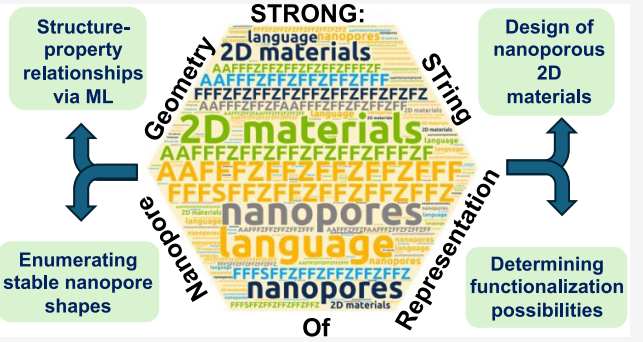
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Machine Learnable Language for the Chemical Space of Nanopores Enables Structure–Property Relationships in Nanoporous 2D Materials

Piyush Sharma, Sneha Thomas, Mahika Nair, and Ananth Govind Rajan*

Cite This: https://doi.org/10.1021/jacs.4c08282
Read Online

ABSTRACT: The synthesis of nanoporous two-dimensional (2D) materials has revolutionized fields such as membrane separations, DNA sequencing, and osmotic power harvesting. Nanopores in 2D materials significantly modulate their optoelectronic, magnetic, and barrier properties. However, the large number of possible nanopore isomers makes their study onerous, while the lack of machine-learnable representations stymies progress toward structure–property relationships. Here, we develop a language for nanopores in 2D materials, called STring Representation Of Nanopore Geometry (STRONG), that opens the field of 2D nanopore informatics. We show that STRONGs are naturally suited for machine learning via recurrent neural networks, predicting formation energies/times of arbitrary nanopores and transport barriers for CO_2, N_2, and O_2 gas molecules, enabling structure–property relationships. The machine learning models enable the discovery of specific nanopore topologies to separate CO_2/N_2, O_2/CO_2, and O_2/N_2 gas mixtures with high selectivity ratios. We also enable the rapid enumeration of unique configurations of stable, functionalized nanopores in 2D materials via STRONGs, allowing systematic searching of the vast chemical space of nanopores. Using the STRONGs approach, we find that a mix of hydrogen and quinone functionalization results in the most stable functionalized nanopore configuration in graphene, a discovery made feasible by expedited chemical space exploration. Additionally, we also unravel the STRONGs approach as ∼1000 times faster than graph theory algorithms to distinguish nanopore shapes. These advances in the language-based representation of 2D nanopores will accelerate the tailored design of nanoporous materials.



INTRODUCTION

Over the last two decades, atomically thin materials that were previously only simulated on computers have come into being.[1] Two-dimensional (2D) materials, including graphene,[2] boron nitride,[3] and transition metal dichalcogenides,[4] have revolutionized various fields. Of note to sustainability, clean water/air, and biological applications, nanoporous 2D materials containing tiny holes, or nanopores, have allowed the atomic-level sieving of atoms, ions, and biomolecules.[5−8] Today, separation processes account for ∼60−70% of the total energy consumption in modern industries.[9,10] Separations in industries are currently carried out by distillation, absorption, or solvent extraction, which are energy-intensive processes, thus leading to a high carbon footprint due to greenhouse gas emissions.[11] The reduced energy requirement, mechanical robustness, and high throughput are some of the benefits of 2D membrane-based separation processes.[12] Additionally, water desalination,[13−15] gas separation,[16−20] and DNA sequencing[21,22] technologies developed using 2D materials have demonstrated the vital role that nanoporous 2D materials can play in our future. In all these applications, extended vacancy defects, i.e., nanopores, significantly modulate the physicochemical[23] and optoelectronic[24] properties of 2D materials. However, to date, 2D nanopores lack a language that is as human-readable as it is machine-interpretable. In this work, we address this significant knowledge gap, thus enabling the development of structure–property relationships for nanoporous 2D materials. We invent a language framework called STring Representation Of Nanopore Geometry (STRONG), that is not only interpretable by humans but also readily learned by computers. We show that this language has several exceptional capabilities impacting the enumeration of nanopores, nanopore discovery, and the determination of stable functional groups at nanopore edges.

Received: June 19, 2024
Revised: October 9, 2024
Accepted: October 10, 2024

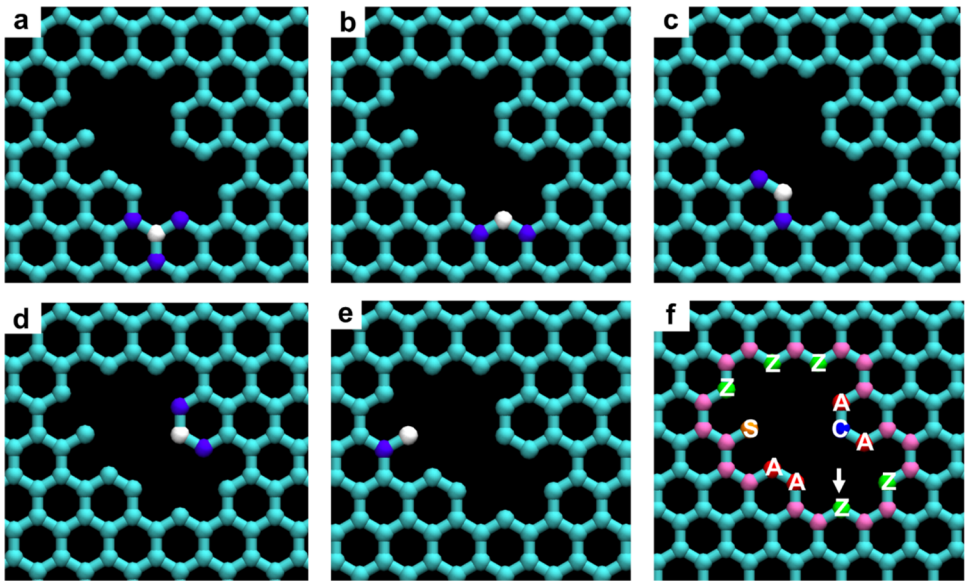


Figure 1. Construction of a STRONG for a nanopore in a 2D hexagonal lattice. (a–e) First, the types of atoms at the nanopore rim are identified: (a) a representative fully bonded atom (in white) having 3 neighbors (blue), (b) a representative zigzag atom (in white) having 2 neighbors (in blue), each of which possesses 3 neighbors, (c) a representative armchair atom (in white) having 2 neighbors (in blue), one of which possesses 3 neighbors, and the other possesses less than 3 neighbors, (d) a representative corner atom (in white) having 2 neighbors (in blue), both of which possess less than 3 neighbors, and (e) a representative singly bonded atom (in white) and its only neighbor (in blue). (f) Subsequently, the nanopore rim atoms (F in magenta, Z in green, A in red, C in blue, and S in orange) are traversed anticlockwise, starting from the atom indicated by the white arrow, to generate the corresponding STRONG.

Pristine 2D materials are impermeable even to the smallest gases, such as hydrogen and helium.[25] Hence, nanosized pores are introduced into the material using techniques such as electron/ion-beam drilling,[26] ozone/oxygen etching,[27] and on-surface chemical reactions,[28,29] naturally resulting in a wide shape and size distribution for the nanopores.[30−33] Additionally, these techniques can introduce various functional groups at the nanopore mouth, significantly impacting the permeability and selectivity of 2D membranes. As the size of a 2D nanopore (quantified by the number of atoms etched from the lattice) increases, the number of possible shapes that can exist increases exponentially.[30] Previous work by Govind Rajan et al. catalogued the unique, most-probable shapes of nanopore isomers in graphene by combining first-principles calculations, stochastic simulations, and chemical graph theory.[30] Recently, Sheshanarayana and Govind Rajan advanced a machine learning (ML) framework to predict formation times and probabilities of arbitrary nanopore shapes in graphene based on various structural features of nanopores.[34] Later, Thomas et al. used the mathematical combinatorics of polyforms to generate a data set of all stable nanopores in graphene.[35] While these studies have greatly enhanced our ability to study hundreds of thousands of nanopores in 2D materials,[36−38] the present work focuses on the use of strings to make tractable the consideration of the massive number of nanopores—to the tune of 11.7 million (as predicted using the combinatorics of polyforms)[35]—that can exist in 2D materials. Indeed, the use of computerized representations and ML for materials design and discovery is promising. For example, using ML, Sorkun et al. discovered functional 2D materials for energy conversion and storage applications.[39] In another study, Xie et al. enabled the controlled synthesis of metal−organic nanocapsules via inferences made using ML techniques.[40] Very recently, Bhattacharya et al. utilized large language models (LLMs) to develop molecular design engines.[41] Accordingly, inspired by the Simplified Molecular-Input Line-Entry System (SMILES) framework for representing organic molecules as strings,[42,43] we report the ML-amenable STRONGs framework for nanoporous structures. The advances made herein have implications for the tailored design of nanoporous 2D materials using ML strategies[44−46] such as reinforcement learning,[47] as reviewed recently by Farimani and colleagues.[48]

RESULTS AND DISCUSSION

Concept of a STRONG. In Figure 1, we introduce the concept of STRONG. Just like letters form the basis of words in any language, in the STRONGs framework, different types of atoms in a 2D hexagonal lattice—fully bonded (F), zigzag (Z), armchair (A), corner (C), and singly bonded (S)—are used to construct the representation of a nanopore. The distinction between these types of atoms can be understood from Figure 1a–e.

As explained therein, an S atom forms only one bond with the nanopore rim, whereas an F atom has three bonds with surrounding atoms. Z, A, and C atoms all form two bonds, but while both neighbors of a Z atom have three neighbors, one neighbor of an A atom has two neighbors, while the other has three neighbors, and both neighbors of a C atom have two neighbors. We note that the unrelaxed structure of a nanopore is used to write the corresponding STRONG. *This is not an issue, because each unrelaxed nanopore structure in a 2D material will correspond to a unique geometry-optimized structure.* In Figure 1f, we show how one can traverse the rim of a nanopore, starting at a particular choice of the first atom, and generate the corresponding STRONG of a nanopore (see the Supporting Information Section 1 for the algorithm to generate STRONGs), i.e., ZFFZFFFACAFFFZFZFFZFFFSFFAAFF. Note that all cyclic permutations of a given STRONG are equivalent, as the choice of the starting atom in a STRONG is immaterial. Thus, we canonicalize all STRONGs by choosing

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

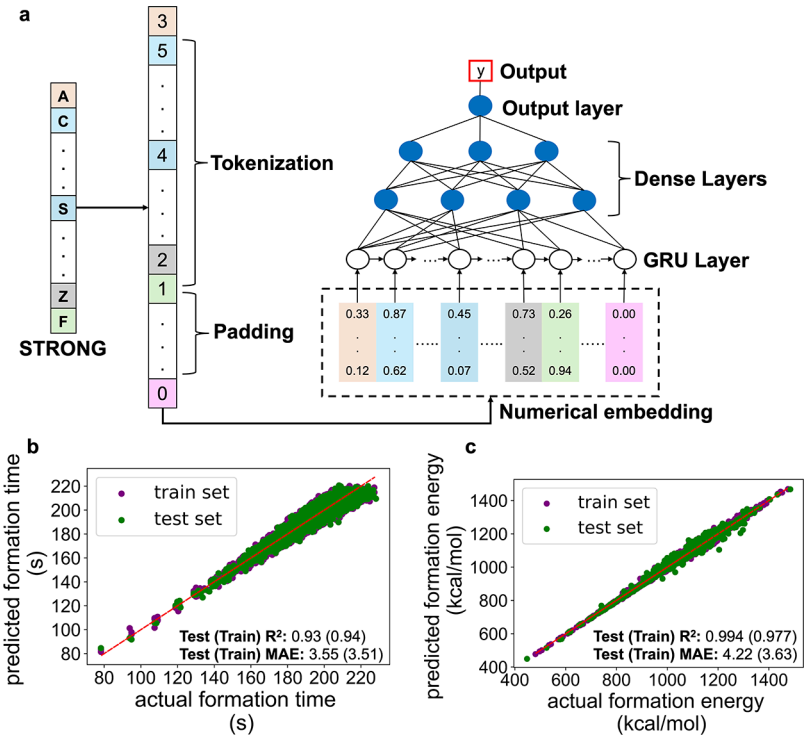


Figure 2. Machine learning (ML) framework using natural language processing (NLP) and prediction of nanopore properties using STRONGs. (a) Recurrent neural network-based (RNN) architecture. The tokenization and numerical embeddings shown are representative. (b) Parity plot for the prediction of the formation time in seconds where the actual formation time is obtained from KMC simulations, and the predicted formation time is from the ML model. (c) Parity plot for the prediction of the formation energy in kcal/mol where the actual formation energy is obtained from molecular mechanics calculations, and the predicted formation energy is from the ML model.

the particular cyclic or reflectional permutation that would appear first in an English-language dictionary. In other words, the canonical form of a STRONG is obtained by sorting all its cyclic/reflectional permutations in alphabetical order and picking the first entry. For instance, the canonical STRONG for the nanopore depicted in Figure 1 is AAFFFSFFZFFZFZFFFACAFFFZFFZFF. This canonical STRONG corresponds to the nanopore formed when the nanopore shown in Figure 1f is reflected in the vertical ($x = 0$) plane, as shown in the Supporting Information Section 2. Before proceeding further, we also demonstrate the extension of the STRONGs approach to other nonhexagonal materials, such as borophene, and quasi-2D materials, such as molybdenum disulfide (MoS_2), in the Supporting Information Section 3.

Advancing Structure–Property Relationships Using Recurrent Neural Network Representations of STRONGs. Researchers working in machine learning and artificial intelligence (AI) have made significant strides in natural language processing[49] (NLP) over the last few decades. Since STRONGs are a language-based representation, they are particularly amenable to using tools such as recurrent neural networks (RNNs)[50] routinely used for NLP. Accordingly, we developed an RNN framework that can capture structure–property relationships for nanopores in 2D materials. Specifically, we focused on creating an ML framework that can predict various experimentally and theoretically valuable metrics—the energy of formation of a nanopore, the time taken to form a nanopore during electron-beam etching,[30,34,51] and the barrier for a CO_2, N_2, and O_2 gas molecule to transport through a nanopore—rapidly and accurately. The methodology to generate the data set for the above properties is outlined in the Methods section. Briefly, kinetic Monte Carlo (KMC) simulations[34,52,53] supported by chemical graph theory[30] are used to obtain the formation times of various nanopores, and molecular mechanics calculations based on accurate reactive/nonreactive force fields are used to determine the energy of formation of nanopores and the barriers for a CO_2, N_2, and O_2 gas molecule passing through hydrogen-functionalized pores. The RNN-based architecture begins with STRONG preprocessing, wherein the characters in the STRONG are tokenized and postpadded with zeros to ensure a consistent length among the tokenized STRONGs. We used TensorFlow[54] for building the RNN-based architecture and a Keras[55] custom tokenizer for tokenization. Each tokenized character is then represented using a numerical embedding, creating dense vectors that capture the semantic and structural information on the nanopore topology via the STRONG. In addition, for predicting the gas transport barrier, the radius of the nanopore in Å multiplied by 100 (to account for decimal variations in radius) is prefixed to each STRONG. The radius is obtained using the concept of a Voronoi diagram[56] by finding the largest circle that can be inscribed inside the unrelaxed nanopore structure. This methodology ensures that the gas molecule is placed at the best active site out of many available sites in a graphene nanopore for gas transport. In the future, the effect of relaxation on the radius of the active sites and the effect of multiple active sites in the nanoporous structures could be explored. The resultant sequential embedded vectors form the input sequence for the RNN. We used gated recurrent units[57] (GRUs) to capture the sequential dependencies within the input. In addition to a

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

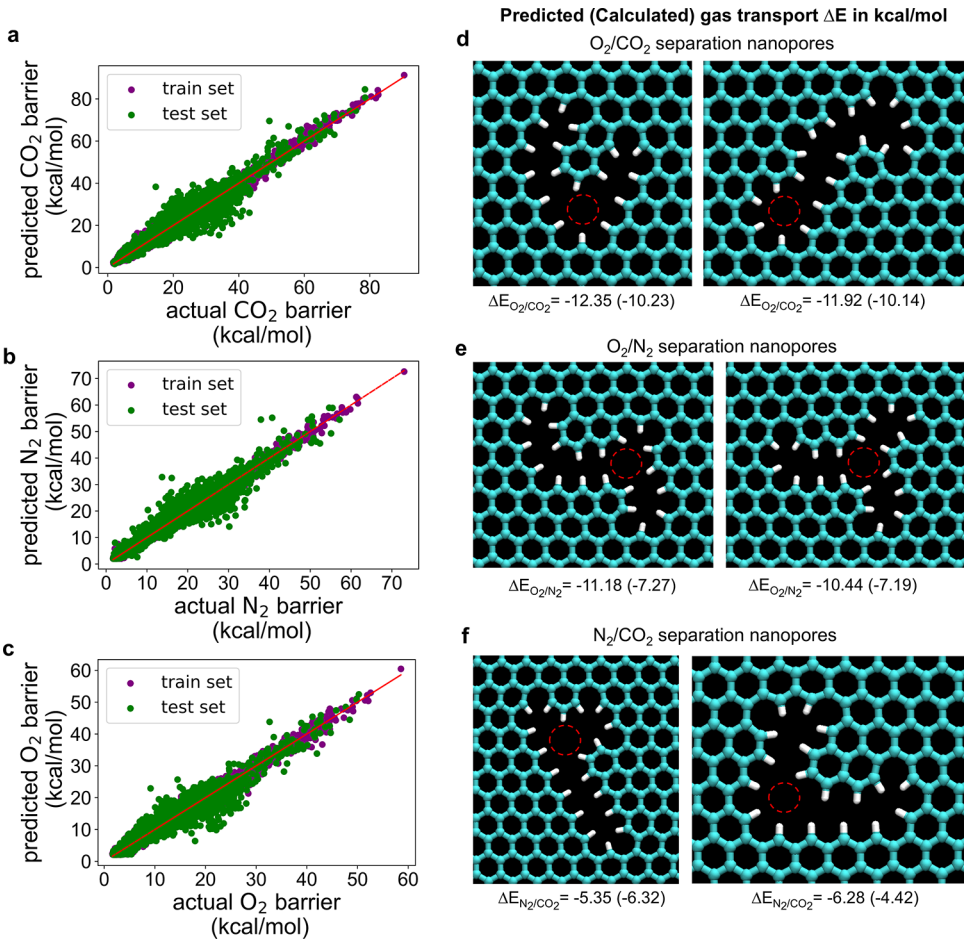


Figure 3. Application of the ML framework to discover candidate nanopore structures to separate O_2/CO_2, O_2/N_2, and N_2/CO_2 gas mixtures. (a–c) Parity plot for the prediction of the CO_2 (a), N_2 (b), and O_2 (c) gas transport barrier in kcal/mol. In each case, the actual gas transport barrier is obtained from molecular mechanics calculations, and the predicted gas transport barrier from the ML model. (d–f) Structures of the best two nanoporous candidates to separate an O_2/CO_2 (d), O_2/N_2 (e), and N_2/CO_2 (f) gas mixture and the difference in the potential energy barrier for transport $\Delta E_{i/j} = E_i - E_j$. Values in parentheses denote the actual barrier differences in kcal/mol, while values outside parentheses indicate ML-predicted barrier differences in kcal/mol. The dashed red circle in panels (d–f) represents the passage for the gas molecules.

GRU layer, we incorporated extra dense layers after the GRU layer to learn more intricate representations and enhance the model's capability to comprehend complex patterns. The STRONG preprocessing and RNN architecture for regression problems is pictorially represented in Figure 2a.

We obtained the optimized RNN-based architecture by performing hyperparameter tuning as discussed in the Methods section. The model's exceptional performance in predicting an arbitrary nanopore's formation time and energy is pictorially represented by the parity plot and train and test metrics in Figure 2b,c, respectively. We obtained *test* R^2 values of 0.93 for formation time prediction and 0.994 for formation energy prediction, with mean absolute errors (MAEs) as low as 3.55 s and 4.22 kcal/mol, respectively. This demonstrates the promise and potential of using RNNs with STRONGs to develop structure–property relationships for nanoporous graphene. As the spread in the parity plots is larger toward the right, we also provide performance metrics by considering solely the data points with longer formation time and larger formation energy in the Supporting Information Section 4. We observe that the model performance remains similar to the model trained with the complete data set. In fact, even training the ML models with solely stable pores (as opposed to pores also including singly bonded atoms) does not deteriorate the model performance, as shown also in the Supporting Information Section 4.

Moving forward, we trained ML models to predict the barrier for a CO_2, N_2, and O_2 molecule passing through an arbitrary stable nanopore. According to Thomas et al., nanopores with no dangling atoms, bonds, or moieties are termed stable nanopores.[35] The data set underlying this ML model was generated using molecular mechanics calculations, as described before (see also the Methods section). The performance of these ML models in predicting the gas transport barrier for CO_2, N_2, and O_2 is pictorially represented by the parity plots shown in Figure 3a–c, respectively. We observed excellent performance of the models, quantified by both train and test metrics, as presented in Table 1. Therein, one sees that the MAE for each gas is lower than 2 kcal/mol. Leveraging the accuracy of this model, we applied it to screen through a comprehensive database of 188,178 stable nanoporous topologies comprising of nanopores formed by removing $N = 3$ to $N = 19$ atoms from the graphene lattice, as enumerated using STRONGs-based rules (see below). We used a criterion of $E < 15$ kcal/mol for the energy barrier for the faster transporting molecule and $|\Delta E| > 10$ kcal/mol for energy difference between the transport barriers of the gases in a gas pair to consider a pore to be a viable candidate, where E

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX–XXX

Table 1. Performance of the Proposed RNN-Based Architecture for the Prediction of CO_2, N_2, and O_2 Gas Transport Barriers Using STRONGs in Terms of the Obtained R^2 Scores and MAE[a]

	CO_2 barrier prediction		N_2 barrier prediction		O_2 barrier prediction	
metric	training set (five folds)	test set	training set (five folds)	test set	training set (five folds)	test set
R^2	0.98	0.95	0.99	0.96	0.98	0.95
MAE	0.95	1.52	0.65	1.18	0.64	1.04

[a]Values for the training set correspond to averages over five validation folds. The unit for the gas transport barrier is kcal/mol.

denotes the energy barrier for gas transport (except for N_2/CO_2, wherein we could only find pores with $|\Delta E| > 5$ kcal/mol). The former criterion translates to a minimum room-temperature transport rate (r) of ~62.7 successful attempts molecule^{-1} s^{-1} pore^{-1} assuming transition-state theory as $r = \frac{k_B T}{h}\exp\left(-\frac{E}{RT}\right)$, where k_B is Boltzmann constant, T is the absolute system temperature (considered to be 298.15 K here), R is the universal gas constant, and h is Planck's constant. The latter criterion corresponds to a minimum selectivity ratio of 2.14×10^7, calculated as $\exp\left(-\frac{\Delta E}{RT}\right)$, where $\Delta E < 0$ represents the difference in the molar transport barrier between the two considered gases in a pair. The ML model trained to predict gas transport barriers led to the discovery of 6342 stable nanopores for separating a O_2/CO_2 gas mixture, 1755 stable nanopores for separating an O_2/N_2 gas mixture, and 1050 stable nanopores for separating a N_2/CO_2 gas mixture. We then carried out molecular mechanics calculations for these nanopores, finding that the best two candidate nanopores provide a selectivity of more than 10^7 for O_2/CO_2 separation, 10^5 for O_2/N_2 separation, and 10^3 for N_2/CO_2 separation. Note that these are idealized predictions based on molecular mechanics calculations. In the future, detailed molecular dynamics simulations considering molecular translation, rotation, and interactions with the pore could be carried out to obtain the actual gas transport rates and selectivities, which could be lower. Indeed, experimental selectivities of the order

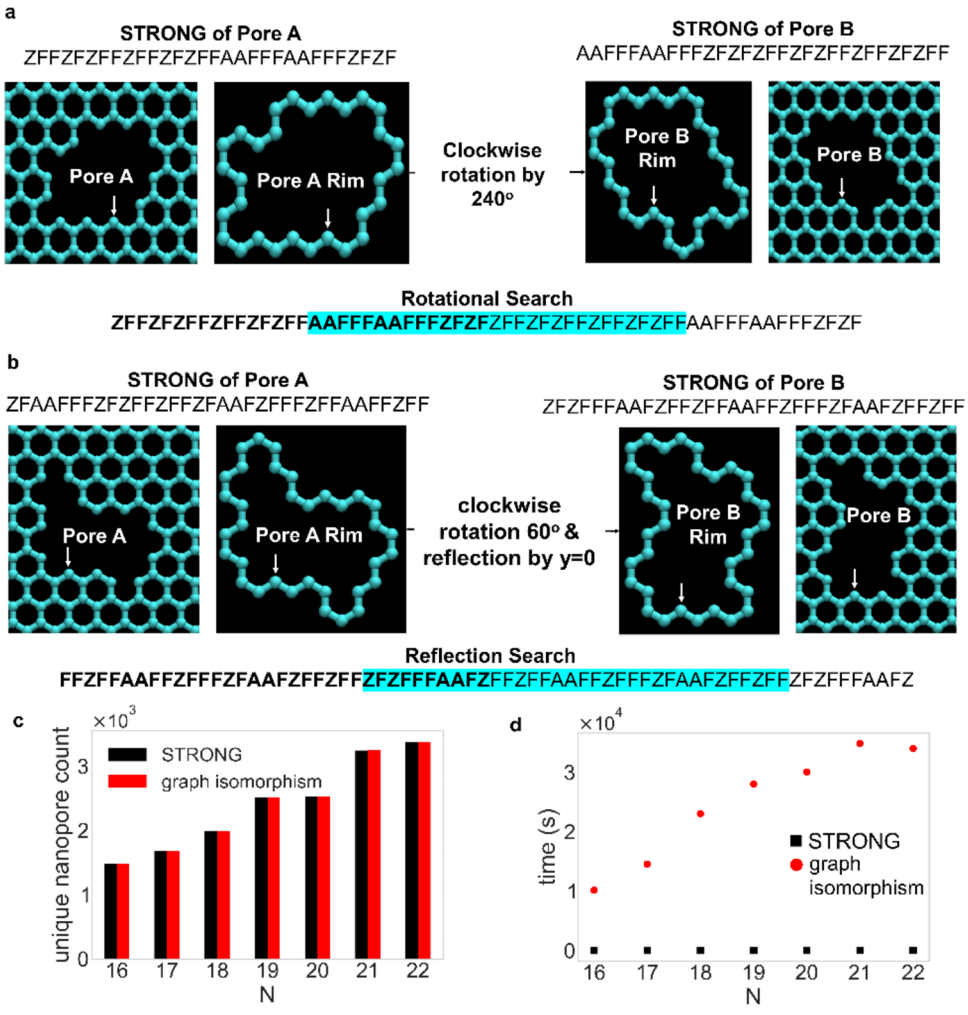


Figure 4. STRONGs algorithm for comparing nanopores using rotational and reflectional symmetry operation, and its evaluation for effectiveness and efficiency with respect to the graph isomorphism algorithm employed on orientation-augmented adjacency matrices.[30] (a) STRONGs comparison algorithm using rotational symmetry operation. (b) STRONGs comparison algorithm using reflectional symmetry operation. The highlighted text depicts the occurrence of the second STRONG in the first STRONG concatenated with itself without (a) and after (b) reversing the first STRONG. (c) Comparison of the STRONGs algorithm with graph isomorphism for accuracy. (d) Comparison of the STRONGs algorithm with graph isomorphism for algorithmic efficiency where N represents the number of atoms removed from the graphene lattice.

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

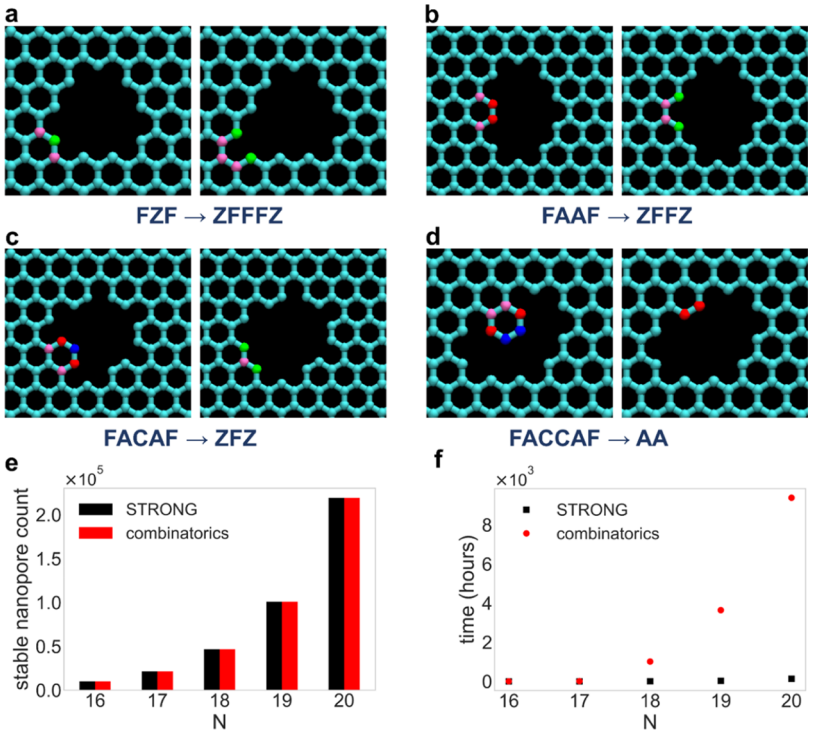


Figure 5. STRONGs algorithm for enumerating stable nanopores using human-knowledge-based rules. (a–d) Implementation of various substring substitutions to generate larger nanopores. (a) FZF to ZFFFZ substitution in a $N = 13$ nanopore. (b) FAAF to ZFFZ substitution in a $N = 14$ nanopore. (c) FACAF to ZFZ substitution in a $N = 13$ nanopore. (d) FACCAF to AA substitution in a $N = 10$ nanopore. (e, f) Comparison of the STRONGs algorithm with the combinatorics-based algorithm[35] in terms of accuracy (e) and computational efficiency (f) for the enumeration of stable nanopores.

of 10^1–10^2 have been observed for N_2/CO_2, O_2/N_2, and O_2/CO_2 gas separation.[5,27] On a related note, our ML framework could also be used to reverse engineer a digital twin model for nanoporous 2D membranes based on experimental data and generative AI tools. It is also important to note that the effect of functionalization and associated transport mechanisms for gas molecules (e.g., selective binding of gas molecules with non-H functional groups) could be further explored in the future.

Note that the computing resources required for molecular mechanics calculations are 4 orders of magnitude higher than what is needed for the machine learning model, thus indicating that it would have been infeasible to carry out molecular mechanics calculations for all the 188,178 stable nanopores to screen promising ones for gas separation. The structures of the two best candidate pores discovered in each case are depicted in Figure 3d–f, respectively, along with the ΔE values determined by the ML model and molecular mechanics calculations. We found a significant correlation between the performance of a pore and its active site's radius and shape. For a gas mixture of O_2/CO_2, we observed that a radius of 1.76 Å for the active site and a spherical region around the active site, as shown in Figure S5, results in the best candidate nanopores to separate the gas mixture. Similarly, we observed that an active site radius of 1.76 Å and a spherical region around the active site, and an active site radius of 1.97 Å with a triangular region around it result in the best nanopores to separate an N_2/CO_2 gas mixture as also shown in Figure S5. We discuss the physical characteristics of the candidate nanopores in detail in the Supporting Information Section 5. A histogram of the calculated difference in gas transport barriers for the predicted nanopores has been plotted in the Supporting Information Section 6. It is noteworthy that, in Figure 3, one sees reasonable agreement (within the limits of the mean absolute errors in Table 1) between the predicted and the actual ΔE values for O_2/CO_2 and N_2/CO_2 gas mixtures, indicating the value of the proposed approach in nanopore discovery for gas separation applications.

Enabling Rapid Comparison of Nanopores in 2D Materials. Not only does the STRONGs approach present notable advantages for ML models, but it also enabled us to create the fastest algorithms presently available for nanopore topology identification.[30,35] Before the current work, nanopores in 2D materials were primarily differentiated using graph-theoretic approaches based on orientation-augmented adjacency matrices.[30] The creation of such matrices is itself a challenge apart from the significant computational bottleneck in applying rotational and reflectional symmetries that exist in 2D material lattices to such representations. In contrast, STRONGs are naturally suited for rapidly differentiating between nanopore shapes. Indeed, the choice of the starting character in a STRONG is immaterial, such that the canonical form of a STRONG will not change even if a pore is rotated by any angle in the lattice. Accordingly, to determine if two nanopores, A and B, are equivalent upon considering rotation operations, we simply check if their (noncanonical) STRONGs are cyclic permutations of each other. This task is accomplished by concatenating the STRONG of one of the nanopores with itself to create an extended STRONG. Subsequently, if the STRONG of the other nanopore is a substring in the extended STRONG, both nanopores are rotationally equivalent. This is because, by simply changing the starting atom, the two STRONGs became identical. Figure 4a provides a visual illustration of this process.

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

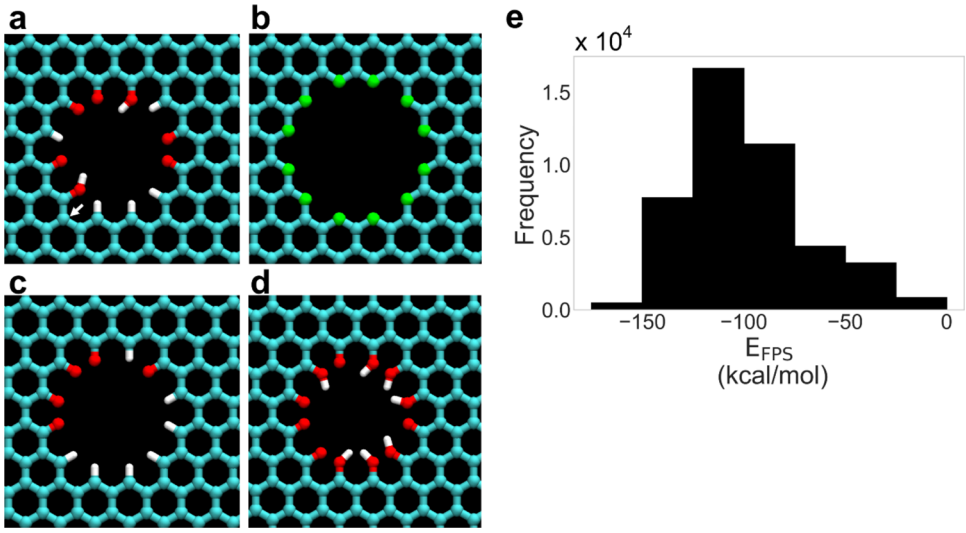


Figure 6. Applications of the STRONGs algorithm for determining unique functionalized nanopore topologies. (a) An example functionalized nanopore configuration with three functional groups—H, OH, and O with the STRONG as indicated in the main text, starting with the white arrow and traversed anticlockwise. (b) N = 24 nanopore containing 12 functional sites (in green). (c, d) The most stable (c) and unstable (d) functionalized nanopore structure out of 44727 functionalized configurations for the nanopore structure shown in panel b. (e) Histogram of the energy of functionalization per functional site (E_{FPS}) for the 44727 functionalized configurations obtained for an N = 24 bare nanopore containing 12 functional sites and 3 functional groups.

Furthermore, to examine the reflectional equivalence of two nanopores, we simply reverse the STRONG of one of the nanopores before comparison with the other STRONG (with a minor modification in some cases; see the Supporting Information Section 7). Figure 4b provides a visual representation of this process. To evaluate the efficiency and effectiveness of the STRONGs algorithm in identifying unique nanopores, we compared the total count of unique nanopores and the computation time required using the graph isomorphism approach employed on orientation-augmented adjacency matrices with the respective values from the STRONGs approach. Figure 4c shows that the STRONGs approach and the graph isomorphism approach yield identical results for the number of isomers as a function of the nanopore size. Furthermore, as seen in Figure 4d, the STRONGs algorithm for comparing nanopores is 1000 times faster than the existing algorithm. We used an AMD Ryzer 5 5500U (1 core, 2.1 GHz) processor and 8 GB RAM with no parallelization to compare nanopores.

Using STRONGs to Enumerate Stable Nanopores and Functionalized Nanopore Topologies Speedily. Despite the significant progress made in comprehending and cataloguing nanopore configurations within graphene, there remains a notable gap in the existing body of knowledge on the enumeration of all feasible stable nanopore shapes and their functionalized configurations in a 2D lattice.[35] To bridge this gap and facilitate the systematic exploration of both *bare* and *functionalized* nanopore topologies, we departed from probabilistic approaches and employed newly developed principles of substring substitution in a STRONG along with the tools for rapidly comparing nanopores, as explained above. According to Thomas et al., stable nanopores are defined as nanopores with no dangling atoms, bonds, or moieties.[35] We adopted the same definition while employing the STRONG algorithm to generate stable nanopores. The algorithm for enumerating stable nanopores using STRONGs begins with a STRONG representing a single-atom vacancy nanopore, i.e., FFFZFFFZFFFZ. Subsequently, we systematically generate STRONGs for nanopores with higher vacancy counts by using new human-knowledge-based rules that we developed in this work (see the Supporting Information Section 8). These rules employ substring substitutions to remove one or more atoms from a STRONG. For example, the removal of a Z atom from the rim of a single-atom vacancy (with the STRONG FF**F**Z**F**FFZFFFZ) transforms the STRONG to FF**ZFFF**ZFFZFFFZ (the bold letters indicate the substring being substituted), replacing the Z atom being removed with two new Z atoms. Based on such rules, for enumerating nanopores with vacancy count N, we utilize STRONGs of nanopores with N-1 to N-4 atoms missing and remove Z, AA, ACA, and ACCA atoms, respectively. We have worked out the requisite STRONG substring substitutions to be FZF with ZFFFZ (Figure 5a), FAAF with ZFFZ (Figure 5b), FACAF with ZFZ (Figure 5c), and FACCAF with AA (Figure 5d) for N-1 to N-4 vacancy STRONGs, respectively. This process is illustrated pictorially in Figure 5a−d to generate STRONGs for larger nanopores from smaller nanopore STRONGs (see the Supporting Information Section 8 for other human-knowledge-based rules that ensure semantically valid STRONGs and remove STRONGs associated with unstable nanopores as defined by Thomas et al.[35]). Incredibly, implementation of our human-knowledge-based rules led to a seventy-fold speedup in stable nanopore generation in comparison to the previous combinatorics-based algorithm, even without the use of parallel computing. The accuracy and computational efficiency of the STRONGs algorithm for generating stable nanopores in contrast to the previous combinatorics-based algorithm are shown in Figure 5e,f, respectively. Note that we have also developed an algorithm to convert a given STRONG to its corresponding XYZ file, which may be further used in molecular simulations, first-principles calculations, or inverse-design approaches using ML in the future. This algorithm is explained in the Supporting Information Section 9.

Additionally, the STRONGs approach cleverly lends itself to predicting unique functionalized configurations of nanopores.

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

This advance is important because the attachment of functional groups at nanopore edges is known to allow tuning of the permeability and selectivity of nanoporous 2D materials for various separations,[58] but so far, there was no computational technique available to enumerate the possible functionalized configurations of nanopores. To achieve this, we developed the concept of a functionalized STRONG wherein functional group formulas are introduced within a STRONG. For example, the STRONG corresponding to the nanopore shown in Figure 6a is "FFZ (H) FZ (H) FFZ (H) FZ (O) FFZ (O) FZ (H) FFZ (OH) FZ (O) FFZ (O) FZ (H) FFZ (O) FZ (OH)" starting with the atom indicated by the white arrow and traversed anticlockwise, where parentheses enclose the functional groups present, and spaces are added to improve readability. Thus, STRONGs can enable navigating through the vast chemical space of functionalized nanopores, just as the SMILES approach enables for molecules.[59]

To generate all unique functionalized nanopores for an unfunctionalized nanopore, as shown in Figure 6b, given the number and type of functional groups, we generate all possible functionalized STRONGs, considering each A/C/Z atom at the nanopore rim to accommodate one functional group. Thereafter, nanopore comparison using STRONGs, as described above, leads to all possible unique, functionalized nanopore configurations. Note that the most common functional groups in nanoporous graphene are hydrogen (H), hydroxyl (OH), and quinone (O).[60−62] Accordingly, the total and unique number of functionalized nanopore configurations considering two and three functional groups are tabulated in Table 2. In this regard, the total number of possible functionalized configurations is n^m where n is the number of functional groups considered, and m is the number of sites available to be functionalized. Thus, corresponding to the nanopore in Figure 6b, the number of configurations considering 2 functional groups is 2^{12}, i.e., 4096, and considering 3 functional groups is 3^{12}, i.e., 531,441, as documented in Table 2.

Table 2. Number of Possible and Unique Functionalized Nanopore Configurations Based on the Implementation of Nanopore Comparison Using the STRONGs Algorithm on Functionalized STRONGs for an $N = 24$ Nanopore (with 12 Functional Sites) Shown in Figure 6b

number of functional groups	number of possible functionalized configurations	number of unique functionalized configurations
2	4096	382
3	531,441	44,727

However, many of these structures would correspond to the same configuration due to rotational and reflectional symmetries in the nanopore. The STRONGs algorithm allows us, for the first time, to determine these duplicate configurations and remove them, thus resulting in a much smaller number of unique functionalized nanopore topologies. For example, out of the 4096 functionalized nanopore configurations considering 2 functional groups, only 382 are unique, and in the 531441 possible configurations with 3 functional groups, only 44,727 are unique! This advance can prove to be instrumental in the systematic study of the transport of species, such as ions, water, gas molecules, and DNA, through functionalized graphene nanopores and in the development of structure−property relationships for functionalized 2D nanopores. Moving forward, to understand the energetics involved in obtaining various functionalized configurations, we utilized all-atom energy calculations, as described in the Methods section. Based on the energy of functionalization per functional site (E_{FPS}), we determined the most stable and unstable functionalized nanopore considering H, OH, and O functional groups having $E_{FPS} = -173.06$ kcal/mol and $E_{FPS} = -3.77$ kcal/mol for an $N = 24$ nanopore with 12 functional sites and 3 types of functional groups, as represented in Figure 6c,d, respectively. We also carried out geometry optimization on these structures using quantum-mechanical density functional theory (DFT) calculations, whose details are provided in the Methods section. The DFT-relaxed structures are shown in the Supporting Information Section 10. Interestingly, the most stable configuration contains 7 *hydrogen* groups and 5 quinone groups, whereas the most unstable configuration contains 7 *hydroxyl* groups and 5 quinone groups, a conclusion that could not have been reached without the massive reduction in computational cost enabled by STRONGs-based enumeration of unique functionalized nanopore topologies. The ability of quinone functional groups to withdraw electrons from the unfunctionalized graphene nanopores can explain how they impart stability to nanoporous graphene. Indeed, many studies have explored the oxidation of pristine graphene sheets, resulting in stable quinone structures.[63,64] However, the reaction mechanisms for these types of chemical functionalization on graphene nanopores must be further explored. On the other hand, the steric hindrance provided by the bulkier hydroxyl functional groups can explain their involvement in the unstable nanopores. Finally, Figure 6e shows the distribution of functionalized nanopores based on the energy of functionalization per functional site. Therein, we observe the spread in the energy of functionalization to be from about −175 kcal/mol per site to −3 kcal/mol per site. This observation further highlights the importance of our proposed approach, wherein not all functionalized configurations are equally stable, and one needs to determine the energetics of functionalization, considering various possible functionalized nanopore configurations. In the absence of STRONGs-based enumeration, the number of functionalized nanopore configurations one would have to consider to determine the spread in the functionalization energies, as well as the most-stable functionalized nanopore configuration would be 531,441, which marks an upper limit of an order of magnitude higher computational cost. More intelligent or random sampling could reduce the computational cost required for determining the spread in functionalization energies. Nevertheless, the STRONGs approach with machine learning could be valuable in the future to study the effect of nanopore functionalization on separation selectivity.

CONCLUSIONS

To conclude, in this work, we created a readable yet machine-interpretable language for representing nanopores in 2D materials. Given the plethora of applications of nanoporous 2D materials, such as DNA sequencing, seawater desalination, carbon capture, and osmotic power harvesting, the creation of such a language can significantly accelerate the nascent field of nanopore informatics and enable the development of structure−property relationships in nanoporous 2D materials. We call the new approach "STring Representation Of Nanopore Geometry" (STRONG). Although we demonstrate

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

the proposed approach using graphene as the prototypical 2D material, the method is readily extendable to other 2D materials, such as hexagonal boron nitride and molybdenum disulfide, by including the atom type (B, N, Mo, or S, for example) of the atoms on the nanopore rim, as part of the STRONG. We demonstrated the extension of the STRONGs approach for nonhexagonal and quasi-2D materials such as borophene and MoS_2, respectively.

The new approach lends itself to using ML-based RNN frameworks to accurately predict nanopore formation energies, formation times, and gas transport barriers, as demonstrated by excellent performance metrics, including the goodness of fit and mean absolute error, on unseen test sets. This will allow the creation of structure−property relationships for nanoporous 2D materials. The ML-based RNN frameworks, in principle, could be extended to other 2D materials and other functional groups apart from H functionalization, and this should be explored in the future. We also applied this novel approach to nanopore discovery, wherein we were able to identify shapes of graphene nanopores which can lead to high selectivity ratios for separating O_2 from CO_2, CO_2 from N_2, and O_2 from N_2. It is important to note here that the extension of the STRONGs approach for developing ML models for other nonhexagonal 2D materials, quasi-2D materials, various types of defects in 2D materials, and multiple layers of 2D materials will require the generation of new simulation data and can be readily investigated once such data is available.

Moreover, the proposed approach massively accelerates the traversal of the nanopore chemical space by allowing the rapid identification of a nanopore's structure in comparison to existing graph-theoretic approaches. We also provide algorithms to convert a nanopore structure to a STRONG and vice versa, which will find use in future ML-based exploration of nanopore structures. Generative AI and large-language models[65] using the STRONGs approach will form promising areas going forward, just as the inverse chemical design enabled by the SMILES framework for molecules.[66] Together with the possibility to speedily enumerate and determine stable nanopores, their unique functionalized configurations given a combination of various chemical functional groups, and the energetics of functionalization, we anticipate that our work will greatly advance the study of 2D nanopores and the design of nanoporous 2D materials for various application areas.

METHODS

Creating a Database of Formation Energies and Times of Graphene Nanopores. Generating a reliable data set is a significant step in developing a machine learning (ML) model. We used the KMC simulation[52] framework employed by Sheshanarayana and Govind Rajan[34] to generate the database of nanopores with various numbers of etched atoms N. Briefly, we employed Gillespie's algorithm[67] to stochastically simulate the time variation of a system wherein multiple events (i.e., the etching of armchair, zigzag, and singly bonded atoms; see the Supporting Information Section 11) can occur at varying rates, as specified in Govind Rajan et al.'s study.[30] A total of 10,000 KMC runs were carried out to generate the database for each value of $4 \leq N \leq 22$, and each simulation was carried out at 500 °C to mimic experimental conditions. Govind Rajan et al. proposed a graph theory approach to differentiate between various shapes or isomers of nanopores in graphene.[30] This approach uses orientation-augmented adjacency matrices to store (and compare) the structure of a nanopore (with another nanopore). The data set thus obtained consisted of 20,812 unique nanopore structures and their corresponding XYZ files. The formation time was the output of each KMC simulation. MATLAB R2021a was used for all the simulations, with each simulation beginning with a monatomic vacancy in graphene. Subsequently, we utilized the unique nanopore shapes generated from the KMC simulations and calculated each nanopore's formation energy. For this work, single-point energy calculations were carried out using the open-source Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) September 2021 package.[68] Interatomic force fields are the key to the accuracy of single-point energy calculations in a molecular mechanics framework; we used in this part of the study a reactive force field (ReaxFF) developed for carbon condensed phases.[69] We first relaxed the structures using the minimize command in LAMMPS with a stopping criteria for relative energy equal to 10^{-4} and force equal to 10^{-6} kcal/mol Å. Thereafter, we calculated the absolute energy of a nanoporous graphene (NPG; E_{NPG}) and pristine graphene sheet (G; E_G). The energy for nanopore formation (NF) is calculated as

$$E_{NF} = \left(E_{NPG} + \frac{N}{N_T}E_G\right) - E_G$$

where E with the appropriate subscript represents the single-point energy, N is the number of atoms etched to form the nanopore, and N_T is the total number of atoms constituting the pristine graphene sheet.

Calculating the Energy of Functionalization of a Graphene Nanopore. We used the reactive force field developed by Chenoweth et al.[70] to calculate the energy of functionalization of a graphene nanopore in the LAMMPS package. The energy of functionalization of a nanopore per functional site (FPS) was calculated as

$$E_{FPS} = \frac{E_{FGN} - E_{BGN} - N_H E_H - N_{OH}E_{OH} - N_O E_O}{N_{sites}}$$

where E_{FGN} denotes the potential energy of a functionalized graphene nanopore (FGN), E_{BGN} indicates the potential energy of the corresponding bare graphene nanopore (BGN), E_H, E_O, and E_{OH} denote, respectively, the potential energies of hydrogen, oxygen, and a hydroxyl group, N_H, N_O, and N_{OH} represent, respectively, the number of H, O, and OH functional groups at the nanopore rim, and N_{sites} indicates the number of sites available for functionalization at the nanopore rim. Note that E_O (E_H) was obtained as half the energy of an oxygen (hydrogen) molecule in vacuum, i.e., $E_O = \frac{1}{2}E_{O_2}$ and $E_H = \frac{1}{2}E_{H_2}$. The energy of a hydroxyl group was determined as $E_{OH} = \left(E_{H_2O} - \frac{E_{H_2}}{2}\right)$, where E_{H_2O} denotes the potential energy of an isolated water molecule.

Calculating the Activation Barriers for Gas Molecule Transport. To train our ML model for gas transport barrier prediction, we developed a database of energy barriers for CO_2, N_2, and O_2 molecule transport through various hydrogen-terminated stable graphene nanopores. The data set consisted of 24,884 stable graphene nanopores chosen randomly from the 188,178 stable nanopores generated using STRONGs-based enumeration (N = 3 to N = 19). The approach used for the energy barrier calculation was automated using shell scripting and is detailed in a schematic shown in the Supporting Information Section 12. The initial system consisted of a graphene nanopore and a gas molecule placed at the pore center (see the atomic schematic for an example N = 10 nanopore in the Supporting Information Section 12), prepared using Packmol[71] and depicted using the Visual Molecular Dynamics (VMD)[72] package. The box size considered in the simulations was 73.92 Å × 64.02 Å in the lateral directions (x and y) and 120 Å in the direction perpendicular to the membrane (z). The center of the nanopore (located at z = 0), where the transporting molecule was placed, was determined by finding the center of the largest circle that can be inscribed inside the unrelaxed nanopore structure using the concept of a Voronoi diagram.[56] The energy of the system was minimized using the LAMMPS package[73] (with stopping criteria for relative energy equal to 10^{-4} and force equal to 10^{-6} kcal/mol Å). Subsequently, the molecule was moved to a vertical distance of z = −1.2 nm while keeping the x and y coordinates fixed. Thereafter, the

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

molecule was iteratively shifted by 0.05 nm and placed at different positions along a straight line, and the system's potential energy was calculated. The last position where the molecule was placed is at z = +1.2 nm, toward the other side of the graphene membrane compared to where the molecule was initially placed. The compute group/group command in LAMMPS was used to calculate the potential energy of interaction between the nanopore and the transporting molecule each time, and the potential energy values were stored. Finally, the energy barrier was calculated as the maximum difference between consecutive minimum and maximum points in the potential energy profile. We compare the molecular mechanics method with first-principles DFT in calculating the transport barrier for the best nanopore identified to separate the O_2/CO_2 gas mixture in the Supporting Information Section 13. We also show in Supporting Information Section 14 that the most stable configuration is when a gas molecule transports perpendicular to the graphene nanopore.

To describe the potential energy of the system, we employed atomistic force fields, wherein bonded interactions involved harmonic bonds, harmonic angles, and dihedral functions, and nonbonded interactions were represented using 12−6 Lennard−Jones (LJ) parameters that captured van der Waals forces and electrostatic interactions via point-charge-based Coulombic potentials. The all-atom Optimized Potentials for Liquid Simulations (OPLS-AA)[74] force field was applied to model the interactions of carbon and hydrogen atoms situated at the periphery of the graphene nanopore, as well as to represent the bonded interactions within the graphene sheet. LJ parameters for graphene were adapted from the study by Cheng and Steele.[75] CO_2 was simulated using the Transferable Potential for Phase Equilibria (TraPPE)[76] model. The Elementary Physical Model 2 (EPM2)[77] was used to determine the bond stretching and bending force constants of CO_2. Note that Yuan et al.[16] employed a similar force field for CO_2, wherein they showed comparable results between the TraPPE and EPM2 force fields, indicating that the differences between the two force fields are minor. The usage of the EPM2 model for bonded interactions, as done in this work, is more accurate as it uses a flexible bond angle potential, whereas, in TraPPE, the bonds and angles are kept rigid. O_2 and N_2 were simulated using the new parameters in the interface force field (IFF).[78] A cutoff distance of 1.4 nm was employed for the LJ interactions. The particle−particle−particle mesh (PPPM) method[79,80] was utilized to capture long-range electrostatic interactions. Periodic boundary conditions were implemented in all spatial directions.

Training of the ML Models Involving STRONGs. We conducted hyperparameter tuning using the GridSearchCV module from the Python sklearn library.[81] This approach allowed us to optimize the model's performance and mitigate the risk of overfitting by employing 5-fold cross-validation on the training set. We followed the standard 75−25 train-test split to prevent poor predictive performance. A fixed seed value was used to ensure the reproducibility of the split. The hyperparameters that were fine-tuned using GridSearchCV included the embedding layer's dimension, the number of gated recurrent units, the number of hidden layers, the number of neurons in each hidden layer, and the activation function. For hyperparameter tuning, we utilized the coefficient of determination (R^2 score) as the evaluation metric. Subsequently, we retrained the model on the training set using the best hyperparameters obtained from GridSearchCV. Note that we used Python version 3.8 to run all the Python codes. The model's performance was then evaluated on the test set, which accounted for 25% of the data set. We used the quadratic loss function, also known as the MSE loss function, to train the ML model. Note that the mean-squared error (MSE) loss is defined as

$$\mathrm{MSE} = \frac{1}{P}\sum_{i=1}^{P}(y_i - \hat{y}_i)^2$$

where P denotes the number of data points, y_i denotes the true value of the ith data point, and $\hat{y}_i$ denotes the predicted value of the ith data point. Further, note that the R^2 value is defined as

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

where $\bar{y}_i$ refers to the mean value of the target variable in the data set. An R^2 value close to 1, in conjunction with low loss function values, indicates good performance. In this regard, the mean absolute error (MAE) reported in the main text is calculated as

$$\mathrm{MAE} = \frac{1}{P}\sum_{i=1}^{P}|y_i - \hat{y}_i|$$

DFT Details. Density functional theory (DFT) calculations were conducted using the Vienna Ab initio Simulation Package (VASP)[82−84] version 5.4.4, utilizing the plane-wave basis set and the projector augmented wave (PAW) method.[85,86] The PAW potentials from the VASP PAW library were employed (PAW_PBE C 08Apr2002, PAW_PBE O 08Apr2002, PAW_PBE H 15Jun2001). For the exchange-correlation energy, the Perdew−Burke−Ernzerhof (PBE)[87] generalized gradient approximation functional was applied. We used a plane-wave energy cutoff of 500 eV and Γ-point sampling due to the large size of the supercell (∼30 Å × 30 Å). To account for van der Waals interactions, Grimme's D3 dispersion correction[88] was included. A vacuum region of 20 Å was added above the 2D plane to prevent interactions between periodic images, and all structures were optimized using the conjugate gradient algorithm until the maximum force on any atom was less than 0.01 eV/Å. Structural configurations for the DFT calculations were created using the Visualization for Electronic and Structural Analysis (VESTA) software.

■ ASSOCIATED CONTENT

Data Availability Statement

The custom codes developed to handle STRONGs in the work using MATLAB and the machine learning algorithms implemented in Python are available online at https://github.com/agrgroup/STRONG. The archived version of this repository can be downloaded from Zenodo at 10.5281/zenodo.13894341.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.4c08282.

Algorithm for generating a STRONG from the XYZ file of a graphene nanopore; explanation of the canonical STRONG of the example nanopore in Figure 1f of the main text; extension of STRONGs to nonhexagonal 2D materials and quasi-2D materials; ML model performance metrics for longer formation time, larger formation energy, and stable nanopores; evaluation of the physical characteristics for the best candidate nanopores to separate gas mixtures; histogram of gas transport barrier difference; subtleties involved in the comparison of two STRONGs via reflectional symmetry; rules to generate semantically valid STRONGs and stable nanopores using STRONG transformations; algorithm for constructing an XYZ file from a STRONG; relaxed functionalized structures through first-principles calculations; details on the etching rates for different atom types; details on the methodology used to determine gas transport barriers through nanopores; comparison of first-principles DFT with molecular mechanics in calculating gas transport barriers through nanopores, and comparison of different confirmational orientations of the gas molecule transporting through a nanopore (PDF)

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Author

Ananth Govind Rajan − *Department of Chemical Engineering, Indian Institute of Science, Bengaluru, Karnataka 560012, India;* orcid.org/0000-0003-2462-0506; Email: ananthgr@iisc.ac.in

Authors

Piyush Sharma − *Department of Chemical Engineering, Indian Institute of Science, Bengaluru, Karnataka 560012, India*

Sneha Thomas − *Department of Chemical Engineering, Indian Institute of Science, Bengaluru, Karnataka 560012, India; Department of Chemical Engineering and Analytical Science, The University of Manchester, Manchester M13 9PL, United Kingdom*

Mahika Nair − *Department of Chemical Engineering, Indian Institute of Science, Bengaluru, Karnataka 560012, India; Division of Sciences, School of Interwoven Arts and Sciences, Krea University, Sri City, Andhra Pradesh 517646, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.4c08282

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the National Supercomputing Mission (NSM) for providing financial support via grant DST/NSM/R&D_HPC_Applications/Extension Grant/2023/16. AGR acknowledges the Infosys Foundation, Bengaluru, for an Infosys Young Investigator grant. We thank the Supercomputer Education and Research Centre at the Indian Institute of Science for computational facilities. AGR acknowledges discussions with Prof. Michael Strano on atom types at the edges of graphene and their use to characterize nanopores.

REFERENCES

(1) Geim, A. K.; Novoselov, K. S. The Rise of Graphene. *Nat. Mater.* **2007**, *6* (3), 183−191.

(2) Novoselov, K. S.; Geim, A. K.; Morozov, S. V.; Jiang, D.; Zhang, Y.; Dubonos, S. V.; Grigorieva, I. V.; Firsov, A. A. Electric Field Effect in Atomically Thin Carbon Films. *Science* **2004**, *306* (5696), 666−669.

(3) Golberg, D.; Bando, Y.; Huang, Y.; Terao, T.; Mitome, M.; Tang, C.; Zhi, C. Boron Nitride Nanotubes and Nanosheets. *ACS Nano* **2010**, *4* (6), 2979−2993.

(4) Wang, Q. H.; Kalantar-Zadeh, K.; Kis, A.; Coleman, J. N.; Strano, M. S. Electronics and Optoelectronics of Two-Dimensional Transition Metal Dichalcogenides. *Nat. Nanotechnol.* **2012**, *7* (11), 699−712.

(5) Huang, S.; Li, S.; Villalobos, L. F.; Dakhchoune, M.; Micari, M.; Babu, D. J.; Vahdat, M. T.; Mensi, M.; Oveisi, E.; Agrawal, K. V. Millisecond Lattice Gasification for High-Density CO_2- and O_2-Sieving Nanopores in Single-Layer Graphene. *Sci. Adv.* **2021**, *7* (9), No. eabf0116.

(6) Heiranian, M.; Farimani, A. B.; Aluru, N. R. Water Desalination with a Single-Layer MoS_2 Nanopore. *Nat. Commun.* **2015**, *6* (1), No. 8616.

(7) Prozorovska, L.; Kidambi, P. R. State-of-the-Art and Future Prospects for Atomically Thin Membranes from 2D Materials. *Adv. Mater.* **2018**, *30* (52), No. 1801179.

(8) Yuan, Z.; He, G.; Li, S. X.; Misra, R. P.; Strano, M. S.; Blankschtein, D. Gas Separations Using Nanoporous Atomically Thin Membranes: Recent Theoretical, Simulation, and Experimental Advances. *Adv. Mater.* **2022**, *34* (32), No. 2201472.

(9) Sholl, D. S.; Lively, R. P. Seven Chemical Separations to Change the World. *Nature* **2016**, *532* (7600), 435−437.

(10) Kiss, A. A.; Smith, R. Rethinking Energy Use in Distillation Processes for a More Sustainable Chemical Industry. *Energy* **2020**, *203*, No. 117788.

(11) Gutiérrez-Guerra, R.; Segovia-Hernández, J. G.; Hernández, S. Reducing Energy Consumption and CO_2 Emissions in Extractive Distillation. *Chem. Eng. Res. Des.* **2009**, *87* (2), 145−152.

(12) Yang, Q.; Su, Y.; Chi, C.; Cherian, C. T.; Huang, K.; Kravets, V. G.; Wang, F. C.; Zhang, J. C.; Pratt, A.; Grigorenko, A. N.; Guinea, F.; Geim, A. K.; Nair, R. R. Ultrathin Graphene-Based Membrane with Precise Molecular Sieving and Ultrafast Solvent Permeation. *Nat. Mater.* **2017**, *16* (12), 1198−1202.

(13) Cohen-Tanugi, D.; Grossman, J. C. Water Desalination across Nanoporous Graphene. *Nano Lett.* **2012**, *12* (7), 3602−3608.

(14) Qian, Y.; Shang, J.; Liu, D.; Yang, G.; Wang, X.; Chen, C.; Kou, L.; Lei, W. Enhanced Ion Sieving of Graphene Oxide Membranes via Surface Amine Functionalization. *J. Am. Chem. Soc.* **2021**, *143* (13), 5080−5090.

(15) Surwade, S. P.; Smirnov, S. N.; Vlassiouk, I. V.; Unocic, R. R.; Veith, G. M.; Dai, S.; Mahurin, S. M. Water Desalination Using Nanoporous Single-Layer Graphene. *Nat. Nanotechnol.* **2015**, *10* (5), 459−464.

(16) Yuan, Z.; Govind Rajan, A.; Misra, R. P.; Drahushuk, L. W.; Agrawal, K. V.; Strano, M. S.; Blankschtein, D. Mechanism and Prediction of Gas Permeation through Sub-Nanometer Graphene Pores: Comparison of Theory and Simulation. *ACS Nano* **2017**, *11* (8), 7974−7987.

(17) Hauser, A. W.; Schwerdtfeger, P. Methane-Selective Nanoporous Graphene Membranes for Gas Purification. *Phys. Chem. Chem. Phys.* **2012**, *14* (38), 13292−13298.

(18) Malekian, F.; Ghafourian, H.; Zare, K.; Sharif, A. A.; Zamani, Y. Recent Progress in Gas Separation Using Functionalized Graphene Nanopores and Nanoporous Graphene Oxide Membranes. *Eur. Phys. J. Plus* **2019**, *134* (5), No. 212.

(19) Wang, L.; Drahushuk, L. W.; Cantley, L.; Koenig, S. P.; Liu, X.; Pellegrino, J.; Strano, M. S.; Scott Bunch, J. Molecular Valves for Controlling Gas Phase Transport Made from Discrete Ångström-Sized Pores in Graphene. *Nat. Nanotechnol.* **2015**, *10* (9), 785−790.

(20) Ying, Y.; Tong, M.; Ning, S.; Ravi, S. K.; Peh, S. B.; Tan, S. C.; Pennycook, S. J.; Zhao, D. Ultrathin Two-Dimensional Membranes Assembled by Ionic Covalent Organic Nanosheets with Reduced Apertures for Gas Separation. *J. Am. Chem. Soc.* **2020**, *142* (9), 4472−4480.

(21) Branton, D.; Deamer, D. W.; Marziali, A.; Bayley, H.; Benner, S. A.; Butler, T.; Di Ventra, M.; Garaj, S.; Hibbs, A.; Huang, X.; Jovanovich, S. B.; Krstic, P. S.; Lindsay, S.; Ling, X. S.; Mastrangelo, C. H.; Meller, A.; Oliver, J. S.; Pershin, Y. V.; Ramsey, J. M.; Riehn, R.; Soni, G. V.; Tabard-Cossa, V.; Wanunu, M.; Wiggin, M.; Schloss, J. A. The Potential and Challenges of Nanopore Sequencing. *Nat. Biotechnol.* **2008**, *26* (10), 1146−1153.

(22) Heerema, S. J.; Dekker, C. Graphene Nanodevices for DNA Sequencing. *Nat. Nanotechnol.* **2016**, *11* (2), 127−136.

(23) Yuan, W.; Chen, J.; Shi, G. Nanoporous Graphene Materials. *Mater. Today* **2014**, *17* (2), 77−85.

(24) Carlsson, J. M.; Scheffler, M. Structural, Electronic, and Chemical Properties of Nanoporous Carbon. *Phys. Rev. Lett.* **2006**, *96* (4), No. 046806.

(25) Berry, V. Impermeability of Graphene and Its Applications. *Carbon* **2013**, *62*, 1−10.

(26) Fischbein, M. D.; Drndić, M. Electron Beam Nanosculpting of Suspended Graphene Sheets. *Appl. Phys. Lett.* **2008**, *93* (11), No. 113107.

(27) Hsu, K.-J.; Villalobos, L. F.; Huang, S.; Chi, H.-Y.; Dakhchoune, M.; Lee, W.-C.; He, G.; Mensi, M.; Agrawal, K. V. Multipulsed Millisecond Ozone Gasification for Predictable Tuning of Nucleation and Nucleation-Decoupled Nanopore Expansion in

Graphene for Carbon Capture. *ACS Nano* **2021**, *15* (8), 13230−13239.
(28) Sarker, M.; Dobner, C.; Zahl, P.; Fiankor, C.; Zhang, J.; Saxena, A.; Aluru, N.; Enders, A.; Sinitskii, A. Porous Nanographenes, Graphene Nanoribbons, and Nanoporous Graphene Selectively Synthesized from the Same Molecular Precursor. *J. Am. Chem. Soc.* **2024**, *146* (21), 14453−14467.
(29) Pawlak, R.; Liu, X.; Ninova, S.; D'Astolfo, P.; Drechsel, C.; Sangtarash, S.; Häner, R.; Decurtins, S.; Sadeghi, H.; Lambert, C. J.; Aschauer, U.; Liu, S.-X.; Meyer, E. Bottom-up Synthesis of Nitrogen-Doped Porous Graphene Nanoribbons. *J. Am. Chem. Soc.* **2020**, *142* (29), 12568−12573.
(30) Govind Rajan, A.; Silmore, K. S.; Swett, J.; Robertson, A. W.; Warner, J. H.; Blankschtein, D.; Strano, M. S. Addressing the Isomer Cataloguing Problem for Nanopores in Two-Dimensional Materials. *Nat. Mater.* **2019**, *18* (2), 129−135.
(31) Cao, Z.; Markey, G.; Farimani, A. B. Ozark Graphene Nanopore for Efficient Water Desalination. *J. Phys. Chem. B* **2021**, *125* (40), 11256−11263.
(32) Liang, L.; Zhou, H.; Li, J.; Chen, Q.; Zhu, L.; Ren, H. Data-Driven Design of Nanopore Graphene for Water Desalination. *J. Phys. Chem. C* **2021**, *125* (50), 27685−27692.
(33) Bondaz, L.; Ronghe, A.; Li, S.; Čerņevičs, K.; Hao, J.; Yazyev, O. V.; Ayappa, K. G.; Agrawal, K. V. Selective Photonic Gasification of Strained Oxygen Clusters on Graphene for Tuning Pore Size in the Å Regime. *JACS Au* **2023**, *3* (10), 2844−2854.
(34) Sheshanarayana, R.; Govind Rajan, A. Tailoring Nanoporous Graphene via Machine Learning: Predicting Probabilities and Formation Times of Arbitrary Nanopore Shapes. *J. Chem. Phys.* **2022**, *156* (20), No. 204703.
(35) Thomas, S.; Silmore, K. S.; Sharma, P.; Govind Rajan, A. Enumerating Stable Nanopores in Graphene and Their Geometrical Properties Using the Combinatorics of Hexagonal Lattices. *J. Chem. Inf. Model.* **2023**, *63* (3), 870−881.
(36) Qiu, H.; Zhou, W.; Guo, W. Nanopores in Graphene and Other 2D Materials: A Decade's Journey toward Sequencing. *ACS Nano* **2021**, *15* (12), 18848−18864.
(37) Fthenakis, Z. G. A Proposed Nomenclature for Graphene Pores: A Systematic Study of Their Geometrical Features and an Algorithm for Their Generation and Enumeration. *Carbon* **2022**, *199*, 508−519.
(38) Grossek, A. S.; Niggas, A.; Wilhelm, R. A.; Aumayr, F.; Lemell, C. Model for Nanopore Formation in Two-Dimensional Materials by Impact of Highly Charged Ions. *Nano Lett.* **2022**, *22* (23), 9679−9684.
(39) Sorkun, M. C.; Astruc, S.; Koelman, J. M. V. A.; Er, S. An Artificial Intelligence-Aided Virtual Screening Recipe for Two-Dimensional Materials Discovery. *Npj Comput. Mater.* **2020**, *6* (1), No. 106.
(40) Xie, Y.; Zhang, C.; Hu, X.; Zhang, C.; Kelley, S. P.; Atwood, J. L.; Lin, J. Machine Learning Assisted Synthesis of Metal−Organic Nanocapsules. *J. Am. Chem. Soc.* **2020**, *142* (3), 1475−1481.
(41) Bhattacharya, D.; Cassady, H. J.; Hickner, M. A.; Reinhart, W. F. Large Language Models as Molecular Design Engines. *J. Chem. Inf. Model.* **2024**, *64* (18), 7086−7096.
(42) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36.
(43) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97−101.
(44) Duan, C.; Nandy, A.; Kulik, H. J. Machine Learning for the Discovery, Design, and Engineering of Materials. *Annu. Rev. Chem. Biomol. Eng.* **2022**, *13* (1), 405−429.
(45) Banik, S.; Loeffler, T.; Manna, S.; Chan, H.; Srinivasan, S.; Darancet, P.; Hexemer, A.; Sankaranarayanan, S. K. R. S. A Continuous Action Space Tree Search for INverse desiGn (CASTING) Framework for Materials Discovery. *Npj Comput. Mater.* **2023**, *9* (1), No. 177.
(46) Hu, Y.; Buehler, M. J. Deep Language Models for Interpretative and Predictive Materials Science. *APL Mach. Learn.* **2023**, *1* (1), No. 010901.
(47) Wang, Y.; Cao, Z.; Farimani, A. B. Efficient Water Desalination with Graphene Nanopores Obtained Using Artificial Intelligence. *Npj 2D Mater. Appl.* **2021**, *5* (1), No. 66.
(48) Cao, Z.; Farimani, O. B.; Ock, J.; Farimani, A. B. Machine Learning in Membrane Design: From Property Prediction to AI-Guided Optimization. *Nano Lett.* **2024**, *24*, 2953−2960.
(49) Hirschberg, J.; Manning, C. D. Advances in Natural Language Processing. *Science* **2015**, *349* (6245), 261−266.
(50) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4* (1), 120−131.
(51) Meyer, J. C.; Eder, F.; Kurasch, S.; Skakalova, V.; Kotakoski, J.; Park, H. J.; Roth, S.; Chuvilin, A.; Eyhusen, S.; Benner, G.; Krasheninnikov, A. V.; Kaiser, U. Accurate Measurement of Electron Beam Induced Displacement Cross Sections for Single-Layer Graphene. *Phys. Rev. Lett.* **2012**, *108* (19), No. 196102.
(52) Voter, A. F. Introduction to the Kinetic Monte Carlo Method. In *Radiation Effects in Solids*; Sickafus, K. E.; Kotomin, E. A.; Uberuaga, B. P., Eds.; NATO Science Series; Springer Netherlands: Dordrecht, 2007; Vol. *235*, pp 1−23.
(53) Govind Rajan, A.; Warner, J. H.; Blankschtein, D.; Strano, M. S. Generalized Mechanistic Model for the Chemical Vapor Deposition of 2D Transition Metal Dichalcogenide Monolayers. *ACS Nano* **2016**, *10* (4), 4330−4344.
(54) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mane, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viegas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2016, arXiv:1603.04467. arXiv.org e-Print archive. https://arxiv.org/abs/1603.04467. (accessed October 10, 2024).
(55) Keras: Deep Learning for Humans. https://keras.io/. (accessed September 06, 2024).
(56) Birdal, T. Maximum Inscribed Circle Using Voronoi Diagram. https://www.mathworks.com/matlabcentral/fileexchange/32543-maximum-inscribed-circle-using-voronoi-diagram. (accessed 2024−10−09).
(57) Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. 2014, arXiv:1412.3555. arXiv.org e-Print archive. http://arxiv.org/abs/1412.3555. (accessed January 15, 2024).
(58) Sint, K.; Wang, B.; Král, P. Selective Ion Passage through Functionalized Graphene Nanopores. *J. Am. Chem. Soc.* **2008**, *130* (49), 16448−16449.
(59) Skinnider, M. A.; Stacey, R. G.; Wishart, D. S.; Foster, L. J. Chemical Language Models Enable Navigation in Sparsely Populated Chemical Space. *Nat. Mach. Intell.* **2021**, *3* (9), 759−770.
(60) O'Hern, S. C.; Boutilier, M. S. H.; Idrobo, J.-C.; Song, Y.; Kong, J.; Laoui, T.; Atieh, M.; Karnik, R. Selective Ionic Transport through Tunable Subnanometer Pores in Single-Layer Graphene Membranes. *Nano Lett.* **2014**, *14* (3), 1234−1241.
(61) Lee, J.; Yang, Z.; Zhou, W.; Pennycook, S. J.; Pantelides, S. T.; Chisholm, M. F. Stabilization of Graphene Nanopore. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111* (21), 7522−7526.
(62) Lalitha, M.; Lakshmipathi, S.; Bhatia, S. K. Defect-Mediated Reduction in Barrier for Helium Tunneling through Functionalized Graphene Nanopores. *J. Phys. Chem. C* **2015**, *119* (36), 20940−20948.
(63) Li, Z.; Zhang, W.; Luo, Y.; Yang, J.; Hou, J. G. How Graphene Is Cut upon Oxidation? *J. Am. Chem. Soc.* **2009**, *131* (18), 6320−6321.

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

(64) Bagri, A.; Grantab, R.; Medhekar, N. V.; Shenoy, V. B. Stability and Formation Mechanisms of Carbonyl- and Hydroxyl-Decorated Holes in Graphene Oxide. *J. Phys. Chem. C* **2010**, *114* (28), 12053−12061.
(65) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging Large Language Models for Predictive Chemistry. *Nat. Mach. Intell.* **2024**, *6* (2), 161−169.
(66) Skinnider, M. A. Invalid SMILES Are Beneficial Rather than Detrimental to Chemical Language Models. *Nat. Mach. Intell.* **2024**, *6*, 437−448.
(67) Gillespie, D. T. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *J. Comput. Phys.* **1976**, *22* (4), 403−434.
(68) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117* (1), 1−19.
(69) Srinivasan, S. G.; van Duin, A. C. T.; Ganesh, P. Development of a ReaxFF Potential for Carbon Condensed Phases and Its Application to the Thermal Fragmentation of a Large Fullerene. *J. Phys. Chem. A* **2015**, *119* (4), 571−580.
(70) Chenoweth, K.; van Duin, A. C. T.; Goddard, W. A. ReaxFF Reactive Force Field for Molecular Dynamics Simulations of Hydrocarbon Oxidation. *J. Phys. Chem. A* **2008**, *112* (5), 1040−1053.
(71) Allouche, A. Software News and Updates Gabedit — A Graphical User Interface for Computational Chemistry Softwares. *J. Comput. Chem.* **2012**, *32*, 174−182.
(72) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33−38.
(73) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. *Comput. Phys. Commun.* **2022**, *271*, No. 108171.
(74) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225−11236.
(75) Cheng, A.; Steele, W. A. Computer Simulation of Ammonia on Graphite. I. Low Temperature Structure of Monolayer and Bilayer Films. *J. Chem. Phys.* **1990**, *92* (6), 3858−3866.
(76) Potoff, J. J.; Siepmann, J. I. Vapor−Liquid Equilibria of Mixtures Containing Alkanes, Carbon Dioxide, and Nitrogen. *AIChE J.* **2001**, *47* (7), 1676−1682.
(77) Harris, J. G.; Yung, K. H. Carbon Dioxide's Liquid-Vapor Coexistence Curve and Critical Properties as Predicted by a Simple Molecular Model. *J. Phys. Chem. A* **1995**, *99* (31), 12021−12024.
(78) Wang, S.; Hou, K.; Heinz, H. Accurate and Compatible Force Fields for Molecular Oxygen, Nitrogen, and Hydrogen to Simulate Gases, Electrolytes, and Heterogeneous Interfaces. *J. Chem. Theory Comput.* **2021**, *17* (8), 5198−5213.
(79) Hockney, R. W.; Eastwood, J. W. *Particle-Particle-Particle-Mesh (P3M) Algorithms, Computer Simulation Using Particles*; CRC Press, 2021.
(80) Toukmaji, A. Y.; Board, J. A. Ewald Summation Techniques in Perspective: A Survey. *Comput. Phys. Commun.* **1996**, *95* (2−3), 73−92.
(81) Pedregosa, F.; Weiss, R.; Brucher, M.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(82) Kresse, G.; Hafner, J. *Ab Initio* Molecular Dynamics for Liquid Metals. *Phys. Rev. B* **1993**, *47* (1), 558−561.
(83) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6* (1), 15−50.
(84) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for *Ab Initio* Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54* (16), 11169−11186.
(85) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50* (24), 17953−17979.
(86) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59* (3), 1758−1775.
(87) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865−3868.
(88) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate *Ab Initio* Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), No. 154104.

https://doi.org/10.1021/jacs.4c08282
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX