventing the graph-based approach. Instead other techniques are required to accelerate materials synthesis. Some recent approaches include ML-

assisted high throughput experimentation[120–122] and *in situ* approaches that characterize[123–125] and manipulate reaction pathways.[126]

Another promising approach for inorganic synthesis planning uses natural language processing (NLP) to extract synthesis data from materials science journal articles.[127,128] Millions of chemistry and materials science articles exist describing synthesis routes, however this data is almost always unstructured, typically in the form of text and tables. NLP aims to understand, interpret, and structure human language.[129] With NLP, it is possible to extract the synthesis information contained within journal articles and convert it to more structured formats suitable for database construction, data mining, and ML. Figure 1-4 shows this extraction flow from scientific journal articles to synthesis databases to synthesis planning tools.

Several software tools exist to extract chemistry and materials science data from the literature using NLP.[130,131] Previous work in the Olivetti group has resulted in "Version 1" of a NLP pipeline that extracts operations, materials, amounts, conditions, and targets from synthesis sections of journal articles.[127,128,132] This pipeline has been successfully applied to study the synthesis conditions and trends of several inorganic systems including metal oxides,[127,128] titania,[127,133] MnO_2,[133] perovskites,[132] metal-insulator transition materials,[134] solid-state battery electrolytes,[135] and alternative cement binders.[136] In addition, other groups have used NLP to study the Curie and Néel temperatures for magnetic materials,[137] dye-sensitized solar cells,[138] nanomaterials,[139] metal-oxide frameworks,[140] and solid-state synthesis.[141,142] These pipelines can be combined with ML and experimental studies to inform the synthesis of novel systems and discover synthesis routes to new materials.

1.4 The Knowledge Gap

Researchers' lack of predictive ability to design synthesis routes aimed at achieving specific zeolite structures is a major bottleneck in zeolite discovery and deployment

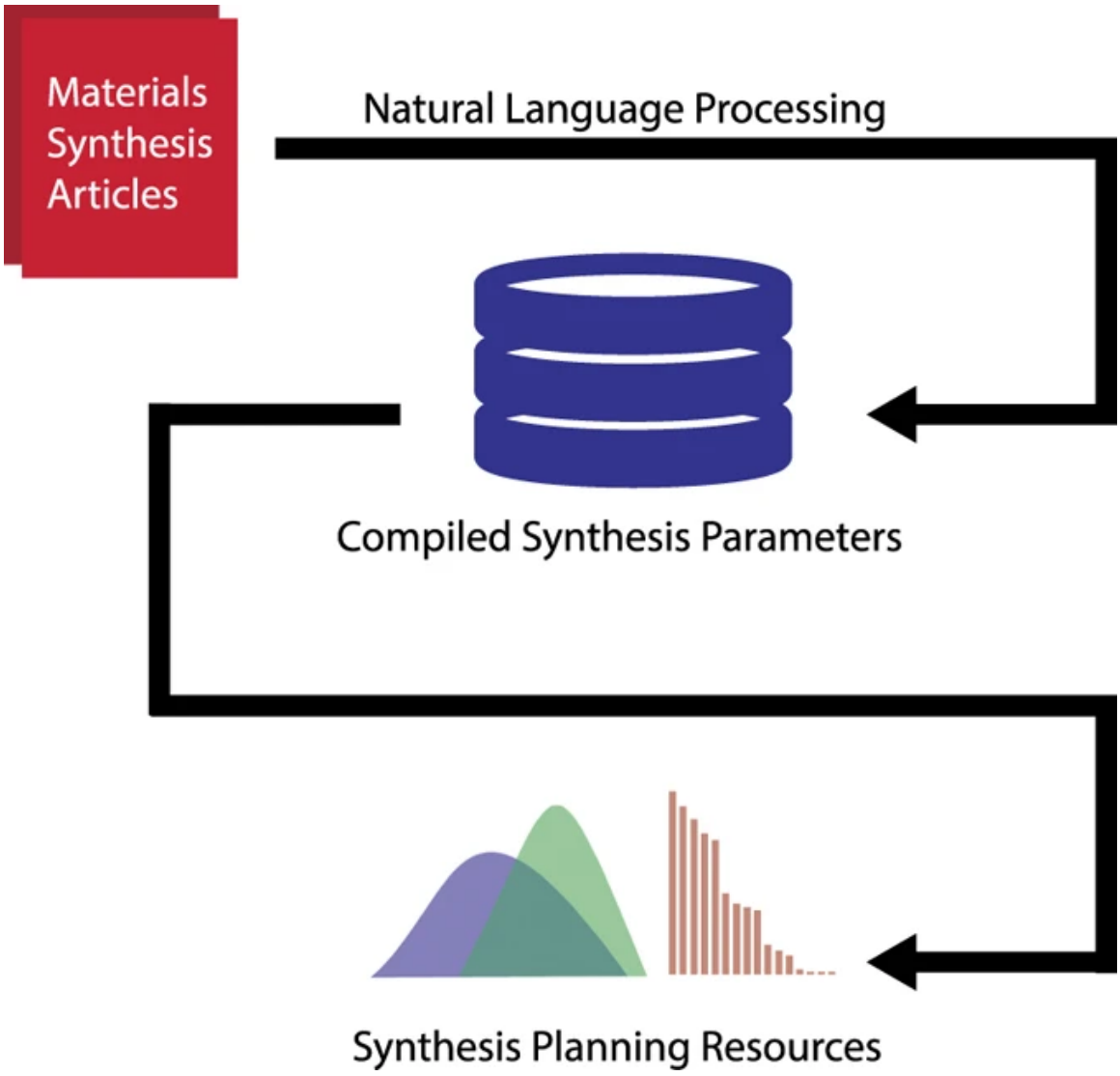


Figure 1-4. Automatic extraction pipeline of synthesis data from journal articles. Reproduced from ref.[128]

due to the knowledge gaps in the relationships between synthesis parameters and structure. While some advanced synthesis strategies for zeolites have been developed in recent years including transition state mimicking,[90,91] interzeolite transformations,[43,143] and high throughput flow synthesis,[144,145] there are still many gaps in the fundamental relationships between synthesis parameters and the resulting zeolite structure. ML shows great promise in modeling materials synthesis, but previous studies that apply ML to the synthesis of zeolites have had a limited effect due to the lack of data and simplification of the modeling. Automatically extracting data from

the literature will increase the zeolite dataset size by several orders of magnitude enabling the study of global zeolite synthesis, highlighting additional relationships between synthesis parameters and zeolite structures while suggesting potential synthesis pathways to new and optimized zeolite materials. This data combined with more complicated ML models will enable study of the complex interactions in zeolite synthesis between the composition, OSDA, reaction conditions, and individual precursors leading to a better understanding of zeolite crystallization.

1.5 Research Plan and Hypotheses

This thesis aims to answer three primary questions related to accelerating the synthesis of zeolite materials:

1. How can zeolite synthesis data be **automatically extracted** on a **large scale**?
2. How can coupling of **data-driven, first principles, and experimental** approaches accelerate understanding of **structure and processing relationships** in zeolite materials?
3. In what ways can this data and discovered relationships be used to **engineer improved zeolite materials**?

In pursuit of answering these questions, this thesis tests several hypothesis that are elaborated on and tested in the following chapters. First in chapter 3, an automatic extraction pipeline for zeolite synthesis data is proposed and tested. Next in chapter 4, relationships between OSDAs and zeolites found in the literature are explored, and models are developed to predict new OSDAs for a given zeolite. Finally in chapter 5, crystallization trends are explored across the inorganic zeolite parameter space, and models are developed to predict regions of the synthesis space that will produce zeolites.

Bibliography

[1] J. Hagen, *Industrial catalysis: a practical approach* (John Wiley & Sons, 2015).

[2] F. Schmidt, in *Basic Principles in Applied Catalysis* (Springer, 2004), pp. 3–16.

[3] M. E. Davis, Nature **417**, 813 (2002).

[4] C. Martínez and A. Corma, Coordination Chemistry Reviews **255**, 1558 (2011).

[5] Y. Li, L. Li, and J. Yu, Chem **3**, 928 (2017).

[6] S. M. Csicsery, Zeolites **4**, 202 (1984).

[7] J. Weitkamp, Solid state ionics **131**, 175 (2000).

[8] A. F. Cronstedt, *Rön och beskrifning om en obekant bärg art, som kallas Zeolites* (1756).

[9] H. d. S. Claire-Deville, Comptes Rendus **54**, 324 (1862).

[10] R. M. Barrer, Journal of the Chemical Society (Resumed) pp. 127–132 (1948).

[11] R. Barrer, L. Hinds, and E. White, Journal of the Chemical Society (Resumed) pp. 1466–1475 (1953).

[12] R. Barrer and C. Marcilly, Journal of the Chemical Society A: Inorganic, Physical, Theoretical pp. 2735–2745 (1970).

[13] R. M. Milton, *Molecular sieve adsorbents* (1959), uS Patent 2,882,243.

[14] R. M. Milton, *Molecular sieve adsorbents* (1959), uS Patent 2,882,244.

[15] R. Barrer and P. Denny, Journal of the Chemical Society (Resumed) pp. 971–982 (1961).

[16] G. T. Kerr and G. T. Kokotailo, Journal of the American Chemical Society **83**, 4675 (1961).

[17] R. L. Wadlinger, G. T. Kerr, and E. J. Rosinski, *Catalytic composition of a crystalline zeolite* (1967), uS Patent 3,308,069.

[18] R. J. Argauer and G. R. Landolt, *Crystalline zeolite zsm-5 and method of preparing the same* (1972), uS Patent 3,702,886.

[19] E. M. Flanigen, J. Bennett, R. Grose, J. Cohen, R. Patton, R. Kirchner, and J. Smith, Nature **271**, 512 (1978).

[20] E. M. Flanigen and R. L. Patton, *Silica polymorph and process for preparing same* (1978), uS Patent 4,073,865.

[21] D. W. Breek, *Zeolite molecular sieves: structure, chemistry and use* (1974).

[22] R. W. THOMPSON, Chemical Engineering Communications **4**, 127 (1980).

[23] B. Lowe, N. MacGilp, and T. Whittam, in *Proceedings of the 5th international conference on zeolites. Heyden, London* (1980), vol. 85.

[24] L. D. ROLLMANN (ACS Publications, 1979).

[25] R. H. Daniels, G. T. Kerr, and L. D. Rollmann, Journal of the American Chemical Society **100**, 3097 (1978).

[26] R. Bell (1989).

[27] F. Roozeboom, H. E. Robson, and S. S. Chan, Zeolites **3**, 321 (1983).

[28] C. Den Ouden, K. Datema, F. Visser, M. Mackay, and M. Post, Zeolites **11**, 418 (1991).

[29] K. D. Schmitt and G. J. Kennedy, Zeolites **14**, 635 (1994).

[30] A. Chatterjee and T. Iwasaki, The Journal of Physical Chemistry A **105**, 6187 (2001).

[31] D. Lewis, C. Freeman, and C. Catlow, The Journal of Physical Chemistry **99**, 11194 (1995).

[32] A. Corma, M. T. Navarro, F. Rey, and S. Valencia, Chemical Communications pp. 1486–1487 (2001).

[33] K. Balkus, Progress in Inorganic Chemistry **50**, 217 (2001).

[34] S. L. Suib, Current Opinion in Solid State and Materials Science **3**, 63 (1998).

[35] G.-Y. Yang and S. C. Sevov, Journal of the American Chemical Society **121**, 8389 (1999).

[36] J. Rocha and M. W. Anderson, European Journal of Inorganic Chemistry **2000**, 801 (2000).

[37] M. Moliner, F. Rey, and A. Corma, Angewandte Chemie International Edition **52**, 13880 (2013).

[38] R. Pophale, F. Daeyaert, and M. W. Deem, Journal of Materials Chemistry A **1**, 6750 (2013).

[39] Y. Li and J. Yu, Chemical reviews **114**, 7268 (2014).

[40] A. Corma, M. Díaz-Cabañas, J. Jiang, M. Afeworki, D. Dorset, S. Soled, and K. Strohmaier, Proceedings of the National Academy of Sciences **107**, 13997 (2010).

[41] T. Ikeda, Y. Akiyama, Y. Oumi, A. Kawai, and F. Mizukami, Angewandte Chemie International Edition **43**, 4892 (2004).

[42] H. Gies, U. Müller, B. Yilmaz, T. Tatsumi, B. Xie, F.-S. Xiao, X. Bao, W. Zhang, and D. D. Vos, Chemistry of Materials **23**, 2545 (2011).

[43] P. Eliášová, M. Opanasenko, P. S. Wheatley, M. Shamzhy, M. Mazur, P. Nachtigall, W. J. Roth, R. E. Morris, and J. Čejka, Chemical Society Reviews **44**, 7177 (2015).

[44] C. Baerlocher, http://www. iza-structure. org/databases/ (2008).

[45] W. Loewenstein, American Mineralogist: Journal of Earth and Planetary Materials **39**, 92 (1954).

[46] R. Xu, W. Pang, J. Yu, Q. Huo, and J. Chen, *Chemistry of zeolites and related porous materials: synthesis and structure* (John Wiley & Sons, 2009).

[47] E. Dempsen (1968).

[48] M. Moshoeshoe, M. S. Nadiye-Tabbiruka, and V. Obuseng, Am. J. Mater. Sci **7**, 196 (2017).

[49] W. Meier and C. Baerlocher, in *Structures and Structure Determination* (Springer, 1999), pp. 141–161.

[50] J. V. Smith, Chemical Reviews **88**, 149 (1988).

[51] P. C. D. C. on Colloid, L. McCusker, F. Liebau, and G. Engelhardt, Microporous and Mesoporous Materials **58**, 3 (2003).

[52] C. S. Cundy and P. A. Cox, Chemical reviews **103**, 663 (2003).

[53] G. Zhu, S. Qiu, J. Yu, Y. Sakamoto, F. Xiao, R. Xu, and O. Terasaki, Chemistry of materials **10**, 1483 (1998).

[54] M. E. Davis and R. F. Lobo, Chemistry of Materials **4**, 756 (1992).

[55] A. W. Burton and S. I. Zones, Stud. Surf. Sci. Catal **168**, 137 (2007).

[56] S. Mintova and V. Valtchev, Microporous and mesoporous materials **55**, 171 (2002).

[57] W. Fan, S. Shirato, F. Gao, M. Ogura, and T. Okubo, Microporous and mesoporous materials **89**, 227 (2006).

[58] S. Alfaro, C. Rodriguez, M. Valenzuela, and P. Bosch, Materials Letters **61**, 4655 (2007).

[59] K. E. Hamilton, E. N. Coker, A. Sacco Jr, A. G. Dixon, and R. W. Thompson, Zeolites **13**, 645 (1993).

[60] R. Castaneda, A. Corma, V. Fornés, F. Rey, and J. Rius, Journal of the American Chemical Society **125**, 7820 (2003).

[61] B. Lok, T. Cannan, and C. Messina, Zeolites **3**, 282 (1983).

[62] L. Shirazi, E. Jamshidi, and M. Ghasemi, Crystal Research and Technology: Journal of Experimental and Industrial Crystallography **43**, 1300 (2008).

[63] B. Modhera, M. Chakraborty, P. Parikh, and R. Jasra, Crystal Research and Technology: Journal of Experimental and Industrial Crystallography **44**, 379 (2009).

[64] P. M. Piccione, C. Laberty, S. Yang, M. A. Camblor, A. Navrotsky, and M. E. Davis, The Journal of Physical Chemistry B **104**, 10001 (2000).

[65] A. Corma and M. E. Davis, ChemPhysChem **5**, 304 (2004).

[66] C. S. Cundy and P. A. Cox, Microporous and mesoporous materials **82**, 1 (2005).

[67] S. Zhdanov, Advances in Chemistry Series p. 20 (????).

[68] M. Nicolle, F. Di Renzo, F. Fajula, P. Espiau, and T. Des Courieres, in *Proceedings from the Ninth International Zeolite Conference* (Elsevier, 1993), pp. 313–320.

[69] C. G. Pope, Microporous and mesoporous materials **21**, 333 (1998).

[70] T. Brar, P. France, and P. G. Smirniotis, The Journal of Physical Chemistry B **105**, 5383 (2001).

[71] R. Aiello, R. Barrer, and I. Kerr (ACS Publications, 1971).

[72] C. S. Cundy, B. M. Lowe, and D. M. Sinclair, Faraday discussions **95**, 235 (1993).

[73] J. Garside and M. B. Shah, Industrial & Engineering Chemistry Process Design and Development **19**, 509 (1980).

[74] J. R. Agger, N. Hanif, and M. W. Anderson, Angewandte Chemie **113**, 4189 (2001).

[75] M. Treacy, I. Rivin, E. Balkovsky, K. Randall, and M. Foster, Microporous and Mesoporous Materials **74**, 121 (2004).

[76] M. W. Deem, R. Pophale, P. A. Cheeseman, and D. J. Earl, The Journal of Physical Chemistry C **113**, 21353 (2009).

[77] J. D. Gale and A. L. Rohl, Molecular Simulation **29**, 291 (2003).

[78] K.-P. Schröder, J. Sauer, M. Leslie, C. Richard, A. Catlow, and J. M. Thomas, Chemical physics letters **188**, 320 (1992).

[79] B. Van Beest, G. J. Kramer, and R. Van Santen, Physical review letters **64**, 1955 (1990).

[80] R. Pophale, P. A. Cheeseman, and M. W. Deem, Physical Chemistry Chemical Physics **13**, 12407 (2011).

[81] L.-C. Lin, A. H. Berger, R. L. Martin, J. Kim, J. A. Swisher, K. Jariwala, C. H. Rycroft, A. S. Bhown, M. W. Deem, M. Haranczyk, et al., Nature materials **11**, 633 (2012).

[82] S. Zones, Microporous and mesoporous materials **144**, 1 (2011).

[83] D. Schwalbe-Koda and R. Gómez-Bombarelli, The Journal of Physical Chemistry C **125**, 3009 (2021).

[84] S. K. Brand, J. E. Schmidt, M. W. Deem, F. Daeyaert, Y. Ma, O. Terasaki, M. Orazov, and M. E. Davis, Proceedings of the National Academy of Sciences **114**, 5101 (2017).

[85] J. E. Schmidt, M. W. Deem, and M. E. Davis, Angewandte Chemie **126**, 8512 (2014).

[86] M. Moliner, P. Serna, Á. Cantín, G. Sastre, M. J. Díaz-Cabañas, and A. Corma, The Journal of Physical Chemistry C **112**, 19547 (2008).

[87] D. Schwalbe-Koda and R. Gómez-Bombarelli, The Journal of Chemical Physics **154**, 174109 (2021).

[88] D. W. Lewis, D. J. Willock, C. R. A. Catlow, J. M. Thomas, and G. J. Hutchings, Nature **382**, 604 (1996).

[89] A. W. Burton, G. S. Lee, and S. I. Zones, Microporous and mesoporous materials **90**, 129 (2006).

[90] E. M. Gallego, M. T. Portilla, C. Paris, A. León-Escamilla, M. Boronat, M. Moliner, and A. Corma, Science **355**, 1051 (2017).

[91] C. Li, C. Paris, J. Martínez-Triguero, M. Boronat, M. Moliner, and A. Corma, Nature Catalysis **1**, 547 (2018).

[92] F. Daeyaert, F. Ye, and M. W. Deem, Proceedings of the National Academy of Sciences **116**, 3413 (2019).

[93] D. Schwalbe-Koda, S. Kwon, C. Paris, E. Bello-Jurado, Z. Jensen, E. Olivetti, T. Willhammar, A. Corma, Y. Román-Leshkov, M. Moliner, et al., Science p. eabh3350 (2021).

[94] D. A. Carr, M. Lach-hab, S. Yang, I. I. Vaisman, and E. Blaisten-Barojas, Microporous and Mesoporous Materials **117**, 339 (2009).

[95] S. Yang, M. Lach-hab, I. I. Vaisman, and E. Blaisten-Barojas, The Journal of Physical Chemistry C **113**, 21721 (2009).

[96] J. D. Evans and F.-X. Coudert, Chemistry of Materials **29**, 7833 (2017).

[97] A. Corma, M. Moliner, J. M. Serra, P. Serna, M. J. Díaz-Cabañas, and L. A. Baumes, Chemistry of materials **18**, 3287 (2006).

[98] J. Manuel Serra, L. Allen Baumes, M. Moliner, P. Serna, and A. Corma, Combinatorial chemistry & high throughput screening **10**, 13 (2007).

[99] K. Muraoka, Y. Sada, D. Miyazaki, W. Chaikittisilp, and T. Okubo, Nature communications **10**, 1 (2019).

[100] P. G. Boyd, Y. Lee, and B. Smit, Nature Reviews Materials **2**, 1 (2017).

[101] S. Chong, S. Lee, B. Kim, and J. Kim, Coordination Chemistry Reviews **423**, 213487 (2020).

[102] K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit, Chemical reviews **120**, 8066 (2020).

[103] Y. Xie, C. Zhang, X. Hu, C. Zhang, S. P. Kelley, J. L. Atwood, and J. Lin, Journal of the American Chemical Society **142**, 1475 (2019).

[104] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, npj Computational Materials **1**, 1 (2015).

[105] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al., APL materials **1**, 011002 (2013).

[106] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, and A. Aspuru-Guzik, The Journal of Physical Chemistry Letters **2**, 2241 (2011).

[107] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nature materials **12**, 191 (2013).

[108] E. Haldoupis, S. Nair, and D. S. Sholl, Journal of the American Chemical Society **134**, 4313 (2012).

[109] B. G. Sumpter, R. K. Vasudevan, T. Potok, and S. V. Kalinin, NPJ Computational Materials **1**, 1 (2015).

[110] B. A. Grzybowski, K. J. Bishop, B. Kowalczyk, and C. E. Wilmer, Nature Chemistry **1**, 31 (2009).

[111] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, arXiv preprint arXiv:1509.09292 (2015).

[112] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, ACS central science **4**, 268 (2018).

[113] P.-C. Kotsias, J. Arús-Pous, H. Chen, O. Engkvist, C. Tyrchan, and E. J. Bjerrum, Nature Machine Intelligence **2**, 254 (2020).

[114] S. V. Ley, D. E. Fitzpatrick, R. J. Ingham, and R. M. Myers, Angewandte Chemie International Edition **54**, 3449 (2015).

[115] M. H. Segler, M. Preuss, and M. P. Waller, Nature **555**, 604 (2018).

[116] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen, ACS central science **3**, 434 (2017).

[117] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, and K. F. Jensen, ACS central science **4**, 1465 (2018).

[118] C. W. Coley, D. A. Thomas, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, et al., Science **365** (2019).

[119] C. W. Coley, L. Rogers, W. H. Green, and K. F. Jensen, Journal of chemical information and modeling **58**, 252 (2018).

[120] P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto, and B. Maruyama, npj Computational Materials **2**, 1 (2016).

[121] Y. Yao, Z. Huang, T. Li, H. Wang, Y. Liu, H. S. Stein, Y. Mao, J. Gao, M. Jiao, Q. Dong, et al., Proceedings of the National Academy of Sciences **117**, 6316 (2020).

[122] N. Szymanski, Y. Zeng, H. Huo, C. Bartel, H. Kim, and G. Ceder, Materials Horizons (2021).

[123] L. Soderholm and J. Mitchell, APL Materials **4**, 053212 (2016).

[124] S. Jiang, M. Hong, W. Wei, L. Zhao, N. Zhang, Z. Zhang, P. Yang, N. Gao, X. Zhou, C. Xie, et al., Communications Chemistry **1**, 1 (2018).

[125] R. Rao, D. Liptak, T. Cherukuri, B. I. Yakobson, and B. Maruyama, Nature materials **11**, 213 (2012).

[126] T. Liang, W. Liu, X. Liu, Y. Li, W. Wu, and J. Fan, Chemistry of Materials (2021).

[127] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, Chemistry of Materials **29**, 9436 (2017).

[128] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, and E. Olivetti, Scientific data **4**, 1 (2017).

[129] C. Manning and H. Schutze, *Foundations of statistical natural language processing* (MIT press, 1999).

[130] L. Hawizy, D. M. Jessop, N. Adams, and P. Murray-Rust, Journal of cheminformatics **3**, 1 (2011).

[131] M. C. Swain and J. M. Cole, Journal of chemical information and modeling **56**, 1894 (2016).

[132] E. Kim, Z. Jensen, A. van Grootel, K. Huang, M. Staib, S. Mysore, H.-S. Chang, E. Strubell, A. McCallum, S. Jegelka, et al., Journal of chemical information and modeling **60**, 1194 (2020).

[133] E. Kim, K. Huang, S. Jegelka, and E. Olivetti, npj Computational Materials **3**, 1 (2017).

[134] A. B. Georgescu, P. Ren, A. R. Toland, S. Zhang, K. D. Miller, D. W. Apley, E. A. Olivetti, N. Wagner, and J. M. Rondinelli, Chemistry of Materials **33**, 5591 (2021).

[135] R. Mahbub, K. Huang, Z. Jensen, Z. D. Hood, J. L. Rupp, and E. A. Olivetti, Electrochemistry Communications **121**, 106860 (2020).

[136] H. Uvegi, Z. Jensen, T. N. Hoang, B. Traynor, T. Aytaş, R. T. Goodwin, and E. A. Olivetti, Journal of the American Ceramic Society **104**, 3042 (2021).

[137] C. J. Court and J. M. Cole, Scientific data **5**, 1 (2018).

[138] C. B. Cooper, E. J. Beard, Á. Vázquez-Mayagoitia, L. Stan, G. B. Stenning, D. W. Nye, J. A. Vigil, T. Tomar, J. Jia, G. B. Bodedla, et al., Advanced Energy Materials **9**, 1802820 (2019).

[139] A. M. Hiszpanski, B. Gallagher, K. Chellappan, P. Li, S. Liu, H. Kim, J. Han, B. Kailkhura, D. J. Buttler, and T. Y.-J. Han, Journal of chemical information and modeling **60**, 2876 (2020).

[140] A. Nandy, C. Duan, and H. J. Kulik, Journal of the American Chemical Society (2021).

[141] O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, and G. Ceder, Scientific data **6**, 1 (2019).

[142] T. He, W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari, and G. Ceder, Chemistry of Materials **32**, 7861 (2020).

[143] C. Li, M. Moliner, and A. Corma, Angewandte Chemie International Edition **57**, 15330 (2018).

[144] T. Yoshioka, Z. Liu, K. Iyoki, A. Chokkalingam, Y. Yonezawa, Y. Hotta, R. Ohnishi, T. Matsuo, Y. Yanaba, K. Ohara, et al., Reaction Chemistry & Engineering **6**, 74 (2021).

[145] Z. Liu, K. Okabe, C. Anand, Y. Yonezawa, J. Zhu, H. Yamada, A. Endo, Y. Yanaba, T. Yoshikawa, K. Ohara, et al., Proceedings of the National Academy of Sciences **113**, 14267 (2016).

Chapter 2

Methodology

2.1 Introduction

This chapter describes the data science tools and techniques necessary for this thesis. It describes terminology, mathematical fundamentals, and algorithms behind these techniques while also describing their practical implementation. While not directly related to answering the questions posed in the thesis, this chapter describes techniques and methodology necessary to understand the thesis work.

2.2 Natural Language Processing

NLP is a sub-field of ML that focuses on gaining insight from human language. In the context of this thesis, NLP is used to extract information from scientific text related to zeolite synthesis.

2.2.1 Text Representation

Before running ML models, text data needs to be converted into a machine-readable format. While many methods exist to accomplish this task,[1–3] this thesis utilizes a class of algorithms called word embeddings. Word embeddings map a corpus of

text to a vector space in which words sharing linguistic similarity share a close proximity in the learned vector space. An example within a materials context are the words "ethanol" and "acetone". Both are commonly observed solvents and appear in very similar linguistic contexts indicating the words will be clustered within the word embedding vector space. Common word embedding algorithms utilized within this thesis include Word2Vec,[4] FastText,[5] ELMO,[6] and BERT.[7] These algorithms are trained in a "semi-supervised" manner that does not require labeled training data making them an efficient method for embedding materials-domain informed features in downstream modeling.[8,9]

2.2.2 Section Identification

Another important NLP task identifies which section of a paper a "chunk" of text belongs. This is a multi-class classification problem with possible labels of "Abstract", "Introduction", "Synthesis", "Non-Synthesis Methods", "Results", "Conclusion", and "null". This thesis takes a hybrid approach to determining which label has the highest probability described in detail in section 3.2.1. The ML component is a recurrent neural network with gated recurrent units.[10] The final layer is a softmax over the possible classes, and the chunk of text's label is assigned based on the highest predicted probability. This model is trained on approximately 1,000 manually annotated paragraphs.[11]

2.2.3 Named Entity Recognition

Named entity recognition (NER) is an NLP subtask that attempts to locate and classify important words and phrases within a chunk of text. In the context of materials synthesis, these words are the important components of the synthesis including the target material, precursors, operations, reaction conditions, amounts of materials, and material properties. NER is a sequence-to-sequence task. The model processes a sentence at a time, taking in the vectorized words and outputting a label for each word. This model uses a similar model architecture as Section

Identification with a recurrent neural network with gated recurrent units[10] and final softmax activation layer. The model is trained on 600 manually annotated synthesis section described in detail in section 3.2.1.

2.3 Machine Learning Applied to Materials Science

2.3.1 Materials Domain Considerations

ML applied in the materials domains presents many unique challenges. Materials synthesis data is often both sparse and scarce,[12] with very large, sparsely populated input spaces and relatively few instances compared with other common application domains of ML. ML models also typically need to be interpretable. It is often not enough to predict a structure or property with a "black box" model. Rather understanding and scientific advancement dictates the need to understand "why" a model makes certain decisions. A final highlighted consideration is the incorporation of domain knowledge into the modeling process. Materials synthesis is governed by thermodynamics and kinetics which provide meaningful relationships and equations. These types of domain considerations can often be incorporated into models through model selection and design as well as featurization schemes to develop physically meaningful models.[13]

2.3.2 Materials Informatics

Simply put, materials informatics is the application of ML to the materials domain. In addition to the unique attributes of materials data science highlighted above, a major component of materials informatics is the featurization of data. Since datasets in materials are typically small, featurization provides the model additional information to help learn useful models with less data. The best featurization scheme depends on the problem and dataset itself. A common approach for inorganic materials represents a composition as a weighted average of its atoms' atomic properties.[14–16] Featurization schemes for materials structures include aggregated

structural properties, crystal structure coordinates, and graph-based representations.[17,18] Organic molecule featurization (often referred to as ChemInformatics) is often based on the three-dimensional structure of the molecule either vectorization schemes including molecular fingerprints[19,20] and WHIM[21] or aggregation of molecular properties such as molecular volume, surface area, charge, etc.[22] Another commonly used organic featurization scheme is to feed the molecules as SMILES[23] strings directly to large neural network models where the molecule is mapped to a continuous latent space.[24]

2.3.3 Generative Modeling

Generative ML models describe the inverse of traditional ML models. Traditional ML models describe $\mathbb{P}(Y|X = x)$ whereas generative ML models describe $\mathbb{P}(X|Y = y)$ where X describes the input space and Y describes the target outcome.[25] Given a target outcome, generative models can be sampled to generate novel inputs from the conditional probability. This situation often more accurately describes the circumstances within synthesis design where the target structure is known and the model can generate synthesis routes conditional on that structure.

One common type of generative model used in this thesis is a conditional variational autoencoder (CVAE).[26] A CVAE model consists of an encoder modeling $\mathbb{P}(z|x)$ that converts input X into a latent space Z and a decoder modeling $\mathbb{P}(X|z, y)$ that converts the latent variable z back to the original input space conditioned on target variable y. This model is trained using two quantities in the loss function, the reconstruction loss which encourages the decoder to reconstruct data similar to the original training data and the Kullback-Leibler divergence (KL-divergence) which encourages the encoder to learn the correct Gaussian distribution over the training data.[27] Once trained, the decoder can be used to generate new input data.

2.3.4 Bayesian Inference

Bayesian inference derives from Bayes Rule which states:

$$\mathbb{P}(H|D) = \frac{\mathbb{P}(D|H)\mathbb{P}(H)}{\mathbb{P}(D)}$$

where $\mathbb{P}(H)$ is the prior probability or the original belief about hypothesis, H, $\mathbb{P}(D|H)$ is the likelihood or the probability of seeing the observed data, D given the current hypothesis, $\mathbb{P}(D)$ is referred to as evidence and does not change with H and is typically ignored in the inference process, and $\mathbb{P}(H|D)$ is the posterior probability and the main quantity of interest. It is the probability of the hypothesis conditioned on the observed data. $\mathbb{P}(H|D)$ is typically a better estimation of a process than $\mathbb{P}(H)$, the original prior distribution, since it also incorporates data into the prediction. $\mathbb{P}(H|D)$ can be updated sequentially as new data becomes available.[28] This approach is used in Chapter 5 to study the crystallization behavior of zeolites.

2.4 Conclusion

This chapter summarizes the important NLP and ML techniques utilized in the following chapters. Rather than formal definitions and mathematical rigor, the intention is to provide context for the methodologies and a solid foundation for understanding the results of this thesis.

Bibliography

[1] Y. Zhang, R. Jin, and Z.-H. Zhou, International Journal of Machine Learning and Cybernetics **1**, 43 (2010).

[2] H. M. Wallach, in *Proceedings of the 23rd international conference on Machine learning* (2006), pp. 977–984.

[3] A. Aizawa, Information Processing & Management **39**, 45 (2003).

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, arXiv preprint arXiv:1301.3781 (2013).

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Transactions of the Association for Computational Linguistics **5**, 135 (2017).

[6] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, arXiv preprint arXiv:1802.05365 (2018).

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, arXiv preprint arXiv:1810.04805 (2018).

[8] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, Chemistry of Materials **29**, 9436 (2017).

[9] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, and E. Olivetti, Scientific data **4**, 1 (2017).

[10] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, arXiv preprint arXiv:1406.1078 (2014).

[11] R. Mahbub, K. Huang, Z. Jensen, Z. D. Hood, J. L. Rupp, and E. A. Olivetti, Electrochemistry Communications **121**, 106860 (2020).

[12] E. Kim, K. Huang, S. Jegelka, and E. Olivetti, npj Computational Materials **3**, 1 (2017).

[13] B. Meredig, *Five high-impact research areas in machine learning for materials science* (2019).

[14] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, npj Computational Materials **2**, 1 (2016).

[15] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, et al., Computational Materials Science **152**, 60 (2018).

[16] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Computational Materials Science **68**, 314 (2013).

[17] T. Xie and J. C. Grossman, Physical review letters **120**, 145301 (2018).

[18] S. Li, Y. Liu, D. Chen, Y. Jiang, Z. Nie, and F. Pan, Wiley Interdisciplinary Reviews: Computational Molecular Science p. e1558 (????).

[19] D. Rogers and M. Hahn, Journal of chemical information and modeling **50**, 742 (2010).

[20] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, arXiv preprint arXiv:1509.09292 (2015).

[21] R. Todeschini and P. Gramatica, SAR and QSAR in Environmental Research **7**, 89 (1997).

[22] T. A. Halgren, Journal of computational chemistry **17**, 490 (1996).

[23] D. Weininger, Journal of chemical information and computer sciences **28**, 31 (1988).

[24] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, ACS central science **4**, 268 (2018).

[25] A. Y. Ng and M. I. Jordan, in *Advances in neural information processing systems* (2002), pp. 841–848.

[26] K. Sohn, H. Lee, and X. Yan, Advances in neural information processing systems **28**, 3483 (2015).

[27] D. P. Kingma and M. Welling, arXiv preprint arXiv:1312.6114 (2013).

[28] J. V. Stone (2013).

Chapter 3

Zeolite Data Extraction

The chapter's content is primarily derived from "A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction" by Zach Jensen et al. appearing in *ACS Central Science* in 2019.[1] The chapter discusses the intersection of automatic and manual data extraction along with data cleaning techniques. This chapter answers the first question guiding this thesis by developing techniques for accurate and efficient data extraction from the scientific zeolite literature.

3.1 Introduction

Before using data driven tools to study zeolite synthesis, zeolite synthesis data must be gathered in sufficient quantity and accuracy to describe the underlying trends in the synthesis. This data comes from experimental results. Previous studies have looked at individual zeolite systems to predict synthesis-structure relationships[2,3] but are limited by a lack of data. Global data driven approaches to study zeolite synthesis requires large amounts of data. Due to historical interest in zeolites both academically and industrial, there is over 60 years of scientific literature regarding zeolite synthesis. However, this data is very unstructured and heterogeneous, located in the text, tables, figures, and supplemental sections of journal articles re-

quiring advanced NLP and text-mining tools to extract and format the data. The goal of data extraction is to compile datasets on zeolite synthesis and use that data to gain insights into zeolites through data mining, illustrated in Figure 3-1. Useful synthesis information for zeolites consists of gel composition, aging and crystallization conditions, precursor selection, OSDA selection, and the resulting zeolite structure. This chapter discusses the extraction of zeolite synthesis data from the scientific literature for use in data driven studies of zeolite synthesis.

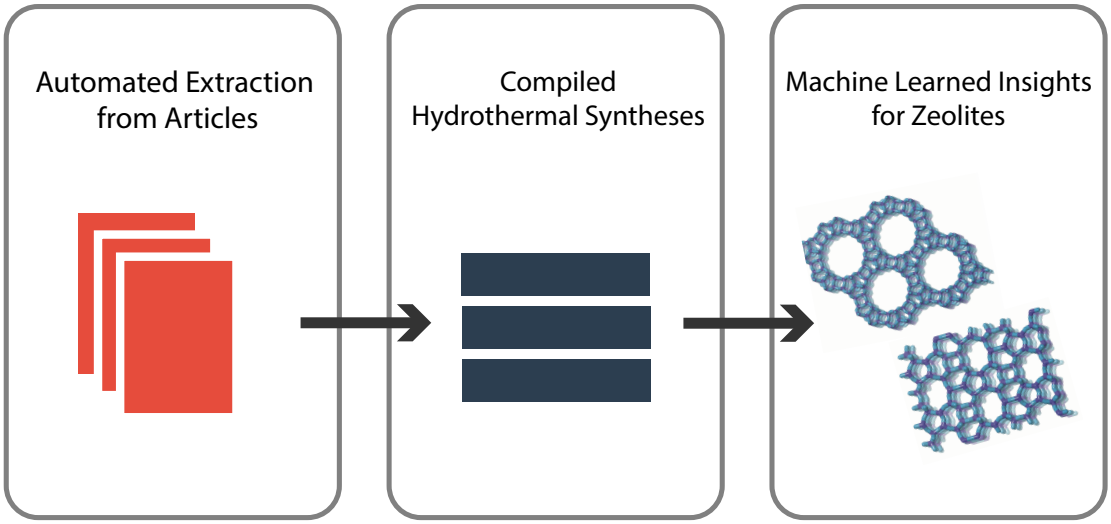


Figure 3-1. Demonstration of the research flow from automated data extraction to compilation of zeolite datasets to discovering insights on zeolite synthesis with data mining and machine learning.

3.2 Automatic Extraction Techniques

Comprehensive data extraction from the literature requires automated techniques to deal with the size of the corpus related to zeolite materials. A literature search using the Scopus data base[4] for the keywords "zeolite", "osda", "molecular sieve", and "aluminophosphate" returns approximately 130,000 journal articles at the time of this thesis. Determining which of these papers contains relevant zeolite synthesis data and extracting that data on realistic timescales requires utilizing the increased speed and pattern recognition of a computer.

3.2.1 Natural Language Processing Pipeline Improvements

A well documented NLP pipeline forms the backbone of the zeolite extraction process.[5–8] This thesis builds on and improves this pipeline in several important ways.

Publisher Specific Text Parsing

The first pipeline step after downloading a journal article is parsing the text from the HTML or XML file. This parsing involves removing unnecessary markup tags and correctly associating section labels with the corresponding paragraphs. Due to inconsistency in the formatting of the HTML or XML file, specific parsing rules for each publisher need to be established separately to increase the completeness of the parsed text.[9]

Hybrid Rule-based/Machine Learning Section Classifier

Section classification is an integral part of the NLP pipeline. It classifies each paragraph in a paper as a abstract, introduction, synthesis, characterization, results, conclusion, or Null type paragraph. This section classification has been improved and is comprised of two components: a rule-based classification built on the section headers of the paragraph and a ML classifier with inputs of the section's word embeddings and a context vector that describes the section's location within the paper. This hybrid model improves on both the accuracy and speed of section classification and is detailed in Mahbub et al.[10]

The classification algorithm first attempts to classify each section based on its section heading e.g. "Introduction", "Experimental", "Results", etc. linked during the parsing phase. The classification rules are designed to have very high precision so any section that has a somewhat ambiguous section header is passed on by the rule-based classification. Some sections cannot be classified with this rule-based approach e.g. sections with "Experimental" as section name which can refer to the synthesis or characterization label and sections from papers without section headings. These sections are classified using a recurrent neural network classifier with

the section's Word2Vec embeddings[11] pretrained on a materials science corpus and a context vector that describes the location of the paragraph within the paper as inputs.

This classifier is trained on approximately 1,000 manually annotated sections and tested on an additional set of 300 sections. The hybrid approach gives an average F1 score of 0.96 across all sections and a F1 score of 0.9 for synthesis sections, an improvement of 0.04 and 0.03 F1 score on all sections and synthesis sections respectively from previous pipeline versions.[8]

Token Annotation Schema

An NER models, referred to as token classification, identifies important aspects of a paper's synthesis section including target materials, lab operations, precursors, and synthesis conditions.[7] Token classification is enabled by annotation, hand-labeling of synthesis recipes that are used to train the NER model. These annotations are expensive, requiring Ph.D. levels of domain knowledge to perform accurately resulting in small datasets. Improving the quality of the annotation and the efficiency of annotation greatly impacts the overall token classification performance.[12] A new annotation procedure has been developed that splits the annotations by task rather than by paper, allowing annotators to focus on a single annotation task, for example all operations within all annotated papers. This approach increases annotation speed and consistency.[13]

3.2.2 Table Extraction

The zeolite literature extensively uses tables in describing experimental synthesis results. Tables are extracted from HTML and XML journal articles into a hierarchical JSON structure suitable for data mining. Rule-based approaches determine the correct position of the table's column and row headers along with any nested headers. Using annotated data from the main NLP pipeline, each word in the row and column headers are classified and the orientation of the table is determined by

the frequency of materials vs properties. "Entities" refer to the materials or samples within the table and "Attribute" refers to a property of that sample. The header that contains the most materials is considered the "Entities" orientation either row or column. Zeolite tables very commonly have samples on one axis and synthesis ratios on the other. Rule-based approaches to recognize the most commonly used synthesis ratios are also employed to determine the table orientation. Once the orientation is set, all the "attribute" values for each "entity" are extracted and stored in an easily accessible, hierarchical structure. Table captions and footers are also extracted as well as and sub/sup-script references within the table. Each reference is linked back to the corresponding footer entry which can be further mined for additional data.[1]

3.2.3 Regular Expression Matching

Some important components of zeolite synthesis have text representations that are difficult for the main pipeline to parse and extract properly. Regular expression (Regex) matching is used to deal with these problematic components. Compositional ratios are the main variables best approached with Regex. Regex is used to search the text of a zeolite paper to scan for commonly used elements (Si, Al, H2O) and common separators (:, /) in compositional ratios. After locating the ratio, the type of numeric value (number, range, or variable) is determined. If the type is a number, it can be assumed that every synthesis route found in the paper will have that value. If the type is a variable (x, y, etc.), typically that means the actual values of each tested sample are given in one of the paper's tables. If the type is a range, it often indicates that the unique synthesis routes are given elsewhere in the paper, either in table or text.[1]

Another usage of Regex is in determining zeolite structures and OSDAs. While the main pipeline can recognize target materials and precursors broadly, it does not know how to differentiate between zeolite materials and precursors specific to zeolite synthesis mainly the OSDA. However, zeolites and OSDAs both have consistent

and predictable naming schemes making Regex suitable for identifying them from other materials. Zeolites structures are given three capital letter codes by the IZA which refers to a unique crystal structure.[14] In addition, individual zeolite materials often follow the naming convention XXX-Y where X refers to capital letters and Y is a number. This consistency in naming can be exploited with Regex to identify zeolites mentions in the literature. OSDAs also follow typical naming patterns for OSDAs. There are many sub-strings that are common across the OSDA literature including "amine", "ammonium", "cyclo" and "N,N". In addition, many OSDAs contain the sub-string Y1,Y2 where both Y1 and Y2 are integers. This type of sub-string is rarely found in inorganic materials.[15]

3.2.4 Automatic Filtering with Domain Knowledge

Manually verifying the accuracy of the extracted data is time consuming so any automatic filtering is very valuable. Basing automatic filters on domain knowledge is a useful and logical choice. Bounds can be set on specific variables or pairs of variables. For example, the molar ratio of the fluoride ion (F^-) and the OSDA is often close to one since the F^- and OSDA charge balance to neutral in the case of a mono-charged OSDA (seen in Figure 3-2.[16–18] Large deviations in this molar ratio may indicate erroneous data extraction. Another example is filtering out zeolite structure names that often appear in other scientific contexts such as "Beta" and "Omega". A final example filters based on the quantity of information extracted from a paper. Both small and large amounts of extracted data can indicate potential errors. Only extracting a single zeolite or precursor can indicate information is missing. Often it indicates additional information regarding the synthesis is located in the Supplementary Information section of the paper which is not directly accessible to the extraction algorithm. Similarly, extracting large amounts of data can also indicate problems especially around OSDAs. The extraction algorithm performs poorly at distinguishing OSDAs from other types of organic molecules. Articles that contain many organic molecules typically are not relevant zeolite papers but rather

focus on the synthesis of many organic molecules.

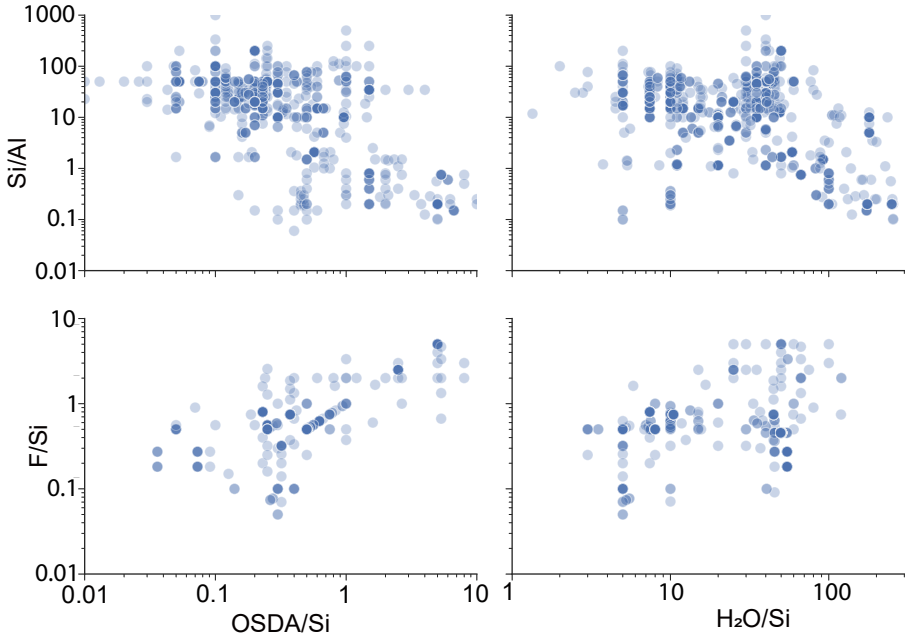


Figure 3-2. Pairwise plot of gel composition data automatically extracted from zeolite tables found in literature.

However, automatic filtering should be handled with caution. Zeolites are traditionally synthesized with Si/Al > 1, OSDA/Si < 1, and H2O/Si > 100 as seen in Figure 3-2, but these limits can be exceeded especially in the case of zeotype systems including aluminophosphates and silicogermanates.[19] Rather than setting rigid, deterministic filters, it is often better in practice to flag unexpected values for manual review.

3.3 Human Computer Interaction

Extracting a dataset with 100% accuracy (relative to a manually extracted dataset) necessitates manually checking the data. Even the best ML models will produce errors. Obtaining datasets with high accuracy while keeping reasonable extraction efficiency high requires optimal interaction between the extraction algorithms and human data checking. An example of the benefits of a similar human-computer

interaction is chess, where average-players with average computers can routinely defeat world-class human players or supercomputers if the average player-computer pairing understands how to optimally interact.[20,21] Similarly, well-designed human computer interaction in data extraction can accomplish more than a human or computer alone.

Optimal cooperation in data extraction requires comprehensive knowledge of both the strengths of extraction algorithms and the extraction domain. Extraction algorithms are extremely fast and reliable at pattern recognition. Combining this speed with domain knowledge allows extremely fast extraction of the necessary synthesis information. However, the computer is very poor at understanding the broader context of the extracted data. This is where the researcher participates supplying the necessary context to make the correct associations within the data.

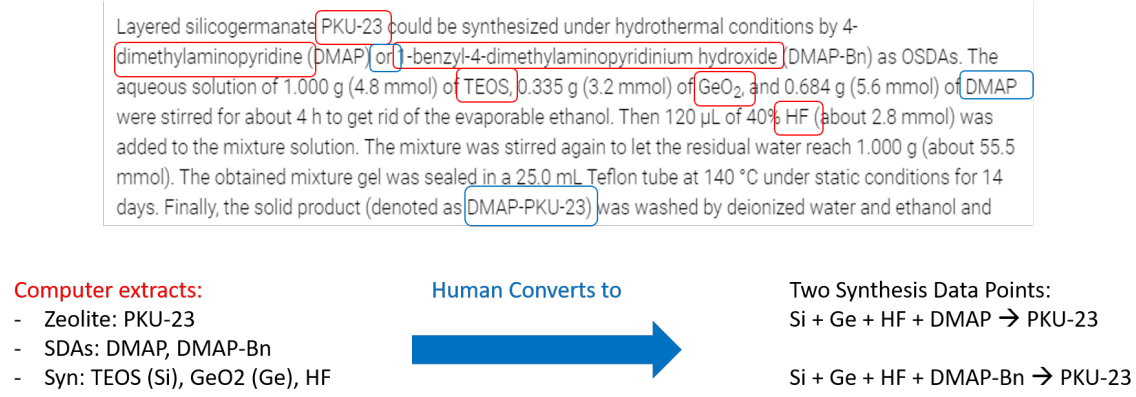


Figure 3-3. Example of workload split in data extraction between the computational extraction algorithm (red) and the human checker (blue). Text comes from ref.[22]

A simple example associates OSDAs with the proper zeolite structure, a simple task when an article only contains a single OSDA and single zeolite but oftentimes that is not the case. Figure 3-3 demonstrates a typical example of this workload split by highlighting a zeolite synthesis paragraph[22] with the tasks performed by the computer versus the human. First, the computer algorithms extracts all the identifiable pieces of the synthesis including the zeolite structure (PKU-23), the OSDAs (dimethylaminopyridine and 1-benzyl-4-dimethylaminopyridinium), and the pre-

cursors (TEOS, GeO_2, HF). However, the computer struggles to understand how the information connects so the researcher provides that context to maintain an accurate dataset. In this example, the researcher determines that the two OSDAs make the same zeolite structure, PKU-23, with the same precursors rather than a different interpretation, i.e. both OSDAs are used in a single, dual-OSDA synthesis route. Using this approach leads to efficient and accurate datasets. Section 3.6 attempts to quantify the effectiveness and efficiency of this approach.

3.4 Data Featurization

Most of the extracted data is not directly suitable for use in data analysis and ML and first needs to undergo featurization. Important data components requiring featurization include the OSDA, zeolite structures, and inorganic precursors.

3.4.1 OSDA Featurization

Multiple OSDA featurization schemes are used in this thesis. The first step in any featurization normalizes all of the literature given OSDA names to canonical SMILES strings. This is done through a combination of Python packages including ChemSpider,[23] PubChem,[24] and CIRpy.[25] SMILES strings form the basis for more advanced featurization.

The first featurization approach is to calculate OSDA features using DFT. SMILES strings are converted to 3-dimensional representations using molSimplify[26] with the Kier Flexibility index[27] and atomic coordinates obtained through force field optimization. The volume and surface area are calculated using ORCA 4.1.[28]

Another featurization scheme accounts for the different conformations of a molecule. For each OSDA, 2,000 gas phase conformers are generated, embedded, and optimized with the MMFF94 force field[29] using the Python package RDkit.[30] Average descriptors across the 2,000 conformers are calculated as well as descriptors cor-

responding to the conformer with minimum energy. Descriptors calculated include surface area, volume, number of rotatable bonds, and molecular charge. Weighted holistic invariant molecular (WHIM) descriptors[31] are also included which compress three-dimensional information about a molecules size, shape, symmetry, and atom distribution into a one-dimensional vector of fixed length. Additionally, the nConf20[32] descriptor of flexibility is also calculated.

The final featurization scheme utilized is neural network featurization used in generative modeling and compression of the OSDA feature space. Each character in a SMILES string is one-hot encoded and fed through three convolutional neural network layers. The representation is then flattened into a one-dimensional vector of fixed size. This vector can then be utilized in generative modeling or as input to traditional machine learning models.

3.4.2 Zeolite Featurization

Zeolite materials extracted from the literature are first normalized to their IZA structural code using string matching with manual confirmation if no match is found. Each structure is then featurized with a variety of structural properties scraped from the IZA website[14] including framework density, maximum ring size, channel dimensionality, maximum included volume of a sphere, accessible volume, maximum channel cross-sectional area, and minimum channel cross-sectional area. For modeling, these features are combined into a single vector. Unsuccessful synthesis products are also extracted and normalized to account for author specificity e.g. "dense" vs "cristobalite". Unsuccessful synthesis products are given a "-1" for each of the properties listed above.

3.4.3 Precursor Featurization

Extracted precursors are normalized into broader buckets that describe which element the precursor is providing to the synthesis e.g. Ludox AS-30 becomes col-

loidal silica. Each synthesis route is represented by a one-hot encoded vector of the shape, number of elements x maximum(precursors per element) that represents all the precursors used for a synthesis route.

3.5 Extracted Datasets

The following section describes five major datasets that were extracted during this thesis. The Germanium zeotype and interzeolite conversion dataset applications are discussed in section 3.7.1 and 3.7.2 respectively. The OSDA-zeolite pair dataset is examined in chapter 4 while the inorganic zeolite and zeolite crystallization datasets are utilized in chapter 5.

3.5.1 Germanium Zeotype Dataset

This data set contains hydrothermal synthesis routes for Germanium-containing zeotypes. Important characteristics include:

- 238 papers, 1,638 synthesis routes
- Successful and Unsuccessful Synthesis Routes (1,214 vs 424)
- Synthesis - quantitative gel composition, crystallization temperature and time
- OSDA - given name, SMILES string, DFT Calculated Features (volume, surface area, Kier flexibility index)
- Zeolite - given name, IZA code, IZA featurization
- Scope - Germanium zeotype literature
- Availability - `https://github.com/olivettigroup/table_extractor`

3.5.2 Interzeolite Conversion Dataset

This data set contains successful interzeolite conversions. Important characteristics include:

- 211 papers, 243 unique interzeolite transformations
- Successful transformations only
- Features - Type of transformation, starting zeolite, transformed zeolite
- Scope - Interzeolite transformations
- Availability - `https://www.nature.com/articles/s41563-019-0486-1?proof=t#Sec9`

3.5.3 OSDA-Zeolite Pair Dataset

This data set contains OSDA-zeolite pairs found in the zeolite literature. Important characteristics include:

- 1,384 papers, 5,663 synthesis routes
- 758 unique OSDA molecules, 205 zeolite structures
- Synthesis - qualitative
- OSDA - given name, SMILES string, Full RDKit conformer featurization
- Zeolite - given name, IZA code, IZA featurization
- Scope - entire zeolite literature
- Availability - `https://github.com/olivettigroup/OSDA_Generator`

3.5.4 Inorganic Zeolite Dataset

This data set contains full hydrothermal synthesis routes across entire zeolite literature. Important Characteristics include:

- 3,096 papers, 23,925 synthesis routes
- Successful and Unsuccessful Synthesis Routes
- Synthesis - quantitative gel composition, aging conditions (time and temperature), crystallization conditions (time, temperature, and rotation)

- OSDA - given name, SMILES string, Full RDKit conformer featurization
- Zeolite - given name, IZA code, IZA featurization, some properties (Si/Al product, percent crystallinity, crystal size)
- Scope - entire zeolite literature
- Availability - To be made public with publication.

3.5.5 Zeolite Crystallization Dataset

This data set contains crystallization data across the zeolite literature. Important Characteristics include:

- 128 papers, 291 crystallization curves, 1,986 data point
- Synthesis - quantitative gel composition, aging conditions (time and temperature), crystallization conditions (time, temperature, and rotation)
- Crystallization - kinetic parameters fit to experimental data, a, b, and k_g
- OSDA - SMILES string
- Zeolite - IZA code, percent crystallinity
- Scope - entire zeolite literature
- Availability - To be made public with publication.

3.6 Quantifying Thesis Question 1

What does it mean to extract zeolite synthesis data automatically? What does it mean to extract zeolite synthesis data on a large scale? Both of these questions are difficult to answer in a quantitative way but make up the fundamental aspects of the first thesis questions. The following are suggested ways of evaluating these questions. All suggestions are rather imprecise and should be interpreted as suggested benchmarks.

As discussed in section 3-3, accurate data extraction requires some manual cleaning. One way to quantify the level and benefit of automation in the extraction process is to compare the time spent cleaning the data compared with a hypothetical situation in which all the extraction is performed manually. Datasets 3.5.3 and 3.5.4 required the majority of the data cleaning which occurred over approximately a five month period. Both datasets were reduced from the original 130,000 papers relating to zeolites to approximately 10,000 by the automatic extraction algorithms. Assuming a typical work week of 40 hours, the dataset cleaning took approximately 850 hours or about 5 minutes per paper. In contrast, if each paper had to be fully read, assuming an average time of 30 minutes would result in about 5,000 hours or 625 eight hour work days to extract the same information. This difference represents approximately an 80% increase in extraction efficiency using automated techniques.

The "large scale" nature of the question is easier to answer. Datasets 3.5.3 and 3.5.4 are the largest known datasets related to zeolite synthesis at 5,663 and 23,925 synthesis routes respectively. For comparison, the latest *Verified Zeolite Synthesis*[33] dataset contains 109 synthesis routes. For additional comparison, a recent effort to extract synthesis conditions automatically resulted in a dataset of approximately 20,000 synthesis routes of solid-state synthesis, a much broader field than zeolites.[9]

A final consideration is the comprehensiveness of the data extraction. Are there ways we can quantify how much data may have been missed? One metric is examining the number of zeolites that end up in the final dataset. 203 of 255 known zeolite structures are found in dataset 3.5.4. Of the missing 52, 16 are natural zeolites with no synthetic analogue, leaving the rest of the 36 or 14% of the known zeolites as missed in the extraction. The majority of the missing zeolites are unique zeotype chemistries with only one or two known synthetic routes. Another way to evaluate comprehensiveness is to attempt to quantify how much additional data may exist that is not captured by the approach. Articles related to zeolite synthesis can be found with a search service such as Web of Science[34] and then compared

to the extracted datasets to determine if the articles are in the dataset, missing from the dataset and contain useful zeolite synthesis information, or missing but correctly omitted from the dataset. By combining articles found in the dataset and articles correctly omitted, a Bernoulli experiment can be performed by sampling the articles from Web of Science. An article sampled from the Web of Science list that contains useful zeolite synthesis information but is missing from the dataset is a success within the context of the Bernoulli experiment. Table 3.1 shows the results of sampling 20 articles randomly from the 1,000 most relevant search results using "zeolite" and "synthesis" as keywords. 7 of the 20 sampled articles (35%) are "missing" from the dataset. However, all but one are not the fault of the extraction algorithm but rather institutional constraints outside of the author's control. Ignoring these uncontrollable elements, only 1 out of 14 sampled articles (7%) are "missing" from the data. Running a statistical test with the null hypothesis that no less than 5% of available articles are missing from the dataset and the alternative hypothesis that greater than 5% of articles are missing results in a p-value of 0.51 indicating the null hypothesis cannot be rejected. Constructing the statistical test with a reversed hypothesis would provide more certainty but requires far more sampling. All of the metrics described in this section are inexact and have author bias built in, however they start to quantitatively answer the first question posed in the thesis.

3.7 Early Applications

This section describes two early applications of extracted zeolite data. These applications and the following studies in the rest of the thesis fall into two broad categories: synthesis-structure predictions and theory/simulation validation.

Table 3.1. Sampling journal articles to determine comprehensiveness of the data extraction of dataset 3.5.4. 20 samples taken randomly from the 1000 most relevant search results in Web of Science[34] looking for "zeolite" and "synthesis" within an article's topics. T&D stands for Text and Data Mining Agreement.

	Title	Result	Comments	Ref
1	CO2 adsorption behavior of microwave..	In Dataset		[35]
2	Co-based MOR/ZSM-5 composite...	Nothing New		[36]
3	An efficient one-pot synthesis of..	Nothing New		[37]
4	Fe3O4@zeolite-SO3H as a magnetically..	Nothing New		[38]
5	Seed-Assisted, OSDA-Free, Solvent-Free..	Missed	No access	[39]
6	Synthesis of a heulandite-type zeolite..	Missed	PDF only	[40]
7	Controlled and rapid growth of MTT..	In Dataset		[41]
8	One pot fusion route for the synthesis..	Missing	No access	[42]
9	Synthesis of Na-A zeolite from 10 A..	Nothing New		[43]
10	Expansion of the ADOR Strategy for..	In Dataset		[44]
11	Synthesis of ZSM-5 zeolite with small..	Missing	No T&D	[45]
12	Hydrogenation of carbon monoxide..	Nothing New		[46]
13	The effect of ultrasound on Na–A..	Missing	No T&D	[47]
14	Fast synthesis of submicron ZSM-5..	Missing		[48]
15	Enhancement of thermal conductivity..	Nothing New		[49]
16	Fast synthesis of thin Silicalite-1 zeolite..	Nothing New		[50]
17	Synthesis of Nanocrystalline MFI..	Nothing New		[51]
18	Synthesis and Characterization of..	Missing	No T&D	[52]
19	Synthesis of Hollow Zeolite Composite..	Nothing New		[53]
20	Silicalite-1 Encapsulated Fe Particles..	Nothing New		[54]

3.7.1 Synthesis-Structure Predictions for Germanium-containing Zeotypes

Germanium addition into zeolites is responsible for the synthesis of many new zeolite structures, especially extra-large zeolite frameworks, over the past two decades.[55] Because of germanium's known correlation with the structure of zeolites, germanium zeotypes are an example that allows for demonstration of the usefulness of the extraction pipeline and basic synthesis-structure predictions across a large composition and OSDA space. Figure 3-4 demonstrates the research process for this section. Data is extracted and combined from multiple sources within a zeolite journal article, then modeled to predict structural properties from the synthesis. This section is taken from Zach Jensen et al.[1]

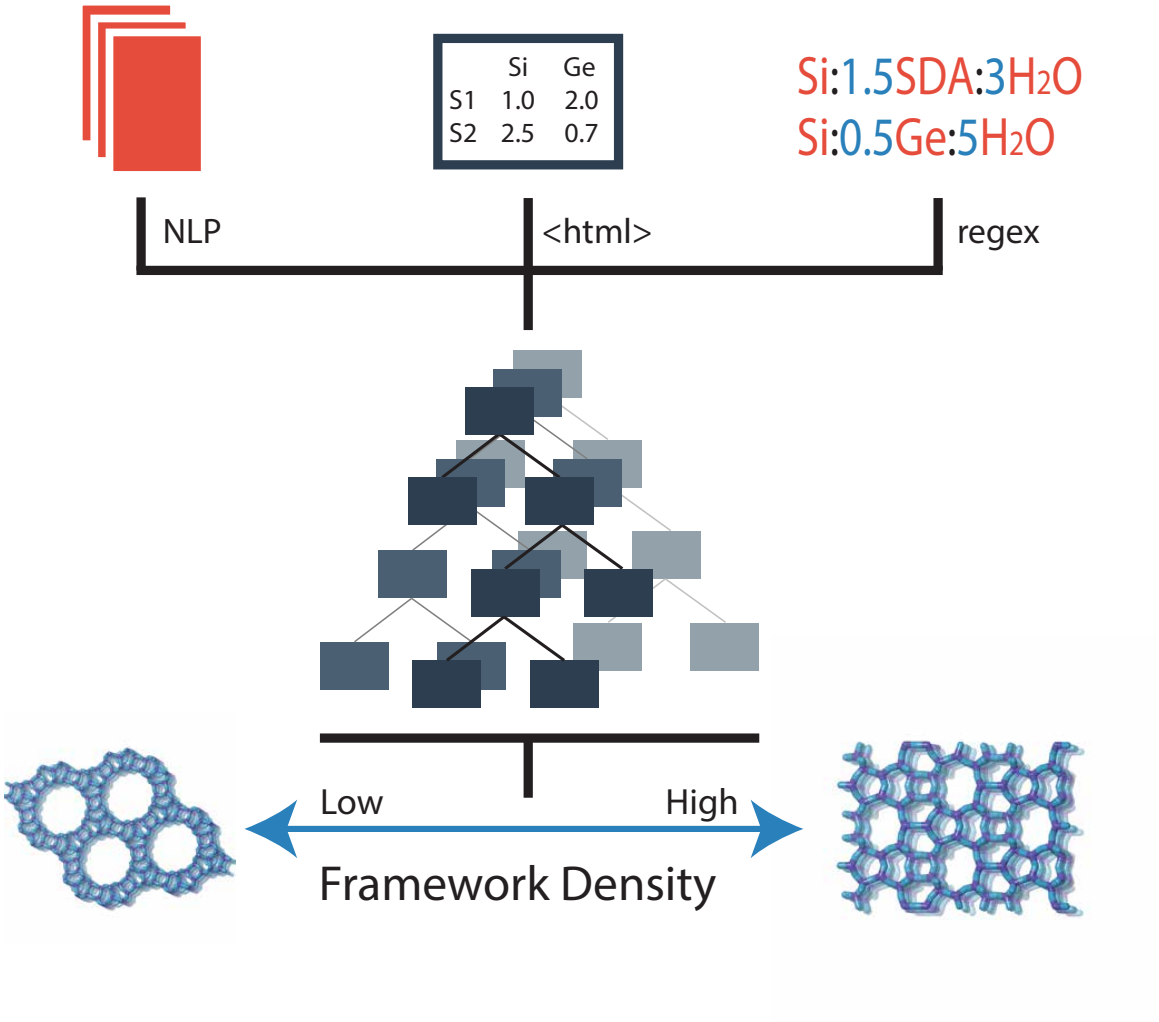


Figure 3-4. Schematic overview of zeolite data engineering including (1) literature extraction from sources such as NLP from body text, parsing of html tables, and regex matching between text and tables, (2) regression modeling, and (3) zeolite structure prediction.

Figure 3-5a shows the wide range of structural variability in germanium-containing zeotypes with medium-, large-, and extra-large pore materials spanning framework densities from 7.5 to 19 T atoms/1000 Å^3. Germanium has a larger nonbonding radius relative to silicon and is capable of forming smaller bond angles with oxygen. Germanium inclusion results in the increased stabilization of small-ring secondary building units (SBUs), including double four-membered rings (D4R), three-membered rings (3MR), and double three-membered rings (D3R).[56,57] These units give rise to zeotype structures with low framework densities and large pores. The extracted data can give rise to new insights including the clustering of extra-large pore structures into three areas corresponding to low, intermediate, and high framework densities (purple triangles, yellow diamonds, and red squares respectively in Figure 3-5a). Materials with framework densities less than 10 T atoms/1000 Å^3 correspond to pure germanium zeotypes referred to as germanates. Materials with densities ranging between 11 and 14 T atoms/1000 Å^3 correspond to structures with some of the biggest pore zeolites including ITQ-33[56] (Ring Size=18) and ITQ-44[58] (Ring Size=18) that have only been obtained with Si/Ge $<$ 4. Extra-large pore zeotypes with framework densities ranging from 15.5 to 16.5 T atoms/1000 Å^3 correspond to Assembly-Disassembly-Organization-Resassembly (ADOR) precursors including UTL and CTH with germanium placed within the D4R units between the siliceous layers.[59,60] These precursors have been exploited to access new zeolite structures by disassembling the interlayer germanium-oxygen bonds and reorganized into a new structure (i.e. the ADOR method).[61,62]

The germanium dataset also demonstrates some correlation between various synthesis elements. Figure 3-5b depicts one of these close relationships between the amounts of germanium and the fluoride ion. The fluoride ion stabilizes small-ring SBUs in a similar fashion to germanium giving rise to a trade-off relationships between the amounts of germanium and fluoride required to stabilize a particular zeolite structure.[64] As a consequence, zeolites with large amounts of germanium can be synthesized with simple OSDAs and small to no amounts of fluoride, although

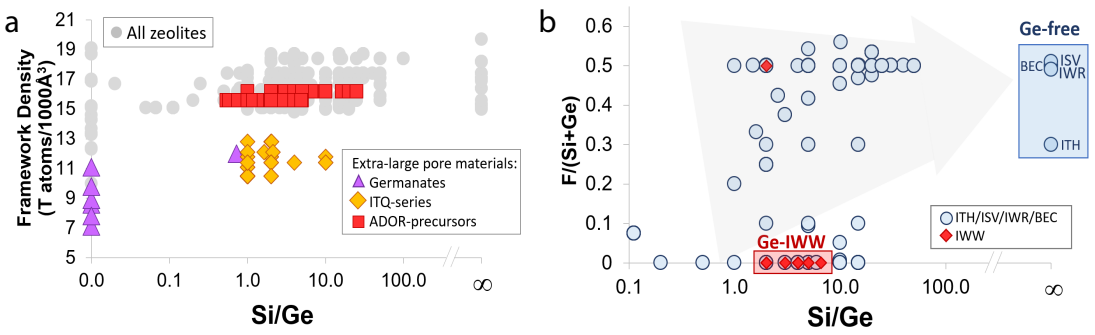


Figure 3-5. Germanium-containing zeolite data extracted with our pipeline. a) Framework density clusters corresponding to different classes of germanium-containing zeolites. b) Tradeoff between Ge content and the amount of F^- ions required to stabilize different zeolites. The three letter codes refer to specific zeolite framework structures defined by the IZA. ADOR is an interzeolite transformation synthesis method.[63]

these materials have lower hydrothermal stability. For examples, BEC and IWR zeolites can be synthesized with Si/Ge < 5 using simple OSDA molecules including tetraethylammonium and hexamethonium without fluoride.[65,66] In contrast, synthesizing more stable forms of zeolite structures with less germanium content requires the use of fluoride often in combination with more complicated, often large OSDAs.[67,68] The data visualization also identifies areas of interest for future study. For example, Figure 3-5b shows several cases for germanium zeotypes including ITQ-22 (IWW) where an OSDA and corresponding synthesis route has not been discovered that yields a germanium-free, high-silica version of the zeolite[69] providing opportunities for researchers to study this topic.

This dataset also allows ML modeling linking the synthesis with the structure of the resulting zeolite. The framework density is modeled as a function of the gel composition, crystallization conditions, and OSDA volume using a random forest ensemble model. Figure 3-6a evaluates the five fold cross validation accuracy of the model through a parity plot where the color hue corresponds to the frequency of data points. The root mean squared error (RMSE) is 0.98 T atoms/1000 Å^3 compared with the standard deviation of framework density in our data of 1.76 T atoms/1000 Å^3. The model's RMSE and r-squared value (see Figure 3-6) indicate the model is

capable of mapping synthesis conditions to the resulting zeolite structure's framework density allowing predictions of synthesis conditions for zeolites with high and low framework densities.

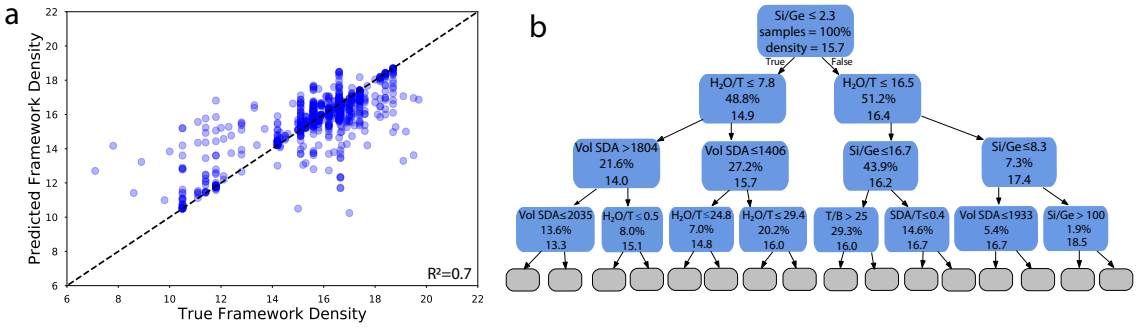


Figure 3-6. Random forest regression model predicting zeolite framework density from synthesis conditions. a) Cross-validation results for the random forest model showing the actual experimental versus model predicted values for framework density. b) A single decision tree regression model trained to predict framework density. Samples values correspond to the percentage of data passing through a node. Density refers to the average framework density value passing through each node. Vol SDA = the volume of the OSDA

Beyond accurately modeling the relationship, tree models also provide human interpretability. Figure 3-6 examines a single decision tree model trained on the germanium dataset to predict framework density. By following the nodes on the tree, synthesis routes to zeolites with specified pore structures can be determined. The top nodes in the tree also correspond to the the more important synthesis parameters for influencing pore size including Si/Ge ratio, H2O/T (T is the sum of all framework elements), and the volume of the OSDA. As validation, most germanium containing zeotypes featuring a low framework density reported in the literature require Si/Ge ratios of 1-2, H2O/T < 5, and large, bulky OSDA molecules, all in good agreement with the model.[70,71] While some of these heuristics may be evident to domain experts, this example represents the first instance of ML decision guidance across the zeolite literature and the potential for modeling synthesis-structure relationships.

3.7.2 Theory Validation for Diffusionless Interzeolite Conversions and Intergrowths

Another valuable use for extracted zeolite data is the validation of theory and simulations. Theory and simulations need to explain and predict observed phenomena to provide value. By comparing against data extracted from the entirety of the zeolite literature, theories can be much more rigorously tested than by single experimental studies while also developing better understanding of their limitations by identifying outliers. This section demonstrates the usefulness of comparing theory with extracted data through a study on interzeolite transformations based on the data extraction contribution to the journal article, Daniel Schwalbe-Koda et al.[72]

Interzeolite conversion is the process of transforming one zeolite into a different zeolite. Often the second zeolite is difficult to synthesize with traditional hydrothermal routes making the interzeolite conversion attractive. There are several different types of interzeolite conversions including competing phases,[73] recrystallization,[74] intergrowth,[75] and diffusionless.[76,77] Despite the usefulness of these transitions,[63,78,79] no clear way existed to predict which zeolite can be converted into another. To address this problem, researchers hypothesized that similarity metrics computed for the graph-based representation of two zeolite structures could predict the potential of interzeolite transformation.[72] To validate the hypothesis, the thesis author searched a corpus of over 70,000 journal articles related to zeolites looking for occurrences of multiple zeolites and interzeolite keywords such as "intergrowth", "topological", "reconstruction", and "ADOR" leading to a dataset of 540 papers that was manually checked to confirm the zeolite pairs and type of transformation.

Figure 3-7 shows the graph similarity metrics, D-measure and SOAP distance, for all of the extracted zeolite pairs broken up into transformation type (CR-Recrystallization, CO-Competing Phases, DL-Diffusionless, IG-Intergrowth). The literature data clearly demonstrates the theory's predictive capabilities for diffusionless and intergrowth transformation while underscoring the lack of predictive ability for recrystalliza-

tion and competing phases. It also found one outlier (red star) which is confirmed as the LTA-IFY diffusionless transformation which requires extremely high pressure (3 GPa).[80] This outlier helps demonstrate the limits of this model under extreme synthesis conditions.

3.8 Conclusion

Data extraction serves a vital role in the study of zeolite synthesis with data driven tools. Without extraction on a large scale from the literature, it is extremely difficult to gain meaningful insights into global zeolite synthesis. This chapter highlights the techniques and innovations within the information extraction space necessary to extract large amounts of zeolite synthesis data from the literature answering question one proposed in this thesis.

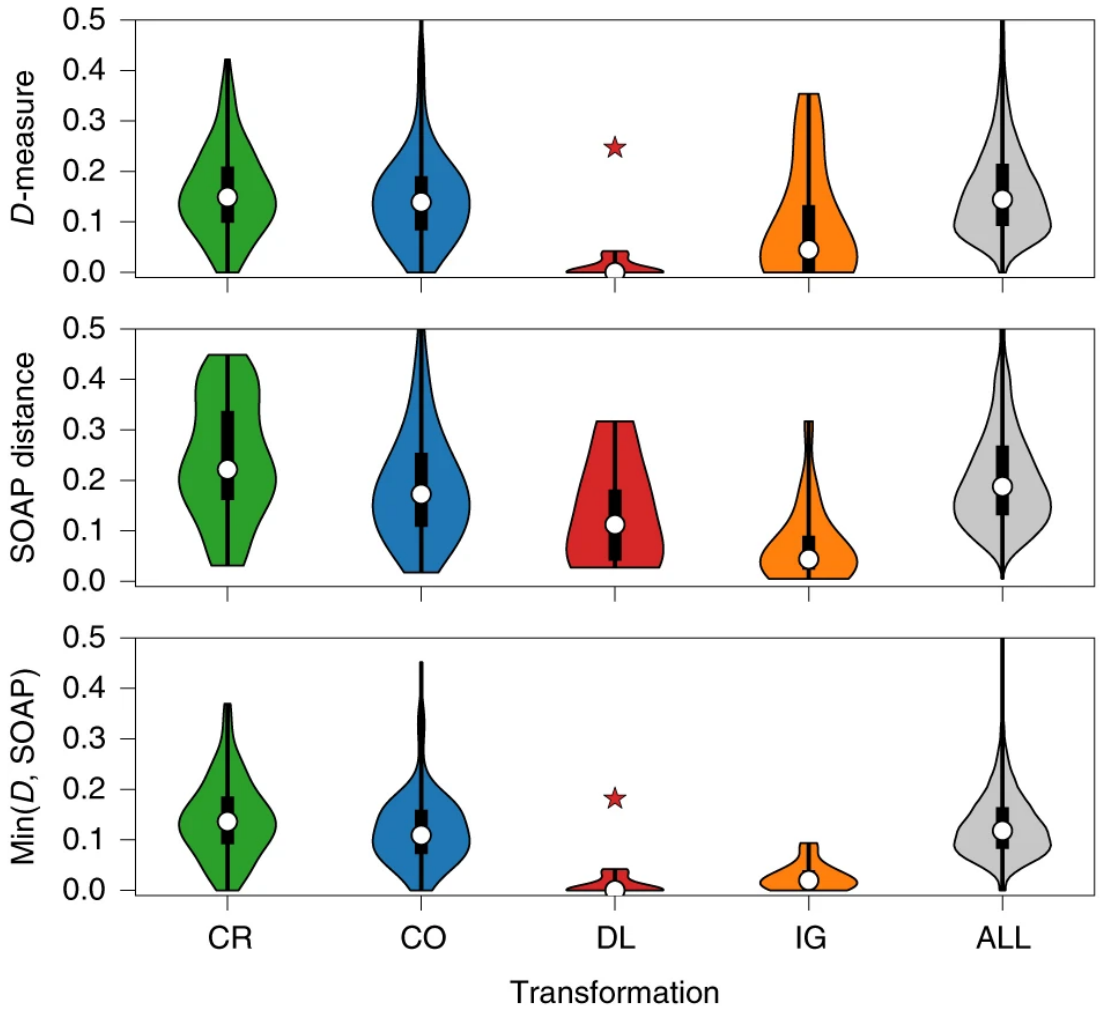


Figure 3-7. Graph similarity (D-measure and SOAP distance) for all of the extracted literature interzeolite transformations. The small range of the distributions for diffusionless (DL) and intergrowth (IG) transformation confirms the theory's ability to predict pairings based on graph similarity. The star represents the only exception found in the literature. Taken from Daniel Schwalbe-Koda et al.[72]

Bibliography

[1] Z. Jensen, E. Kim, S. Kwon, T. Z. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, and E. Olivetti, ACS central science **5**, 892 (2019).

[2] A. Corma, M. Moliner, J. M. Serra, P. Serna, M. J. Díaz-Cabañas, and L. A. Baumes, Chemistry of materials **18**, 3287 (2006).

[3] J. Manuel Serra, L. Allen Baumes, M. Moliner, P. Serna, and A. Corma, Combinatorial chemistry & high throughput screening **10**, 13 (2007).

[4] F. Boyle and D. Sherman, The Serials Librarian **49**, 147 (2006).

[5] E. Kim, Ph.D. thesis, Massachusetts Institute of Technology (2019).

[6] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, Chemistry of Materials **29**, 9436 (2017).

[7] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, and E. Olivetti, Scientific data **4**, 1 (2017).

[8] E. Kim, Z. Jensen, A. van Grootel, K. Huang, M. Staib, S. Mysore, H.-S. Chang, E. Strubell, A. McCallum, S. Jegelka, et al., Journal of chemical information and modeling **60**, 1194 (2020).

[9] O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, and G. Ceder, Scientific data **6**, 1 (2019).

[10] R. Mahbub, K. Huang, Z. Jensen, Z. D. Hood, J. L. Rupp, and E. A. Olivetti, Electrochemistry Communications **121**, 106860 (2020).

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, arXiv preprint arXiv:1301.3781 (2013).

[12] S. Mysore, Z. Jensen, E. Kim, K. Huang, H.-S. Chang, E. Strubell, J. Flanigan, A. McCallum, and E. Olivetti, arXiv preprint arXiv:1905.06939 (2019).

[13] T. O'Gorman, Z. Jensen, S. Mysore, K. Huang, R. Mahbub, E. Olivetti, and A. McCallum, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2021).

[14] C. Baerlocher, http://www. iza-structure. org/databases/ (2008).

[15] Z. Jensen, S. Kwon, D. Schwalbe-Koda, C. Paris, R. Gómez-Bombarelli, Y. Román-Leshkov, A. Corma, M. Moliner, and E. A. Olivetti, ACS central science **7**, 858 (2021).

[16] P. Barrett, M. Camblor, A. Corma, R. Jones, and L. Villaescusa, The Journal of Physical Chemistry B **102**, 4147 (1998).

[17] M. A. Camblor, A. Corma, and S. Valencia, Journal of Materials chemistry **8**, 2137 (1998).

[18] S. I. Zones, R. J. Darton, R. Morris, and S.-J. Hwang, The Journal of Physical Chemistry B **109**, 652 (2005).

[19] B. M. Lok, C. A. Messina, R. L. Patton, R. T. Gajek, T. R. Cannan, and E. M. Flanigen, Journal of the American Chemical Society **106**, 6092 (1984).

[20] T. Cowen and G. Kasparov, *Garry kasparov on ai, chess, and the future of creativity (ep. 22)*, Medium (2017).

[21] J. Bridle, The Guardian **15** (2018).

[22] X. Wang, A. Firdous, L. Xu, P. Chen, F. Liao, J. Lin, and J. Sun, Crystal Growth & Design **19**, 2272 (2019).

[23] H. E. Pence and A. Williams, *Chemspider: an online chemical information resource* (2010).

[24] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, et al., Nucleic acids research **47**, D1102 (2019).

[25] M. Swain, Matt Swain's Blog (2012).

[26] E. I. Ioannidis, T. Z. Gani, and H. J. Kulik, *molsimplify: A toolkit for automating discovery in inorganic chemistry* (2016).

[27] L. B. Kier, Quantitative Structure-Activity Relationships **8**, 221 (1989).

[28] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, Journal of cheminformatics **3**, 1 (2011).

[29] T. A. Halgren, Journal of computational chemistry **17**, 490 (1996).

[30] G. Landrum, Google Scholar There is no corresponding record for this reference (2016).

[31] R. Todeschini and P. Gramatica, SAR and QSAR in Environmental Research **7**, 89 (1997).

[32] J. G. Wicker and R. I. Cooper, Journal of chemical information and modeling **56**, 2347 (2016).

[33] H. Robson and S. Mintova, *Verified synthesis of zeolitic materials* (Gulf Professional Publishing, 2016).

[34] C. Analytics, *Web of science* (2017).

[35] H.-S. You, H. Jin, Y.-H. Mo, and S.-E. Park, Materials Letters **108**, 106 (2013).

[36] S. Cheng, B. Mazonde, G. Zhang, M. Javed, P. Dai, Y. Cao, S. Tu, J. Wu, C. Lu, C. Xing, et al., Fuel **223**, 354 (2018).

[37] B. Das, B. Ravikanth, R. Ramu, and B. V. Rao, Chemical and pharmaceutical bulletin **54**, 1044 (2006).

[38] M. Kalhor, Z. Zarnegar, F. Janghorban, and S. A. Mirshokraei, Research on Chemical Intermediates **46**, 821 (2020).

[39] P. Zhang, S. Li, P. Guo, and C. Zhang, Waste and Biomass Valorization **11**, 4381 (2020).

[40] S. Khodabandeh and M. E. Davis, Chemical Communications pp. 1205–1206 (1996).

[41] O. Muraza, I. A. Bakare, T. Tago, H. Konno, A.-l. Adedigba, A. M. Al-Amer, Z. H. Yamani, and T. Masuda, Chemical engineering journal **226**, 367 (2013).

[42] S. K. Kirdeciler and B. Akata, Advanced Powder Technology **31**, 4336 (2020).

[43] D. Novembre, B. Di Sabatino, and D. Gimeno, Clays and Clay Minerals **53**, 28 (2005).

[44] V. Kasneryk, M. Shamzhy, M. Opanasenko, P. S. Wheatley, S. A. Morris, S. E. Russell, A. Mayoral, M. Trachta, J. Čejka, and R. E. Morris, Angewandte Chemie International Edition **56**, 4324 (2017).

[45] M. Liu, N. J. Guan, S. H. Xiang, J. X. Zhang, S. Y. Liu, and S. Q. Liu, Chinese Chemical Letters **10**, 519 (1999).

[46] T. Ishihara, H. Iwakuni, K. Eguchi, and H. Arai, Applied catalysis **75**, 225 (1991).

[47] D. Vaičiukynienė, L. Jakevičius, A. Kantautas, V. Vaitkevičius, and V. Vaičiukynas, Revista Română de Materiale/Romanian Journal of Materials **50**, 471 (2020).

[48] S. Han, Y. Liu, C. Yin, and N. Jiang, Microporous and Mesoporous Materials **275**, 223 (2019).

[49] E. J. Hu, D.-S. Zhu, X.-Y. Sang, L. Wang, and Y.-K. Tan (1997).

[50] N. Xu, L. Kong, Y. Zhang, X. Kong, M. Wang, X. Tang, D. Meng, Y. Zhang, et al., Journal of Membrane Science **611**, 118361 (2020).

[51] R. Srivastava, N. Iwasa, S.-i. Fujita, and M. Arai, Chemistry–A European Journal **14**, 9507 (2008).

[52] E. A. Hildebrando, C. G. B. Andrade, C. A. F. d. Rocha Junior, R. S. Angélica, F. R. Valenzuela-Diaz, and R. d. F. Neves, Materials Research **17**, 174 (2014).

[53] J. Zheng, Q. Zeng, J. Ma, X. Zhang, W. Sun, and R. Li, Chemistry letters **39**, 330 (2010).

[54] S. Cheng, B. Mazonde, G. Zhang, M. Javed, C. Amoo, Y. Shi, K. Guo, M. Yao, C. Lu, G. Yang, et al., ChemistrySelect **3**, 13632 (2018).

[55] J. Li, A. Corma, and J. Yu, Chemical Society Reviews **44**, 7112 (2015).

[56] A. Corma, M. J. Diaz-Cabanas, J. L. Jorda, C. Martinez, and M. Moliner, Nature **443**, 842 (2006).

[57] A. Corma, M. Díaz-Cabañas, J. Jiang, M. Afeworki, D. Dorset, S. Soled, and K. Strohmaier, Proceedings of the National Academy of Sciences **107**, 13997 (2010).

[58] J. Jiang, J. L. Jorda, M. J. Diaz-Cabanas, J. Yu, and A. Corma, Angewandte Chemie **122**, 5106 (2010).

[59] J. H. Kang, D. Xie, S. I. Zones, S. Smeets, L. B. McCusker, and M. E. Davis, Chemistry of Materials **28**, 6250 (2016).

[60] A. Corma, M. J. Díaz-Cabañas, F. Rey, S. Nicolopoulus, and K. Boulahya, Chemical communications pp. 1356–1357 (2004).

[61] W. J. Roth, P. Nachtigall, R. E. Morris, P. S. Wheatley, V. R. Seymour, S. E. Ashbrook, P. Chlubná, L. Grajciar, M. Položij, A. Zukal, et al., Nature chemistry **5**, 628 (2013).

[62] E. Verheyen, L. Joos, K. Van Havenbergh, E. Breynaert, N. Kasian, E. Gobechiya, K. Houthoofd, C. Martineau, M. Hinterstein, F. Taulelle, et al., Nature materials **11**, 1059 (2012).

[63] P. Eliášová, M. Opanasenko, P. S. Wheatley, M. Shamzhy, M. Mazur, P. Nachtigall, W. J. Roth, R. E. Morris, and J. Čejka, Chemical Society Reviews **44**, 7177 (2015).

[64] H. Kessler, J. Patarin, and C. Schott-Darie, ChemInform **26**, no (1995).

[65] A. Corma, M. T. Navarro, F. Rey, J. Rius, and S. Valencia, Angewandte Chemie **113**, 2337 (2001).

[66] R. Castaneda, A. Corma, V. Fornés, F. Rey, and J. Rius, Journal of the American Chemical Society **125**, 7820 (2003).

[67] M. Moliner, P. Serna, Á. Cantín, G. Sastre, M. J. Díaz-Cabañas, and A. Corma, The Journal of Physical Chemistry C **112**, 19547 (2008).

[68] A. Cantín, A. Corma, M. J. Diaz-Cabanas, J. L. Jordá, and M. Moliner, Journal of the American Chemical Society **128**, 4216 (2006).

[69] A. Corma, F. Rey, S. Valencia, J. L. Jordá, and J. Rius, Nature materials **2**, 493 (2003).

[70] J. Sun, C. Bonneau, Á. Cantín, A. Corma, M. J. Díaz-Cabañas, M. Moliner, D. Zhang, M. Li, and X. Zou, Nature **458**, 1154 (2009).

[71] J. Jiang, J. L. Jorda, J. Yu, L. A. Baumes, E. Mugnaioli, M. J. Diaz-Cabanas, U. Kolb, and A. Corma, Science **333**, 1131 (2011).

[72] D. Schwalbe-Koda, Z. Jensen, E. Olivetti, and R. Gómez-Bombarelli, Nature materials **18**, 1177 (2019).

[73] H. Robson, *Verified synthesis of zeolitic materials* (Gulf Professional Publishing, 2001).

[74] C. Li, M. Moliner, and A. Corma, Angewandte Chemie International Edition **57**, 15330 (2018).

[75] T. Willhammar and X. Zou (2013).

[76] A. Alberti, G. Cruciani, and A. Martucci, American Mineralogist: Journal of Earth and Planetary Materials **102**, 1727 (2017).

[77] A. Alberti and A. Martucci, in *Studies in Surface Science and Catalysis* (Elsevier, 2005), vol. 155, pp. 19–43.

[78] K. Honda, M. Itakura, Y. Matsuura, A. Onda, Y. Ide, M. Sadakane, and T. Sano, Journal of nanoscience and nanotechnology **13**, 3020 (2013).

[79] B. Marler and H. Gies, European Journal of Mineralogy **24**, 405 (2012).

[80] J. L. Jordá, F. Rey, G. Sastre, S. Valencia, M. Palomino, A. Corma, A. Segura, D. Errandonea, R. Lacomba, F. J. Manjón, et al., Angewandte Chemie **125**, 10652 (2013).

Chapter 4

Organic Zeolite Synthesis Planning

This chapter's content derives from "Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks" by Zach Jensen et al. appearing in *ACS Central Science* in 2021.[1]. The chapter discusses the application of data mining, chemical informatics, and generative neural networking to the organic aspects of zeolite synthesis planning. The chapter aims to partially answer thesis question two and three by advancing the understanding around OSDA design and creating models to generate novel OSDA-zeolites pairs.

4.1 Introduction

OSDA molecules play a crucial role in zeolite synthesis. They can provide different effects within the synthesis from charge balancing and space filling to a templating, lock-and-key relationship[2] resulting in a wide range of OSDA specificity with some OSDAs able to crystallize many different zeolites while others only direct the formation of a limited number of phases. The size, flexibility, hydrophilicity, and charge of the OSDA, among other factors, all play an important role in zeolite crystallization kinetics and phase specificity making *a priori* predictions of suitable OSDA-zeolite pairs very challenging.[3–5] Two main options for prediction existed before this the-

sis: experimental heuristics and density functional theory (DFT) simulations. Experimental heuristics are inexact, making prediction and OSDA design from them extremely difficult (see 4.2).[3] Simulation approaches are very expensive and time consuming requiring DFT and molecular dynamics simulations to suggest candidate OSDAs for a specific zeolite.[6–8] Beyond cost, these simulations are limited to a single zeolite system and focus only on pure silica zeolite systems.

This chapter aims to add a third, data driven option to OSDA design. The author examines a comprehensive dataset of experimental OSDA-zeolite pairs found in the literature and uses structural descriptors of the OSDA molecules to explain the relationships observed between OSDAs and specific zeolite structures. The chapter also contains a generative neural network modeling approach that moves beyond explaining the literature data and moves towards generating novel OSDA-zeolite pairs.

4.2 Characteristics of Literature OSDAs

The OSDA-zeolite pairs dataset (see 3.5.3) contains OSDA molecules, the resulting zeolite structures formed, and the elements used in the gel chemistry. It contains 758 distinct OSDA molecules and 205 zeolite phases found in the literature between the years 1966 to 2020.

Figure 4-1 summarizes some of the important properties possessed by literature OSDAs. Figure 4-1a shows the average conformer molecular volume distribution of the OSDAs. These values range from approximately 30 to 1000 Å^3. Larger OSDAs are related to the synthesis of zeolites rather than aluminophosphate (AlPO) zeotypes which agrees with experimentally observed lack of correlation between the OSDA and the pores/cages of an AlPO material.[3] Large-pore AlPO materials have limited stability compared to their aluminosilicate counterparts which mostly precludes studies using bulky OSDAs in their synthesis.

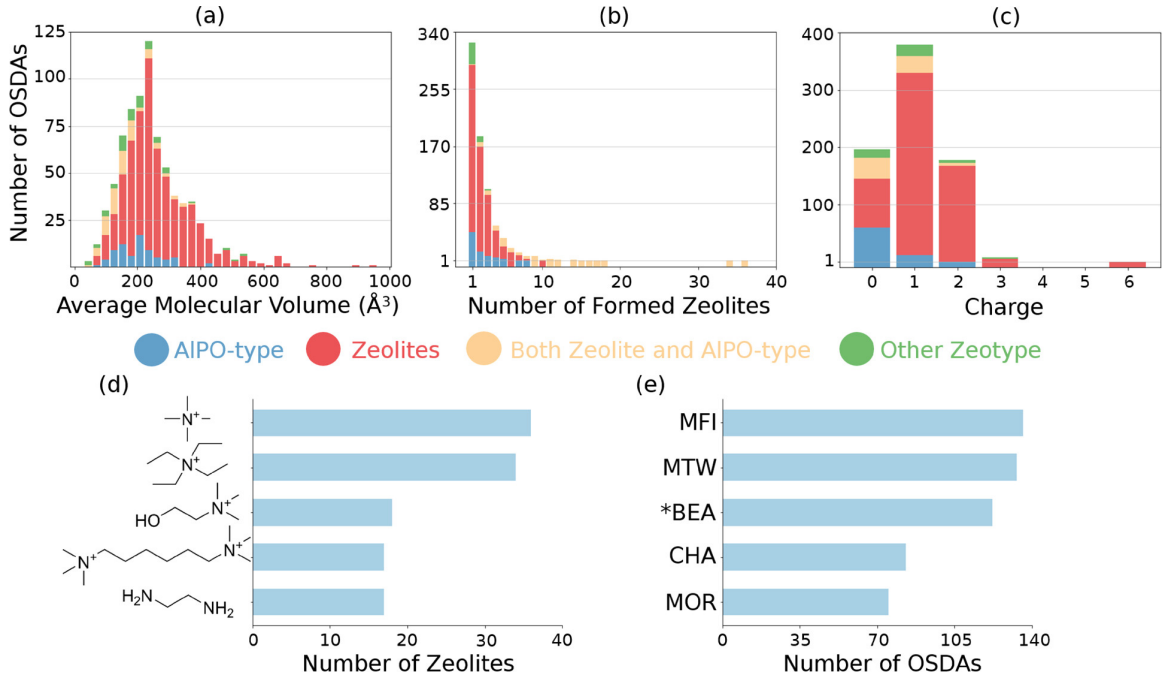


Figure 4-1. Overview of literature OSDAs. (a–c) Average conformer molecular volume, OSDA specificity, and charge distributions for all OSDAs in the data set. (d) Shows the five OSDAs known to make the most zeolite structures. (e) Shows the five zeolites that can be made with the most OSDAs.

OSDA specificity is also an important consideration in selecting OSDA-zeolite pairs. Figure 4-1b shows the majority of the ODSAs have high specificity, producing fewer than 5 zeolite phases, while some outliers are capable of making more than 20 phases. These lower specificity OSDAs are typically small and simple alkylammonium cations, such as tetramethylammonium (TMA), tetramethylammonium (TEA), and hexamethonium shown in Figure 4-1d. These low specificity molecules typically act as space-filling molecules that provide charge balance to the framework and generally do not provide a true templating effect. Some feature high flexibility with many rotatable bonds, such as hexamethonium, which allows different conformations of the molecule to act as an OSDA for different zeolites. The zeolite structure also plays a role in determining OSDA specificity. The literature data shows that some zeolites including MFI, MTW, *BEA, CHA, and MOR (Figure 4-1)e can be made with a large number of OSDA molecules. These zeolites are among the most widely used industrial application (along with FAU and FER),

therby having more research efforts to improve their physicochemical properties and cost effectiveness.[9]

Ionic charge number and distribution within an OSDA plays an important role in the nucleation and crystallization processes. The charges, often together with alkali cations, position the negatively charged heteroatoms (Al, B) in specific framework positions. Heteroatom position can drastically alter the catalytic properties of the materials.[10–12]. In zeolite synthesis, most OSDAs contain one or two positive charges, generally in the form of mono- or dicationic species (Figure 4-1c).[3–5] In contrast, AlPO materials are preferentially synthesized using neutral amines as OSDAs (blue bar in 0 charge in Figure 4-1c, which are protonated in the neutral or acidic media of the typical AlPO material synthesis gel.

The literature data also allows testing of conventional heuristics in the field to determine if they possess predictive power. Experimental heuristics connect the OSDA size with increasing zeolite pore size and increasing OSDA rigidity with increasing specificity or formation of fewer zeolite phases.[3] Figure 4-2 examines these two heuristics across the dataset by looking at framework density (lower framework density, higher zeolite pore size) against average conformer molecular volume for the OSDA-zeolite pairs in (a) and the number of zeolites formed against nConf20[13] for each molecule which measures a molecule's flexibility (higher nCon20, higher flexibility) in (b). It is clear from the lack of observed relationship that these heuristics to not provide predictive power to design new OSDAs for specific zeolites and more advanced informatics are needed to understand the relationship.

4.3 Correlation between OSDAs and Zeolite Structures

4.3.1 WHIM Descriptors

Weighted holistic invariant molecular (WHIM)[14] descriptors contain information about the size, shape, symmetry, and atom distribution of a molecule and are depen-

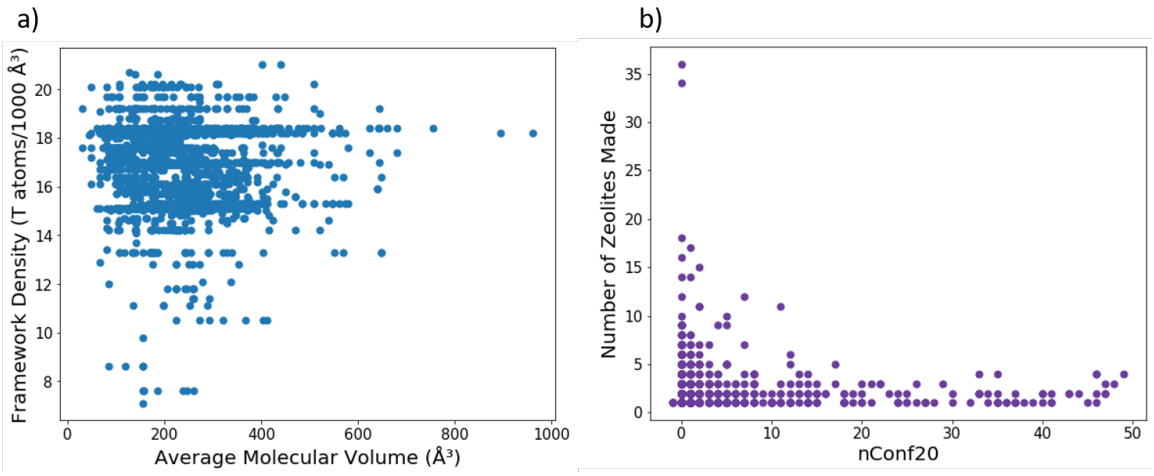


Figure 4-2. Simple relationships describing conventional heuristics used in zeolite synthesis. The lack of correlation indicates more advanced featurization is required to understand relationships between OSDAs and zeolites. (a) Framework density vs average conformer OSDA volume (b) Number of zeolites formed vs nConf20.

dent on its three-dimensional conformation. Different conformations can have drastically different WHIM representations depending on the flexibility of the molecule. For example, a long linear molecule can either stretch out or fold, giving two different three-dimensional representations, demonstrated in Figure 4-3. This varying three-dimensional representation for each molecule is accounted for by calculating the average conformer WHIM descriptor.

WHIM is high-dimensional descriptor (114 length), so principal component analysis (PCA) is used to reduce the dimensionality and enable visualization. The first principal component (PCA 1) accounts for 58% of the variance and correlates with the volume of the molecule. The second and third principal component axes account for 15% and 13% of the variance respectively. Correlations between OSDAs and zeolites can be visualized using the first three principal component axes.

4.3.2 Correlation in Cage-based Zeolites

Cage-based zeolites have strong correlation between the three-dimensional structure of the OSDA and the shape of the cage. Five cage-based, small-pore zeolites, LEV, CHA, AEI, LTA, and AFX (Figure 4-4a), are selected and examined through

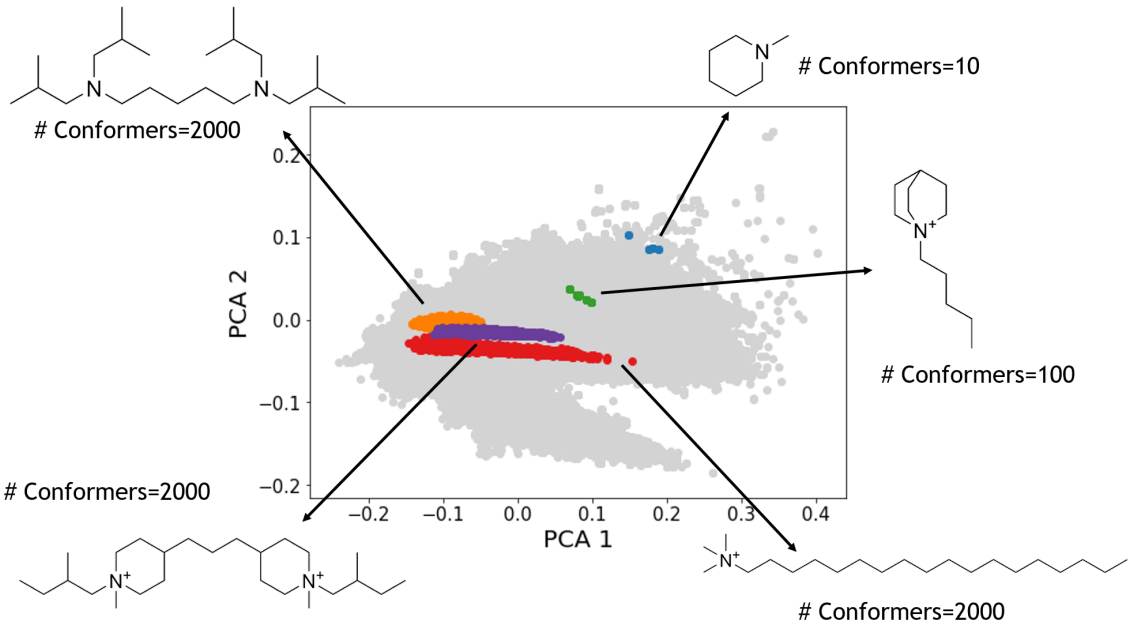


Figure 4-3. Examples of conformer effects for several selected OSDAs. The conformers are plotted in the WHIM space compressed into two-dimensions through principal component analysis. The OSDAs are selected to represent a variety of both flexible and inflexible molecules.

WHIM featurization and PCA analysis in Figure 4-4b,c. Gel composition also affects the relationship between OSDAs and zeolites so the visualized data is limited to only conventional zeolite chemistry.

For these five zeolites, Figure 4-4 shows each zeolite is associated with specific and distinct OSDA characteristics. Differences are observed between locations of the clusters, particularly in PCA 2 and PCA 3, likely due to the differences in the zeolite's cage size and shape which require different molecular structures. The location of the clusters within the compressed WHIM space is meaningful demonstrated by the overlap between CHA and AEI (blue diamonds and red circles in Figure 4-4. This suggests the OSDAs used to synthesize these two structures are structurally similar. In fact, some of the molecules can be used to make either CHA or AEI depending on the synthesis conditions. This relationship is explained by the structural similarity of AEI and CHA including a cage-like three-dimensional small-pore system and identical framework density (15.1 T/1000 $Å^3$). However, the cavities are

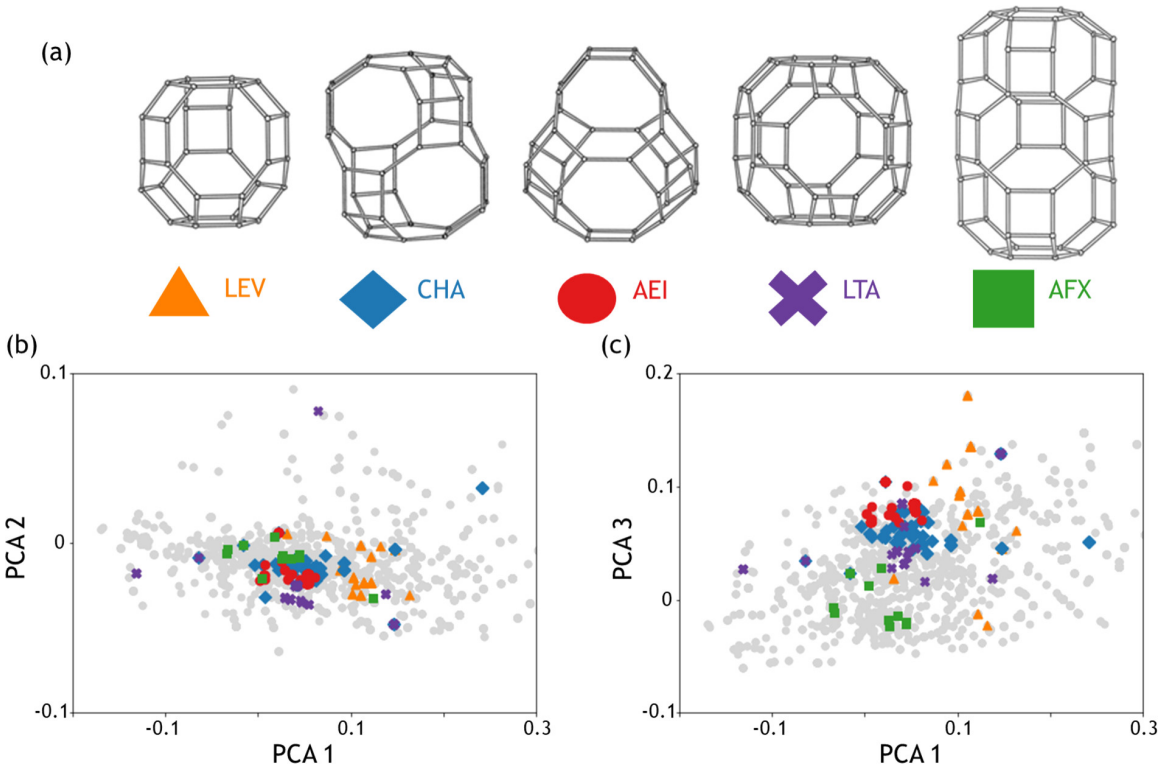


Figure 4-4. Principal component analysis (PCA) WHIM vector representation of OSDA molecules used in five cage-based small-pore zeolite systems. PCA 1, 2, and 3 represent the first three principal component axes. The gray points represent all of the OSDAs extracted from the literature.

specific to each zeolite, elongated and symmetrical for CHA (11.7 x 10.2 Å) and basket-cage-like for AEI (12.6 x 11.2 Å) which explains why some OSDAs preferentially facilitate the crystallization of either CHA or AEI selectively.

The zeolite structure and stability also plays a role in the affect of the OSDA. The OSDAs for LTA show larger variability across the compressed WHIM space than those of the other zeolite structure (purple crosses in Figure 4-4). The synthesis of high-silica LTA preferentially uses large aromatic molecules,[15,16] while low-silica LTA utilizes small organic molecules including tetramethylammonium and diethyldimethylammonium which act as pore fillers in combination with alkali cations. In contrast, CHA and AEI have significantly reduced cluster variance compared with LTA indicating the need for OSDAs to provide a true templating effect across the range of known chemistry.

4.3.3 Correlation in Large-pore Zeolites

To examine the generalization of the WHIM approach, selected large-pore zeolites (FAU, EMT, BEC, ISV, CON) are also examined in Figure 4-5. Similar to the cage-based, small-pore zeolites, the relationship in the WHIM space also depends on the zeolite. FAU and EMT do not exhibit well defined regions due to difficulty in finding a suitable template for their large cavities and the role inorganic species play in the formation of these phases.[17–19] Other large pore systems including BEC, IWV, ISV, and CON show much better defined structural clusters in the WHIM space indicating the importance of the OSDA in these systems. This generalization to large-pore systems helps illustrate the robustness of the WHIM space approach.

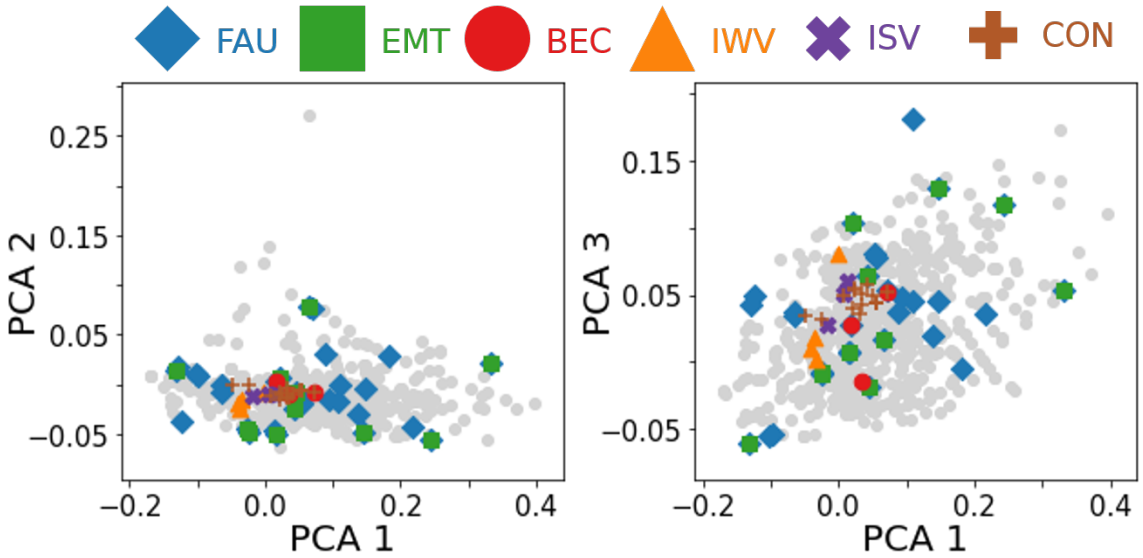


Figure 4-5. PCA WHIM vector representation of OSDA molecules used in six large-pore zeolite systems.

4.4 Novel OSDA-Zeolite Pair Generation

Beyond explaining the OSDA-zeolite pairs observed in the literature, this thesis also aims to predict and generate new OSDA-zeolite pairs. This objective is towards answering thesis question 3 by generating new OSDAs for hypothetical zeolites and better OSDAs for existing zeolites.

4.4.1 Generative Neural Network for novel OSDA prediction

A generative neural network model published by Kotsias et al.[20] is adapted to suggest alternative molecules for use as OSDAs. The model is trained on the literature data to output an OSDA's SMILES string given a zeolite phase and gel chemistry as input, shown in Figure 4-6. This model requires a large quantity of data to train a useful model, which is enabled by size and scope of the automated data extraction.[21]

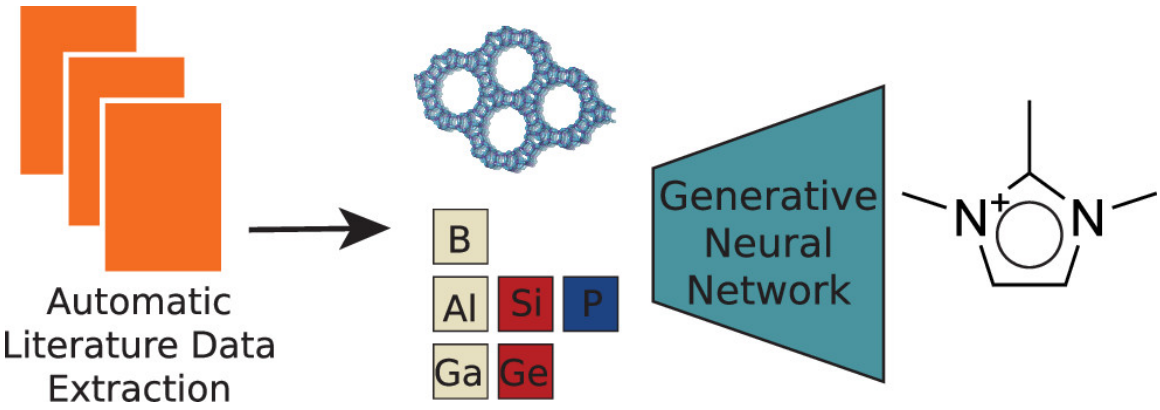


Figure 4-6. Schematic of the generative neural network modeling process.

This generative neural network borrows heavily in both architecture and training protocol from Kotsias et al.[20] The zeolite structure and gel chemistry inputs are fed through six dense layers of 256 units with ReLU activation followed by three unidirectional LSTM layers consisting of 256 units and ending with a feedforward dense layer with 35 units with a softmax activation. Batch normalization is used on the first dense and LSTM layers. The model is trained for 100 epochs on a variety of train/test splits to test various aspects of the generative model (see 4.4.2) using the "teacher's forcing method."[22] Adam algorithm is used for optimization with a batch size of 128 and an exponentially decaying learning rate starting at 10^{-3} and ending at 10^{-6}. Up to 100 different noncanonical versions of each OSDA's SMILES string are generated to augment the dataset, resulting in approximately 150,000 training data points depending on train/test split. This type of data augmentation has shown increased accuracy in organic molecule generative models.[23]

4.4.2 Model Metrics

To test the generative modeling approach, multiple models are trained using different train/test splits:

1. Training and test split is chosen at random with 80% of the data used to train and 20% used to test.

2. All data points resulting in CHA are isolated in the test set. Number of training points-5,398, Number of testing points-265(5%).

3. All data points resulting in AEI are isolated in the test set. Number of training points-5,555, Number of testing points-108(2%).

4. All data points are used to train the model.

Splits 1, 2, and 3 are used to evaluate model performance while split 4 is used to look at the case studies in sections 4.4.3 and 4.4.4. Holding out an entire zeolite structure from training (splits 2 and 3) test the model's capability of suggesting new OSDA candidates for previously unseen zeolites and can confirm the model is not memorizing OSDA/zeolite pairs, which can occur when randomly splitting. CHA and AEI are chosen due to their cage-like structure which correlates well with OSDA structure (4.3.2), industrial relevance, and presence of enough data to construct a large enough test set for benchmarking.

Evaluating generative models is difficult due to the lack of well-defined metrics.[24] Fortunately, the organic generative modeling space has benchmark platforms, Molecular Sets (MOSES),[25] that can be adopted to OSDA generation to evaluate the general quality of the model. Table 4.1 shows the MOSES benchmarks for models trained on train/test splits 1, 2, and 3. The reference values refer to a model trained to predict organic molecules from their properties for drug discovery and design using the same generative model architecture.[20] While the reference modeling task is different, comparing it with OSDA models gives an idea of state-of-

the-art performance on a related learning task. Comparing to the baseline, our models perform better in fraction of valid molecules (Valid) indicating the model generates real, physical molecules at a very high rate. The fraction of generated molecules not present in the training set (Novelty) is also comparable to the baseline for two of the three train/test splits indicating that the model generates novel molecules not present in the training data. Scaffold similarity (Scaff), similarity to nearest neighbors (SNN), and internal diversity (IntDiv) also exceed or match the baseline performance demonstrating the model generates structurally similar and closely related molecules while also remaining chemically diverse. The model under performs the baseline in the number of unique molecules in 1,000 generations (Unique@1k). This is because molecules suitable for use as OSDAs encompass a much small subspace than organic chemistry as a whole. Since the model is trained to generate OSDA-like molecules, the search space is much smaller than in the benchmark task and therefore the number of unique molecules is expected to be lower. A final benchmark considered is the ability to generated the exact molecule used in the test set (Reconstructibility). While high reconstructibility may appear beneficial, training a model to optimize reconstructibility can produce highly deterministic models not suitable to the generation task. Overall, the model performs better or comparable to the baseline across the many of the examined metrics and across the three train/test splits demonstrating the model's overall ability and potential application on new zeolite systems.

The negative log-likelihood (NLL) of sampling different molecules can also be used to evaluate model performance (Figure 4-7). The closer the NLL distribution is to zero, the model deterministic the model becomes. All three of the train/test splits have smaller mean NLL (3.9, 3.3, and 4.1) than our benchmark model[20] (15.9) reinforcing the idea of OSDAs encompassing a smaller search space. Differences in distributions between the training and testing NLLs can indicate overfitting especially if the training distribution is closer to zero, and therefore more deterministic.[26] Figure 4-7 does not indicate that there is significant overfitting in any of the

Table 4.1. Benchmarking using the MOSES[25] standard for several different models trained on different train/test splits. Upward arrows indicate that higher scores are better. The different data splits are described in the main text.

Metrics		Random	CHA Out	AEI Out	Reference[20]
Valid	↑	0.994	0.993	0.999	0.881
Unique@1k	↑	0.114	0.019	0.028	0.996
FCD	↓	2.468	4.593	9.683	7.981
SNN	↑	0.566	0.610	0.365	0.341
Frag	↑	0.850	0.824	0.766	0.920
Scaff	↑	0.387	0.196	0.061	0.094
IntDiv	↑	0.891	0.877	0.852	0.845
Novelty	↑	0.835	0.741	0.537	0.878
Reconstructibility	-	0.013	0.158	0.125	<0.001

train/test splits.

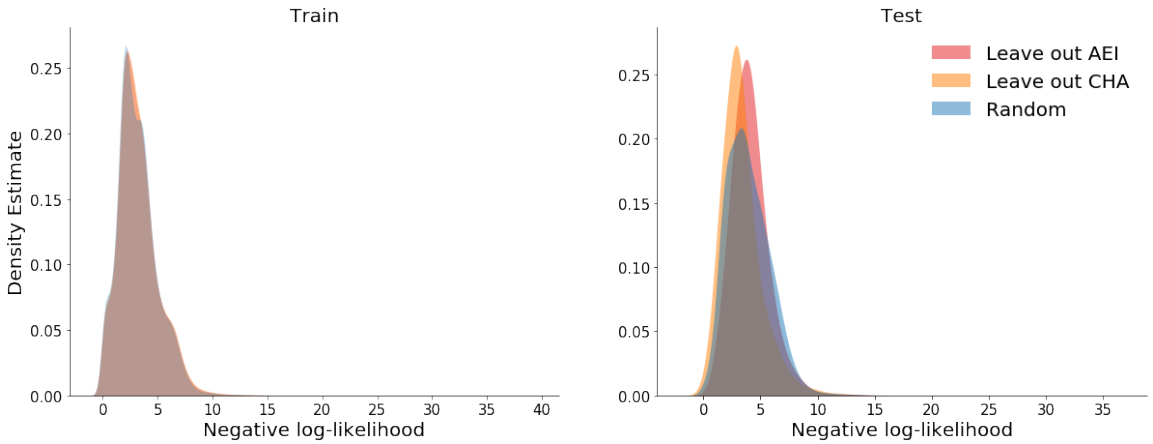


Figure 4-7. The NLL values for the test sets of three different models, random, leave out CHA, and leave out AEI. Differences in training and test set distributions can be an indication of model overfitting which is not observed for our models. Distributions closer to zero correspond to more deterministic output while the variance of the distribution relates to the uniformity of sampling the chemical space.

It is also important to demonstrate the model's ability to generate different molecules for different zeolite inputs rather than generalizing across all zeolite structures. Molecules generated for SFW are compared with molecules generated for LAU. LAU is structurally very different than SFW, having a higher framework density (18 T/1000 Å^3), a 1-dimensional, 10-membered ring channel, and no composite building units in common with SFW. LAU is also chemically distinct, typically being

synthesized as an M-(Al/Ga)PO (M=Co, Mn, Zn, Fe)-type material[27,28] while SFW is a conventional zeolite.[29,30] There is a clear difference in the WHIM distributions of the molecules generated for the two systems indicating the model's abililty to distinguish between the structures during prediction (Figure 4-8a). Figure 4-8b shows the distributions of the minimum distance in the WHIM space to a known SFW or LAU OSDA. OSDAs for LAU with conventional zeolite chemistry are also generated to compare the effect chemistry has on the model. As expected, having similar chemistry shifts the generated distributions closer together although they are are still distinct.

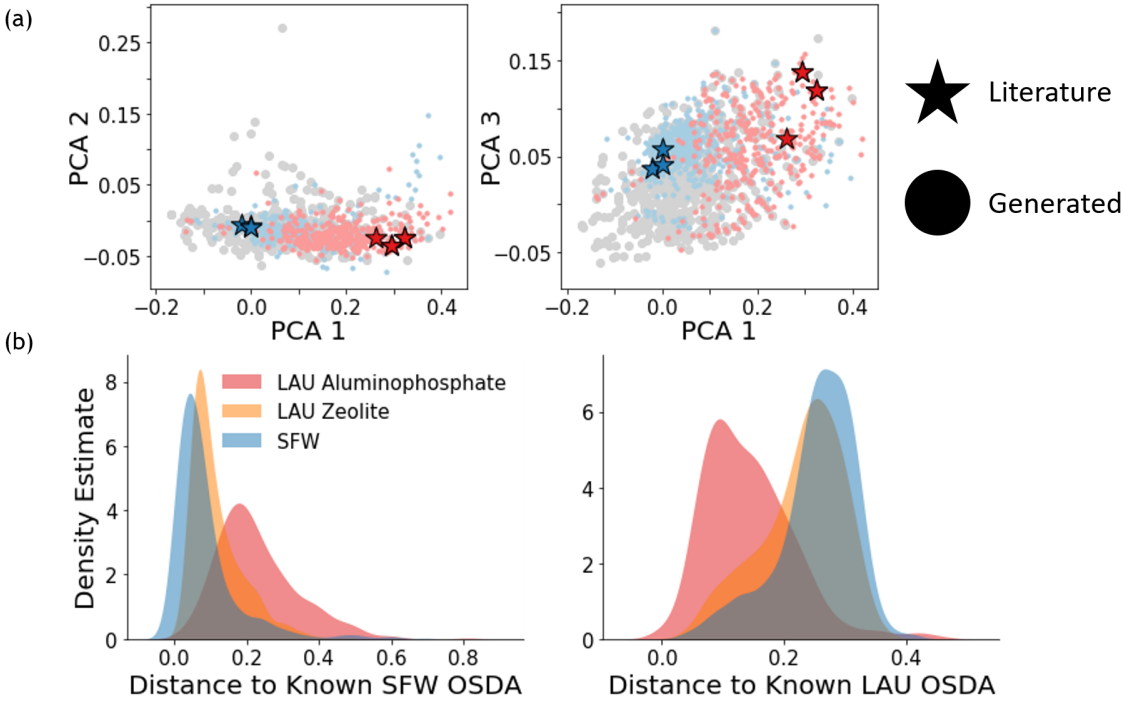


Figure 4-8. Differences in Distributions between SFW and LAU generated OSDAs. a) shows the differences in WHIM distributions between SFW OSDAs generated with zeolite chemistry and LAU OSDAs generated with M-AlPO (M=Co, Fe, Zn, Mn) chemistry. b) Distributions to the nearest SFW and LAU literature OSDA in the WHIM space for generated molecules with SFW zeolite, LAU aluminophosphate, and LAU zeolite seed conditions. These results indicate the model generates different molecule distributions for zeolites that are structurally very different. They also indicates that chemistry plays an important role in the generated molecules where similar chemistry indicates more similar distributions.

It is also important to understand the model's limitations. Figure 4-9 shows the

binding energies of OSDAs generated for SFW and OSDAs from the entire zeolite literature. These distributions are very similar, indicating the model may have a limited ability to predict OSDAs specific to each zeolite system. However, the model matches the literature distribution, containing molecules known to be suitable OSDAs. These results, taken together with the discussion above about SFW and LAU, demonstrate the model's ability to generate different OSDA suggestions by injecting chemical noise into the OSDA space while matching the performance of known literature OSDAs. It also indicates that generated molecules may have potential as OSDAs for several similar zeolite systems.

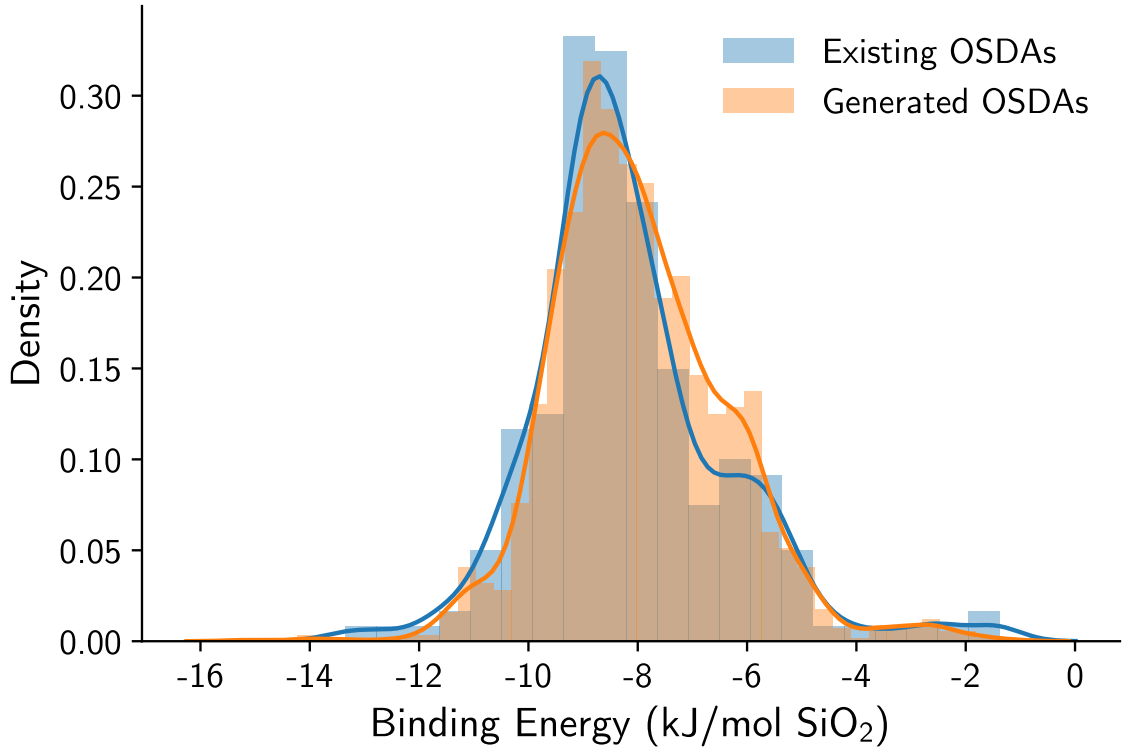


Figure 4-9. Distribution of binding energy with SFW for the generated molecules and all literature OSDAs (for all zeolites). Matching the literature indicates our model is able to inject chemical noise into the OSDA space although it casts doubt on the model's ability to distinguish OSDAs for specific zeolite systems.

4.4.3 Model Test Case Study: CHA

The first case study examines CHA, a cage-based, small-pore zeolite featured in 4.3.2, due to its industrial relevance. 408 unique OSDAs are generated for CHA from a total of 10,000 samples drawn from the model using the CHA structure and a variety of zeolite gel chemistries as the zeolite and chemistry inputs respectively.

The generated OSDAs are filtered by comparing to the industry standard for CHA, N,N,N-trimethyladamantammonium (TMAda). 57 of the 468 generated OSDA molecules fall within an ellipsoid of centered around TMAda spanning 5% of the range along the first three principal component within the PCA-reduced WHIM space (Figure 4-10a,b) 11 additional OSDAs reported to synthesize CHA and 24 other OSDAs reported for other zeolite structures also fall within this TMAda-centered ellipsoid. Organic molecules within this ellipsoid are expected to be structurally similar to TMAda and therefore may be a suitable alternative OSDA for CHA.

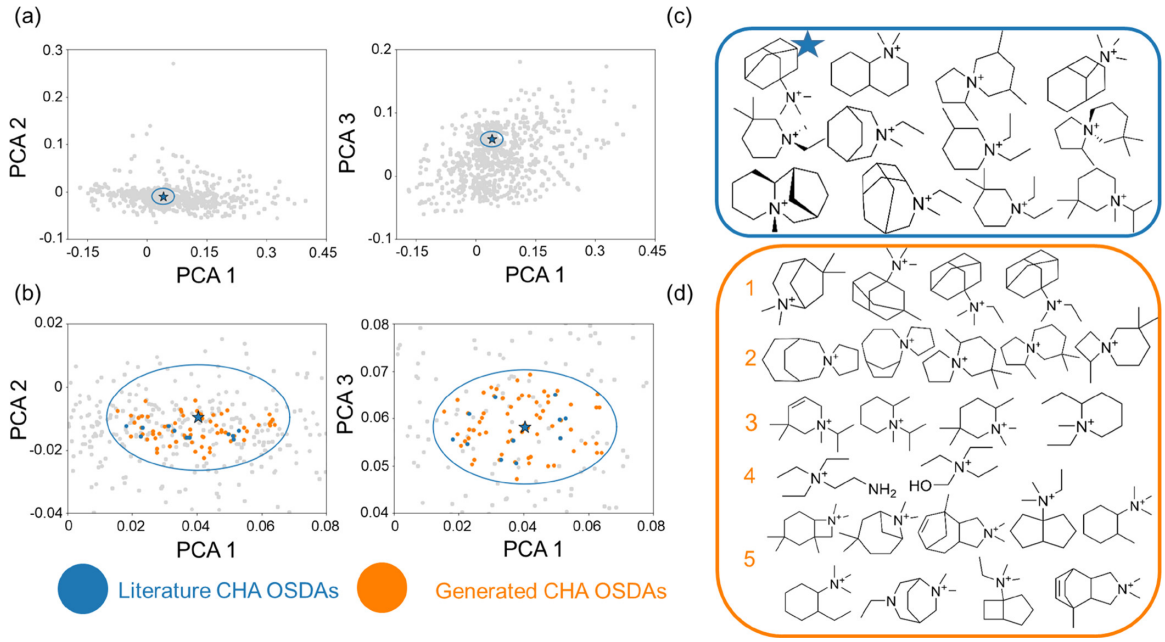


Figure 4-10. Comparing literature OSDAs and generated OSDAs of a CHA zeolite. (a) Shows the position of TMAda (shown with the blue star) relative to the rest of the OSDAs in the PCA WHIM space. (b) A zoomed in view of the ellipse surrounding it. (c) The blue square contains literature CHA OSDAs that fall within the ellipse. (d) The orange square contains examples of generated OSDAs for CHA that fall within the ellipse.

Figure 4-10 shows the filtering process and some of the resulting generated organic molecules suggested for CHA (Figure 4-10d). Qualitatively, the generated OSDAs contain many similar features as the OSDAs found in the literature for CHA (Figure 4-10c). For instance, different adamantyl-type, rigid molecules are predicted (row 1 in Figure 4-10d), in good agreement TMAda, considered as the most effective template to stabilize the CHA cavity.[31–33] Beyond adamantyl-type molecules, different alkyl-substituted spiro and piperidinium molecules are generated (row 2 and 3 respectively in Figure 4-10d) presenting similar structural features as some reported CHA OSDAs. In addition, two simple tetraalkylammonium cations are generated (row 4 in Figure 4-10d) which are similar to a recent report that uses tetraethylammonium to synthesize CHA in its silicoaluminate form.[34]. The model also generated other categories of molecules not directly seen in the literature (row 5 in Figure 4-10) but that possess commonly observed features including a single positively charged nitrogen atom and cyclic structures. These molecules demonstrate the model's ability to add domain and data-informed chemical noise into the OSDA space that allows intelligent prediction of potential OSDA candidates.

4.4.4 Model Test Case Study: SFW

The second case study examines SFW zeolite, a much less studied zeolite than CHA with only three known literature OSDAs that has high potential impact as a catalyst for NO_x.[29,30] SFW is structurally similar to CHA being cage-based with the *gme* cage replacing the *cha* cage and the same framework density (15.1 T/1000 $Å^{-3}$). This simulates testing the model on a hypothetical zeolites while allowing verification through the limited literature examples. Because there are fewer known molecules to compare with, molecular mechanic simulations calculate the binding energy of the generated molecules with the SFW framework[35,36] to gauge the model's predictive performance.

These simulations demonstrate that many of the generated molecules are suitable candidates for SFW. 60% of the generated OSDAs have binding energies within the

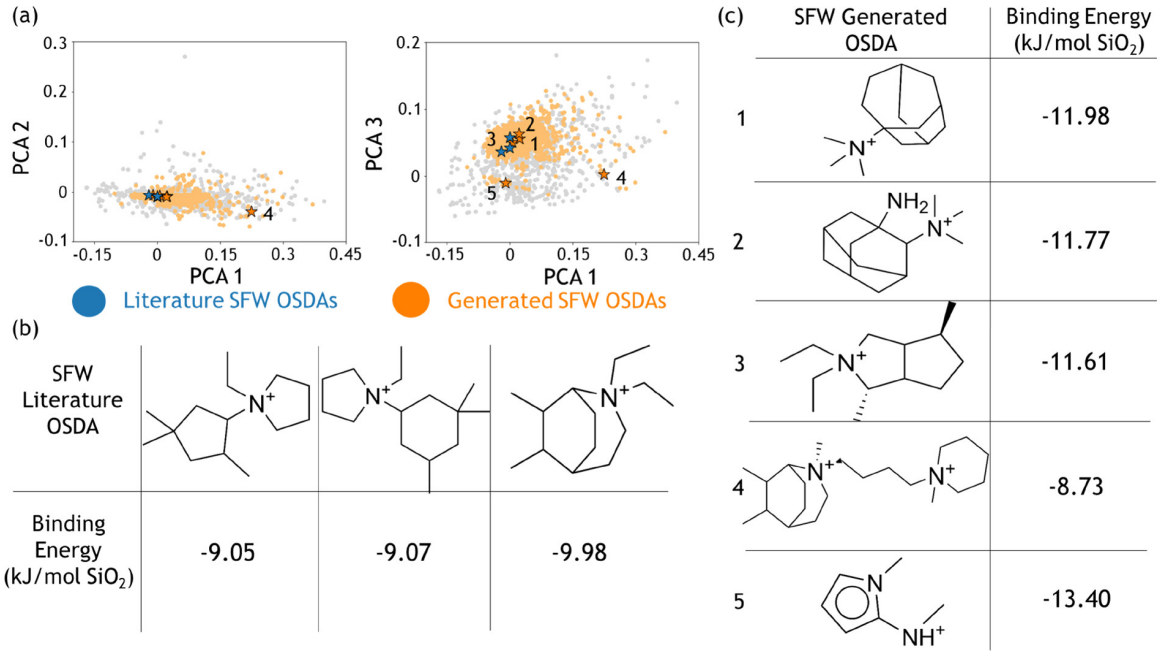


Figure 4-11. OSDAs for SFW obtained from literature and generated by our model. (a) PCA-reduced WHIM locations for the three OSDAs known to make SFW (blue stars) and five selected molecules generated by our model (orange stars). (b) Minimum conformer binding energy with SFW for the three literature OSDAs. (c) Binding energy with SFW for the five selected generated molecules.

range of the literature SFW OSDAs (-9.98 to -7.48 kJ/mol SiO2). Additionally, 7% have lower binding energies than the known OSDAs. Figure 4-11a shows the generated OSDAs for SFW in the PCA-reduced WHIM space. The blue stars represent the literature OSDAs that synthesize SFW, N-ethyl-N-(2,4,4-trimethylcyclopentyl)pyrrolidinium, N-ethyl-N-(3,3,5-trimethylcyclohexyl)pyrrolidinium, and N,N-diethyl-5,8-dimethylazonium bicyclo[3.2.2]nonane, while the orange points represent the generated molecules. Figure 4-11b shows the structure and binding energy with SFW for each of the three literature OSDAs. Five of the generated molecules are shown in Figure 4-11c and analyzed further. Molecules 1, 2, and 3 are structurally similar to the literature OSDAs and have even lower binding energies. As with the literature OSDAs, two of these molecules fit inside the SFW cage.[29,30] These binding energies demonstrate the relationship between distance in the WHIM space and OSDA potential. It also demonstrates the model's ability to generate similarly structured molecules with known OSDAs with potentially greater templating ability. Molecules

4 and 5 are chosen due to their strong binding energies while being structurally dissimilar to the known OSDAs. Molecule 4 is significantly larger than the literature OSDAs indicating that a single, well-fitting OSDA per cage could also provide a strong templating effect towards SFW. Molecule 5 is significantly smaller than the literature OSDAs, requiring packing more molecules into the cage. These two molecules demonstrate the model's ability to suggest molecules that are dissimilar from existing OSDAs, potentially finding new OSDA families and providing additional value beyond taking distance metrics in the WHIM space.

4.5 Conclusion

In summary, this chapter highlights characteristics of OSDAs found in the zeolite literature, correlates OSDA molecules with specific zeolites through WHIM featurization to partially answer thesis questions two. It also predicts new OSDA-zeolite pairs through generative modeling partially answering thesis question three by providing researchers with tools that are potentially useful for accelerating OSDA design for zeolites.

Bibliography

[1] Z. Jensen, S. Kwon, D. Schwalbe-Koda, C. Paris, R. Gómez-Bombarelli, Y. Román-Leshkov, A. Corma, M. Moliner, and E. A. Olivetti, ACS central science **7**, 858 (2021).

[2] B. Lok, T. Cannan, and C. Messina, Zeolites **3**, 282 (1983).

[3] R. F. Lobo, S. I. Zones, and M. E. Davis, Journal of inclusion phenomena and molecular recognition in chemistry **21**, 47 (1995).

[4] M. Moliner, F. Rey, and A. Corma, Angewandte Chemie International Edition **52**, 13880 (2013).

[5] A. Burton, Catalysis Reviews **60**, 132 (2018).

[6] S. K. Brand, J. E. Schmidt, M. W. Deem, F. Daeyaert, Y. Ma, O. Terasaki, M. Orazov, and M. E. Davis, Proceedings of the National Academy of Sciences **114**, 5101 (2017).

[7] F. Daeyaert, F. Ye, and M. W. Deem, Proceedings of the National Academy of Sciences **116**, 3413 (2019).

[8] M. Moliner, P. Serna, Á. Cantín, G. Sastre, M. J. Díaz-Cabañas, and A. Corma, The Journal of Physical Chemistry C **112**, 19547 (2008).

[9] S. Zones, Microporous and mesoporous materials **144**, 1 (2011).

[10] J. Dědeček, E. Tabor, and S. Sklenak, ChemSusChem **12**, 556 (2019).

[11] B. C. Knott, C. T. Nimlos, D. J. Robichaud, M. R. Nimlos, S. Kim, and R. Gounder, Acs Catalysis **8**, 770 (2018).

[12] C. Li, A. Vidal-Moya, P. J. Miguel, J. Dedecek, M. Boronat, and A. Corma, ACS Catalysis **8**, 7688 (2018).

[13] J. G. Wicker and R. I. Cooper, Journal of chemical information and modeling **56**, 2347 (2016).

[14] R. Todeschini and P. Gramatica, SAR and QSAR in Environmental Research **7**, 89 (1997).

[15] A. Corma, F. Rey, J. Rius, M. J. Sabater, and S. Valencia, Nature **431**, 287 (2004).

[16] B. W. Boal, J. E. Schmidt, M. A. Deimund, M. W. Deem, L. M. Henling, S. K. Brand, S. I. Zones, and M. E. Davis, Chemistry of Materials **27**, 7774 (2015).

[17] T. Chatelain, J. Patarin, M. Soulard, J. Guth, and P. Schulz, Zeolites **15**, 90 (1995).

[18] T. Chatelain, J. Patarin, E. Brendle, F. Dougnier, J. Guth, and P. Schulz, in *Studies in Surface Science and Catalysis* (Elsevier, 1997), vol. 105, pp. 173–180.

[19] J. Dhainaut, T. Daou, A. Chappaz, N. Bats, B. Harbuzaru, G. Lapisardi, H. Chaumeil, A. Defoin, L. Rouleau, and J. Patarin, Microporous and mesoporous materials **174**, 117 (2013).

[20] P.-C. Kotsias, J. Arús-Pous, H. Chen, O. Engkvist, C. Tyrchan, and E. J. Bjerrum, Nature Machine Intelligence **2**, 254 (2020).

[21] S. Mohapatra, T. Yang, and R. Gómez-Bombarelli, Nature Machine Intelligence **2**, 749 (2020).

[22] R. J. Williams and D. Zipser, Neural computation **1**, 270 (1989).

[23] J. Arús-Pous, S. V. Johansson, O. Prykhodko, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen, and O. Engkvist, Journal of cheminformatics **11**, 1 (2019).

[24] N. Chen, A. Klushyn, R. Kurle, X. Jiang, J. Bayer, and P. Smagt, in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2018), pp. 1540–1550.

[25] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, et al., Frontiers in pharmacology **11**, 1931 (2020).

[26] J. Arús-Pous, T. Blaschke, S. Ulander, J.-L. Reymond, H. Chen, and O. Engkvist, Journal of cheminformatics **11**, 1 (2019).

[27] X. Song, J. Li, Y. Guo, Q. Pan, L. Gan, J. Yu, and R. Xu, Inorganic chemistry **48**, 198 (2009).

[28] F. O. Gaslain, K. E. White, A. R. Cowley, and A. M. Chippindale, Microporous and mesoporous materials **112**, 368 (2008).

[29] T. M. Davis, A. T. Liu, C. M. Lew, D. Xie, A. I. Benin, S. Elomari, S. I. Zones, and M. W. Deem, Chemistry of Materials **28**, 708 (2016).

[30] D. Xie, L. B. McCusker, C. Baerlocher, S. I. Zones, W. Wan, and X. Zou, Journal of the American Chemical Society **135**, 10519 (2013).

[31] S. I. Zones, *Zeolite ssz-13 and its method of preparation* (1985), uS Patent 4,544,538.

[32] E. M. Gallego-Sánchez, C. Li, C. Paris, N. Martín-García, J. Martínez-Triguero, M. Boronat Zaragoza, M. Moliner Marin, and A. Corma Canós, Chemistry-A European Journal **24**, 14631 (2018).

[33] J. R. Di Iorio, S. Li, C. B. Jones, C. T. Nimlos, Y. Wang, E. Kunkes, V. Vattipalli, S. Prasad, A. Moini, W. F. Schneider, et al., Journal of the American Chemical Society **142**, 4807 (2020).

[34] N. Martín, M. Moliner, and A. Corma, Chemical Communications **51**, 9965 (2015).

[35] D. Schwalbe-Koda and R. Gómez-Bombarelli, The Journal of Chemical Physics **154**, 174109 (2021).

[36] D. Schwalbe-Koda and R. Gómez-Bombarelli, The Journal of Physical Chemistry C **125**, 3009 (2021).

Chapter 5

Inorganic Zeolite Synthesis Planning

This chapter examines the rest of the zeolite synthesis space outside of OSDA design. It uses a number of data science techniques to answer questions two and three of the thesis, How can coupling of data-driven, first principles, and experimental approaches accelerate understanding of structure and processing relationships in zeolite materials? and In what ways can this data and discovered relationships be used to engineer improved zeolite materials?, by providing insight into how synthesis variables affect the crystallization of zeolites. It provides a number of ML models that can be used to help plan and expedite zeolite synthesis.

5.1 Introduction

While the previous chapter joins a growing body of research aimed at advancing and accelerating OSDA designs,[1–3] there has been much less attention paid to exploring and understanding the influence of other synthesis parameters using similar simulation and data driven tools. OSDA design plays a crucial role in determining the achievable subset of zeolite structures, but the majority of OSDAs are not ex-

clusively selective for a single zeolite structure.[3,4] This lack of selectivity requires design around these additional synthesis parameters, referred to in this thesis as "inorganic" variables, to synthesize the desired phase. Many studies have examined aspects of the zeolite synthesis space individually including compositional gel ratios (Si/Al, Na/Si, OSDA/Si, H_2O/Si, etc),[5–8] aging conditions,[9–11] crystallization conditions,[12–14] and precursor selection[15–17] for specific OSDA system, but general knowledge of these effects is lacking.

Zeolite crystallization is a complex process frustrating efforts to decouple the effects of many varying inorganic parameters. Most zeolites are metastable with energies approximately 30 kJ mol^{-1} greater than the ground state found in quartz.[18,19] This energetic disadvantage makes synthesizing a crystalline, porous zeolite rather than a dense crystalline or amorphous aluminosilicate uncertain and delicate. When a zeolite is formed, the crystallization process typically results in a characteristic "S" curve of crystallization time versus percent crystallinity that can be fit with experimental data.[20–24] Equation 5.1 is a common expression for the crystallization curve, referred to as the Gualtieri model.[25]

$$\alpha = \frac{1}{1 + exp\{-\frac{t-a}{b}\}} * \{1 - exp(-(k_g t)^n)\} \tag{5.1}$$

This equation treats the relationship between crystallinity (α) and crystallization time (t) as combination of distinct nucleation and crystallization phenomena. The left hand side of the equation corresponds to the number of nuclei in the system at time t, modeled as a cumulative Gaussian distribution. The right side is the common Avrami equation for crystal growth.[26] The model constants are fit using experimental data and have physical meaning within zeolite crystallization. The constant a is the inverse of the nucleation rate with $k_n = \frac{1}{a}$, while b is the standard deviation in the Gaussian nucleation probability equation and is related to the nucleation mechanism, k_g is the growth rate constant, and n is the growth dimensionality usually inferred from an SEM micrograph. This equation has been applied to numerous ex-

perimental studies in the porous materials space[27–30] although no sizable datasets of kinetic parameters exist for zeolites.

Data science has shown promise in studying reaction outcomes and modeling synthesis in both zeolites and chemistry more broadly. Many efforts have been made in the organic domain to classify reaction outcomes[31–33] and predict crystallinity.[34,35] Efforts in the zeolite field have primarily centered around suitable ODSA design,[1,3,36] although predicting reaction outcomes from the inorganic variables has also been successful in narrowly defined synthesis spaces including OSDA-free synthesis,[37] germanium-zeotypes,[38] ZSM-43,[39] LTA,[40] ITQ-21,[41] and zeolite Beta.[42] However, all of these studies are limited in scope to a specific zeolite sub-domain and stop short of predicting the full crystallization behavior due to lack of data and simplistic modeling choices. To expand the scope of crystallization modeling, more comprehensive data is needed. These data are provided by text extraction techniques highlighted in Chapter 3.

This chapter studies the impact of the inorganic synthesis variables on the crystallization behavior and reaction products in zeolite synthesis. The author models the probability of crystallization for a synthesis route and uses model interpretability to gain insight into the factors that determine the reaction outcome. The author then examines how the crystallization behavior of zeolites progresses by fitting kinetic equations to extracted crystallization curves and then modeling the behavior of unseen synthesis routes using a combined ML/Bayesian inference approach. Finally, the author generates potential synthesis routes for specific OSDA-zeolite pairs using generative modeling.

5.2 Characteristics of Zeolite Synthesis Data

Dataset 3.5.4 is the main dataset used in this chapter. This dataset contains comprehensive synthesis data on zeolite synthesis including gel composition, reaction conditions (aging, crystallization, pH, reactor size), precursors, and OSDAs. It also

includes the result of the zeolite synthesis including the zeolite materials and structures (or lack thereof) formed for each synthesis route and, in some instances, zeolite properties including Si/Al ratio in the product, crystal size, percent crystallinity, and BET surface area. The dataset consists of 23,925 synthesis routes from 3,096 journal articles spanning the years 1966-2021. It contains data on 921 unique OSDA molecules, 233 zeolite structures, and 1,022 unique materials. The extracted gel composition contains 51 different gel components including Si, Al, P, Na, K, F, Ge, Ti, B, Ga, V, OSDA, H_2O, and additional solvents.

Figure 5-1a-d shows a breakdown of some common observations across the dataset. Figure 5-1a shows the most commonly observed zeolite structures. Aggregating all amorphous and dense crystalline phases accounts for approximately 25% of the dataset. The most common zeolite is MFI, unsurprising due to the academic and industrial relevance of several important materials with the MFI structure including ZSM-5, silicalite-1, and TS-1. The remaining nine consist of other industrially important and well-studied zeolites often with multiple zeotype chemistries including CHA, *BEA, AFI, and FAU.[43] Figure 5-1b shows the most common zeotype chemistries. Conventional zeolite synthesis (Si-Al or pure Si framework) makes up the majority of the data. Other common zeotypes including aluminophosphates (AlPO), germanosilicates, titanosilicates, borosilicates, and other metal containing structures (Fe, Co, V, Zn, Sn, etc) also have a significant amount of the synthesis routes. Figure 5-1c shows the breakdown of the number of OSDAs used in a synthesis route. Using one type of OSDA molecule (1) is by far the most common, accounting for approximately 80% of the data. OSDA-free (0) and dual OSDA (2) synthesis account for most of the remaining synthesis routes while triple OSDA (3) systems are rarely observed. Figure 5-1d shows the most commonly observed precursor materials utilized in zeolite synthesis. NaOH is the most observed precursor, used as a source of Na and the OH^- mineralizer. Others in the top 10 include TEOS and Ludox AS-40 as Si sources, GeO2 for Ge, HF and NH_4F for F^- anion, phosphoric acid for P in AlPO systems, and sodium aluminate and aluminum isopropoxide

(AIP) for Al. There is a large variety in precursors with 44 unique source for Si and 32 for Al.

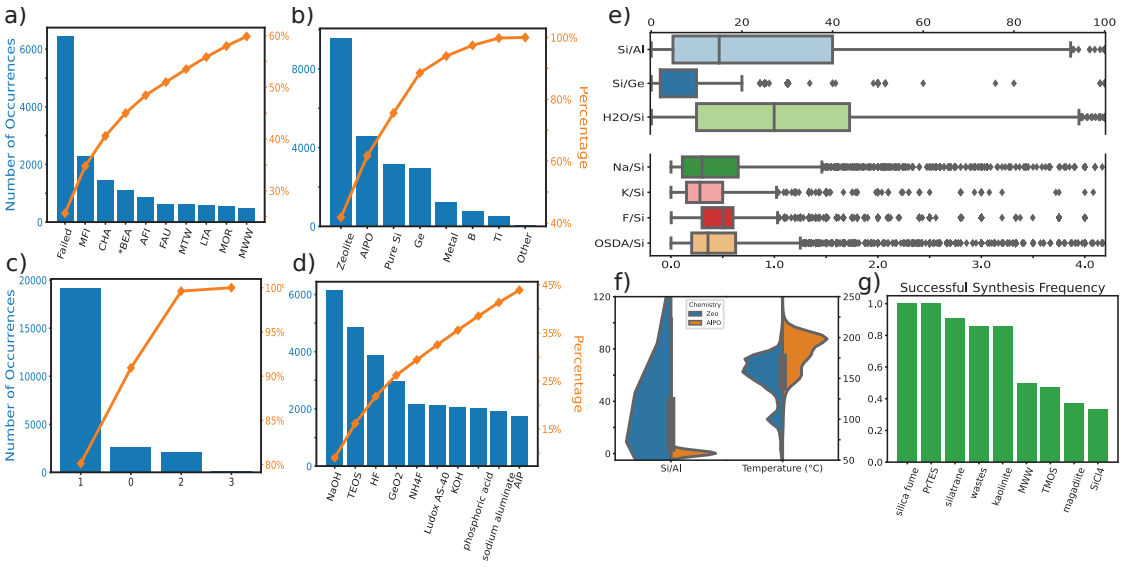


Figure 5-1. Overview of the extracted zeolite synthesis dataset. a) Most commonly observed zeolite structures. b) Most commonly observed chemistries. c) Number of synthesis routes that utilize that number of OSDAs. d) Most commonly observed precursors. e) Observed ranges for several import gel composition variables. f) Difference between conventional zeolite chemistry and AlPO-type observed in the Si/Al ratio and crystallization temperature. g) Frequency of successful synthesis starting from different Si precursors

Figure 5-1e shows the distribution and range of several important gel compositional ratios including Si/Al, Si/Ge, H_2O/Si, Na/Si, K/Si, F/Si, and OSDA/Si. Common Si/Al values typically range from 5 to 40 although a significant number of the synthesis routes take place above or below this range. While conventional zeolite synthesis typically occurs with Si/Al > 1, values below 1 exist in the dataset due to the presence of AlPO and other zeotype synthesis routes.[44] The bottom four ratios represent the common inorganic and organic structure directing agents compositional ratios along with the F^- mineralizer. As is seen, these ratios with Si are typically below 1 but outliers do exist. OSDA-free synthesis often uses an abundance of cations including Na and K to provide a structure directing effect in the absence of an OSDA.[45–48] F^- and OSDA abundance often occurs in germanosilicates where both the F^- and OSDA play a crucial role in the formation of large pore zeolites.[49,50]

Visualizing trends allows basic insight into the data. Figure 5-1f shows the difference in the Si/Al ratio and crystallization temperature between conventional zeolite and AlPO synthesis routes. As expected, the Si/Al ratio is significantly higher for conventional zeolites due to Lowestein's rule.[51] AlPO synthesis often follows a formula of Si_xAlP_{1-x} with x varied from 0 to 1.[44] This explains the high density of AlPO synthesis routes with low Si/Al ratios. A difference is also observed in the crystallization temperatures. Conventional zeolites synthesis generally occurs at lower temperatures than AlPO synthesis with very few AlPO synthesis routes below 130°C. In contrast, many stable zeolite phases can be synthesized at 100°C which is observed by the second density peak. Additionally, we look at trends within the extracted precursors. Figure 5-1g shows the normalized frequency that a Si precursor results in a successful zeolite synthesis from the extracted data. The left five-most precursors are the most likely to be successful while the right four precursors the least likely across the 44 Si precursors. While no direct thermodynamic or kinetic information can directly be taken from this analysis, it is useful to examine this type of relationship from a frequentist approach to help guide precursor choice.

5.3 Modeling Crystallization Probability

This section describes modeling the crystallization probability i.e. trying to answer "if" a zeolite will form at specific synthesis conditions or if the product will be an amorphous or dense crystalline phase. The crystallization probability should not be confused with the percent crystallinity of a zeolite sample which is the focus of section 5.4. Crystallization probability is modeled as a classification problem between "successful" and "failed" synthesis routes.

5.3.1 "Failed" Synthesis Data

The lack of negative or "failed" data presents major problems in data driven synthesis and science as a whole. This data gap leads to positive results bias[52] where

the vast majority of data available in the literature contains positive results biasing perceptions of chemistry and hindering scientific progress.[53,54] Having access to negative results can greatly advance understanding of a system and accelerate materials discovery.[55,56] Fortunately, zeolites are a rare sub-domain where researchers often publish negative results, amorphous or dense crystalline phases, alongside successful zeolite crystallization. This negative data is captured through the data extraction process and then can be modeled to determine whether a synthesis route will be successful. The inclusion of negative data in the zeolite field does not remove bias from the data however. Certainly there is still positive results bias within the zeolite literature somewhat masked by the inclusion of some negative data. This bias is hard to quantify but needs to be kept in mind when interpreting results in this thesis chapter.

Synthesis routes from dataset 3.5.4 are converted into either "successful" or "failed" based on the text-extracted products. If the extracted products contain a word from a manually curated list of negatives including "amorphous", "dense", "quartz", etc, the synthesis is considered "failed". All other syntheses are "successful" including multiphase zeolite products and "zeolite-like" products that do not have official IZA structure codes such as ITQ-21 and ASU-14. This choice also affects the overall analysis and other rational choices exist such as classifying multiphase zeolite products as "failed." Overall, there are 19,275 synthesis routes with 13,537 "successful" and 5,738 "failed" synthesis used for the classification model. Figure 5-2 shows the breakdown of different subgroups within these two class labels. "Sucessful" makes up the two most common subgroups, single zeolite and multiphase products. The rest are "failed" syntheses. Among the "failed" syntheses, "amorphous" is by far the most common product followed by "dense" and a long tail of much less frequent intermediate phases including "layered" and "lamellar" and dense crystalline phases including "quartz" and "tridymite" for zeolite synthesis and "berlinite" for AlPO synthesis. "Magadiite" is a sodium silicate mineral sometimes used as a zeolite precursor indicating that a zeolite never formed with that synthesis route.

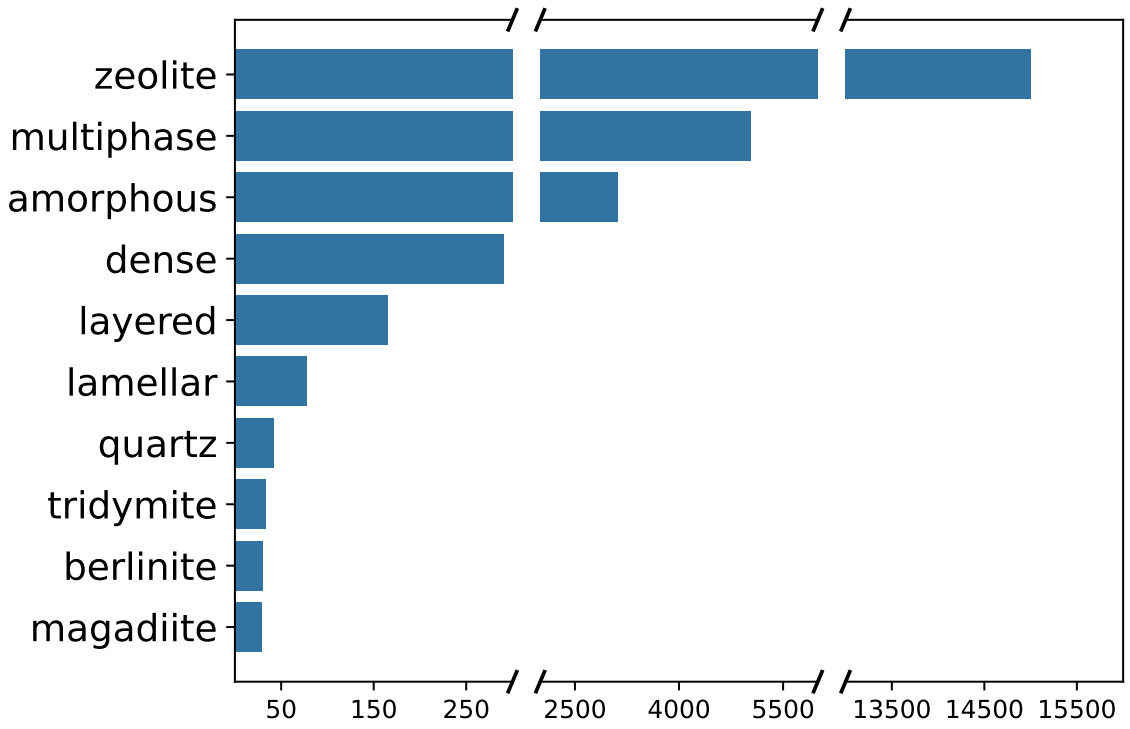


Figure 5-2. Breakdown of the most common "Successful" and "Failed" extracted products in dataset 3.5.4

5.3.2 Model Description, Optimization, and Performance

This data is used in a classification model to predict crystallization probability, whether a zeolite synthesis will be successful. Figure 5-3 shows the model schematic. Each synthesis route consists of four input categories: gel composition, reaction conditions, precursors, and OSDA. Gel composition is the normalized molar ratios of each synthesis element. There are 51 different gel components including Si, Al, P, Na, K, F, Ge, Ti, B, Ga, V, OSDA, H_2O, and additional solvents. Gel composition is a sparse input since most gel components are zero for any individual synthesis route. Reaction conditions comprise aging time, aging temperature, crystallization time, crystallization temperature, rotating or static reactor, seeding behavior, and presence of microwaves. Precursors are featurized according to section 3.4.3 and converted into a continuous two-dimensional representation by an autoencoder. OSDAs are featurized according to section 3.4.1 using the neural network method. Each OSDA (up to three) utilized in the synthesis is converted to a two dimen-

sional latent representation with an autoencoder and combined to give a six length representation for each OSDA system. These four inputs feed into a classification algorithm that predicts the probability of a successful zeolite crystallization. Model probability above 0.5 is interpreted as "successful".

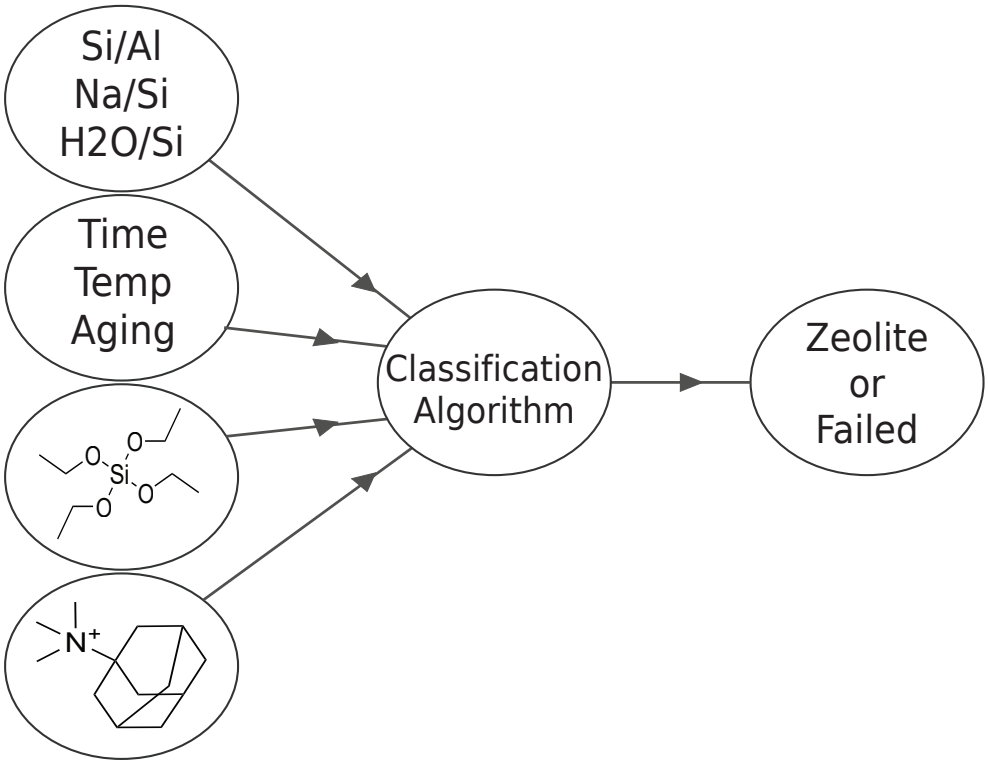


Figure 5-3. Schematic of classification approach to modeling crystallization probability.

Two train/validation/test split schemes are used for optimization and performance evaluation of the algorithm. The first splits 10% of the data randomly into test set and an additional 10% of the remaining data into validation set for optimization of the model's hyperparameters. The remaining data comprises the training data. The second split holds out 10% of the unique OSDAs systems in the data as the test set and an additional 10% of OSDAs for the validation. This split ensures that the model is tested on never before seen OSDA systems, mimicking performance on a newly discovered OSDA. Both of these splits incorporate a 10-fold cross validation loop so every data point is tested in exactly one fold. The results are aggregated

together to examine classifier performance.

Within each fold, five different classification algorithms are trained to compare performance: a dummy classifier, logistic regression, deep neural network, random forest, and XGBoost. The algorithms are optimized using a Bayesian optimization search algorithm[57,58] to find the optimal set of hyperparameters maximizing the macro F1 score across the validation set. Each hyperparameter space is unique to the specific classification algorithm. For example, random forest hyperparameters include the number of trees in the ensemble, classes weighting scheme, and minimum samples to consider for an ending leaf.

Figure 5-4 shows the classification performance of the five different classification algorithms evaluated after optimization for the two different train/test splits. The results displayed are the result of aggregating all predictions over the ten-fold cross validation. As expected, the random split performs better as the model potentially sees more information about the OSDA system of each synthesis route. However, even in the OSDA split, all the algorithms are substantially better than the dummy classifier (random guessing based on distribution) indicating the model is capable of providing value on unseen OSDAs. Of the five classification algorithms, the tree models, random forest and XGBoost, have the best performance on both train/test splits with an area under the curve (AUC) score of 0.89 and 0.9 respectively for random split and 0.68 and 0.71 for the OSDA split. These metrics indicate these algorithms can provide robust predictions both in an interpolation context and on new, previously unseen systems.

5.3.3 Model Interpretability

Beyond providing accurate predictions, the classification model can also provide insight into why it makes certain predictions and the relative importance of each of the synthesis variables. For each prediction, Shapley (Shap) values are calculated which use a game theory approach to determining each feature's contribution to the

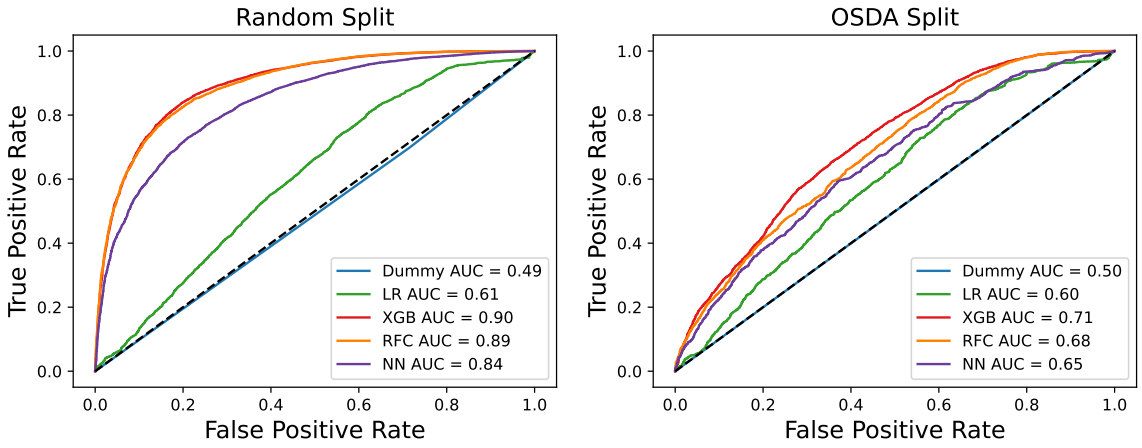


Figure 5-4. Classification performance across different train/test splits and algorithm types for the crystallization probability model.

model prediction.[59–61] These values can provide insight into the crystallization process. Shap value calculations are done on the XGBoost model using one randomly held out test set for all figures.

Figure 5-5a shows the Shap values for all synthesis routes sorted by the maximum absolute impact on the the model output. The color represents the relative value of that feature value. Crystallization time has the largest impact of any synthesis variable. There are a number of low crystallization time values that have a significant negative impact on the synthesis result. This result agrees with intuition as nucleation and growth often takes longer timescales for zeolite crystallization. Aging time also has a strong impact on model output although there is not a easily identifiable trend for high or low aging time values. Crystallization temperature is another important synthesis variable that generally agrees with intuition. Higher crystallization temperatures are generally correlated positively with crystallization percentage except for a scattering of synthesis routes with high crystallization temperature but very negative impact on the crystallization probability. Too high of temperatures results in dense crystalline phases[62] which may be responsible for these results. Finally, several compositional variable have interesting trends as well. F, Ge, and Na are all important gel components that can play a stabilizing role in certain zeolite structures.[63–65] As such, higher values are generally associated with higher crystal-

lization probability. H_2O shows the opposite trend with high values associated with lower crystallization probability due to the difficulty in synthesizing zeolites at low concentrations.[66]

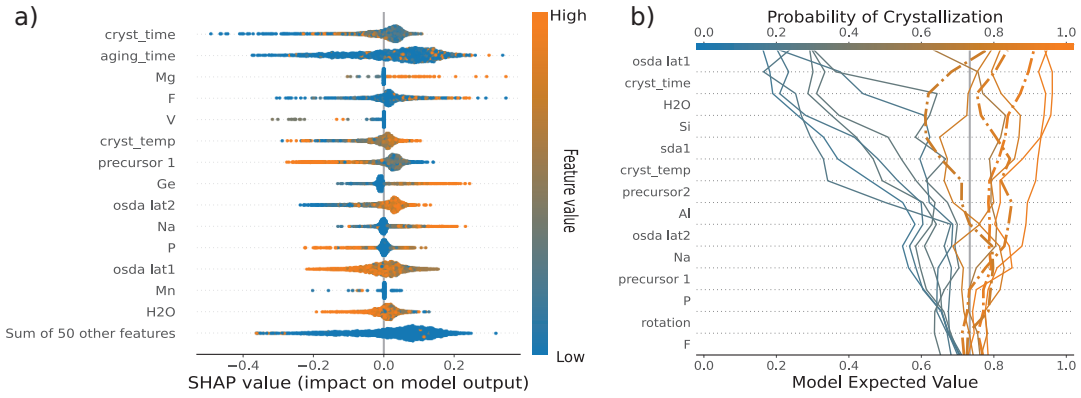


Figure 5-5. High level visualization that demonstrate particular features' impacts on the crystallization probability model. a) Top 15 features ranked by maximum absolute impact on a sample. b) Impact of specific features on 15 selected samples from the test set. Dotted lines indicate the sample was misclassified by the model.

The decision plot also provides an interesting visualization for examining synthesis routes with Shap values. The plot accounts for the impact of individual features by tracing the path from the expected value of the data to the model's output crystallization probability. Solid lines correspond to correct predictions from the model while dashed lines are errors. Figure 5-5b shows this plot for fifteen synthesis route. Obvious trends affecting crystallinity across many samples will show up in this type of plot. The disorganized and random behavior observed between the synthesis routes indicates there are many different routes to achieving crystallinity in zeolites suggesting it is probably beneficial to examine subsets of the zeolite synthesis space and even individual synthesis routes.

5.3.4 Test Case: TMAda

The complex nature of the synthesis space often necessitates examining synthesis routes on a individual basis to gain insight. This thesis examines one specific OSDA system, N,N,N-trimethyladamantammonium (TMAda), although the following type

of analysis could be applied to any subset of the zeolite synthesis space. TMAda is chosen as a parallel to section 4.4.3, the industrial relevance of its main zeolite product CHA, and the abundance of both successful and failed synthesis routes in the data.

Figure 5-6 shows the most important synthesis variables' influence on four selected TMAda synthesis routes. The Shap values for each synthesis route are presented as arrows with orange as a positive influence on crystallization probability and blue a negative impact. While whether a high or low value for each synthesis parameter is causing the effect cannot be directly interpreted from these plots, it is often possible to infer from zeolite domain knowledge. Figure 5-6a shows a successful CHA synthesis route[67] correctly predicted by the model to be successful with high confidence (0.915). Several important synthesis variables, Al content, crystallization time, the precursors (FAU zeolite), and rotation all have a positive influence on the crystallization. The only negative is the OSDA amount with a minor influence. Figure 5-6b shows a failed synthesis resulting in an amorphous phase[65] also predicted correctly by the model (0.308). This system is a pure Si framework with K cations. The lack of Al has a strong negative influence on the prediction along with the precursor selection (Fumed silica and KOH). FAU zeolite is typically a better precursor than different types of amorphous silicas[68] which helps explain the positive precursor influence in a and negative in b. This underscores the complexity of zeolite synthesis and how Shap values provide tools to start unraveling these relationships for individual synthesis routes.

Synthesis routes with less predicted certainty are also potentially interesting. Figure 5-6c shows a synthesis route correctly predicted but right on the edge of the classification threshold (0.495) resulting in a mixed AFI and dense crystalline phase.[69] Examining the Shap values potentially gives an indication of how improvements could yield a pure zeolite phase. From the Shap values, the model would predict a crystalline phase except for OSDA content. This synthesis route uses a very low OSDA/Si ratio, 0.03. Raising the ratio could be a way to accomplish a pure zeo-

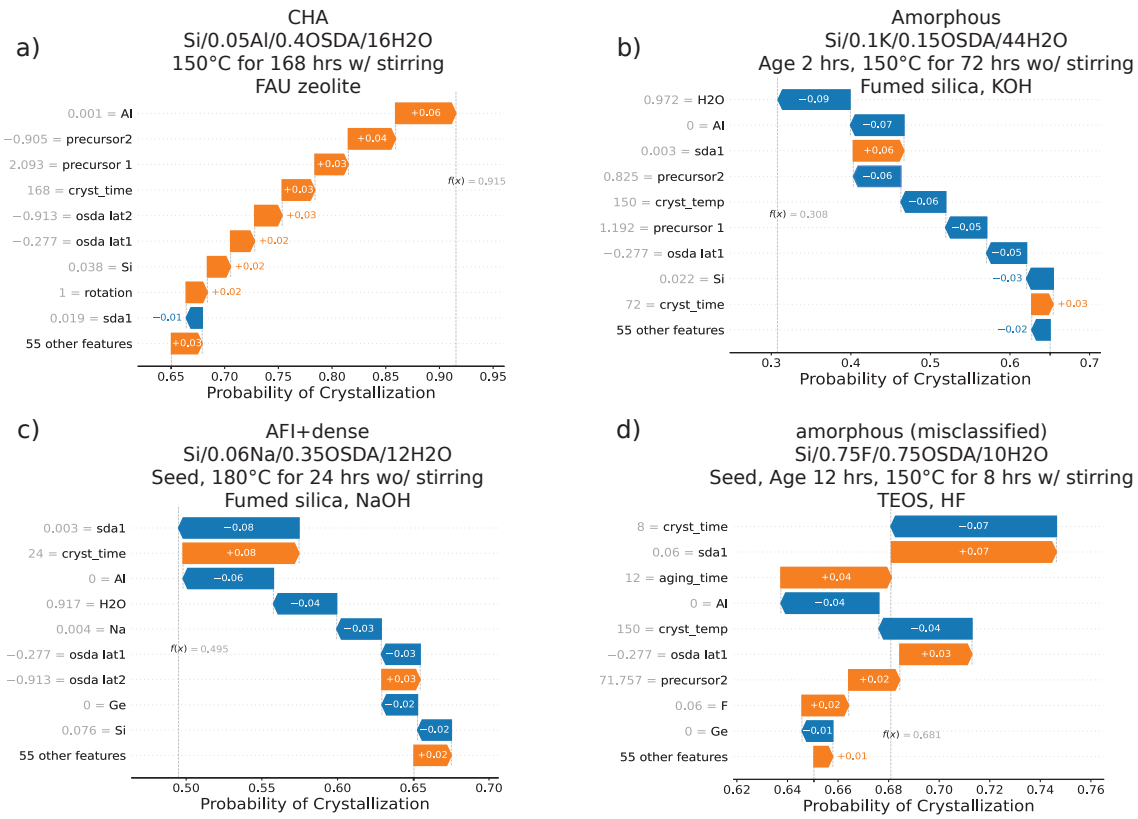


Figure 5-6. Shap values for four synthesis routes using TMAda. a) Successful CHA synthesis route.[67] b) A failed amorphous route.[65] c) A failed synthesis route resulting in AFI mixed with dense crystalline phase.[69] This route is very close to the prediction threshold. d) A failed amorphous synthesis route misclassified as a successful synthesis.[70]

lite phase. Figure 5-6d shows an incorrectly classified failed synthesis route that the model predicts as successful.[70] The Shap values provide insight into which synthesis variables confuse the model. The crystallization time for this route is low, only 8 hours. The same synthesis route at 24 and 96 hours yield a pure zeolite phase (STT) indicating the model does not put enough negative impact on the low crystallization time for this synthesis route.[70]

Aggregating the results from these four synthesis routes provides insight as well, although with only four points the results need to be investigated further. In all four systems, having Al in the framework is a positive on crystallization probability while a pure Si framework is a substantial negative. This reinforces the idea that

aluminum can help stabilize zeolite frameworks.[71,72] Crystallization temperature also has interesting commonalities between synthesis routes. When the crystallization temperature is 150°C (b and d), it has a considerable negative effect on crystallization probability. For the higher temperature systems, it does not seem to play a large role. This could indicate that zeolites in the TMAda system are best synthesized with a higher crystallization temperature.

5.4 Crystallization Curve Modeling

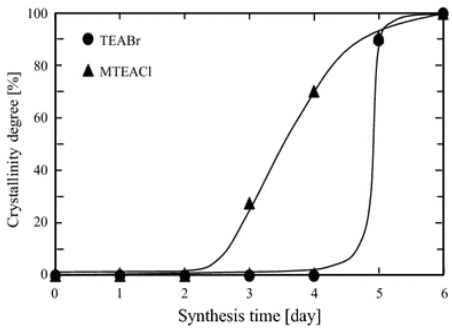
The previous section investigated crystallization probability, attempting to answer "if" a zeolite will form. This section informs "how" a zeolite will crystallize by examining the evolution of crystallizinity with time. It examines zeolite crystallization behavior as a function of the synthesis variables and combines multiple types of crystallization data into a Bayesian inference framework.

5.4.1 Crystallization Curve Data

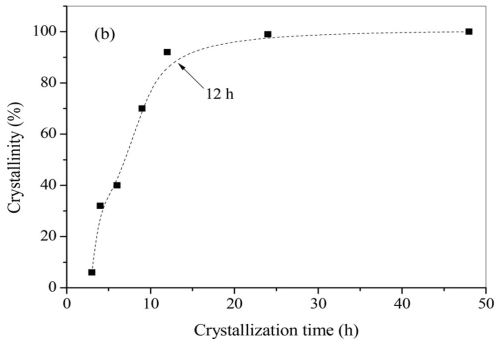
Crystallization data exists in two forms in the literature, as quantitative data forming the characteristic "S" curve typically found in article Figures and qualitative data that describes the outcome of a synthesis route such as "Amorphous", "Dense", "ZSM-5", and "SSZ-13+amorphous" typically found in text and tables. A huge difference in data volume between the quantitative and qualitative crystallization data exists due to considerable characterization efforts required to quantify the entire crystallization curve and the considerable efforts required to extract that data from article figures. The quantitative data consists of 291 crystallization curves (dataset 3.5.5 while the qualitative data, found in dataset 3.5.4, is orders of magnitude larger with approximately 20,000 synthesis routes but without detailed crystallization information. However, using a Bayesian inference approach, both types of data can be used to create a model that predicts the crystallization behavior. This usage of two datasets with different levels of data quality is often called multi-

fidelity data and has been used effectively in the materials space to study property correlation and crystallinity.[73–75]

a)



b)



c)

Table 1. Selection of the Most Representative Hydrothermal Syntheses Performed under Static Conditions at 170°C for 14 Days in the (Si, Ge) System with DMAP as the OSDA[a]

	molar synthesis mixture compositions				
sample	Si/Ge	DMAP/T[b]	HF/T[b]	H_2O/T[b]	products[c,d]
1	0.4:0.6	0.5	0.5	8	Q + A
2	0.6:0.4	0.5	0.5	8	Q + A + IM-18
3	0.75:0.25	0.5	0.5	8	IM-18 + A + Q

Figure 5-7. Demonstration of the two types of crystallization data found in the literature. a) Quantitative data from ref.[76] b) Quantitative data from ref.[77] c) Qualitative data from ref.[78]

5.4.2 Gualtieri Model Fit

The first step in the modeling process determines a, b, and k_g for each extracted curve by fitting the extracted experimental data with the Gualtieri model. The fit is performed across the extracted crystallization curves using the SciPy package[79] with a dogbox implementation of least squares optimization.[80,81] 0.001 to infinity acts as the bounds for all three kinetic parameters. n is assumed to be 3 unless that results in a poor fit in which case n = 1 and 2 are computed in addition. Poorly fit curves (low R2 score) are revisited manually to improve the parameters.

Figure 5-8 shows the fitting process with data from ref.[76] and the R2 values for

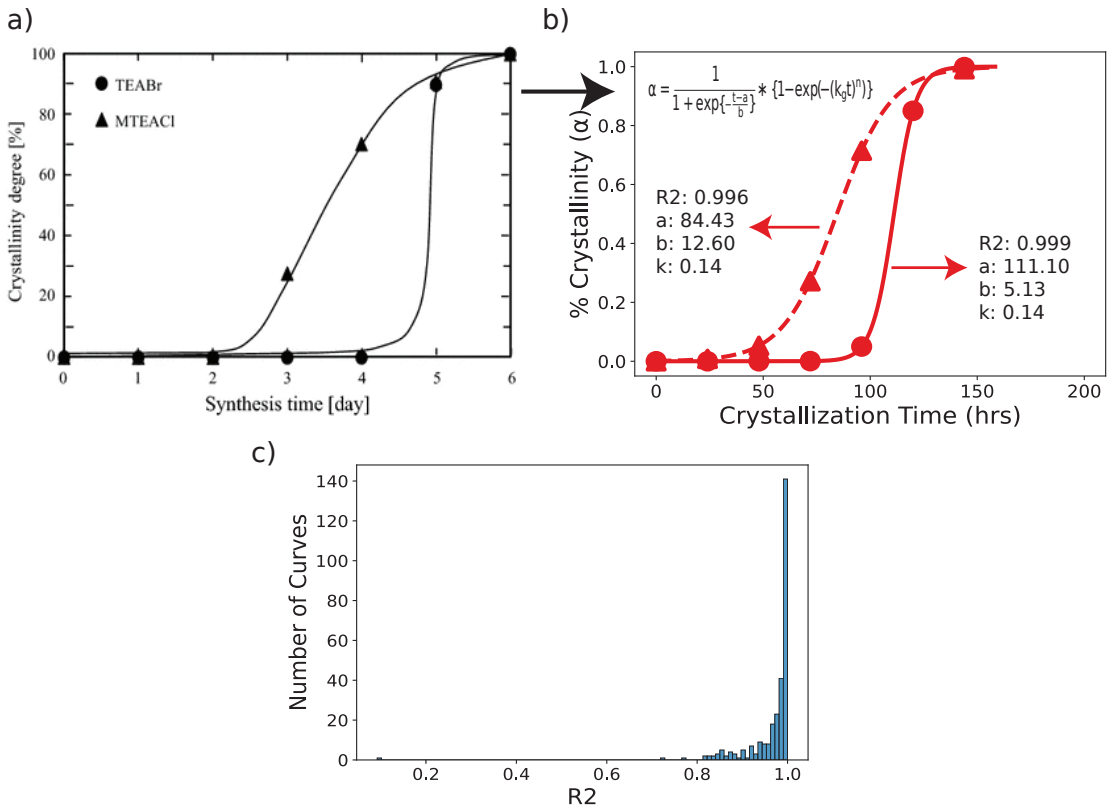


Figure 5-8. Crystallization scheme and results a) Experimental data from ref.[76]. b) Fitted experimental results using the Gualtieri model. c) R2 scores from all 291 fits.

the data overall.The experimental data fits very well to the Gualtieri model with a median R2 of 0.990 and R2 for 90% of curves above 0.90. Figure 5-9 shows the a, b, and k_g values for all the extracted curves. All three parameters have dense clusters at low values with some large outliers although b and k_g display this much more strongly than a. These values are used as the "true" parameters in the following modeling.

Since all subsequent modeling efforts are based on the parameters extracted from this fitting process, it is worth discussing potential problems with fitting the data and with the Gualtieri model more generally. The first set of problems are numerical. Due to the difficulty in characterization of the early stages of nucleation,[82] there are often not enough data points at low α values to accurately determine a and b resulting in a range of a and b values that do not affect the goodness of fit (R2)

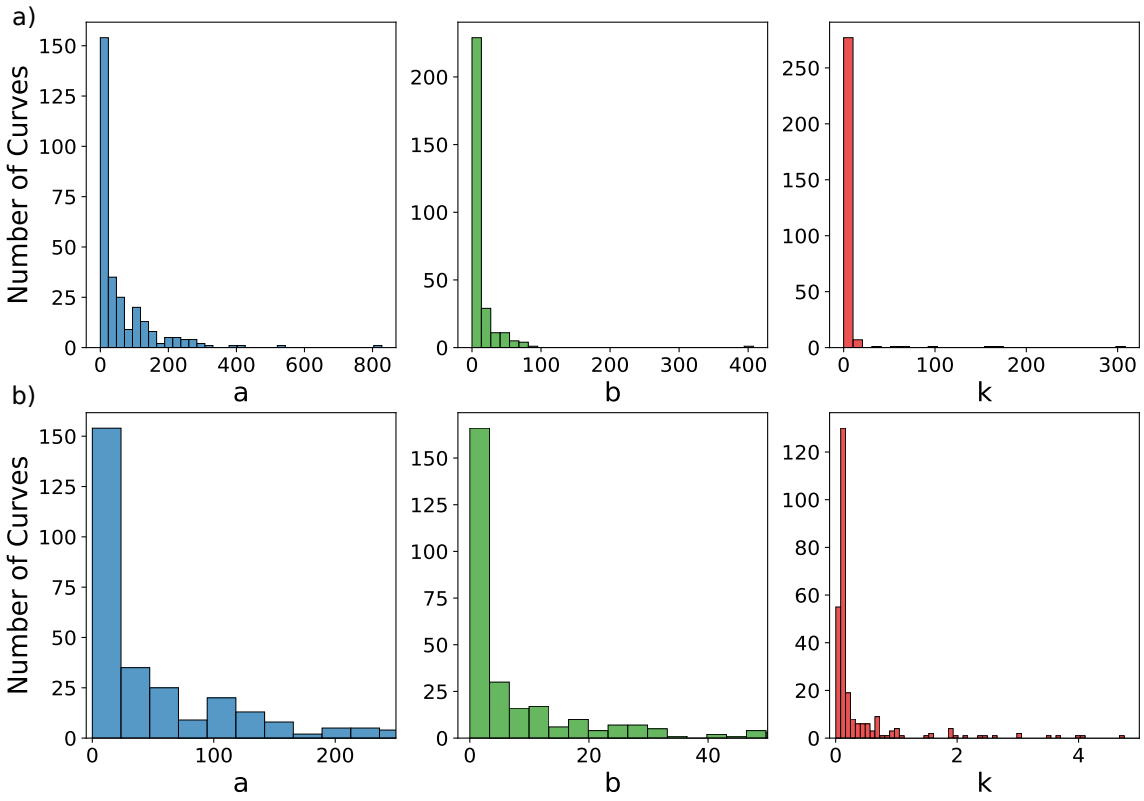


Figure 5-9. a, b, and k_g histograms for the 291 extracted curves. a) Full histograms. b) Zoomed in on dense areas.

to the data. Instead of evaluating trends in the parameters themselves, it is often better to calculate properties of the curves numerically such as crystallization time to α= 0.1, 0.5, 0.9 and difference in time between α= 0.1 and α= 0.9. The second set of problems involve the Gualtieri model and its assumptions. The assumption of a normal distribution to model the probability of nucleation may be too simplistic for certain types of synthesis routes. Synthesis routes that utilize seeds, other crystalline zeolites as precursors, or age prior to crystallization may already have nuclei present at the start of the crystallization step.[83] Other probability distributions such as an exponential distribution may better describe the nucleation behavior for such synthesis routes.

5.4.3 Crystallization Prior Modeling

The first modeling step predicts a, b, and k_g as a function of the synthesis route using ML. Difficulties arise since data is very limited (291) and zeolite synthesis is a very complex feature space. This input spans similar variables as the previous section with gel chemistry, reaction conditions, precursors, and OSDAs with an additional variable, the zeolite structure undergoing crystallization represented with a compressed latent representation learned with an autoencoder similar to the precursors and OSDA inputs. The gel composition is also compressed to a five length vector using PCA to reduce the dimensionality of the input space. The first five principal components account for 95% of the variance in the gel composition space. Three separate models are trained with the same input space for a, b, and k_g independently. The ML model is a random forest model that utilizes jackknife variance estimates to quantify uncertainty.[84,85] This uncertainty estimate is important for the Bayesian inference described in the subsequent section. Since the three parameters are related, the model's hyperparameters are tuned simultaneously using the area between the current predicted curve and the true curve as the optimization function. Through the modeling section, the ML and Bayesian inference predictions are compared to randomly sampling a, b, and k_g values from their underlying distributions described in Figure 5-9.

Figure 5-10 shows the results from the ML model running a leave-one-out cross validation. As hypothesized, the model performs poorly on this small dataset with R2 scores of 0.21, 0.03, and -0.19 for a, b, and k_g respectively. However, the model does perform better than randomly sampling the distributions which gives R2 values of -1.45, -0.27, and -1.98 for a, b, and k_g. This performance difference is also highlighted by comparing the difference in area between the true and predicted curves. This metric tests the models ability to mimic the shape and location of the crystallization curve more broadly rather than looking at each kinetic parameter individually. The ML-learned prior has a statistically significant lower distribution relative to the random sampling with a p-value of 2.7E-7 using a Kolmogorov-Smirnov test.[86]

This metric demonstrates again the value provided by the ML model as opposed to random sampling.

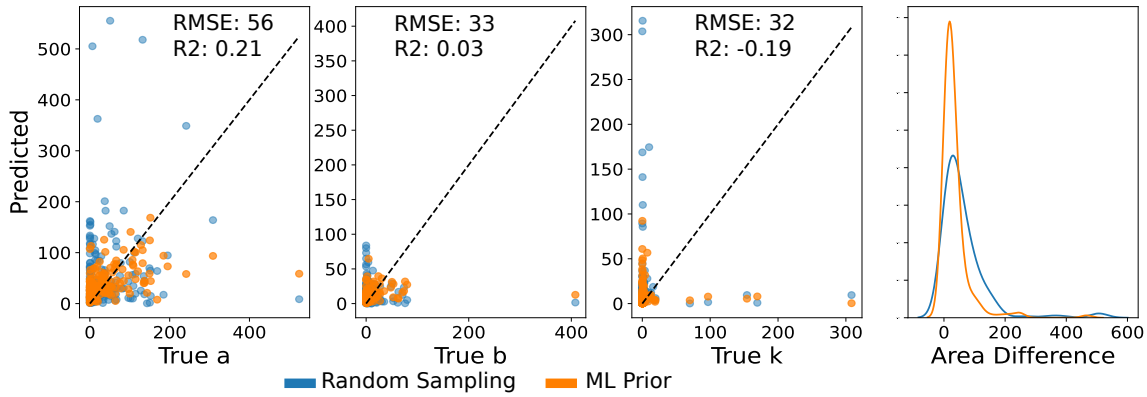


Figure 5-10. True versus predicted values for a, b, and k_g from the ML model and randomly sampled. The area difference is difference in area between the predicted curve and the true curve for the two different schemes.

5.4.4 Bayesian Inference for Posterior Estimate

As expected due to the difficult nature of the data and input space, the ML prediction alone provides a poor estimate of the crystallization curves although better than simply sampling the underlying distribution. Luckily, there is another source of data, the qualitative data that can help improve predictions. This data is incorporated into a Bayesian inference framework shown in Equation 5.2.

$$\mathbb{P}(a, b, k_g|t, \alpha, syns) \propto \mathbb{P}(t, \alpha|a, b, k_g, syns) * \mathbb{P}(a, b, k_g|syns) \tag{5.2}$$

The right most term in Equation 5.2 is the prior distribution or the original estimate of a, b, and k_g without incorporating the qualitative data. The previous section ML modeling acts as our prior or the probability of a, b, and k_g conditioned on the synthesis route with the model prediction as the mean of a normal distribution and the uncertainty estimate as the standard deviation. The qualitative data, crystallization time (t) and α, are incorporated in the likelihood distribution, the middle term, which quantifies the probability of t and α given the predicted a, b,

and k_g values. The α value of each qualitative point is estimated based on its text representation. For pure zeolites, amorphous samples, amorphous+zeolite, and zeolite+amorphous, $\alpha \simeq$ 1.0, 0, 0.33, and 0.66 respectively. Only synthesis route with a single observed zeolite structure are considered. The prior and likelihood are combined into the posterior, the probability of a, b, and k_g conditioned on both the synthesis route and the qualitative data. The posterior calculations are performed using PyMC3.[87] Figure 5-11 demonstrates this process on data from ref.[76]

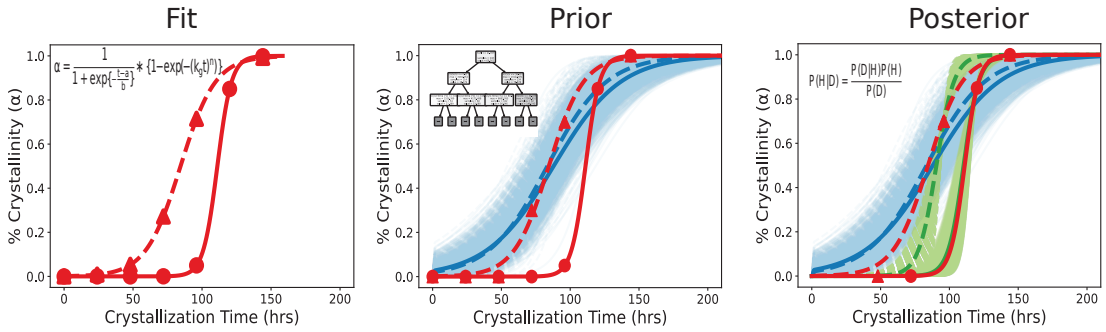


Figure 5-11. Progression of crystallization modeling from Gualtieri fit to ML prior modeling to posterior estimation using qualitative data.

Figure 5-12 shows the prediction results for the posterior estimates of a, b, and k_g. The posterior is sampled 1,000 times with the mean used as the prediction of a, b, and k_g. The posterior predicts a and b better than the prior although interestingly worse for k_g. Similar to the prior modeling, the posterior performs worst on the prediction of k_g. This trouble could be from the structure of the extracted k_g with several very large outliers seen in Figure 5-9. A similar area comparison also reveals the posterior improves the prediction. The posterior area difference distribution is lower than the prior distribution with a p-value of 2.4E-8. These results indicate that including the posterior calculation improves the ability to predict crystallization curves although it is still quite challenging.

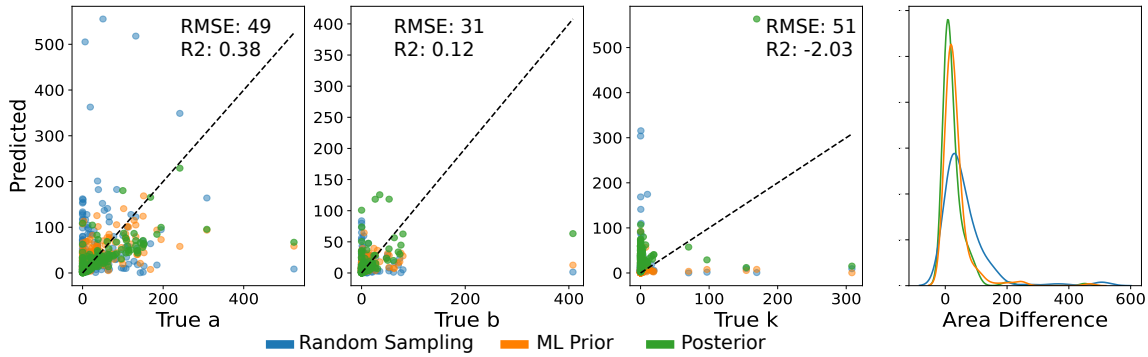


Figure 5-12. Comparison to true versus predicted parameters for the posterior estimate.

5.4.5 Crystallization Trends Across Data

This Bayesian inference framework is run across all unique syntheses resulting in a pure phase zeolite structure compiling a dataset of approximately 10,000 predicted crystallization curves. Rather than compare against a, b, and k_g directly, it is better to numerically compute values from the predicted crystallization curves. Two important values are the crystallization time to 10% crystallinity ($\alpha = 0.1$) which can loosely be considered the nucleation time and the crystallization time from 10% crystallinity to 90% crystallinity which can be considered the growth period. Similar to the findings in the previous section, the relationships between the crystallization behavior is very complex making evaluation on specific systems typically necessary. However, there are some general observations that can be made.

Figure 5-13 demonstrates some of these general trends in both nucleation and growth for several synthesis conditions. Aging time shows a clear trend with high aging times correlated with both fast nucleation and growth conditions. Aging typically helps order the gel phase and pre-promote nucleation so this finding agrees with intuition.[10,83,88] Other synthesis conditions like the use of microwaves and seed also has a strong impact on increasing nucleation speed. Microwaves appear to also speed the growth process while seeding appears to play a much smaller role in growth than nucleation. Other variables such as rotation, appear to have a much smaller impact on specific nucleation and growth values at least in a general

context.

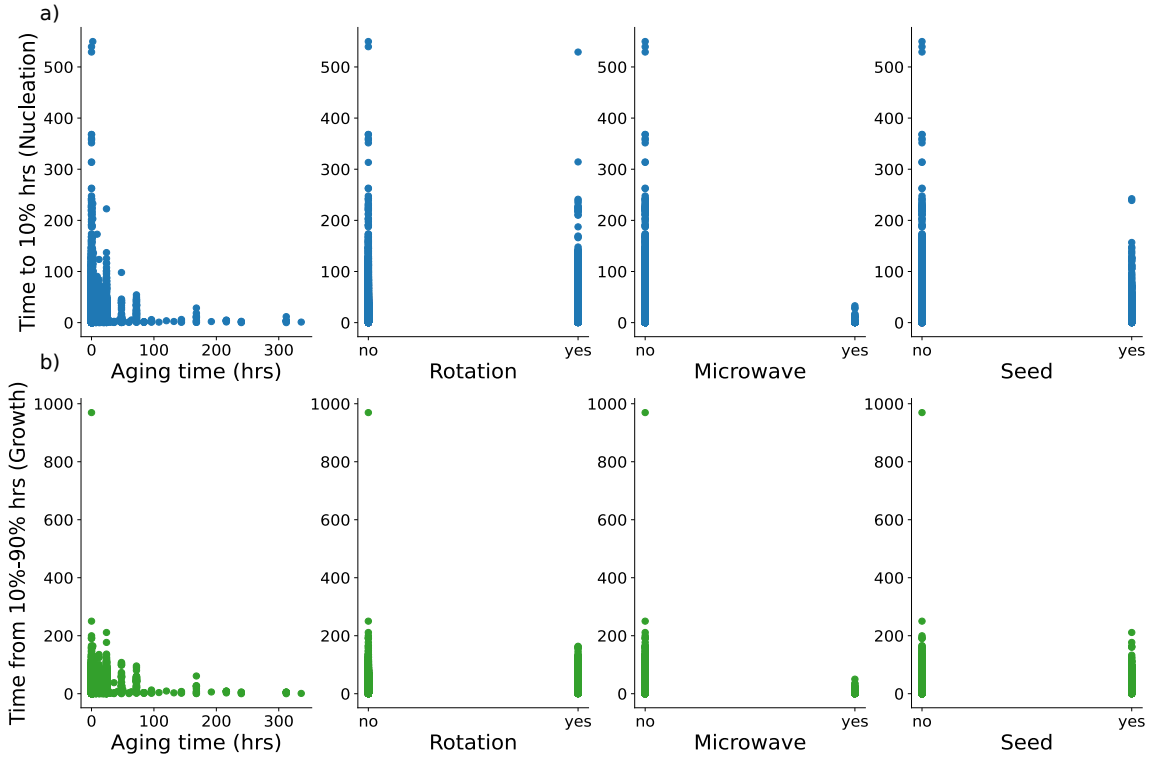


Figure 5-13. General trends between crystallization parameters and synthesis conditions. a) Synthesis variables compared with the time it takes the synthesis to reach 10% crystallinity. b) Synthesis variables compared with the time it takes for the system to growth from 10% to 90% crystallinity.

Taken together with the previous section on modeling crystallization probability, researchers can use this Bayesian modeling approach to gain insight into specific zeolite systems of interest. It looks at the "how" aspect of zeolite crystallization, predicting the behavior from a synthesis system known to crystallize into a zeolite. Another interesting aspect of Bayesian inference is the iterative nature. Data can continuously added until the prediction reaches suitable accuracy and enough knowledge is gained about the system. This type of Bayesian model could be incorporated with experiments to target specific areas of zeolite synthesis with unknown kinetics.

5.5 Generative Modeling of Inorganic Conditions

The previous two sections helped explain the crystallization behavior of a given zeolite synthesis route. One remaining question is what synthesis route should be used to make a specific zeolite. This section provides tools to help answer that question through the use of generative modeling to predict potential synthesis routes given a specific zeolite and OSDA pair.

5.5.1 Model Description and Optimization

The generative model is a series of CVAEs that each predict an aspect of the synthesis route conditioned on the given OSDA-zeolite pair as well as the previous CVAE output as shown in Figure 5-14. First, the model generates the synthesis components best suited for the OSDA-zeolite pair. This limits the chemical space by narrowing to a specific zeotype such as conventional Si-Al zeolites, AlPO, or germanosilicates. It also predicts additional synthesis components such as alkali cations and F^- as a mineralizer and binary reaction variables such as aging, reactor rotation, and seed usage. Second, the model predicts the gel composition conditioned on the selected synthesis components. Third, the model predicts the reaction conditions such as aging behavior, crystallization time, and crystallization temperature conditioned on the gel composition. Finally, the model predicts suitable precursors. Each model is conditioned on an OSDA-zeolite pair featurized as described in section 3.4.1 and 3.4.2. The sequential nature of the modeling process has precedent in predicting reaction conditions[89] and has significant benefits compared to an all-in-one model. Sequential modeling allows for easier, independent optimization of each model. It also allows users control over which parts of the synthesis process to model and incorporate their own domain expertise and intution into the model. For example, a researcher may have strong intuition that an OSDA-zeolite pair is best suited for AlPO chemistry. Rather than generating the synthesis components, they can input their intuition about the chemical environment directly into step 2 of the model and generate suitable gel composition, reaction conditions,

and precursors. This flexibility to input domain knowledge directly would not exist in an all-in-one model.

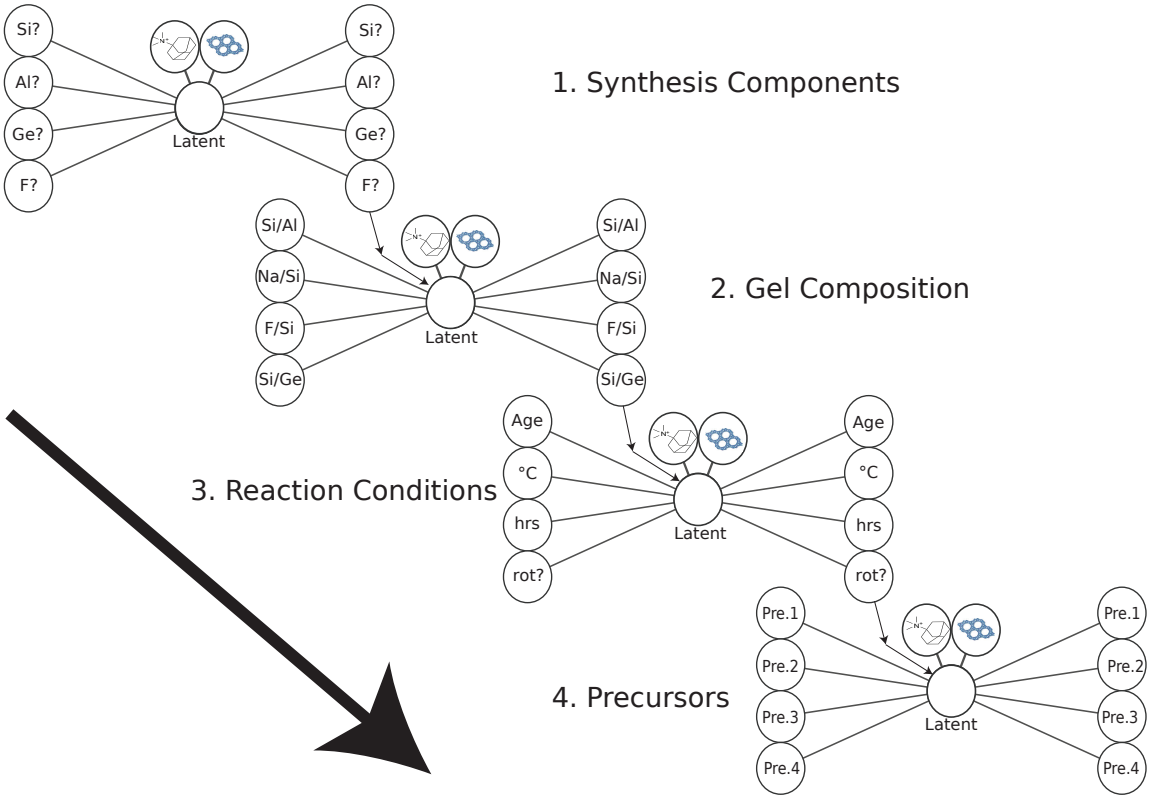


Figure 5-14. Schematic of the sequential CVAE models that comprise the inorganic conditions generative model.

Each of the CVAE models are deep neural networks utilizing a multi-term loss function consisting of a reconstruction and KL divergence term. For synthesis components and precursors, the reconstruction loss term is binary crossentropy while mean squared error is used for gel composition and reaction conditions. For synthesis components, gel composition, and reaction conditions, both the encoder and decoder are comprised of two fully connected, dense layers. The precursor model, due to its three-dimensional representation, utilizes three convolutional layers in the encoder and three recurrent, GRU layers for the decoder. All layers utilize the Relu activation function except for the ouptut layer of each model which is treated as a hyperparameter. Each model is trained independently using the default Adam optimizer with a batch size of 64 for 100 epochs. Training uses early stopping

criteria by monitoring loss on a held out validation set with patience value of 10. The models are implemented in Keras v2.2.4 with TensorflowGPU v1.15.0 as the backend and trained using two NVIDIA Titan Xp GPUs.

Synthesis components, gel composition, and reaction conditions are optimized over beta (relative weighting of reconstruction versus KL divergence), the latent space dimensionality, prior standard deviation, and final layer activation function. Precursors has additional hyperparameters connected to the convolutional and recurrent layer including the recurrent dimension, convolutional window, and convolution filters. Defining a metric to optimize over is difficult because the the hyperparameter beta is included in the loss function. Instead, performance metrics are devised based on overall reconstruction and probability. This performance metric is a weighted average of F1/RMSE values multiplied by the Wasserstein distance of the generated versus true distribution for each feature. This metric punishes the model for being far away from the test points as well as not recreating the literature distribution for each feature overall. Each model is optimizing using Bayesian optimization for 20 iterations with this metric as a function of the hyperparameters.

5.5.2 Model Performance

To evaluate model performance, a ten-fold cross validation is run with 10% of the OSDA systems withheld from each fold as a test set. These predictions are aggregated together and compared to the actual literature distributions for each part of the synthesis route. Each aspect of the synthesis route can be evaluated separately because of the sequential nature of the model.

Figure 5-15 evaluates the synthesis component part (Step 1) of the model. Because of the probabilistic nature of generating predictions, the best practice is typically to aggregate predictions to examine the most probable generated values. For each test point, 1,000 samples are generated. The most commonly generated synthesis components are then compared to the actually utilized synthesis components. Fig-

ure 5-15a shows a histogram of the percentage of overlap between the most commonly generated and true synthesis components. For example, if the true synthesis route uses Si, Al, Na, and H_2O and the four most frequently generated synthesis components are Si, Na, K, and H_2O, the overlap is 3 out of 4 (0.75). The majority of synthesis routes have at least 50% of the components correctly predicted with the plurality between 80-90%. Figure 5-15b shows the most common correctly identified synthesis components. All of the top five are very common especially "sda1" and "H_2O" with only OSDA-free synthesis and a small amount of solvothermal synthesis routes not having these variables.

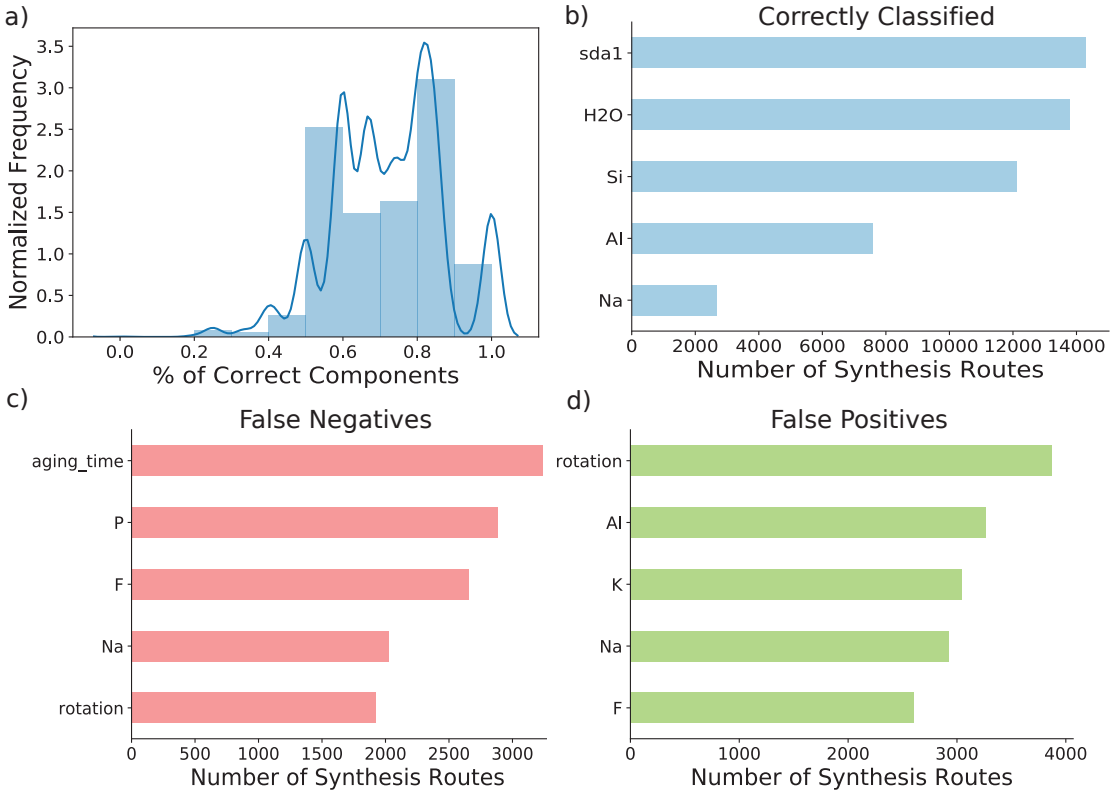


Figure 5-15. Overview of the synthesis component aspect of the generative model. a) Histogram of percentage of synthesis components correctly classified for each synthesis route. b) The most common correctly classified synthesis components. c) The most common synthesis components that are missed by the model d) The most common incorrectly generated components from the model.

Figure 5-15c-d show some of the model's common mistakes. Figure 5-15c shows the most common literature synthesis components that the model fails to generate.

Aging time is sparsely recorded in the literature with many authors not providing adequate data which may be limiting model performance. P, F, and Na are concerning since AlPO systems behave quite differently than conventional zeolites while F and Na often play a structure directing role in addition to mineralizing and charge balance.[63,65] Figure 5-15d shows the most common incorrectly generated synthesis components. Interestingly, there is some overlap between false negatives and false positives with rotation, Na, and F appearing in multiple lists. Rotation is another sparsely reported variable. The model assumes any synthesis route without a reported rotation occurred at static conditions which may lead to erroneous results. This may be an example of bias in the reporting of synthesis data affecting model results.

Figure 5-16 examines the performance of the model on gel composition (Step 2). Several important compositional ratios are examined including Si/Al, Al/P, K/Si, H_2O/Si, Si/Ge, Na/Si, F/Si, and OSDA/Si. Blue distributions are the extracted literature values and orange distributions are from the CVAE model. Ideally, the two distributions for each ratio should be approximately equal. We observe relatively good agreement between the literature and generated distributions for the majority of the ratios. The least agreement comes from Si/Ge which may be partially explained by the lack of understanding about the effect of Ge on zeolite structure.[90,91]

Figure 5-17 examines the model performance on the two most important reaction conditions (Step 3): crystallization time and crystallization temperature. The model appears to match the literature distributions reasonably well although the peaked nature of the literature distributions is not reflected and most apparent for crystallization temperature where the generated distribution smooths out the peaks observed in the literature values. This smoothing phenomenon is not necessarily bad as the peaked nature of the literature data is due to bias rather than underlying kinetic factors. Researchers are much more likely to choose 5 or 10°C increments rather than test on a continuous temperature scale. Instead the model learns a

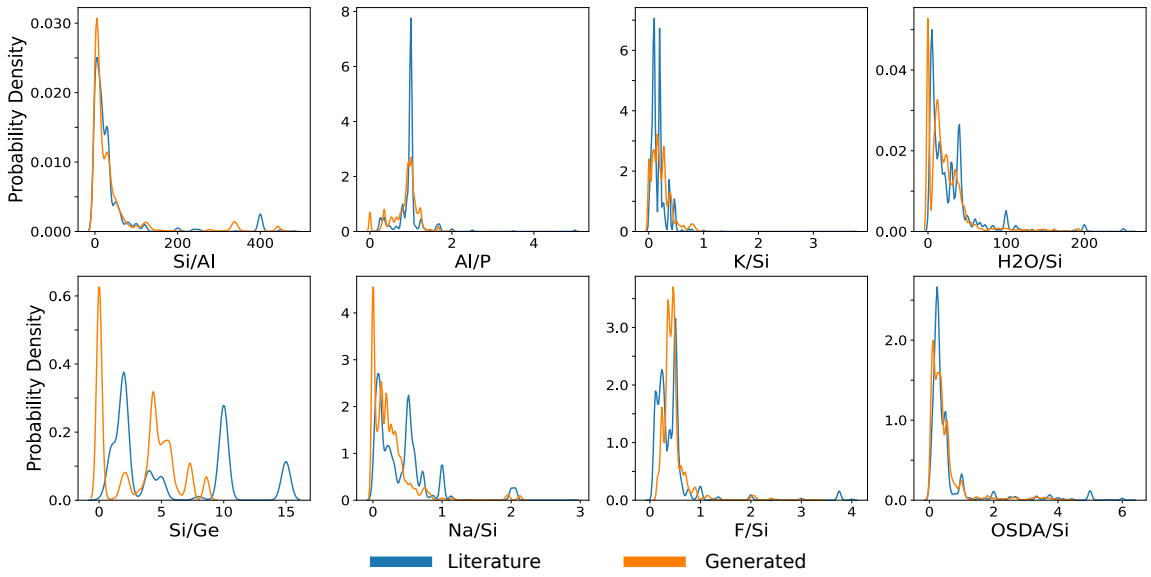


Figure 5-16. Aggregated performance of several important gel composition ratios.

similar type distribution smoothed over the crystallization temperature space, potentially resulting in predictions that are rooted more in underlying kinetic factors than human bias.

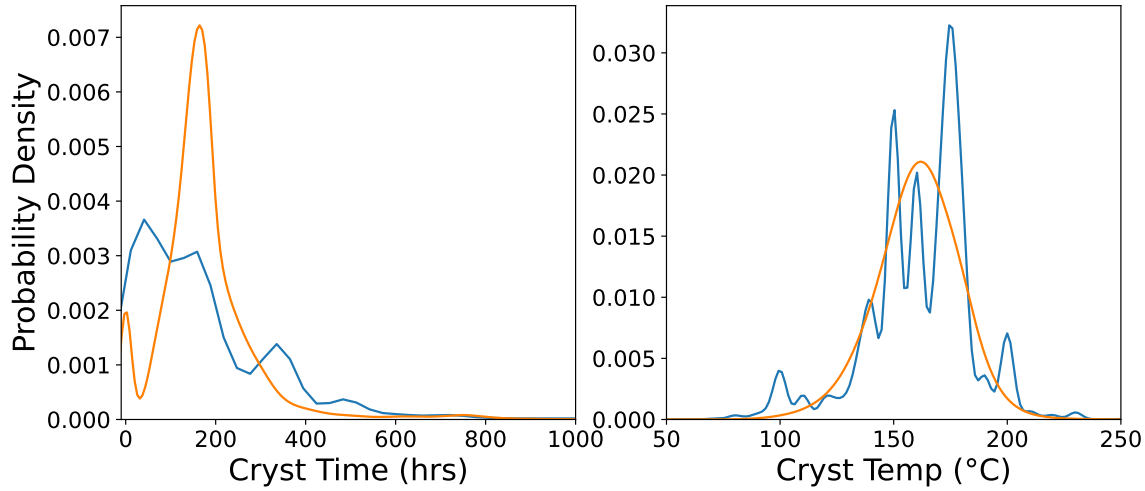


Figure 5-17. Aggregated performance of crystallization time and temperature.

Precursor generation is the final piece of the generative pipeline. Figure 5-18a and b show the most common precursors sets generated by the model and observed in the literature respectively (Step 4). A number of the most common precursor sets overlap including "HF, TEOS", "colloidal silica, sodium aluminate", "$Al(NO_3)_3$,

NaOH, fumed silica", "GeO_2, HF, TEOS", and "aluminum isopropoxide, phosphoric acid" indicating the model generates similar precursor distributions from the literature data. Figure 5-18c and d show the model performance on Si and Al precursors. Si and Al precursors are typically the most important choice[92] as most other synthesis components have only one or two popular precursor choices while Si and Al have 44 and 32 respectively. For each synthesis route, 100 different precursor sets are generated and aggregated based on the provided synthesis component. Starting with the most popular generated precursor, the sorted precursor list is traversed until the exact match from the literature is found. In this scheme, "first" indicates the most popular generated precursor was the literature precursor, "second" indicates the second most popular generated precursor matched the literature and so on for "third" through "fifth." "More/none" indicates the literature precursor is not in the top five most frequently generated precursors. Figure 5-18c and d indicate the model's generations mimic the literature used precursors with high accuracy for both Si and Al precursors. A plurality of generated synthesis routes match the literature precursor as the most frequently generated precursor for both Si and Al while 85% and 80% fall within the top five most commonly generated for Si and Al respectively. This demonstrates the accurate performance of the precursor piece of the model.

While matching the literature distribution for synthesis variables is important, it by itself does not demonstrate suitable model performance. The model must also be capable of generating tailored predictions for specific OSDA-zeolite systems, rather than just mimicking the underlying literature distribution of synthesis variables. To test this scenario, the model generates synthesis variables for each specific OSDA system found in the literature. The average root mean squared error (RMSE) and Wasserstein distance between each test point and the OSDA system's distribution is calculated. These values are compared to the RMSE and Wasserstein distance of randomly associated OSDA systems and generated points. Figure 5-19 shows this process for the highlighted synthesis variables. A model that successfully generates

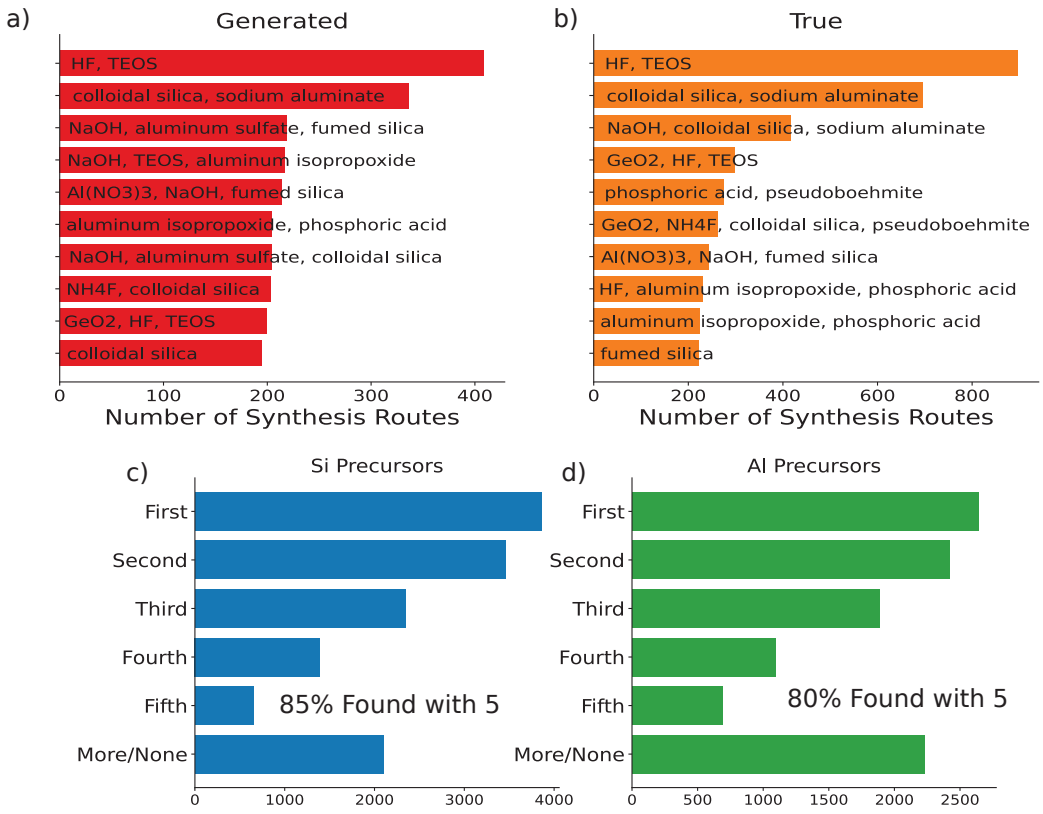


Figure 5-18. Examining the performance of the generative precursor model. a) Most commonly generated sets of precursors. b) The most common precursor sets found in the literature. c) Ranking the accuracy of the most common Si precursors predicted by the model. d) Ranking the accuracy of the most common Al precursors predicted by the model.

predictions tailored for each set of OSDA conditions should have a majority of RMSEs and Wasserstein distances less than (to the left) of the randomized value (thick black lines). All of the variables examined have a majority of systems with lower RMSE and Wasserstein distance values than the randomized sample indicating our model is generating good predictions suitable for each OSDA system. Si/Al and H_2O/Si are the worst performing with approximately 90% and 80% respectively of the OSDA systems outperforming the random baseline for both RMSE and Wasserstein distance. The rest of the variables shown have 100% of the OSDA systems with better RMSE and Wasserstein distances than the baseline. This indicates the model learns to generate synthesis routes, especially gel composition for specific OSDAs very well.

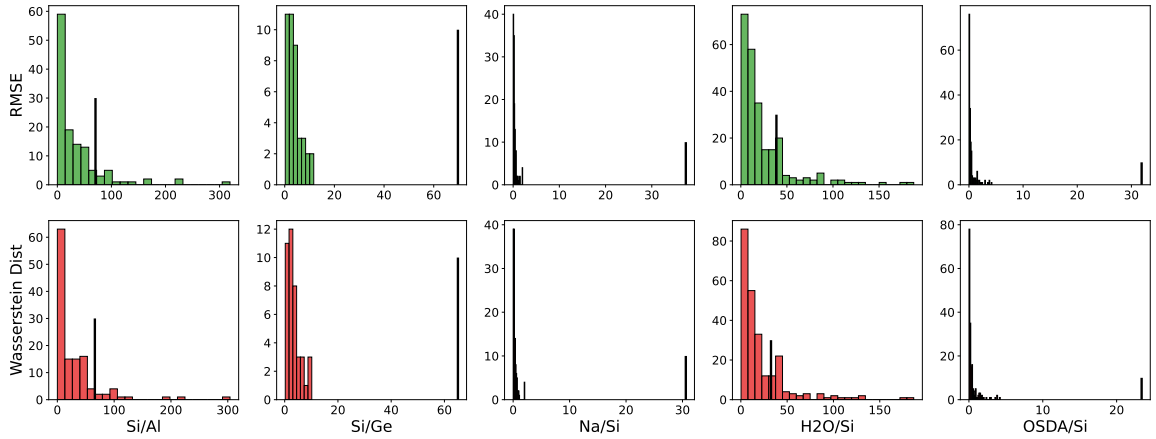


Figure 5-19. Examining the performance of important compositional ratios conditioned on specific OSDA systems.

Figure 5-20 shows a similar analysis on crystallization time and temperature. The model has good performance on crystallization temperature with approximately 85% of predictions better than the baseline. Crystallization time though performs very poorly with only about 35% of predictions beating the baseline. One potential reason for poor performance could be the inexact nature of crystallization time. Typically, zeolite synthesis takes on the order of days making the reporting relatively inaccurate. The literature also contains a wide range of synthesis times from only a couple of minutes for microwave and accelerated crystallization systems to several months. The model developed in section 5.4.1 could also be utilized to help get better predictions of necessary crystallization times.

5.5.3 Novel Zeolite Structures

To test the model's capability to assist in novel zeolite syntheis, the model's predictions on several newly realized zeolite systems are examined. These systems represent three different design scenarios. PTO, PTY, and ETV are new zeolite structures with new OSDAs, representing synthesizing a new, hypothetical zeolite with a newly designed OSDA.[93–95] SOV and YFI are new zeolite structures with previously used OSDA for a different structure, representing the synthesis optimization within an OSDA system to yield a new structure.[96,97] CHA is an old zeolite structure

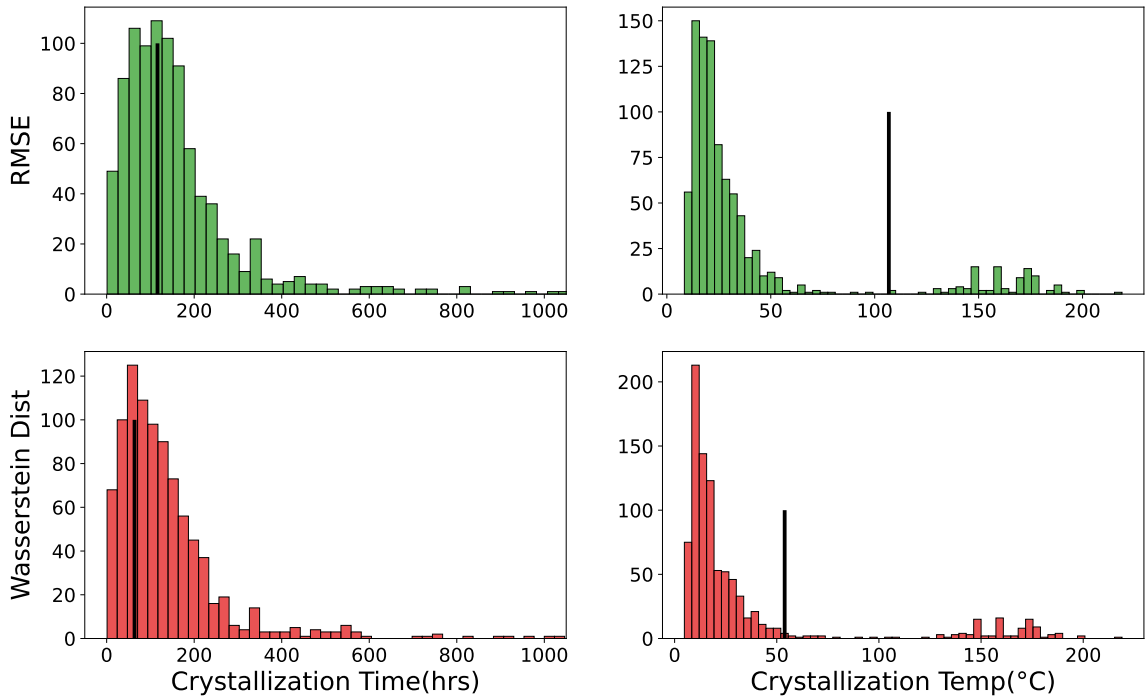


Figure 5-20. Examining the performance crystallization time and temperature conditioned on specific OSDA systems.

with new OSDA representing the optimization of the OSDA utilized in the synthesis of an existing zeolite structure.[3] We train a CVAE model, withholding the systems described above, then generate the synthesis routes and compare with the known synthesis routes. Figure 5-21a shows this process for ETV.

Since the model runs sequentially, there are multiple options regarding which sets of predictions to use in subsequent generation steps. To simplify the process, only three are considered termed "True", "Greedy", and "All". Starting from the generation of the synthesis components, "True" refers to using the known components used in the synthesis i.e. for ETV Si, Al, Na, OSDA, and H_2O. This skips the model prediction of these elements and represents the case where a researcher already knows the type of zeolite chemistry for the system. "Greedy" and "All" both involve using the model to make predictions about the synthesis components. "Greedy" only takes the most probable set of synthesis elements whereas "All" passes each set generated by the model. While this selection step could occur at each step in the model, each

step beyond the synthesis component prediction takes the "All" approach as to not have a large combinatorial space of selection options.

Figure 5-21c shows the results of generating synthesis components for each of the six selected systems. 1,000 synthesis element sets are generated for each system and aggregated to find the most frequently generated synthesis components. The % of Correct Elements is then taken as the overlap between the most frequently predicted components and the actual components used in the synthesis. All six selected systems have at least 60% of their correct synthesis components predicted by the model. PTO has 7/8 correctly predicted while YFI and CHA both have 5/6 predicted correctly. The worst, SOV at 60%, is synthesized as a germanosilicate potentially indicating again the germanium-zeotype interactions are difficult to predict.

To evaluate the gel composition and crystallization conditions, a range is defined around each true value and probability is calculated based on how many generated values fall within the range. The ranges are ±5 for Si/Al and Si/Ge, ±0.1 for Na/Si, K/Si, F/Si, and OSDA/Si, ±10 for H20/Si, ±24 hrs for crystallization time, and ±10°C for crystallization temperature. Figure 5-21b shows these results averaged across the six systems for each algorithmic split. "True" algorithm appears to have the best performance indicating the potential benefit of combining researchers' knowledge with the model. Combining the three algorithms gives the best performance. Time appears hard to correctly predict following the similar theme across multiple models.

Finally, precursors are generated for these systems and compared with the literature precursors. 1,000 sets of precursors per system are generated and aggregated. Figure 5-21d shows results for Si and Al precursor prediction. Each bar represents how frequently that precursor is predicted relative to others providing the same element, either Si or Al. Across the nine different predictions (SOV does not require Al), six of them have the correct precursor within the four most frequently generated precursors out of 44 for Si and 32 for Al. For the systems where precursors were not

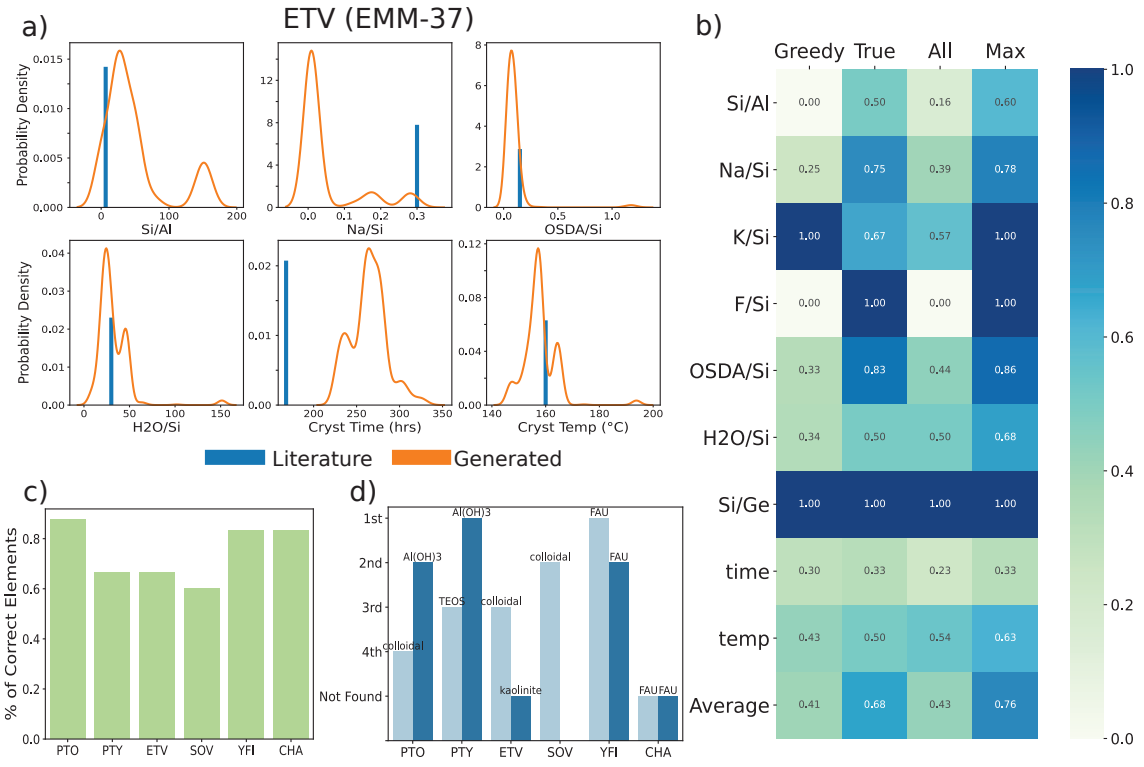


Figure 5-21. Examining the model performance on the six most recently synthetically confirmed zeolite structures. a) Predicted values vs the real synthesis value for ETV.[95] b) Percentage of generated synthesis routes that fall within a user defined range around the actually used synthesis route for important selected variables and the three algorithm choices. c) Percentage of correctly predicted synthesis components for each of the six zeolites. d) Ranking the most common silicon and aluminum precursors for each of the six zeolites.

found, ETV was made using kaolinite as the Al source, a relatively uncommon precursor. CHA with the new OSDA is made using FAU zeolite as the precursor rather than an amorphous silicon source. This is also rather uncommon as CHA can be made with other, more conventional precursors including silica sol and fumed silica and may indicate the model is slightly overconfident in the ability of the OSDA or other inorganic structure directing agents to form the CHA structure. Overall this demonstrates the model's ability to suggest precursors that are suitable for new zeolite systems.

5.6 Conclusion

In summary, this chapter examines the rest of the zeolite synthesis space outside of OSDA design termed "inorganic" aspects through a data driven lens. It models the probability of zeolite crystallization enabled through the reporting of negative data in the zeolite literature and uses Shap values to interpret the impact of different synthesis variables on the crystallization probability. It also combines high quality, quantitative crystallization data with qualitative data to model the crystallization behavior of zeolites. These two tasks help answer question two of the thesis by correlating synthesis variables with the outcomes and process of zeolite crystallization and providing tools for researchers to examine any known system in the zeolite literature. Finally, the chapter also provides a generative model capable of suggesting inorganic synthesis conditions for an OSDA-zeolite pair. This model helps answer thesis question three by learning from the literature data and providing a tool for researchers to aid in zeolite design.

Bibliography

[1] F. Daeyaert, F. Ye, and M. W. Deem, Proceedings of the National Academy of Sciences **116**, 3413 (2019).

[2] M. Moliner, P. Serna, Á. Cantín, G. Sastre, M. J. Díaz-Cabañas, and A. Corma, The Journal of Physical Chemistry C **112**, 19547 (2008).

[3] D. Schwalbe-Koda, S. Kwon, C. Paris, E. Bello-Jurado, Z. Jensen, E. Olivetti, T. Willhammar, A. Corma, Y. Román-Leshkov, M. Moliner, et al., Science p. eabh3350 (2021).

[4] R. F. Lobo, S. I. Zones, and M. E. Davis, Journal of inclusion phenomena and molecular recognition in chemistry **21**, 47 (1995).

[5] A. Simon-Masseron, J. Marques, J. M. Lopes, F. R. Ribeiro, I. Gener, and M. Guisnet, Applied Catalysis A: General **316**, 75 (2007).

[6] C. S. Blackwell, R. W. Broach, M. G. Gatter, J. S. Holmgren, D.-Y. Jan, G. J. Lewis, B. J. Mezza, T. M. Mezza, M. A. Miller, J. G. Moscoso, et al., Angewandte Chemie International Edition **42**, 1737 (2003).

[7] R. Mostowicz, F. Testa, F. Crea, R. Aiello, A. Fonseca, and J. B. Nagy, Zeolites **18**, 308 (1997).

[8] O. V. Shvets, N. Kasian, A. Zukal, J. Pinkas, and J. Čejka, Chemistry of Materials **22**, 3482 (2010).

[9] D. Ginter, A. Bell, and C. Radke, Zeolites **12**, 742 (1992).

[10] S. Alfaro, C. Rodriguez, M. Valenzuela, and P. Bosch, Materials Letters **61**, 4655 (2007).

[11] Y. Wu, X. Ren, and J. Wang, Microporous and mesoporous materials **116**, 386 (2008).

[12] H. Zhang, B. Xie, X. Meng, U. Müller, B. Yilmaz, M. Feyen, S. Maurer, H. Gies, T. Tatsumi, X. Bao, et al., Microporous and mesoporous materials **180**, 123 (2013).

[13] X. Zhang, D. Tang, M. Zhang, and R. Yang, Powder Technology **235**, 322 (2013).

[14] I. Güray, J. Warzywoda, N. Bac, and A. Sacco Jr, Microporous and mesoporous materials **31**, 241 (1999).

[15] R. Li, A. Chawla, N. Linares, J. G. Sutjianto, K. W. Chapman, J. G. Martínez, and J. D. Rimer, Industrial & Engineering Chemistry Research **57**, 8460 (2018).

[16] N. Martín, M. Moliner, and A. Corma, Chemical Communications **51**, 9965 (2015).

[17] M. Kumar, R. Li, and J. D. Rimer, Chemistry of Materials **28**, 1714 (2016).

[18] M. W. Deem, R. Pophale, P. A. Cheeseman, and D. J. Earl, The Journal of Physical Chemistry C **113**, 21353 (2009).

[19] R. Pophale, P. A. Cheeseman, and M. W. Deem, Physical Chemistry Chemical Physics **13**, 12407 (2011).

[20] B. M. Lowe, Zeolites **3**, 300 (1983).

[21] R. W. Thompson and A. Dyer, Zeolites **5**, 202 (1985).

[22] S. Bosnar, T. Antonić-Jelić, J. Bronić, I. Krznarić, and B. Subotić, Journal of crystal growth **267**, 270 (2004).

[23] F. Di Renzo, F. Remoué, P. Massiani, F. Fajula, F. Figueras, and C. T. Des, Zeolites **11**, 539 (1991).

[24] C. Den Ouden and R. Thompson, Industrial & engineering chemistry research **31**, 369 (1992).

[25] A. Gualtieri, Physics and Chemistry of Minerals **28**, 719 (2001).

[26] M. Avrami, The Journal of chemical physics **7**, 1103 (1939).

[27] M. Ermer, J. Mehler, B. Rosenberger, M. Fischer, P. S. Schulz, and M. Hartmann, ChemistryOpen **10**, 233 (2021).

[28] J. M. Garcia-Garfido, J. Enríquez, I. Chi-Durán, I. Jara, L. Vivas, F. J. Hernández, F. Herrera, and D. P. Singh, ACS omega **6**, 17289 (2021).

[29] S. D. Bagi, A. S. Myerson, and Y. Román-Leshkov, Crystal Growth & Design (2021).

[30] B. Seoane, J. M. Zamaro, C. Tellez, and J. Coronas, CrystEngComm **14**, 3103 (2012).

[31] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen, ACS central science **3**, 434 (2017).

[32] F. Pereira, CrystEngComm **22**, 2817 (2020).

[33] J. G. Wicker and R. I. Cooper, CrystEngComm **17**, 1927 (2015).

[34] A. Ghosh, L. Louis, K. K. Arora, B. C. Hancock, J. F. Krzyzaniak, P. Meenan, S. Nakhmanson, and G. P. Wood, CrystEngComm **21**, 1215 (2019).

[35] C. Brown, D. Maldonado, A. Vassileiou, B. Johnston, and A. Florence (2020).

[36] Z. Jensen, S. Kwon, D. Schwalbe-Koda, C. Paris, R. Gómez-Bombarelli, Y. Román-Leshkov, A. Corma, M. Moliner, and E. A. Olivetti, ACS central science **7**, 858 (2021).

[37] K. Muraoka, Y. Sada, D. Miyazaki, W. Chaikittisilp, and T. Okubo, Nature communications **10**, 1 (2019).

[38] Z. Jensen, E. Kim, S. Kwon, T. Z. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, and E. Olivetti, ACS central science **5**, 892 (2019).

[39] G. Raman, ChemistrySelect **6**, 10661 (2021).

[40] S. Ghanbari and B. Vaferi, Materials Science-Poland **35**, 486 (2017).

[41] A. Corma, M. Moliner, J. M. Serra, P. Serna, M. J. Díaz-Cabañas, and L. A. Baumes, Chemistry of materials **18**, 3287 (2006).

[42] J. Manuel Serra, L. Allen Baumes, M. Moliner, P. Serna, and A. Corma, Combinatorial chemistry & high throughput screening **10**, 13 (2007).

[43] S. Zones, Microporous and mesoporous materials **144**, 1 (2011).

[44] B. M. Lok, C. A. Messina, R. L. Patton, R. T. Gajek, T. R. Cannan, and E. M. Flanigen, Journal of the American Chemical Society **106**, 6092 (1984).

[45] A. Duan, T. Li, H. Niu, X. Yang, Z. Wang, Z. Zhao, G. Jiang, J. Liu, Y. Wei, and H. Pan, Catalysis today **245**, 163 (2015).

[46] L. Zhang and Y. Huang, Journal of Porous Materials **22**, 843 (2015).

[47] B. Yilmaz, J. Warzywoda, and A. Sacco Jr, Journal of crystal growth **271**, 325 (2004).

[48] E.-P. Ng, G. K. Lim, G.-L. Khoo, K.-H. Tan, B. S. Ooi, F. Adam, T. C. Ling, and K.-L. Wong, Materials Chemistry and Physics **155**, 30 (2015).

[49] X. Ren, J. Liu, Y. Li, J. Yu, and R. Xu, Journal of Porous Materials **20**, 975 (2013).

[50] L. A. Villaescusa, P. A. Barrett, and M. A. Camblor, Angewandte Chemie International Edition **38**, 1997 (1999).

[51] W. Loewenstein, American Mineralogist: Journal of Earth and Planetary Materials **39**, 92 (1954).

[52] A. Mlinarić, M. Horvat, and V. Šupak Smolčić, Biochemia medica **27**, 447 (2017).

[53] F. Song, S. Parekh, L. Hooper, Y. K. Loke, J. Ryder, A. J. Sutton, C. Hing, C. S. Kwok, C. Pang, and I. Harvey, Health technology assessment **14**, 1 (2010).

[54] X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist, et al., Nature **573**, 251 (2019).

[55] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, Nature **533**, 73 (2016).

[56] S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou, and B. Smit, Nature communications **10**, 1 (2019).

[57] J. Snoek, H. Larochelle, and R. P. Adams, Advances in neural information processing systems **25** (2012).

[58] J. Bergstra, D. Yamins, and D. Cox, in *International conference on machine learning* (PMLR, 2013), pp. 115–123.

[59] L. S. Shapley, *17. A value for n-person games* (Princeton University Press, 2016).

[60] S. M. Lundberg and S.-I. Lee, in *Proceedings of the 31st international conference on neural information processing systems* (2017), pp. 4768–4777.

[61] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, Nature machine intelligence **2**, 56 (2020).

[62] Z. Liu, K. Okabe, C. Anand, Y. Yonezawa, J. Zhu, H. Yamada, A. Endo, Y. Yanaba, T. Yoshikawa, K. Ohara, et al., Proceedings of the National Academy of Sciences **113**, 14267 (2016).

[63] H. Kessler, J. Patarin, and C. Schott-Darie, ChemInform **26**, no (1995).

[64] J. Li, A. Corma, and J. Yu, Chemical Society Reviews **44**, 7112 (2015).

[65] C. Gittleman, A. Bell, and C. Radke, Catalysis letters **38**, 1 (1996).

[66] S. Z. Patuwan and S. E. Arshad, Materials **14**, 2890 (2021).

[67] L. V. Dang, S. T. Le, R. F. Lobo, and T. D. Pham, Journal of Porous Materials **27**, 1481 (2020).

[68] S. Goel, S. I. Zones, and E. Iglesia, Chemistry of Materials **27**, 2056 (2015).

[69] L. Tang, K.-G. Haw, P. He, Q. Fang, S. Qiu, and V. Valtchev, Inorganic Chemistry Frontiers **6**, 3097 (2019).

[70] M. Miyamoto, T. Nakatani, Y. Fujioka, and K. Yogo, Microporous and Mesoporous Materials **206**, 67 (2015).

[71] D. E. Akporiaye, I. M. Dahl, H. B. Mostad, and R. Wendelbo, The Journal of Physical Chemistry **100**, 4148 (1996).

[72] R. Simancas, A. Chokkalingam, S. P. Elangovan, Z. Liu, T. Sano, K. Iyoki, T. Wakihara, and T. Okubo, Chemical Science (2021).

[73] R. Batra, G. Pilania, B. P. Uberuaga, and R. Ramprasad, ACS applied materials & interfaces **11**, 24906 (2019).

[74] S. Venkatram, R. Batra, L. Chen, C. Kim, M. Shelton, and R. Ramprasad, The Journal of Physical Chemistry B **124**, 6046 (2020).

[75] C. Chen, Y. Zuo, W. Ye, X. Li, and S. P. Ong, Nature Computational Science **1**, 46 (2021).

[76] T. Taniguchi, Y. Nakasaka, K. Yoneta, T. Tago, and T. Masuda, Catalysis Letters **146**, 666 (2016).

[77] X. Yang, Y. Liu, X. Li, J. Ren, L. Zhou, T. Lu, and Y. Su, ACS Sustainable Chemistry & Engineering **6**, 8256 (2018).

[78] M. O. Cichocka, Y. Lorgouilloux, S. Smeets, J. Su, W. Wan, P. Caullet, N. Bats, L. B. McCusker, J.-L. Paillaud, and X. Zou, Crystal Growth & Design **18**, 2441 (2018).

[79] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al., Nature Methods **17**, 261 (2020).

[80] C. Voglis and I. Lagaris, in *WSEAS International Conference on Applied Mathematics* (2004), vol. 7.

[81] J. Nocedal and S. Wright, New York (2006).

[82] M. K. Choudhary, R. Jain, and J. D. Rimer, Proceedings of the National Academy of Sciences **117**, 28632 (2020).

[83] C. S. Cundy and P. A. Cox, Microporous and mesoporous materials **82**, 1 (2005).

[84] J. Ling, M. Hutchinson, E. Antono, S. Paradiso, and B. Meredig, Integrating Materials and Manufacturing Innovation **6**, 207 (2017).

[85] S. Wager, T. Hastie, and B. Efron, The Journal of Machine Learning Research **15**, 1625 (2014).

[86] F. J. Massey Jr, Journal of the American statistical Association **46**, 68 (1951).

[87] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, PeerJ Computer Science **2**, e55 (2016).

[88] A. Á. B. Maia, R. N. Dias, R. S. Angélica, and R. F. Neves, Journal of Materials Research and Technology **8**, 2924 (2019).

[89] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, and K. F. Jensen, ACS central science **4**, 1465 (2018).

[90] A. Corma, M. Díaz-Cabañas, J. Jiang, M. Afeworki, D. Dorset, S. Soled, and K. Strohmaier, Proceedings of the National Academy of Sciences **107**, 13997 (2010).

[91] A. Corma, F. Rey, S. Valencia, J. L. Jordá, and J. Rius, Nature materials **2**, 493 (2003).

[92] G. Kiihl, Verified Synthesis of Zeolitic Materials p. 19 (2001).

[93] D. Jo, Y. Zhang, J. H. Lee, A. Mayoral, J. Shin, N. Y. Kang, Y.-K. Park, and S. B. Hong, Angewandte Chemie **133**, 6001 (2021).

[94] D. Jo and S. B. Hong, Angewandte Chemie **131**, 13983 (2019).

[95] E. Kapaca, A. Burton, E. Terefenko, H. Vroman, S. C. Weston, M. Kochersperger, M. Afeworki, C. Paur, L. Koziol, P. Ravikovitch, et al., Inorganic chemistry **58**, 12854 (2019).

[96] Y. Luo, S. Smeets, Z. Wang, J. Sun, and W. Yang, Chem. A Eur. J. **25**, 2184 (2019).

[97] N. Nakazawa, T. Ikeda, N. Hiyoshi, Y. Yoshida, Q. Han, S. Inagaki, and Y. Kubota, Journal of the American Chemical Society **139**, 7989 (2017).

Chapter 6

Outlook and Conclusions

6.1 Future Outlook

This thesis represents a small piece of a rapidly expanding sub-domain of materials science aimed at accelerating synthesis. At the completion of this thesis, there are continuing efforts to expand upon the themes outlined including incorporating data science, simulations, and experiments within a circular pipeline. There are also a number of recent advances and frontiers in the application of NLP to the materials domain that provide potentially fruitful research endeavors. Finally, it may be productive to reflect on the nature of publication itself and discuss paths to better sharing of data between researchers.

6.1.1 Accelerated Zeolite Synthesis Planning

After decades of trial and error, the past decade has seen significant progress in accelerating zeolite design and development. Many tools have been developed including hypothetical zeolite databases,[1] models and platforms for OSDA selection,[2,3] new synthesis methods,[4] and high throughput flow synthesis platforms.[5,6] One major remaining question is how best to combine all of these tools in optimal ways to accomplish the ultimate goal, at will zeolite synthesis of any zeolite

structure.

An increasing body of evidence suggests that ML and data science can provide the missing link in zeolite synthesis.[7] ML is a potential solution to determining feasibility and expanding hypothetical zeolite structures, understanding interactions between OSDA and zeolites, and increasing the accuracy and speed of zeolite characterization. Succinctly, some of the main challenges are representation learning for the OSDA and zeolite, mimicking binding energy calculations for non-Si-only frameworks, and prediction of the zeolite formed given synthesis conditions.

One aspect where this thesis work could be expanded is in the prediction of specific zeolite structures given a synthesis route. This work predicts whether a zeolite is formed but stops short of predicting what phase that zeolite will be. This is a very challenging problem mainly because of the complexity of zeolite representation. Zeolites are often represented as combinations of structural properties[8] or graph-based representations[9] making regression-based models to predict structure necessarily complex and multi-output. Predicting a specific phase could also be treated as a classification problem although this presents the problems of a very large number of potential classes with over 250 known zeolites and the inability to generalize to novel, hypothetical zeolites. Further efforts into this task will need to content with this complex zeolite representation without sacrificing the ability to generalize. Recent efforts to develop metrics of templating ability for OSDA-zeolite pairs[3] will most likely be crucial in compressing the potential zeolite output space down to only the possible subset of zeolite structures allowing the ML to focus on predicting on the smaller synthesis space.

Another promising route to understanding zeolite kinetics is high-throughput synthesis. Using active learning that builds on the concept of Bayesian inference,[10] the synthesis space can be efficiently explored, mapping the zeolite phase space and also providing important kinetic insight. One problem however is the output space of zeolite synthesis is complex. It is hard to define a single quantity typically needed

for an optimization process. Research is needed into the best ways to represent the zeolite structural space within the context of optimization algorithms.

Finally, the extracted datasets provide many opportunities to look at interesting problems within the zeolite domain including:

- The general effect of using crystalline zeolite precursors instead of amorphous silica source on structure and crystallization. This behavior is often described as a recrystallization interzeolite transition and is poorly explained by theories that describe other types of interzeolite transitions.[9]
- Ge-free synthesis for structures that only exist as Ge-zeotypes expanding on previous work[11] to predict conditions necessary to synthesis pure Si versions of structures like IWW.
- Understanding the important relationships in industrial relevant zeotypes such as titanosilicates.[12]
- Expanding the datasets to include catalytic properties of the zeolite to start linking synthesis directly with the desired properties beyond structure.
- Expand the kinetic analysis in section 5.4 to solid-state transformations within zeolites to predict the phase progression through time.

6.1.2 Use of NLP in Inorganic Materials Synthesis

NLP continues to make impact when applied in the materials domain. Recent innovations and trends including transformer models with attention mechanisms[13,14] which should improve the text vectorization, section classification, and named entity tasks important for information extraction. However, several challenges exist within the information extraction space that NLP and ML research could impact.

The current, most important problem in materials information extraction is extracting and linking information from different parts of an article. This problem occurs

across varying length scales within an article. On a small length scale, current NER and information associating algorithms only span the context of a single sentence. This sentence-only-context is problematic for tasks such as identifying target materials where often sentence level context is inadequate or associating conditions that are listed in a succeeding sentence to its operation. On a larger scale, difficulty in associating various pieces of data bottlenecks automatic extraction. Ideally, algorithms would identify individual samples in a paper and associate all relevant data including composition, synthesis conditions, material structure, microstructure, and material properties. However, all these pieces of information are located in different locations within an article making the association very difficult. This problem is similar to the "Co-reference" task in the NLP domain and is notoriously difficult.[15,16] However, it will be necessary to solve this problem to automatically extract comprehensive materials datasets with all relevant quantities.

Another issue centers around information extraction from sources even more unstructured than text, namely Supplementary Information (SI) sections. The vast majority of publishers structure SI sections of articles as PDFs with essentially no structure. Authors are free to structure SI sections however they would like, creating enormous challenges in extracting that information. Currently, most of the data found within SI sections needs to be extracted manually which can create huge bottlenecks in the information extraction process especially as synthesis sections are often relegated to the SI. Advances in extraction from PDFs as well as publisher reforms around the structure and file formats of SI sections could greatly speed up the extraction process by removing a troublesome road block.

A final important frontier in ML applied to materials literature highlighted in this thesis is image processing. Many articles present results in figures, often in the form of graphs, schematics, or micrographs. Application of image processing techniques on this source of data in a similar fashion as NLP to scientific text could great enhance the scope and size of information extraction. Many researchers understand the scope and opportunity this problem presents and progress has been made in

extraction from specific types of images.[17,18] However, many questions remain regarding generalizing extraction to vastly different types of images found across the entire materials literature.

6.1.3 Rethinking Publishing and Data Communication

Upon reflection, this thesis exposes a flaw within the publishing and sharing of scientific data and results. In building the datasets, data was first collected by a researcher performing experiments who then wrote text describing that data. Then extraction tools are used to convert this text back into data similar to the original form. At both steps, converting the data to text and converting the text back to data, information is lost and corrupted. This framing highlights the inefficiency of the current process.

Most industries as well as academia have realized the importance of aggregating, curating, and storing data. A similar realization in the publishing industry could drastically improve data driven research and the scientific process as a whole. While writing articles in prose is undoubtedly very valuable, the publication process should also require publication of the data behind the written article. Different scientific domains could have different templates of what data is required. This data publication would reduce the problem of data sparsity[19] and inconsistency of specific reported quantities.[20] It could also reduce and eventually replace the need for complicated extraction algorithms greatly improving availability and accessibility for aggregating data and performing research. Direct publication of data would also increase confidence in the peer review process and help with reproducibility of scientific studies.

6.2 Conclusions

In this thesis, the synthesis of zeolite materials is studied through data science techniques including NLP, materials informatics, generative modeling, and Bayesian in-

ference. Five datasets are extracted over the length of this thesis each corresponding to specific aspects of zeolite synthesis. To the author's knowledge, these datasets are the largest and most comprehensive related to zeolite synthesis and are made publicly available to the research community for use to guide experimental studies, validate theories/simulations, and supply data driven research. These datasets are used to study zeolite synthesis including correlating synthesis with structural properties, examining the relationships between OSDA templates and zeolites, and studying the effect of synthesis parameters on zeolite crystallization. Finally, models are also created and made public to predict new potential OSDA-zeolite pairs, generate hydrothermal synthesis conditions for an OSDA-zeolite pair, classify of successful zeolite synthesis, and estimate the crystallization curve for a given zeolite and synthesis environment. Hopefully the resources and insights developed in this thesis will help facilitate zeolite design and improve data driven synthesis planning more broadly.

Revisiting the questions posed at the beginning of this this thesis:

1. How can zeolite synthesis data be **automatically extracted** on a **large scale**?

    - Adaptation of NLP and text mining tools specifically to target relevant aspects of zeolite literature.

    - Automatic filtering based on domain knowledge and researcher-computer interaction to optimize extraction accuracy and efficiency.

    - Results in the largest collection of known zeolite synthesis data with approximately an 80% improvement in extraction efficiency.

2. How can coupling of **data-driven, first principles, and experimental** approaches accelerate understanding of **structure and processing relationships** in zeolite materials?

    - WHIM representation of OSDAs reveals clusters related to specific of ze-

olite structures.

- Negative data, ML modeling, and Shap values for interpretability provide insight into the probability of zeolite crystallization for any synthesis route.
- Multi-fidelity data and Bayesian inference model the crystallization behavior of zeolites.

3. In what ways can this data and discovered relationships be used to **engineer improved zeolite materials**?
    - Generative model predicts suitable OSDA candidates given a zeolite.
    - Generative model predicts suitable reaction chemistry, hydrothermal conditions, and precursors given an OSDA-zeolite pair.

Bibliography

[1] R. Pophale, P. A. Cheeseman, and M. W. Deem, Physical Chemistry Chemical Physics **13**, 12407 (2011).

[2] F. Daeyaert, F. Ye, and M. W. Deem, Proceedings of the National Academy of Sciences **116**, 3413 (2019).

[3] D. Schwalbe-Koda, S. Kwon, C. Paris, E. Bello-Jurado, Z. Jensen, E. Olivetti, T. Willhammar, A. Corma, Y. Román-Leshkov, M. Moliner, et al., Science p. eabh3350 (2021).

[4] P. Eliášová, M. Opanasenko, P. S. Wheatley, M. Shamzhy, M. Mazur, P. Nachtigall, W. J. Roth, R. E. Morris, and J. Čejka, Chemical Society Reviews **44**, 7177 (2015).

[5] A. Corma, M. Moliner, J. M. Serra, P. Serna, M. J. Díaz-Cabañas, and L. A. Baumes, Chemistry of materials **18**, 3287 (2006).

[6] D. E. Akporiaye, I. M. Dahl, A. Karlsson, and R. Wendelbo, Angewandte Chemie International Edition **37**, 609 (1998).

[7] M. Moliner, Y. Román-Leshkov, and A. Corma, Accounts of chemical research **52**, 2971 (2019).

[8] Z. Jensen, S. Kwon, D. Schwalbe-Koda, C. Paris, R. Gómez-Bombarelli, Y. Román-Leshkov, A. Corma, M. Moliner, and E. A. Olivetti, ACS central science **7**, 858 (2021).

[9] D. Schwalbe-Koda, Z. Jensen, E. Olivetti, and R. Gómez-Bombarelli, Nature materials **18**, 1177 (2019).

[10] D. Packwood, *Bayesian Optimization for Materials Science* (Springer, 2017).

[11] Z. Jensen, E. Kim, S. Kwon, T. Z. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, and E. Olivetti, ACS central science **5**, 892 (2019).

[12] H. Xu and P. Wu, Chinese Journal of Chemistry **35**, 836 (2017).

[13] D. Hu, in *Proceedings of SAI Intelligent Systems Conference* (Springer, 2019), pp. 432–448.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, in *Advances in neural information processing systems* (2017), pp. 5998–6008.

[15] J. H. Clark and J. P. González-Brenes, Language and Statistics II Literature Review **14** (2008).

[16] K. Clark and C. D. Manning, in *Empirical Methods on Natural Language Processing* (2016), URL `https://nlp.stanford.edu/pubs/clark2016deep.pdf`.

[17] E. Schwenker, W. Jiang, T. Spreadbury, N. Ferrier, O. Cossairt, and M. K. Chan, arXiv preprint arXiv:2103.10631 (2021).

[18] T. N. Nguyen, Y. Guo, S. Qin, K. S. Frew, R. Xu, and J. C. Agar, npj Computational Materials **7**, 1 (2021).

[19] E. Kim, K. Huang, S. Jegelka, and E. Olivetti, npj Computational Materials **3**, 1 (2017).

[20] H. El-Bousiydy, T. Lombardo, E. Primo, M. Duquesnoy, M. Morcrette, P. Johansson, P. Simon, A. Grimaud, and A. Franco, Batteries & Supercaps (2021).

Chapter 7

Appendix

This appendix briefly summarizes some practical tips and lessons learned over the course of the thesis.

7.1 Practical Tips for Information Extraction

NLP improvements do not necessarily map well to materials information extraction: Innovations within the NLP community are vital for improving information extraction from materials text. However, not every innovation and performance increase in the NLP literature maps neatly to materials text data. The NLP domain typically develops models on standard datasets that are both large and well-formatted whereas materials data is typically neither. NLP models will typically perform worse when applied to materials text. Extraction speed is another consideration for materials that is rarely considered when developing NLP models. Caution should be used when deciding to implement state-of-the-art NLP models to determine the performance increase on materials text and whether a potential decrease in extraction efficiency is justified by that performance gain.

The more "sample specific" the data is, the more challenging the extraction: As described in section 3.3, computers are very good at extracting specific values

and much worse at associating those values correctly. By considering this difference when forming hypotheses, research plans can be developed that minimize the amount of associating or manual data cleaning. Domains where only one material is synthesized per paper is ideal from an extraction context since all quantities can be grouped together. Obviously this comes at the expense of data quantity. Often insights can be gained from aggregating data such as examining all extracted temperatures without consideration of the specific sample. These types of extractions are much easier but can limit insight making it domain and context dependent on how "sample" focused the data should be.

Material synonyms present a challenge: A major challenge faced in this thesis is standardizing different names for materials and chemicals so that all different text versions map to the same object. This is especially pertinent for organic molecules which have multiple naming conventions. Finding a common representation is essential. For organic molecules this is the SMILES string although this is also not unique for each molecule. Canonical rules exist for SMILES, but is necessary to use the same software package to convert to canonical form to ensure consistency. Some molecules are going to evade software packages which typically requires manual grouping.

7.2 Practical Tips for ML in the Materials Domain

Start simple and build up in complexity: This phrase should be a guiding principle for ML in general but especially in materials science. Materials data, especially small datasets, are often not well-suited for complex state-of-the-art ML models. Often linear models with domain engineered features will out perform complex models with far easier implementation and shorter training times. Only when simple models lack predictive ability should more complex models be used.

Test sets need to reflect the model intention: Often in ML, test sets are drawn randomly from the data, the underlying assumption being the train and test sets

have identical distributions. This is rarely the case in materials since ML models are typically used to discover new materials or phenomena. This means randomly held out tests sets typically over perform the "real" model performance on its given task. Test sets usually need to mimic some form of extrapolation behavior such as holding out an entire class of materials or holding out the top X% of a certain property. Evaluating on these types of test sets will give much more accurate descriptions of model performance.