

Smaller than the Unit Cell: Smallest Repeating Units of Zeolite Frameworks

Akhilesh Gandhi, M. M. Faruque Hasan*

Artie McFerrin Department of Chemical Engineering, Texas A&M University
College Station, TX 77843-3122, USA.

Abstract

Representations of crystal frameworks are important for structure description, design, and discovery of new frameworks. Zeolite frameworks with cartesian coordinates-based representation of atomic positions in three dimensions are convenient for human perception but are computationally inefficient due to the large information required to describe the framework. We exploit the Hamiltonian representation of crystal frameworks that incorporates a graph-theoretic approach for efficiently capturing relative atomic positions and connectivity. We introduce a new building block, namely the smallest repeating unit (SRU), that utilizes fewer T-nodes to describe a zeolite in comparison to the traditional unit cell. The Hamiltonian graph-based representation is both invertible and scalable in the sense that it only uses topologically distinctive T-nodes, thereby significantly reducing the description space. We also develop algorithmic and optimization-based approaches to identify SRUs of large crystallographic frameworks. SRU identification is formulated as a special instance of traveling salesman problem (TSP). Overall, we describe the SRUs of 158 existing zeolites with up to 4 times reduction in T-nodes and over 10,000 hypothetical zeolite frameworks. For example, Chabazite framework can be represented using only 12 T-nodes in the Hamiltonian graph representation as opposed to 36 T-nodes in the traditional unit cell. One additional benefit of SRU representation is that it provides a systematic and efficient approach to generate and analyze all plausible cation (e.g., Aluminum) substituted frameworks for different Si/Al ratios. We envision this representation to also open new avenues for the design and discovery of novel nanoporous materials.

Keywords: Framework representation, Hamiltonian graphs, unit cell, zeolites

1 Introduction

Zeolites are microporous crystalline materials with a wide range of industrial, medical, environmental, and microelectronics applications.^{1,2} They are used as catalysts and adsorbents³ due to their cage-like porous structures and their ability to trap cations for ion exchange. The design and discovery of new zeolites have been an active field of interest since the 1940s.^{4,5} Many computational methods, such as molecular modeling, charge calculations, force fields, and GCMC simulations, have been developed and used for zeolite characterization and screening.⁶

The crystal structure of pure-silica frameworks contains several tetrahedral nodes (T-nodes) of Si with O as connecting atoms. Changing the chemical composition of these frameworks by replacing

*Correspondence concerning this article should be addressed to M.M. Faruque Hasan at hasan@tamu.edu, Tel.: 979-862-1449.

some of the Si with Al or other tetrahedral atoms induces a noticeable change in the properties of zeolites. This affects the affinity of the framework to other cations that can be exploited to customize ion-exchange properties and different adsorbent surfaces. Given the three-dimensional geometry of these tetrahedral atoms, rings, cages, channels, and pores are generated that lead to diverse frameworks. To date, there exists over 241 distinct pure-silica frameworks that have been listed in the International Zeolite Association (IZA-SC) database.⁷ It is important to note here that, of these 241, only 229 are frameworks are built with all tetrahedral nodes.

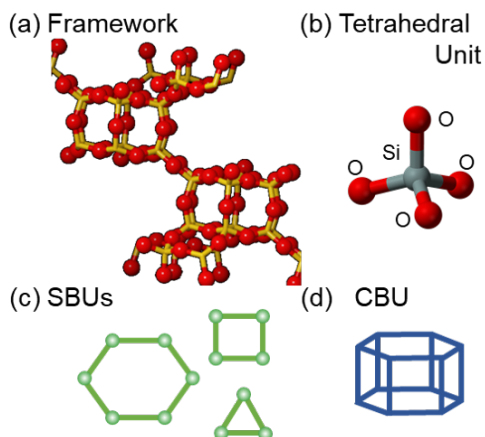


Figure 1: Representation of large zeolite frameworks (a) Example framework for Chabazite, (b) tetrahedral unit of Si and O as the basic unit of the framework, (c) examples of secondary building units (SBU) with only Si atoms, (d) composite building unit (CBU) observed in Chabazite with only Si atoms on vertices.

To describe complex frameworks such as zeolites, it is essential to develop a representation that can efficiently capture the details of the lattice. Crystal lattices are often represented by the unit cell that is defined by a parallelepiped shape within which the spatial positions of atoms are specified. While describing any new crystal framework, it is sufficient to define the parameters defining the unit cell and the coordinates of the atoms within the unit cell. The unit cell is defined as the smallest unit having the full symmetry of the crystal structure that can be repeated.⁸ The constraint is that it has to be a parallelepiped, defined by the cell edges and the angles between the edges $(a, b, c, \alpha, \beta, \gamma)$, as well as the coordinates of the atoms. Another notable representation is the Smooth Overlap of Atomic Position (SOAP)^{9,10} Representations that utilize the advantages of machine learning approaches are an active field of interest.¹¹

Unit cells have been used to generate many potential zeolite frameworks. Using the parameters that define a unit cell, Monte Carlo methods have led to several feasible structures.¹² The result of these simulations has been of great use in identifying hypothetical frameworks and over 5 million hypothetical zeolite frameworks have been defined to date.¹³ The advent of computational power and new algorithms have led to many potential zeolite frameworks.¹⁴ These approaches employ secondary building units (SBU), composite building units (CBU), and the difference in ring sizes (Figure 1).¹⁵ These building blocks can be used to build new structures and verify the feasibility

of hypothetical structures. They can also be used for inverse design of crystalline structures.^{16,17} Identifying these key units within a framework has been used in other representations including the Simplified Molecular-Input Line-Entry System (SMILES).¹⁸ Determination of largest free sphere in the lattice is done using Delaunay tessellation and Voronoi networks.¹⁹ Using tessellation techniques, tiling notations on the faces of the structures have been developed using TOPOS program package and further applied to address the zeolite conundrum.^{20,21} The smooth overlap of atomic positions (SOAP) representation has been applied to several hypothetical zeolite frameworks to accurately capture structure-property relations.¹⁰ Characterization of portals, channels, and cages within zeolites are done using Markov chain models.²² Graph theory has been applied to study network properties of zeolite frameworks. For example, some zeolites are observed to have Hamiltonian cycles within their crystallographically independent nodes.^{23–26} Natural tiling of periodic networks in zeolites²⁷ and machine learning approaches²⁸ are also used to obtain insights into the zeolite frameworks.^{29,30}

Development of different descriptors for zeolite frameworks has been fueled by their utility towards predicting properties of the frameworks. The SOAP representation originally proposed in 2013 has been developed further and has been used in machine learning algorithms towards predicting properties such as volume and energy of frameworks.³¹ The machine learning models can predict certain properties of the framework with the input of the distances and the angles of the neighboring T-atoms. While this approach is definitely useful, the SRU representation is different from the SOAP representation in the form that it serves a different purpose of addressing the design of zeolite frameworks. The SRU representation retains the distances and angles within the descriptor while also obtaining certain insights by representing the framework using graph theory. That being said, it is also possible to apply the SOAP representation to the SRU as an advantage since it is a reduced representation which can further be used for property prediction. Further in the results, we explain how this representation is better at capturing the Al-substitutions and generating new compositions following Loewenstein’s rule. Thus, currently the SRUs are posed from a design perspective in contrast to the application of the SOAP representation using machine learning for property prediction. The natural building units (NBUs) or natural tilings also aim to address the structure of the framework using packing units and individual packing units are then used to identify the feasibility of the framework.²¹

The fundamental principle behind most of these approaches relies on determining the number of T-atoms (Si) in the unit cell and their coordinates, the structural orientation of the atoms, pore sizes, and cell dimensions. The unit cell captures the details of the lattice by defining the symmetry for the T-atoms across 3-d planes. However, for a given zeolite framework, the number of T-atoms that are needed to obtain a unit cell can be much greater than required to capture the repetition. This is because more T-atoms are required to satisfy the constraint of the unit cell to be parallelepiped.

The systematic characterization of geometric configurations, chemical compositions, and structural properties have generated new approaches towards the design and screening of new zeolites and Metal-Organic frameworks (MOFs).^{17,32,33} Zeolites have been classified based on the size and orientation of the rings present within their structures. Topological descriptors have been developed for zeolites that aim to take advantage of machine learning,^{34–37} optimization,³⁸ and algorithmic approaches.¹⁶ These approaches have an inherent dependence on enumeration that leads to a large combinatorial,

101 symmetric, and degenerate design space.

102 To this extent, we propose a new representation of zeolite frameworks where we capture the
 103 symmetry of the cell and details such as the ring sizes. We call this Smallest Repeating Unit (SRU).
 104 An SRU is unique for each zeolite and thus captures the essence of the framework and connectivity
 105 within it. It is different from the unit cell since it has fewer tetrahedral nodes. By analyzing the SRU
 106 structures and by studying their properties, we claim that the SRUs brings a unique perspective to the
 107 design and characterization of zeolites. SRUs drastically reduce the details that are needed to define
 108 zeolite frameworks (see Figure 2). The lattices can now be defined using fewer T-nodes, along with a
 109 connectivity matrix.

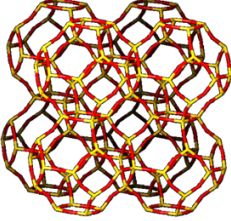

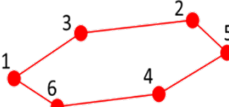
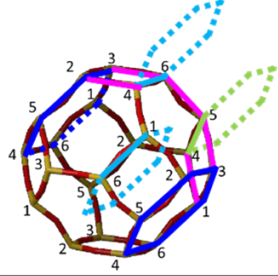
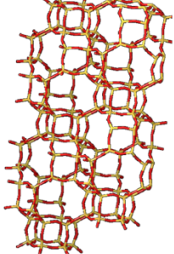
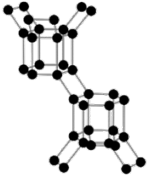
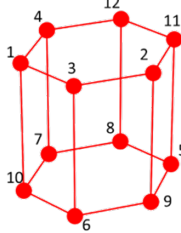
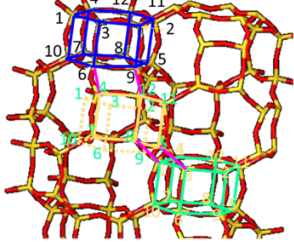
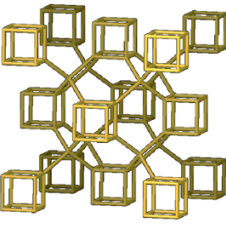
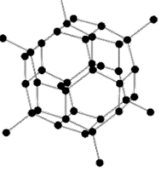
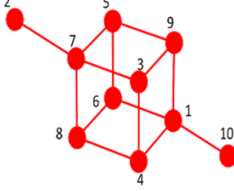
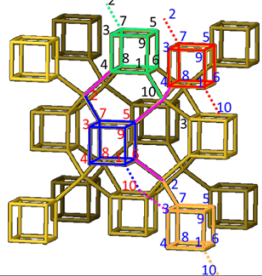
Zeolite	Framework	Unit Cell	SRU	Lattice represented using SRU as building units
SOD				
CHA				
AST				

Figure 2: Zeolite frameworks, unit cells, and SRUs. Using SOD, CHA, and AST as examples, we see that SRU based representation of zeolite frameworks require fewer T-nodes if the connectivity rules are defined appropriately.

110 In this work, we describe the methods to identify SRUs for zeolite frameworks. We also check
 111 whether the graphs of the SRUs are Hamiltonian (more in Section 2). A Hamiltonian cycle is said to
 112 exist in a graph if one can traverse every node of the graph exactly once before returning to the initial
 113 node. We propose two methodologies for identifying SRUs. The first is based on an optimization
 114 model that is formulated as a special instance of the traveling salesman problem (TSP). The second

is an algorithmic approach that employs backtracking and a depth-first search for the Hamiltonian.

The rest of the article is structured as follows. A brief explanation of what Hamiltonian graphs are and how these graphs are useful in the representation of zeolite frameworks is given in Section 2. The methods to identify SRUs are described in Section 3. This includes a step-by-step procedure starting from the CIF file available on the IZA website to finally generate the SRUs. We present the results and discuss the implications of SRUs in Section 4 before concluding.

2 Smallest Repeating Unit (SRU) and Hamiltonian Representation

Hamiltonian paths are graphs where one can traverse all nodes exactly once without re-using a connection. A Hamiltonian cycle, on the other hand, exists when the first and last nodes of a Hamiltonian path are also connected. Identification of Hamiltonian cycles is an NP-complete problem and has been studied extensively in the literature. Several algorithmic, mathematical programming and heuristic approaches have been suggested to reduce the complexity of the problem.^{39–41} The necessary and sufficient conditions for the identification of a Hamiltonian cycle are also known.⁴² Enumerative algorithms guarantee a solution at the cost of computational complexity since they explore all possible paths.^{43,44} The importance of the property of a graph being Hamiltonian is essential to several problems including the TSP. Different linear and mixed-integer linear programming (LP and MILP) formulations of the TSP can be applied to identify the Hamiltonian cycles in a graph.^{45,46}

Hamiltonian cycles allow us to identify the largest rings in zeolite frameworks. In this context, Hamiltonian graphs provide an avenue to break down the complex frameworks into simpler modules, which we call SRUs. The SRU is defined using only topologically independent T-nodes in their respective multiplicities of occurrence in the framework and forming a Hamiltonian path. Essentially, the SRU is a collection of connected T-nodes that adhere to the ratio of topological independent T-nodes in the lattice. The connectivity matrix is the governing rule that defines how these SRUs are to be connected at every T-node to generate the lattice. The matrix value in position (i, j) defines if T-nodes from categories i and j are connected in the lattice. The unit cell enforces parallelepiped structures but it does not capture the connections between T-nodes. The Hamiltonian graph representation, on the other hand, is a visual way to capture the information provided in a connectivity matrix. For a given zeolite framework, its SRU is the smallest and single building block that can be repeatedly used to construct the framework. The copies of the SRU block are connected by a corresponding connectivity matrix. An example is shown in Figure 3. Figure 3a is the SOD framework while Figure 3b demonstrates the position of the SRU in the framework. Figure 3c has the connectivity matrix of size 6 since the SOD framework has 6 topologically independent T-nodes that make up the SRU. Figure 3e, the SRU in red, node 1 is connected to nodes 3 and 6 of the same SRU and nodes 2 and 4 of other SRUs. The repetitive use of this connectivity rule (as shown in Figure 3e) and the SRUs leads to generating an entire zeolite framework.

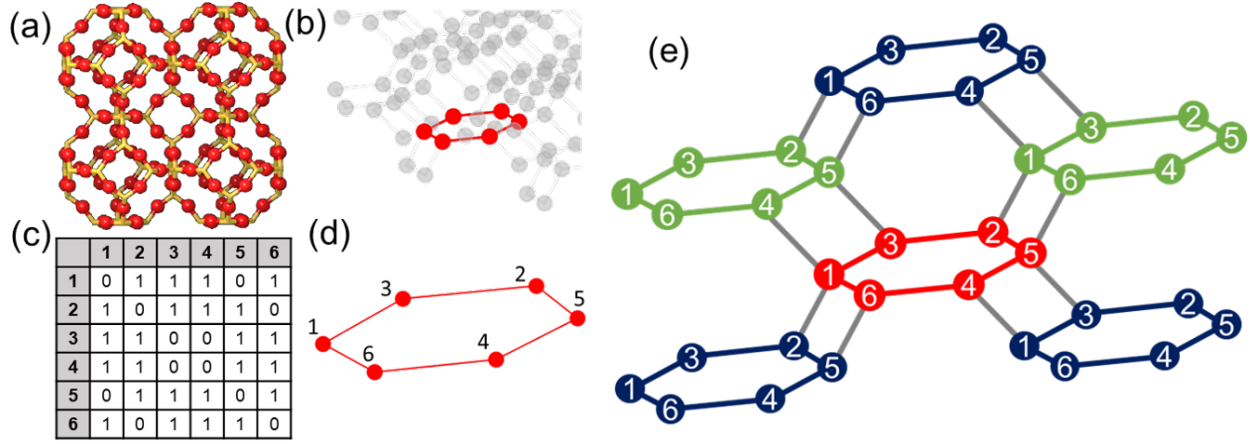


Figure 3: Sodalite (SOD) framework and its SRU. (a) SOD framework, (b) SRU shown in the framework, (c) connectivity matrix that defines the connectivity rules in SOD SRU, (d) SOD SRU with labeled T-nodes, and (e) SOD framework generation using SRUs following the connectivity matrix. The three colors used are illustrative of three different layers in the lattice.

In the case of Chabazite(CHA), the SRU requires only 12 T-nodes shown in Figure 4. On the other hand, 36 T-nodes are necessary to represent the framework using the traditional approach of the unit cell. In Figure 4b, the matrix is of size 12 because of the number of T-node categories in CHA as also seen in Figure 4a. In the connectivity matrix, node 12 is connected to nodes 4, 11, and 8 of the same SRU and node 10 of another SRU. The connectivity matrix should have the sum of each row and the sum of each column has to be equal to four due to the tetrahedral nature. The Hamiltonian graph representation is another way of capturing the information in the connectivity matrix. The max sized Hamiltonian is captured in the circumference of this notation (shown in blue) while the smaller rings are also captured in red. The six-membered ring of CHA is also captured by the T-nodes set [1, 4, 12, 11, 2, 3].

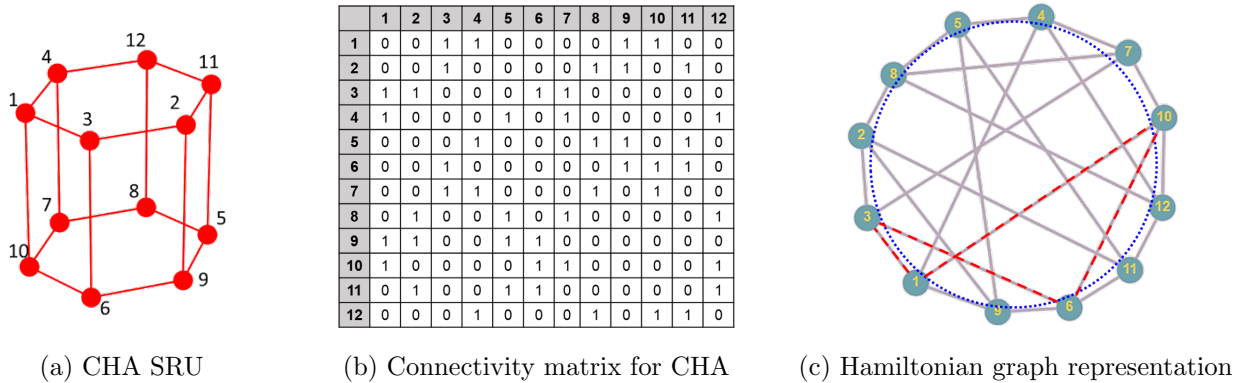


Figure 4: Representing the SRU of zeolite framework CHA using a Hamiltonian graph. (a) The CHA SRU, (b) connectivity matrix for the SRU structure, (c) Hamiltonian graph representation of the connectivity matrix, where the largest Hamiltonian cycle is the circumference of the network.

3 SRU identification

Here we describe how to identify the SRU of a given crystalline framework. We start from the standard Crystallographic Information File (CIF) and obtain all nodes and the connectivity between the nodes. We then classify nodes based on their directional tetrahedral structure followed by categorizing them such that all nodes of the same class are together. Next, we obtain the multiplicity of each class within the unit cell to ensure that the proposed SRU contains the appropriate number of T-atoms from each class. This data is then used towards the identification of the SRU. These steps are also summarized in Figure 5.

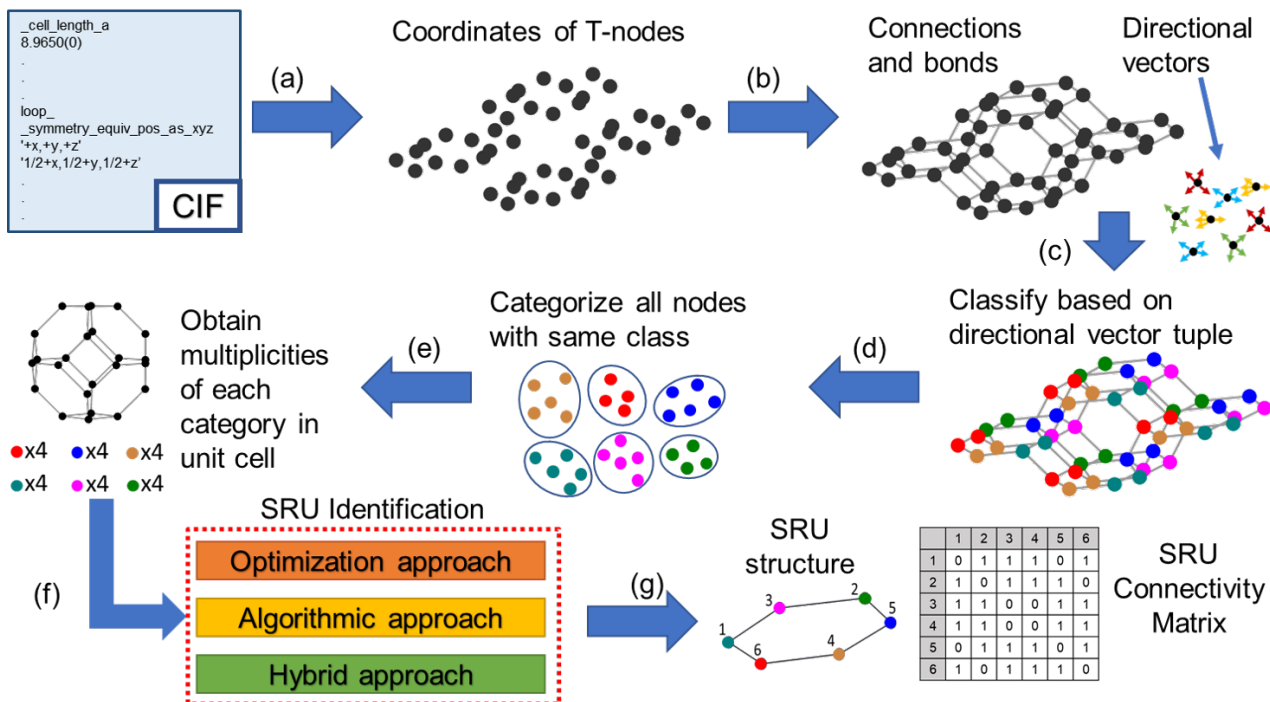


Figure 5: SRU identification methodology summarized. (a) xyz coordinates are generated from the CIF file, (b) connections are obtained between Si-Si neighbors, (c) T-nodes are classified into different categories using directional vector tuples, (d) T-nodes with the same class are categorized, (e) multiplicity of each class is calculated based on the occurrence in the traditional unit cell, (f) all these data are then used to obtain the SRU structures and connectivity matrices(g).

The CIF file contains data for a framework in a compressed fashion that uses planes of symmetry within the unit cell. We first obtain the CIF files from the IZA Database⁷ and generate the *xyz* coordinates for the atoms within the lattice. We then perform a nearest neighbor search on the lattice to identify the connectivity between nodes.

We use the coordinates of the T-atoms and O-atoms(oxygen) given in the CIF file to generate the unit cell. This computation is performed in fractional coordinates of the cell dimensions. After executing all symmetry operations on the T-atoms and O-atoms, we remove atoms with positional duplicity, i.e., overlapping fractional coordinates. The fractional coordinates are generated on the

176 vectors along with the unit cell that may not be aligned with the orthogonal xyz axis. Thus, one needs
 177 to take care of the angles between the vectors of the unit cell and consider the components of each
 178 vector on each axis while converting from the fractional coordinates to the orthogonal xyz system.
 179 For identifying the unique structural T-nodes, we need a sufficiently large lattice. Thus, we consider
 180 a system of $3 \times 3 \times 3$ unit cells. Since we only need T-atoms to identify unique structural T-nodes, we
 181 discard the O-atoms and save the xyz coordinates of all T-atoms in this lattice.

182 3.1 Nearest neighbor search within the lattice

183 Having obtained xyz coordinates, our goal is to identify the neighbors of each T-node. To maintain
 184 consistency of notation with graph theory, we refer to each T-node as a node. If node i is within a
 185 certain distance of node j then we say that these two nodes are connected. This distance may change
 186 for different zeolites because of the bond length and the strain on the bond angles to ensure minimum
 187 energy for stability. For zeolite frameworks, we choose a lower bound on the distance of 2.5\AA and an
 188 upper bound of 3.4\AA . These bounds are approximated from the upper and lower bounds of the Si-O-Si
 189 bond distances and angles.

190 Ideally, we would like to find and compare the distances between all possible pairs of nodes but
 191 this has a computational complexity of $\mathcal{O}(n^2)$, where n is the total number of nodes in the lattice.
 192 For larger zeolites, where n is more than 1000, the computational demand is too high. Therefore,
 193 we perform a grid search method over the pairs of nodes. For every node, we consider a cube with
 194 the node at the center and the length of the cube to be twice the maximum distance permitted for
 195 the bond, i.e., 6.8\AA . This reduces the computational complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$, thus making it
 196 possible to perform this step on larger zeolite frameworks.

197 3.2 Directional vectors, vector tuples, and classification of T-nodes

198 After generating the connections between nodes, we obtain directional vectors from node i to node j
 199 and associate this vector with node i . For nodes that are not on the boundary of the lattice, each node
 200 i has exactly four such unit vectors associated with it due to the tetrahedral nature of the T-nodes.
 201 We store this tuple of four-unit vectors mapped to node i and perform this operation for all nodes.
 202 For nodes lying on the boundary, there may not exist all four-vectors. Therefore, we remove them
 203 from our consideration for the next step. We then categorize together all nodes that have the same
 204 directional unit vector tuple. In other words, we group the nodes that have all four of their directional
 205 unit vectors the same. This ensures that all nodes that were considered earlier are either removed
 206 from consideration for not having four neighbors or are classified based on their unit directional vector
 207 tuple.

208 To understand the filtering of nodes, we demonstrate with the case of CHA in Table 1, beginning
 209 with 288 T-nodes. Some of these T-nodes being on the boundary of the $3 \times 3 \times 3$ only have 3 neighbors
 210 and thus only 3 directional vectors. After removing them from consideration, a total of 172 T-nodes
 211 are left with a tuple of four directional vectors. If they are classified based on uniqueness of the tuple,
 212 12 classes are identified and all of these 172 T-nodes can be classified into these 12 classes.

Table 1: Classification of T-nodes of CHA. After filtering, 172 T-nodes have been classified into 12 categories from the initial 288 T-nodes.

Class	Node numbers in class
1	5, 6, 9, 10, 13, 14, 15, 16, 19, 23
2	27, 28, 29, 30, 31, 32, 33, 34, 37, 38, 39, 40, 41, 45, 47
3	49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 69
4	73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 93
5	101, 102, 105, 106, 109, 110, 111, 112, 115, 119
6	123, 124, 125, 126, 127, 128, 129, 130, 133, 134, 135, 136, 137, 139, 141
7	147, 148, 149, 150, 151, 152, 159, 160, 162, 166
8	171, 172, 173, 174, 175, 176, 177, 178, 181, 182, 183, 184, 188, 190, 192
9	193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 212, 216
10	217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 236, 240
11	243, 244, 245, 246, 247, 248, 255, 256, 258, 262
12	267, 268, 269, 270, 271, 272, 273, 274, 277, 278, 279, 280, 282, 284, 288

For zeolites with moderate to large unit cells, there may present smaller substructures that occur multiple times within the unit cell. For example, if 3 nodes from class 1 and only 1 node from class 2 are present in the traditional unit cell, we want to ensure that the proposed SRU has the same multiplicity. Note that when we speak of repeating units within a unit cell, we are looking for only translational similarity, since rotational symmetry exists for all T-nodes with different orientations of the tetrahedral atom. This ensures that when the lattice is generated using SRU, it captures each type of node the number of times it occurs. A clear understanding of incorporating multiplicity can be understood from the lattice image for zeolite MWF, shown in Figure 6. The blue square marked over the structure is the unit cell for MWF. The positions marked by red circles occur more often than the one with a purple triangle. The ratio of red circles to purple triangles remains to be 5:1. To ensure that this ratio is maintained, multiplicity needs to be considered while defining the SRU. In principle, this is equivalent to adhering to the pigeonhole principle⁴⁷ that ensures that there are not extra nodes of any class or shortage of any nodes from a class when we reconstruct the framework using the SRU.

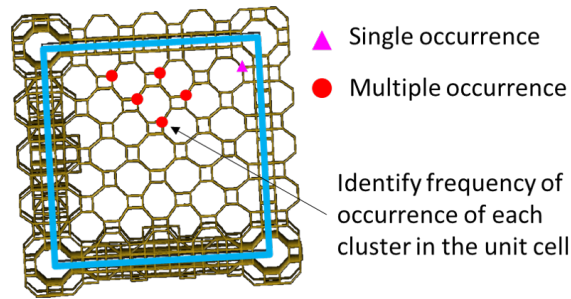


Figure 6: Multiplicities in MWF zeolite framework. Here, the traditional unit cell is outlined in blue. Different structural patterns are present with different frequency of occurrence.

3.3 Connectivity matrix

We define the connectivity matrix as a rule that governs how the SRU structures are placed and connected as a network for lattice generation. After classifying and obtaining all nodes within each class, we want to obtain the connectivity matrix for this network of nodes. To generate the connectivity matrix, we create a matrix of size $U \times U$, where U is the total number of unique classes. We add a value of 1 to the matrix position (l, m) if T-nodes of class u_l and u_m are connected. There may be classes of T-nodes that are connected more than once to another class, which will make the value of the corresponding element in the connectivity matrix to be greater than 1. We will discuss this phenomenon in detail in Section 4. We investigate if the matrix generated in this fashion is Hamiltonian. Several smaller Hamiltonian cycles may exist within this matrix due to the different portals and rings present in the zeolite frameworks. In this work, we check if the entire matrix contains a max-sized Hamiltonian cycle of a certain size, and whether this size is equal to the number of T-nodes in an SRU.

3.4 SRU structure identification

Towards the identification of the SRU structure, the goal is to utilize the multiplicity of each of the classes of T-nodes to come up with an SRU such that using this SRU and the connectivity matrix as the governing rule, one can generate the entire lattice. The SRU and the connectivity matrix together are capable of capturing the complexity of the zeolite in a notation that is smaller compared to the unit cell.

Let $\mathcal{N} = \{1, \dots, N\}$ be the set of total nodes in consideration, $\mathcal{U} = \{1, \dots, U\}$ be the set of classes identified, and $L_u \subset \mathcal{N} \quad \forall u \in \mathcal{U}$, where L_u contains the subset of nodes $n \in \mathcal{N}$ that belong in class $u \in \mathcal{U}$. We also define the set $\mathcal{T} = \{1, \dots, T\}$ where T is the total number of nodes in the SRU. Note that T is known because of the multiplicity obtained of the topologically unique T-nodes. Let $M_{n,n}$ represent the connectivity between nodes, $CM_{u,u}$ be the Connectivity Matrix denoting the connectivity between classes of T-nodes, and m_u denotes the multiplicities of the number of nodes for each class. Given these sets, the SRU identification problem is to identify a structure from the n nodes forming a Hamiltonian path using edges in $M_{(n,n)}$. The final structure should have connectivity of classes between the first and the last node in $CM_{(u,u)}$.

3.4.1 Mathematical programming approach

Here, we describe a MILP formulation for obtaining a compact solution for the SRU. Using the above declared parameters and sets, we declare the following variables for the MILP formulation:

- y_i : 0-1 binary variable to denote if node $i \in \mathcal{N}$ is selected to be a part of SRU,
- $z_{i,j}$: 0-1 continuous variable to capture if the edge between node i
and node j is selected to be a part of SRU,
- $h_{i,l}$: 0-1 binary variable to capture if node i belongs to SRU in position $l \in \mathcal{T}$,

fl_u : 0-1 binary variable to denote if the first node in SRU belongs to class u ,

ll_u : 0-1 binary variable to denote if the last node in SRU belongs to class u ,

$FL_{u,u'}$: 0-1 binary variable to denote if the first node belongs to class u ,

and the last node belongs to class u' .

257

The SRU identification model is as follows:

$$\max \sum_{i=1}^N \sum_{j=i+1}^N z_{i,j}, \quad (1)$$

$$\text{s.t.} \quad \sum_{i \in L_u} y_i = m_u, \quad u \in \mathcal{U}, \quad (2)$$

$$\sum_{i=1}^N y_i = T, \quad (3)$$

$$z_{i,j} \leq y_i M_{i,j}, \quad i \neq j, \quad i \in \mathcal{N}, j \in \mathcal{N}, \quad (4)$$

$$z_{i,j} \leq y_j M_{i,j}, \quad i \neq j, \quad i \in \mathcal{N}, j \in \mathcal{N}, \quad (5)$$

$$z_{i,j} \leq M_{i,j}(1 - y_i - y_j), \quad i \neq j, \quad i \in \mathcal{N}, j \in \mathcal{N}, \quad (6)$$

$$\sum_{l=1}^T h_{i,l} = 1, \quad i \in \mathcal{N}, \quad (7)$$

$$\sum_{i=1}^N h_{i,l} = 1, \quad l \in \mathcal{T}, \quad (8)$$

$$h_{i,l-1} + h_{j,l} \leq 1 + z_{i,j} \quad i \in \mathcal{N}, j \in \mathcal{N}, l \in \mathcal{T} \quad (9)$$

$$\sum_{u=1}^U \sum_{i \in L_u} u \times h_{i,1} = \sum_{u=1}^U u \times fl_u, \quad (10)$$

$$\sum_{u=1}^U fl_u = 1, \quad (11)$$

$$\sum_{u=1}^U \sum_{i \in L_u} u \times h_{i,T} = \sum_{u=1}^U u \times ll_u, \quad (12)$$

$$\sum_{u=1}^U ll_u = 1, \quad (13)$$

$$FL_{u,u'} \leq fl_u, \quad u \in \mathcal{U}, u' \in \mathcal{U}, \quad (14)$$

$$FL_{u,u'} \leq ll_{u'}, \quad u \in \mathcal{U}, u' \in \mathcal{U}, \quad (15)$$

$$FL_{u,u'} \geq fl_u + ll_{u'} - 1, \quad u \in \mathcal{U}, u' \in \mathcal{U}, \quad (16)$$

$$\sum_{u=1}^U \sum_{u'=1}^U FL_{u,u'} \times CM_{u,u'} = 1, \quad (17)$$

$$y_i, h_{i,l}, fl_u, ll_u, FL_{u,u'} \in \{0, 1\},$$

$$0 \leq z_{i,j} \leq 1 \in \mathbb{R}^{N \times N}.$$

The goal of the optimization model is to identify a Hamiltonian path in the lattice that would represent the SRU. The nodes selected for SRU should be connected such that they satisfy the condition of a Hamiltonian cycle. Equation 1 represents the objective that maximizes the connectivity for an SRU to ensure a densely packed SRU. A compact structure helps in perceiving the SRU with the maximum number of edges. Equation 2 ensures that the number of nodes selected in the SRU from a class is equal to the multiplicity m_u for that class. The total number of selected nodes in the SRU must equal to a specific number T , which is enforced by Equation 3. This total number is the sum of all multiplicities over all classes that maintain the ratio observed in the unit cell. This constraint may seem redundant over Equation 2 but has been observed to assist the solver (CPLEX) to converge faster. Equations 4, 5, and 6 assign $z_{i,j}$ a value of 1 if edge (i, j) is included in the final SRU solution and to a value of 0 otherwise. $M_{i,j}$ is a parameter table which is assigned a value of 1 if nodes i and node j are connected in the lattice.

The binary variable $h_{i,l}$. $h_{i,l}$ takes the value of 1 for node i if it represents the l^{th} position visited in the Hamiltonian path. The construction of the Hamiltonian path is formulated as a special instance of TSP. The traveling salesman does not need to visit all nodes but visits only a specific number (multiplicity) of each class in his route, and node i is visited in the l^{th} position of the traveling sequence. Equations 7 and 8 ensure that this condition is met. Equation 9 ensures that the path defined by $h_{i,l}$ is indeed a Hamiltonian path. Equations 10, 11, 12, 13 assign values to the variables fl_u and ll_u which denote the class of the first node and last node of the path. Equation 17 imposes the condition that the SRU contains a Hamiltonian cycle by making the first and last selected nodes' classes to be connected.

3.4.2 Algorithmic approach

In this case, we use a backtracking algorithm⁴⁴ to avoid enumerating all solution possibilities. The detailed algorithmic process is explained in Figure 7 with the help of a flowchart. The algorithmic approach starts from a single node in the lattice and follows the connectivity for the next node in search of a Hamiltonian path with a depth-first search. The Hamiltonian cycle is enforced between the class of the first node and the last node in this explored path. We explore the neighbors recursively until we reach a solution or when the multiplicity threshold is violated. Using such a recursive approach, we effectively enumerate the best paths.

When the search for structure begins, the algorithm loads the information related to all nodes, classes, connectivity matrix, and multiplicity of each class in the framework. From this framework of nodes, the algorithm starts at (i) base node n , (ii) the total length t initialized at T , and (iii) parent tree that is initialized with a null value. The next check is to verify if the number of feasible solutions has already been obtained. The algorithm proceeds to expand the neighbors of node n that are given by $[b_w, b_x, b_y, b_z]$, and the class of these nodes $[l_w, l_x, l_y, l_z]$. As a depth-first approach, we expand on any one of these neighbors. The neighbor that is expanded on, is stored in B , and its class is stored in L .

If L violates the multiplicity of its class m_u , the search continues on the next neighbor. After finding a neighbor that satisfies multiplicity, B is appended to a list of the children of n . This branch

is recorded for future considerations. If the total length explored at this stage is equal to the desired length of T , a check is performed for the first and last node in this branch. The solution is stored if the Hamiltonian cycle criteria is fulfilled and discarded otherwise. In the case that the length is not equal to T , all other neighbors are expanded and stored for future expansion if needed. This list of recorded solutions operates in a last-in-first-out manner when restarting a search from a discarded solution.

Once all neighbors are examined, (i) the base node is updated to the current neighbor B , (ii) the length of the remaining tree to be explored is updated to $t - 1$, and (iii) the node n is appended to the parent list. The parent list contains the solution explored until now. The recursive search is then performed with the new base node and the desired length to be explored reduced by 1 unit. The search stops when the algorithm reports the desired number of SRU structure solutions.

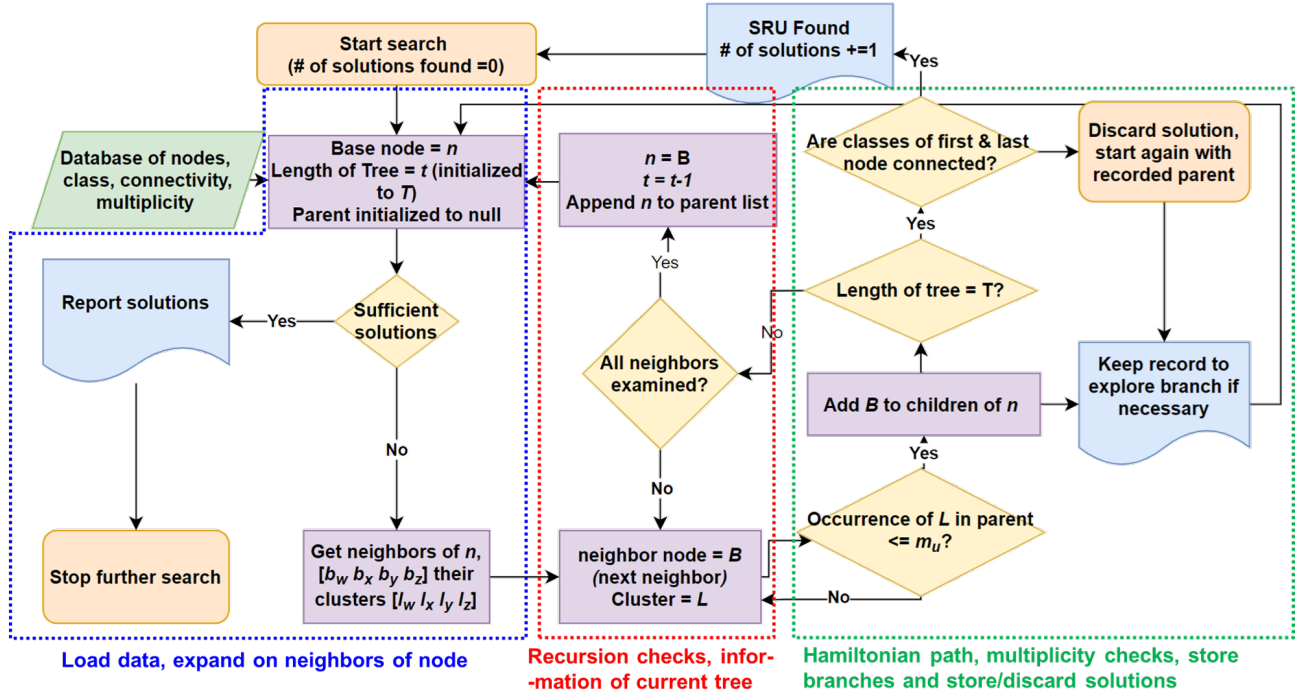


Figure 7: Flowchart of the proposed algorithm for generating SRU solutions using backtracking approach in a depth-first search.

Figure 8 illustrates the working of the algorithm for the case of Sodalite. The algorithm involves expanding on the base node that is updated at every recursive call. The node number is in the circle, while the superscript is the class of that node. The blue dashed line denotes no expansion on that path since it leads to the parent. The orange color nodes cannot be expanded, since they violate the multiplicity of class in the solution. For the case of SOD, there are 6 classes, each with a multiplicity of 1. Thus resulting in a value of 6 for T . The connectivity matrix is previously given in Figure 3c.

The algorithm begins with node 10 of class 6, and its neighbors are [67, 118, 172, 274]. A randomly selected node is expanded on, in this case, node 67 (class 4) followed by node 148 (class 5). It should be noted here that node 10 cannot be expanded as it is present in the parent of the current tree. When considering neighboring nodes of 148, we encounter node 41, which cannot be explored since the class

of node 41 (class 6) already exists in the parent tree i.e. node 10 (class 6). At every step, the parent is updated and the base node is reset. The algorithm ends when the total number of nodes in the parent is equal to T , and the class of the last node is connected to the class of the first node in the tree. The green path highlights the Hamiltonian path where the classes of the first and last nodes are connected in the connectivity matrix, thus leading to a Hamiltonian cycle.

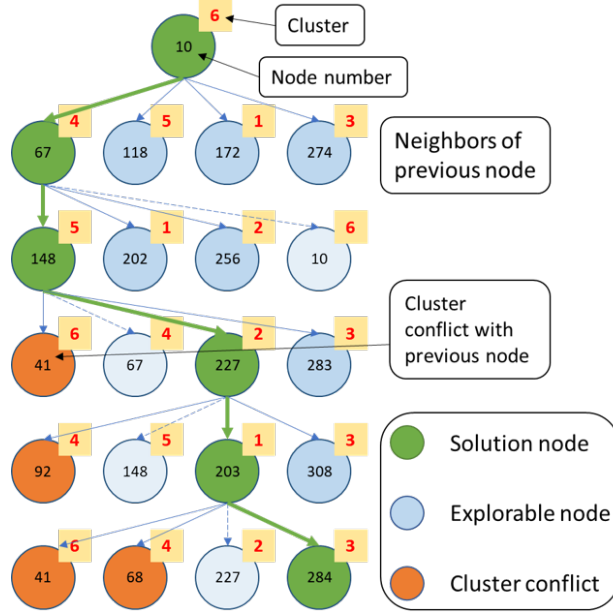


Figure 8: Illustrating the algorithmic approach that explains how node tracking works on SOD framework. Beginning with node 10, the algorithm recursively iterates local neighbors using depth first search with unique class of T-nodes and eventually identifies the green path as the solution.

3.4.3 Hybrid Approach

The hybrid approach combines both the optimization model and the algorithmic approach to gain the benefits of each. The original optimization model (Equations 1–17) is complete in the sense that it generates the final feasible solutions satisfying the Hamiltonian cycle constraints (Equations 7–17). While the model is succinct, the required number of binary variables and constraints to satisfy the Hamiltonian nature of the SRUs can increase the computation. In the hybrid approach, we relax the Hamiltonian constraints to generate a feasible SRU configuration with smallest number of T-nodes. Therefore, the relaxed optimization model in the hybrid approach includes only Equations 1–6. The obtained SRU connectivity matrix is checked for the Hamiltonian cycle using a part of the original algorithmic approach. This guarantees that the solution generated by the relaxed optimization model is Hamiltonian. Rather than obtaining the complete solution from both approaches, the hybrid approach exploits the strengths of each and generates the complete SRU solution. After that we find the Hamiltonian cycle using a part of the algorithmic approach. We use the backtracking algorithm starting with the first element of the connectivity matrix and keep appending nodes as the algorithm explores the path. If the explored path size matches with the size of the matrix, we check for connectivity of the first and the last nodes of the path to ensure the existence of Hamiltonian cycle. Otherwise, the

backtracking continues towards exploring a feasible path. Overall, the hybrid approach reduces the computation time drastically.

4 Results and Discussion

In comparison to unit cells that require six parameters (a , b , c , α , β , γ) and coordinates for the T-nodes, SRUs utilize fewer T-nodes and have been observed to have Hamiltonian properties. SRUs do not need to be parallelepiped. With well-defined connectivity, structures with fewer T-atoms exist that can uniquely represent the zeolite framework. A total of 229 zeolites were studied as listed on the IZA Database,⁷ and connectivity matrices were obtained for 180 zeolites. Within these 180 zeolites, 179 were observed to have a full-sized Hamiltonian cycle, i.e., the size of the largest Hamiltonian cycle was equal to the size of the matrix. The SRU of zeolite AFY was observed with no Hamiltonian cycle. Further investigation is needed to elucidate why this is the case. The SRU for AFY consists of 17 T-nodes, and a total of 214,321 possible paths for Hamiltonian have been explored. Not all zeolites have to be Hamiltonian cycles of the full size. An interesting observation is that most of the zeolites (179 out of 180) have SRUs with connectivity among nodes that are Hamiltonian cycles. These SRUs and their connectivity matrices are reported in the Supporting Information.

We analyzed the reduction in T-nodes that SRU representation has to offer. SRUs require less than half of the T-nodes compared to unit cells for most of the zeolites studied. While some SRUs have non-intuitive structures of selected T-nodes, the use of the fewest possible number of T-nodes is the greatest advantage of the SRU-based representation of crystalline frameworks. From Figure 9a, we observe that the SRU representation reduces the number of T-nodes necessary to describe the zeolite framework by 2, 3, and in some cases even 4 times in magnitude (Figure 9b). Several zeolite SRUs have the number of T-nodes same as their original unit cells but offer insights into their nature through their Hamiltonian graphs. The detailed reduction for each zeolite is given in Table 2.

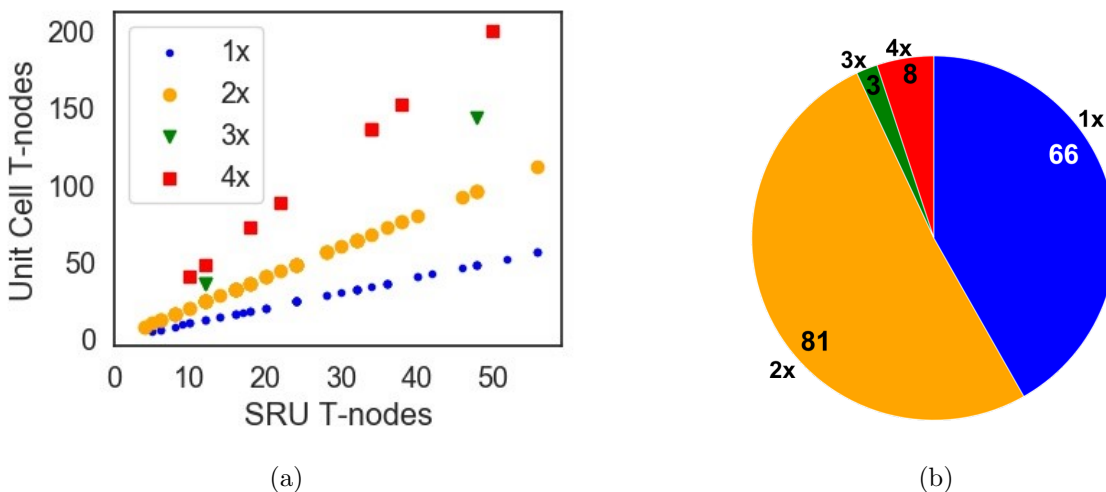


Figure 9: Reduction in the number of T-nodes that are required to describe a zeolite framework using the proposed SRU in comparison with the traditional unit cell (a) Number of T-nodes in SRU vs unit cell, (b) Summary of reduction in T-nodes provided by SRUs.

Table 2: Number of T-nodes in zeolite SRUs. The value corresponding to each zeolite is the number of T-nodes in the SRU. A factor of 1x, 2x, 3x, and 4x is observed between the unit cell and SRU T-nodes.

Zeolites with T-node no reduction													
PTY	10	AFV	30	PON	24	SOS	24	BEC	32	MTT	24	NPT	36
PWW	33	OWE	16	ASV	20	DFT	8	MOV	12	VET	17	UOZ	40
AWW	24	BRE	16	LTL	36	LTJ	16	SFS	56	CGS	32	IRR	52
CZP	24	LTA	24	SSY	28	GME	24	BPH	28	JBW	6	MAZ	36
BOF	24	ZON	32	SFE	14	LIO	36	AVL	42	JNT	32	IFY	48
OFF	18	THO	10	MRT	24	EDI	5	IFO	32	ERI	36	LOV	18
SBN	10	MEP	46	SFF	32	ETV	14	CAN	12	ATT	12	AFI	24
ETR	48	AFS	56	SVV	36	JOZ	20	CSV	22	NPO	6	OSO	9
RRO	18	AFR	32	MEI	34	JSN	16	PUN	36	AFX	48	PWO	20
JSW	48	UOS	24	SAV	48								
Zeolites with T-node reduction factor 2x													
EWO	12	RWR	16	MFS	18	SGT	32	MER	16	RWY	24	GOO	16
ITH	28	CAS	12	AFN	16	ATN	8	BOZ	46	GIS	8	SAF	32
SZR	18	MTF	22	TON	12	EPI	12	SAS	16	PCR	30	IHW	56
ITW	12	CDO	18	JRY	12	WEI	10	SAO	28	USI	20	_CON	28
OSI	16	KFI	48	NAT	10	OKO	34	TER	40	LAU	12	RTE	12
ATS	12	NSI	6	AEL	20	BOG	48	ITE	32	BIK	6	ACO	8
SOD	6	RTH	16	OBW	38	NAB	5	APC	16	SFN	16	STF	16
MEL	48	IFR	16	PHI	16	SFH	32	GON	16	ABW	4	EZT	24
AWO	24	CFI	16	UFI	32	SOF	20	AEN	24	APD	16	MON	8
ANA	24	IWR	28	FER	18	MOR	24	DAC	12	YUG	8		
AHT	12	VSV	18	DON	32	BCT	4	AEI	24	SFO	16		
MTW	14	ATV	12	RHO	24	AFO	20	UTL	38	AET	36		
Zeolites with T-node reduction factor 3x													
CHA	12	SBT	48	ATO	12								
Zeolites with T-node reduction factor 4x													
MTN	34	STI	18	EEI	50	NES	34	UEI	34	NON	22	IWV	38
AST	10												

We proposed two different approaches to identify the SRU structures. Both approaches satisfy the criteria of their ability to recreate the lattice using SRUs and the connectivity matrix. The difference arises in the number of edges or the compactness of the SRU structures. Figure 10 shows the differences in SRU structures obtained from the optimization and the algorithmic approach, respectively, for Chabazite. The number of edges in SRU obtained using the optimization model (Equations 1-17) is 18 (global maxima), which takes several minutes to solve, whereas the solutions from the algorithmic approach (Figure 7) are obtained within a fraction of a second. The optimization model is large

and involves many discrete variables, making it difficult to solve due to the inherent symmetry and the presence of multiple global optimal solutions. The SRUs are defined as the building units of the zeolites and these are contained in numbers in the lattice. This poses challenges for the solvers since no branch can be cut off while searching for the global optima.

To understand the cause of symmetry, consider an example of an SRU that has T-nodes from three classes (denoted by 1, 2, and 3). If a Hamiltonian path as [123] is a solution, so are [231] and [312] due to Hamiltonian cycle constraints imposed. Since these are just the classes of the structure, a Hamiltonian cycle exists for any three nodes (denoted by i in the model) along the path that is defined by classes [12312...123]. An algorithmic approach was developed to overcome this limitation. The drawback of the algorithmic approach, however, is that we may not obtain a compact structure for the SRU that would ease visualization. The advantage is that it allows us to obtain potentially many feasible solutions in a quick time that may be used to initialize the solver for solving the optimization model where solver time is a bottleneck.

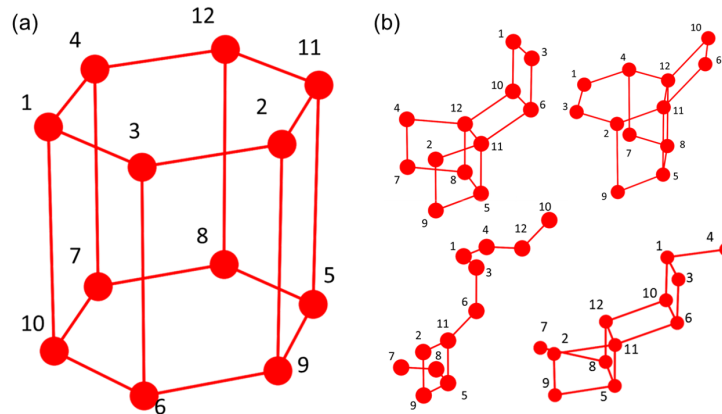


Figure 10: Differences in Chabazite SRU structures obtained from optimization and algorithmic approaches. (a) The optimization model generated a solution with maximum connectivity, while (b) the algorithmic approach generated several sub-optimal but feasible solutions.

We can also now comment on the uniqueness of the solutions of SRUs. As discussed above, the optimization model gives the most compact structure (which could have multiplicity as well), whereas the algorithmic approach generates several feasible solutions. Irrespective of whether the solution is just a feasible solution or optimal, the SRU structure obtained is capable of generating the lattice. Each feasible structural solution has the same connectivity matrix. For Chabazite, we see in Figure 11 that irrespective of the solution, the final lattice generated using the connectivity matrix generates the CHA framework. Even though the structures obtained are different, they obey the rule of having exactly 1 node from each of the classes from 1 through 12. This adheres to the multiplicity defined earlier since, in the unit cell, 3 of each of the 12 occur adding up to the 36 T-nodes.

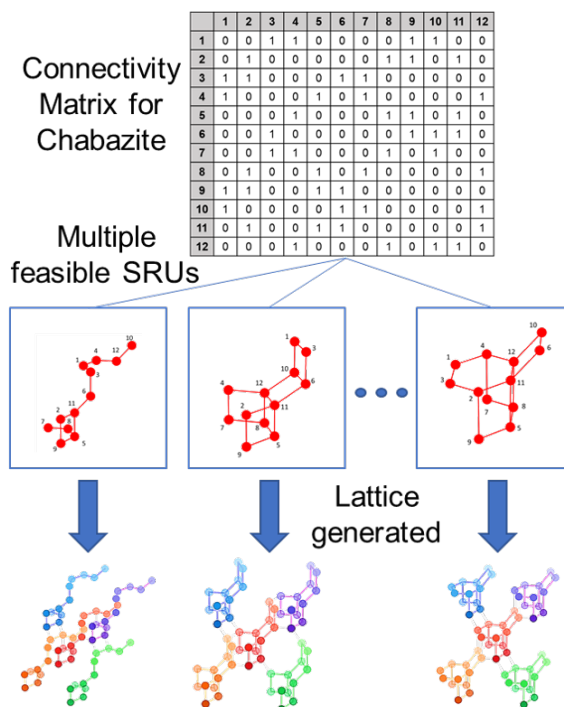


Figure 11: Multiple feasible SRU structures obtained for CHA from the same connectivity matrix demonstrating the non-uniqueness of SRU structures.

The isomorphic nature of the connectivity matrix of an SRU is another important issue to consider. Two graphs are isomorphic if renaming the vertices in a different order generates the same matrix. Note that two different solutions for the same framework are obtained from the same connectivity matrix. The isomorphic nature is not the same as the multiplicity of the solutions. If two different frameworks having different structures have matrices that can be relabeled to give the same matrix, they would be called isomorphic. This property would be a defining factor in the determination of the upper bound of possible SRUs that can be designed for a given number of T-nodes. As shown in Figure 12, the network graphs for DFT and MON are not the same at first glance. However, under the following mapping $[(1, 1), (2, 5), (3, 7), (4, 2), (5, 3), (6, 4), (7, 8), (8, 6)]$, where the pair is represented as vertex label (DFT, MON), they are essentially the same. Even though their connectivity matrices are isomorphic, their SRUs are different (see Supporting Information).

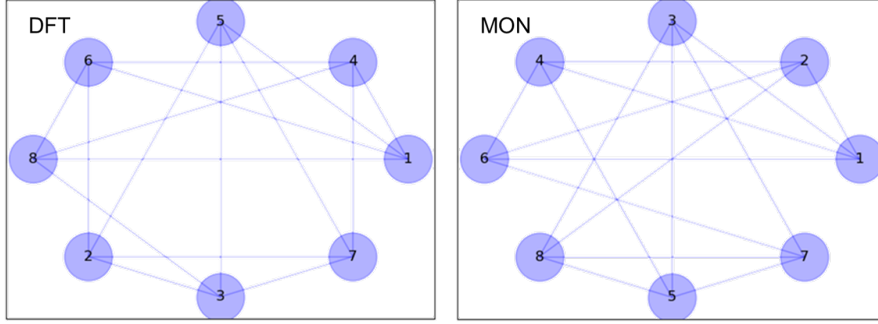


Figure 12: Hamiltonian graphs for DFT and MON. Under the mapping [(1, 1), (2, 5), (3, 7), (4, 2), (5, 3), (6, 4), (7, 8), (8, 6)], they are essentially the same graphs and thus isomorphic.

Additional rules that can be used to determine the possible number of unique matrices are given below. For a $n \times n$ matrix M , where each element is denoted by $M_{i,j}$, the following rules hold for the number of tetrahedral T-nodes. Equations 18 and 19 ensure that each node in the connectivity matrix has exactly 4 adjacent nodes.

$$\sum_{i=1}^n M_{i,j} = 4 \quad j \in \mathcal{N}, \quad (18)$$

$$\sum_{j=1}^n M_{i,j} = 4 \quad i \in \mathcal{N}. \quad (19)$$

The distribution of SRU size for the zeolites studied is given in Figure 13a. This information combined with the upper bound obtained for zeolites can direct the search for new zeolite materials for a given number of T-nodes. The SRU representation can be potentially used for new framework prediction. The rules of the connectivity matrices themselves limit the number of possible structural solutions thus giving an upper bound of such designs. For inverse design of new nanoporous frameworks, such as zeolites, we need to design new SRUs. This will involve constructing new connectivity matrices for fixed number of topologically independent T-nodes. If the size of a new SRU is denoted by N , then we need to design a new connectivity matrix of size $N \times N$ while satisfying the following general conditions: (i) the matrix is symmetric since the node connectivity are symmetric, (ii) diagonal elements of the matrix is zero for pure silica frameworks, (iii) the matrix elements take the values of 0, 1, 2, or 3. Although no observation has been made in this work for a connectivity value of three, the possibility of such a structure should not be eliminated, and (iv) the sums of all the elements in each row and each column are both equal to four for pure silica frameworks. Based on our observation, one can also include that the connectivity matrix should be Hamiltonian, to restrict the already large combinatorial design space. Furthermore, we can reduce the computation by imposing that the designed matrix is isomorphically unique. Given these rules, the number of feasible matrices of size $N \times N$ is finite, which provides with an upper bound towards the existence of frameworks. It is important to note here that the inverse design of zeolites requires both the design of connectivity matrix and 3-D geometric configuration. While we layout the rules for connectivity matrix, the geometric aspect of the SRU needs further work.

Observing the current hypothetical zeolites and then identifying the potential gap where certain types of frameworks might not have been observed can also be beneficial. To that end, a sample of 10,570 hypothetical zeolite frameworks from the Atlas of Prospective Zeolite Structures (formerly known as Database of Hypothetical Zeolites Structures, or Deem Database)⁴⁸ is studied in terms of their SRUs using the hybrid approach. The distribution of SRU sizes in terms of the number of T-nodes is shown in Figure 13b. Interestingly, we observe that most of the frameworks have SRUs with 40 or less number of T-nodes.

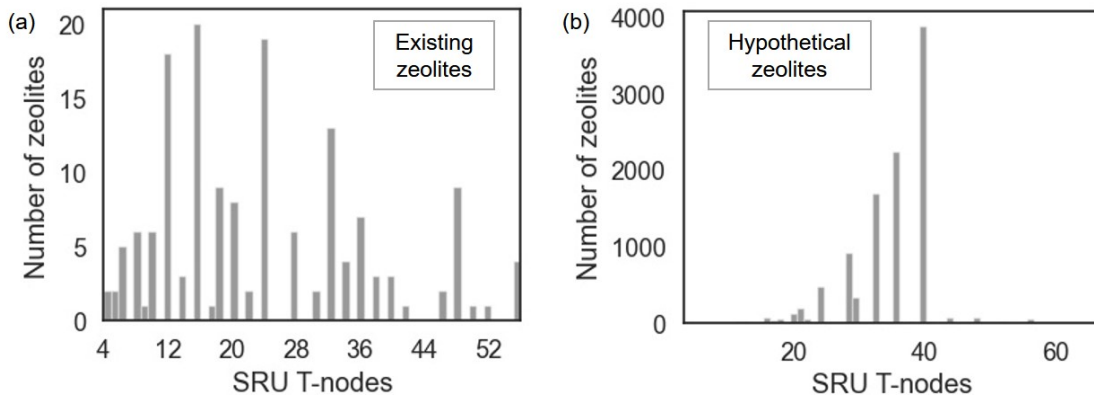


Figure 13: Distribution of SRU sizes: (a) existing 158 zeolite frameworks, and (b) 10,570 hypothetical zeolite structures.

Earlier, we referred to the possibility of some matrix elements having a connectivity value of 2. This can be associated with an AB type of symmetry of layers within the zeolite. The connectivity matrix of ATN, as shown in Figure 14, is an example of such a situation. The SRU has r T-nodes (Figure 14c) While each T-node is connected with four other T-nodes, due to the tetrahedral structure, some of these T-nodes are connected to two T-nodes of the same class. For example, as shown in Figure 14d, node 1 is connected to node 8 of two different SRUs. This is possible when the SRU has a small length in one of the dimensions. The connectivity of 2 is obtained in a manner that one of the nodes is within one SRU and the other node (of the same category) is from another SRU. This can be seen from Figure 14e, where class 7 of red SRU is connected to nodes of class 2 of both red and blue SRUs.

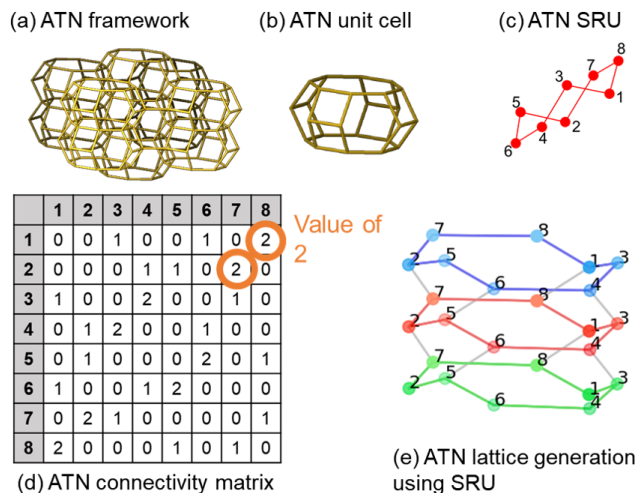


Figure 14: Understanding the observed special feature in some frameworks. (a) ATN framework represented using T-nodes, (b) traditional unit cell for ATN, (c)SRU generated for ATN, (d) connectivity matrix for ATN with highlighted special value, and (e) ATN SRUs stacked using connectivity matrix explaining the observed feature.

Aluminum substitution in place of Silicon in the zeolite framework changes the acidic nature of the zeolite. This Al substituted Bronsted acid contains certain cations to balance the charge thus introducing an acidic nature. Several possible sites exist for Al substitutions which will lead to different properties of the zeolites and these Al substitutions are governed by the Loewenstein's rule which states that there should not exist any Al-O-Al bond in the framework.⁴⁹ Though there are exceptions, the Loewenstein's rule holds for most zeolites.⁵⁰ Determining the possible locations of Al substitutions using the unit cell representation is computationally expensive for at least two reasons. First, there are more candidate positions for substitution which exponentially increases the number of possibilities. Second, it is nontrivial to enforce the Loewenstein's rule for large number of T-atoms within a unit cell. On the contrary, the SRU representation provides a clear advantage due to its smallest size and number of unique T-atoms. Furthermore, the connectivity matrices provide a systematic way to generate all possible Al substitutions under the assumption that the translational symmetry of Al and Si in the zeolite framework also holds within the SRU.²³

To describe how one can use the connectivity matrix of an SRU to generate different Al-substituted frameworks, we first introduce the concept of node index (NI), which is a number assigned to each T-atom based on the types of atoms (Si or Al) that constitute the T-atom and its connected neighbors. For a pure silica framework, the NI for each Si T-atom is always four, because it is connected with four other Si T-atoms. However, when a T-atom is substituted by Al, the node index value changes. The NI for a central Si increases with +1 if it is connected to another Si and decreases by -1 if it is connected to Al. For a central Al atom, another -2 is added to distinguish it from a central Si atom. Since each atom is connected with four other atoms, this gives rise to five other possible NI values, as shown in Figure 15. For example, if one of the four connected atoms is substituted by Al, then the node index of the central Si T-atom is two ($= 1+1+1-1$). If two of the four connected atoms are

substituted by two Al, then the NI of the central Si T-atom is zero ($= 1+1-1-1$), and so on. Therefore, the six possible NI values are 4, 2, 0, -2, -4, and -6. The advantage of NI is that we can use it to enforce the Loewenstein's rule. Specifically, a T-atom can be Al-substituted if and only if its NI is four.

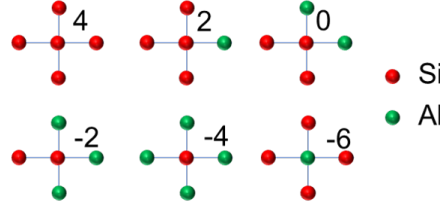


Figure 15: The six possible scenarios of Al-substitution and the corresponding node index values for each T-atom in a zeolite framework.

Using the NI and the Loewenstein's rule, one can enumerate all possible Al substitutions for a connectivity matrix. In a pure silica framework, each element of the matrix can have a value of either zero or one. Figure 16a shows such an example for CHA, where the Si atom in T-node #1 is connected with Si atoms in T-nodes #3, 4, 9 and 10, thereby giving a NI value of 4. However, in the case of Al substitution, we assign -1 to denote that a Si T-atom is connected to Al-substituted T-atom. If the T-atom itself is Al, then we assign -2 to the diagonal element. This is illustrated in Figure 16b, where the Si atom in T-node #1 is substituted by Al. If we want to substitute another Si atom using a second Al atom, then the candidate nodes are #2, 5, 6, 7, 8, 11 and 12. These are the nodes that allows a second substitution without violating the Loewenstein's rule. If we select node #2, then the resultant connectivity matrix is given in Figure 16c. Note that if we select a different node than node #2, then the connectivity matrix will be different. We also notice that, after the first two substitutions in nodes #1 and 2, the candidate nodes for the third substitution are T-nodes #5, 6, 7, and 12.

	1	2	3	4	5	6	7	8	9	10	11	12	NI
1	0	0	1	1	0	0	0	0	1	1	0	0	4
2	0	0	0	0	0	0	0	1	1	0	1	0	4
3	1	1	0	0	0	1	1	0	0	0	0	0	4
4	1	0	0	0	1	0	1	0	0	0	0	1	4
5	0	0	0	1	0	0	0	1	1	0	1	0	4
6	0	0	1	0	0	0	0	1	1	1	1	0	4
7	0	0	1	1	0	0	0	1	0	1	0	0	4
8	0	1	0	0	1	0	1	0	0	0	0	1	4
9	1	1	0	0	1	1	0	0	0	0	0	0	4
10	1	0	0	0	1	1	0	0	0	0	0	1	4
11	0	1	0	0	1	1	0	0	0	0	0	1	4
12	0	0	0	1	0	0	0	1	0	1	1	0	4
NI	4	4	4	4	4	4	4	4	4	4	4	4	4

(a) No Al substitutions

	1	2	3	4	5	6	7	8	9	10	11	12	NI
1	-2	0	-1	-1	0	0	0	0	-1	-1	0	0	-6
2	0	0	1	0	0	0	0	1	1	0	1	0	4
3	-1	1	0	0	0	1	1	0	0	0	0	0	2
4	-1	0	0	0	1	0	1	0	0	0	0	1	2
5	0	0	0	1	0	0	0	1	1	0	1	0	4
6	0	0	1	0	0	0	0	0	1	1	1	0	4
7	0	0	1	1	0	0	0	1	0	1	0	0	4
8	0	1	0	0	1	0	1	0	0	0	0	1	4
9	-1	1	0	0	1	1	0	0	0	0	0	0	2
10	-1	0	0	0	0	1	1	0	0	0	0	1	2
11	0	1	0	0	1	1	0	0	0	0	0	1	4
12	0	0	0	1	0	0	0	1	0	1	1	0	4
NI	-6	4	2	2	4	4	4	4	2	2	4	4	4

(b) Al substitution at node 1

	1	2	3	4	5	6	7	8	9	10	11	12	NI
1	-2	0	-1	-1	0	0	0	0	-1	-1	0	0	-6
2	0	-2	-1	0	0	0	0	0	-1	-1	0	-1	-6
3	-1	-1	0	0	0	1	1	0	0	0	0	0	0
4	-1	0	0	0	1	0	1	0	0	0	0	1	2
5	0	0	0	1	0	0	0	1	1	0	1	0	4
6	0	0	1	0	0	0	0	0	1	1	1	0	4
7	0	0	1	1	0	0	0	1	0	1	0	0	4
8	0	-1	0	0	1	0	1	0	0	0	0	1	2
9	-1	-1	0	0	1	1	0	0	0	0	0	0	0
10	-1	0	0	0	0	1	1	0	0	0	0	1	2
11	0	-1	0	0	1	1	0	0	0	0	0	1	2
12	0	0	0	1	0	0	0	1	0	1	1	0	4
NI	-6	-6	0	2	4	4	4	2	0	2	2	4	4

(c) Al substitution at node 1,2

Figure 16: CHA connectivity matrix with node index highlighted after additional substitution of Al. (a) No Al substituted, node index is 4 for all thus allowing the possibility of Al substitution at any site. (b) Al is substituted at the first node thus impacting all node indices, only node 2, 5, 6, 7, 11, and 12 are possible to be substituted if node 1 is Al following the Loewenstein rule. (c) Node 2 is substituted with Al thus reducing the further substitution of Al only to nodes 5, 6, 7, and 12.

Table 3: Al substitutions in the zeolite framework LTL.

Al substitutions	Si/Al ratio	Number of possible structures
1	35.00	36
2	17.00	558
3	11.00	4,896
4	8.00	26,925
5	6.20	97,212
6	5.00	235,054
7	4.14	381,888
8	3.50	412,596
9	3.00	288,952
10	2.60	124,944
11	2.27	30,240
12	2.00	3,148
Total		1,606,449

480 Following the above strategy, we can generate all plausible frameworks with given number of Al
 481 substitutions. For example, the possible frameworks modifications with Al substitutions for LTL
 482 framework are presented in Table 3. LTL has an SRU with 36 T-nodes. If three of these 36 Si T-nodes
 483 are substituted by Al with a Si/Al ratio of 11, then the possible number of frameworks are 4,896.
 484 In total, there are 1,606,449 cases possible. While this number is high, the SRU-based framework
 485 generation approach allows us to enumerate all of them in a systematic way. Note that, for a given
 486 number of Al atoms to be substituted in the framework, we calculate the Si/Al ratio assuming the
 487 translational symmetry of Al and Si in the zeolite framework hold within the SRU. Interestingly,
 488 we can have at most 12 substitutions. Beyond this, a violation of the Loewenstein’s rule occurs.
 489 Since SRUs are the smallest structural units, we can only generate the framework descriptions (i.e,
 490 connectivity matrices) for Si/Al ratios with uniform distribution of Al. Sato²³ computed the number
 491 of possible Al substitutions along the same way we describe above. However, Sato’s work was based
 492 on the traditional unit cell, while we use SRUs.

493 There are many frameworks which may have the same number of T-nodes. However, the number
 494 of possible Al substitutions can still be different, if the frameworks are not isomorphic. Given two
 495 isomorphic networks, the possible number of Al substitutions following the Loewenstein rule are the
 496 same. To illustrate this, we compute the number of substitutions possible in several frameworks with
 497 SRU size 12. The results are given in Table 4 . The Si/Al ratio of these zeolites vary between 1 and
 498 11. We observe that CHA, AHT and ATV can have 150 possible Al substitutions, which suggests
 499 that these frameworks are isomorphic. One the other hand, EWO and TON can have 182 possible
 500 substitutions, while CAN and ATS can have 195 possible substitutions. Further investigation on the
 501 feasibility of these structures may be carried out using molecular simulation towards stability of the
 502 structures as done in identifying the hypothetical zeolite database.

Table 4: Possible number of configurations with specified Al substitutions in zeolite frameworks with SRU with 12 T-nodes.

Zeolite	Al substitutions						Total
	1	2	3	4	5	6	
CHA	12	42	52	30	12	2	150
EWO	12	48	76	42	4	0	182
CAN	12	48	76	45	12	2	195
ATS	12	48	76	45	12	2	195
CAS	12	48	72	32	0	0	164
TON	12	48	76	42	4	0	182
JRY	12	42	52	22	4	0	132
EPI	12	42	52	18	0	0	124
ATO	12	48	76	45	12	2	195
MVY	12	48	76	48	12	2	198
UEI	12	42	52	26	4	0	136
LAU	12	42	48	17	0	0	119
DAC	12	42	48	14	0	0	116
AHT	12	42	52	30	12	2	150
ATV	12	42	52	30	12	2	150

5 Conclusion

In this work, we used Hamiltonian graphs to find the smallest repeating units (SRUs) of zeolite frameworks that require fewer T-nodes than the traditional unit cells. We developed two approaches to locate SRUs, namely (i) a mathematical programming-based optimization approach motivated by the traveling salesman problem formulation, and (ii) an algorithmic approach based on backtracking and depth-first search. A hybrid approach combining both of these to improve computational speed has also been proposed. To demonstrate the advantages of this novel representation, we analyzed 158 zeolite frameworks from the IZA-SC database and over 10,000 hypothetical zeolite frameworks. The number of T-nodes required to describe these frameworks is reduced by 50-75% in comparison to their traditional unit cells for 92 of the 158 zeolite frameworks. The SRUs for all 158 zeolite frameworks constructed using structurally independent T-nodes obey the properties of Hamiltonian graphs. Due to its Isomorphic nature, multiple SRU configurations are found for the same zeolite framework. Using the conditions of Hamiltonian and Isomorphism, one can obtain an upper bound on the number of potential frameworks possible and thus guide the search towards new zeolite frameworks. It should be noted that the optimization formulation is a large-scale problem with inherent symmetry. As a final remark, there exists several more hypothetical zeolite frameworks where this approach can be applied and can be utilized in predicting possible Al substituted frameworks.

Acknowledgment

Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing. The authors gratefully acknowledge funding support from the U.S. National Science Foundation (NSF CAREER award CBET-1943479) and the American Chemical Society Petroleum Research Fund (ACS PRF 58764-DNI9). They also thank Akhil Arora for several suggestions towards improving the exposition of this manuscript.

Supporting Information

SRU for 158 zeolite frameworks. (PDF)

Connectivity matrices for 158 zeolite frameworks. (XLSX)

Literature Cited

1. Davis, M. E. Ordered porous materials for emerging applications, *Nature* **2002**, *417*, 813–821.
2. Weckhuysen, B. M.; Yu, J. Recent advances in zeolite chemistry and catalysis, *Chem. Soc. Rev.* **2015**, *44*, 7022–7024.
3. Pérez-Ramírez, J.; Christensen, C. H.; Egeblad, K.; Christensen, C. H.; Groen, J. C. Hierarchical zeolites: enhanced utilisation of microporous crystals in catalysis by advances in materials design, *Chem. Soc. Rev.* **2008**, *37*, 2530–2542.
4. Li, J.; Corma, A.; Yu, J. Synthesis of new zeolite structures, *Chem. Soc. Rev.* **2015**, *44*, 7112–7127.
5. Barrer, R. M. Synthesis of a zeolitic mineral with chabazite-like sorptive properties, *J. Chem. Soc.* **1948**, pages 127–132.
6. Yang, Q.; Liu, D.; Zhong, C.; Li, J.-R. Development of computational methodologies for metal–organic frameworks and their application in gas separations, *Chem. Rev.* **2013**, *113*, 8261–8323.
7. Baerlocher, C.; McCusker, L. B. Database of zeolite structures <http://www.iza-structure.org/databases>, (accessed Nov 19, 2020).
8. West, A. R. *Basic solid state chemistry*; John Wiley and Sons Incorporated, 1999.
9. Rosenbrock, C. W.; Homer, E. R.; Csányi, G.; Hart, G. L. Discovering the building blocks of atomic systems using machine learning: application to grain boundaries, *NPJ Comput. Mater.* **2017**, *3*, 1–7.
10. Helfrecht, B. A.; Semino, R.; Pireddu, G.; Auerbach, S. M.; Ceriotti, M. A new kind of atlas of zeolite building blocks, *J. Chem. Phys.* **2019**, *151*, 154112.

- 550 11. Moliner, M.; Romágn-Leshkov, Y.; Corma, A. Machine learning applied to zeolite synthesis: The
551 missing link for realizing high-throughput discovery, *Acc. Chem. Res.* **2019**, *52*, 2971–2980.
- 552 12. Earl, D. J.; Deem, M. W. Toward a database of hypothetical zeolite structures, *Ind. Eng. Chem.*
553 *Res.* **2006**, *45*, 5449–5454.
- 554 13. Deem, M. W.; Pophale, R.; Cheeseman, P. A.; Earl, D. J. Computational discovery of new
555 zeolite-like materials, *J. Phys. Chem. C* **2009**, *113*, 21353–21360.
- 556 14. Pophale, R.; Cheeseman, P. A.; Deem, M. W. A database of new zeolite-like materials, *Phys.*
557 *Chem. Chem. Phys.* **2011**, *13*, 12407–12412.
- 558 15. Blatov, V.; O’keeffe, M.; Proserpio, D. Vertex-, face-, point-, schläfli-, and delaney-symbols in
559 nets, polyhedra and tilings: recommended terminology, *CrystEngComm* **2010**, *12*, 44–48.
- 560 16. Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.;
561 Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; others. Inverse design of nanoporous crystalline
562 reticular materials with deep generative models, *Nature Mach. Int.* **2021**, pages 1–11.
- 563 17. Li, Y.; Yu, J. New stories of zeolite structures: their descriptions, determinations, predictions,
564 and evaluations, *Chem. Rev.* **2014**, *114*, 7268–7316.
- 565 18. Weininger, D. SMILES, a chemical language and information system. 1. introduction to
566 methodology and encoding rules, *J. Chem. Inf. Model.* **1988**, *28*, 31–36.
- 567 19. Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and tools for
568 high-throughput geometry-based analysis of crystalline porous materials, *Microporous Mesoporous*
569 *Mater.* **2012**, *149*, 134–141.
- 570 20. Blatov, V. Nanocluster analysis of intermetallic structures with the program package topas, *Struct.*
571 *Chem.* **2012**, *23*, 955–963.
- 572 21. Blatov, V. A.; Ilyushin, G. D.; Proserpio, D. M. The zeolite conundrum: why are there so many
573 hypothetical zeolites and so few observed? a possible answer from the zeolite-type frameworks
574 perceived as packings of tiles, *Chem. Mater.* **2013**, *25*, 412–424.
- 575 22. First, E. L.; Gounaris, C. E.; Wei, J.; Floudas, C. A. Computational characterization of zeolite
576 porous networks: an automated approach, *Phys. Chem. Chem. Phys.* **2011**, *13*, 17339–17358.
- 577 23. Sato, M. Hamiltonian graph representation of zeolite frameworks and Si, Al ordering in the
578 framework, *J. Math. Chem.* **1991**, *7*, 341–352.
- 579 24. Delgado-Friedrichs, O.; Hyde, S. T.; O’Keeffe, M.; Yaghi, O. M. Crystal structures as periodic
580 graphs: The topological genome and graph databases, *Struct. Chem.* **2017**, *28*, 39–44.
- 581 25. Witman, M.; Ling, S.; Boyd, P.; Barthel, S.; Haranczyk, M.; Slater, B.; Smit, B. Cutting materials
582 in half: a graph theory approach for generating crystal surfaces and its prediction of 2D zeolites,
583 *ACS Cent. Sci.* **2018**, *4*, 235–245.

- 584 26. Boyd, P. G.; Woo, T. K. A generalized method for constructing hypothetical nanoporous materials
585 of any net topology from graph theory, *CrystEngComm* **2016**, *18*, 3777–3792.
- 586 27. Blatov, V. A.; Delgado-Friedrichs, O.; O’Keeffe, M.; Proserpio, D. M. Three-periodic nets and
587 tilings: natural tilings for nets, *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2007**, *63*, 418–425.
- 588 28. Lach-hab, M.; Yang, S.; Vaisman, I.; Blaisten-Barojas, E. Novel approach for clustering zeolite
589 crystal structures, *Mol. Inf.* **2010**, *29*, 297–301.
- 590 29. Foster, M.; Rivin, I.; Treacy, M.; Friedrichs, O. D. A geometric solution to the largest-free-sphere
591 problem in zeolite frameworks, *Microporous Mesoporous Mater.* **2006**, *90*, 32–38.
- 592 30. Haldoupis, E.; Nair, S.; Sholl, D. S. Pore size analysis of > 250000 hypothetical zeolites, *Phys.*
593 *Chem. Chem. Phys.* **2011**, *13*, 5053–5060.
- 594 31. Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments, *Phys. Rev. B* **2013**,
595 *87*, 184115.
- 596 32. Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q.
597 Large-scale screening of hypothetical metal–organic frameworks, *Nat. Chem.* **2012**, *4*, 83.
- 598 33. Martin, R. L.; Smit, B.; Haranczyk, M. Addressing challenges of identifying geometrically diverse
599 sets of crystalline porous materials, *J. Chem. Inf. Model.* **2012**, *52*, 308–318.
- 600 34. Krishnapriyan, A. S.; Haranczyk, M.; Morozov, D. Topological descriptors help predict guest
601 adsorption in nanoporous materials, *J. Phys. Chem. C* **2020**, *124*, 9360–9368.
- 602 35. Park, C. W.; Wolverton, C. Developing an improved crystal graph convolutional neural network
603 framework for accelerated materials discovery, *Phys. Rev. Mater.* **2020**, *4*, 063801.
- 604 36. Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and
605 interpretable prediction of material properties, *Phys. Rev. Lett.* **2018**, *120*, 145301.
- 606 37. Ryan, K.; Lengyel, J.; Shatruk, M. Crystal structure prediction via deep learning, *J. Am. Chem.*
607 *Soc.* **2018**, *140*, 10158–10168.
- 608 38. Collins, S. P.; Daff, T. D.; Piotrkowski, S. S.; Woo, T. K. Materials design by evolutionary
609 optimization of functional groups in metal-organic frameworks, *Sci. Adv.* **2016**, *2*, e1600954.
- 610 39. Dei, V. G.; Klinz, B.; Woeginger, G. J. Exact algorithms for the hamiltonian cycle problem in
611 planar graphs, *Oper. Res. Lett.* **2006**, *34*, 269–274.
- 612 40. Deogun, J. S.; Steiner, G. Polynomial algorithms for hamiltonian cycle in cocomparability graphs,
613 *SIAM Journal on Computing* **1994**, *23*, 520–552.
- 614 41. Seeja, K. Hybridham: A novel hybrid heuristic for finding hamiltonian cycle, *J. Optim.* **2018**,
615 *2018*.

- 616 42. Gould, R. J. Advances on the hamiltonian problem—a survey, *Graphs and Combinatorics* **2003**,
617 19, 7–52.
- 618 43. Martello, S. Algorithm 595: An enumerative algorithm for finding hamiltonian circuits in a directed
619 graph, *ACM Trans. Math. Software* **1983**, 9, 131–138.
- 620 44. Rubin, F. A search procedure for hamilton paths and circuits, *J. Assoc. Comp. Machinery* **1974**,
621 21, 576–580.
- 622 45. Diaby, M. The traveling salesman problem: a linear programming formulation, *WSEAS, Transac.*
623 *Math.* **2006**, 6.
- 624 46. Orman, A.; Williams, H. P. A survey of different integer programming formulations of the travelling
625 salesman problem, *Optim., Econ. Financ. Anal.* **2006**, 9, 93–108.
- 626 47. Herstein, I. N. *Topics in Algebra*; John Wiley & Sons, 2006.
- 627 48. Atlas of prospective zeolite structures <http://www.hypotheticalzeolites.net>, (accessed Jan 9, 2021).
- 628 49. Loewenstein, W. The distribution of aluminum in the tetrahedra of silicates and aluminates,
629 *American Mineralogist: Journal of Earth and Planetary Materials* **1954**, 39, 92–96.
- 630 50. Fletcher, R. E.; Ling, S.; Slater, B. Violations of löwenstein’s rule in zeolites, *Chem. Sci.* **2017**, 8,
631 7483–7491.

