Knowledge-informed generation of organic structure-directing agents for zeolites using ChatGPT towards human-machine collaborative molecular design

Shusuke Ito[1], Koki Muraoka[1,*], Akira Nakayama[1,*]

[1]*Department of Chemical System Engineering, The University of Tokyo, Tokyo 113-8656, Japan*

Abstract

Designing organic molecules lies at the heart of solving numerous chemistry-related challenges, necessitating effective collaboration between human intuition and computational power. This study demonstrates how general-purpose Large Language Models (LLMs) such as GPT-4 can facilitate the design of potent molecules, leveraging feedback from experiments and empirical knowledge through natural language. We used this approach to design organic structure-directing agents (OSDAs) that guide the crystallization of zeolites. A computational workflow was developed, wherein the LLM proposed novel OSDAs to stabilize targeted zeolites. The suggested candidates underwent evaluation through empirical screening criteria and atomistic simulation. Feedback was then provided to the LLM in natural language to refine subsequent proposals, thus progressively enhancing the proposed OSDAs and promoting the exploration of chemical space. The predicted candidates encompassed experimentally validated OSDAs, structurally analogous ones, and novel ones with superior affinity scores, underscoring the robust capability of the LLM. The collaborations between humans and machines, utilizing natural language as the communication interface, hold potential for application in other molecular design tasks, including drug design.

Introduction

Designing organic molecules is the fundamental approach to solving many chemistry-related problems[1,2]. This involves identifying the optimal point within the vast expanse of chemical space, sometimes called a needle-in-a-haystack problem[3], balancing multiple parameters, including chemical activities or physical properties[4]. Traditionally, the design of organic molecules has been the exclusive domain of experimental scientists[5]. The complexity of the molecular design often necessitates costly trial-and-error methodologies and typically requires the expertise of seasoned organic chemists[6]. To resolve this issue with the leverage of big data, various de novo molecular design methodologies have been developed[7,8]. While these algorithms can process a lot of molecules that are unachievable by experiments, human intuition still surpasses computationally generated molecules[9,10]. The ideal story would be the effective collaboration of human and de novo molecular design workflow, which has been gathering attention recently[11].

One promising platform to realize effective human-machine collaboration would be large language models (LLMs)[12–14]. LLMs are machine learning models that have been trained on a diverse range of internet text. They are proficient in generating human-like text based on the given text[15]. Some of the most advanced LLMs, such as GPT-3 and GPT-4 developed by OpenAI[15] as well as Bard or Gemini developed by Google[16,17], have shown remarkable performance in a wide array of tasks, including translation, question answering, and even generating creative content[15]. These models have gained considerable attention due to their capacity to engage in diverse discussions with humans, ranging from general topics

to specialized subjects[18]. The application of LLMs presents a significant opportunity to propel scientific research forward, providing innovative techniques for exploring and interpreting complex data and theoretical concepts[19,20]. Because LLMs allow users to interact easily via natural language, it is natural to assume that they have the potential to become a great platform for realizing human–machine collaboration for designing targeted molecules.

The implementation of LLMs in the fields of materials science and chemistry has already begun. Recent studies have demonstrated the application of LLMs in guiding the synthesis of metal–organic frameworks, circumventing costly trial-and-error experimentation[19,20]. This highlights the potential of LLMs in materials synthesis. Bajorath et al. have developed a generative chemical language model to predict highly potent compounds from less potent ones as input for drug discovery[21]. Priyakumar et al. enabled a transformer-decoder model named MolGPT inspired by GPT models to generate drug-like molecules[22]. While their studies employed highly customized language models specifically trained to generate chemical compounds, we postulated that general-purpose LLMs such as GPT and Gemini would perform well in designing potent molecules, without further extensive training.

Our target for molecular design is organic structure-directing agents (OSDAs) for zeolites. OSDAs, typically quaternary ammonium cations, facilitate the crystallization of zeolites[23]. Zeolites are porous crystalline aluminosilicates with diverse polymorph structures, each with unique channels and cavities. OSDAs can stabilize the specific shape of the inner structure of zeolites[24,25], enabling the crystallization of zeolites, including previously unknown ones[26,27]. Furthermore, the replacement of OSDAs with cheaper ones is also an interesting topic to study[28], as a significant fraction of the cost of the production of zeolites is occupied by OSDAs[29].

There are molecular design algorithms for the prediction of potent OSDAs for zeolites[30–34] and some of them are confirmed experimentally. One of the first de novo molecular design algorithms of OSDAs was reported by Lewis et al., where a stochastic algorithm lets compounds grow in the cages of **CHA** and **MFI** zeolites[30]. To predict the synthesizable OSDAs, Deem et al. used a genetic algorithm for purchasable reagents and well-documented chemical reactions[31]. Recently, a nature-inspired meta-heuristic approach was employed to perform multi-objective optimization for affinity and cost of OSDAs for syntheses of targeted zeolites[32]. A generator based on a self-attention mechanism and long and short-term memory networks was used to design potent OSDAs[33].

In this study, we developed a workflow using a general-purpose, pre-trained LLM, GPT-4. Our methodology involves feedback to GPT-4 based on the affinity scores and empirical screening criteria to mimic the human-machine collaboration using the natural language, fostering improved subsequent suggestions, and ultimately achieving the prediction of highly potent OSDAs. This workflow demonstrates the possibility of the human-machine collaboration of the molecular design, which could open new avenues in developing pharmaceuticals and other complex molecular compounds.

Results & Discussion

GPT-4 in Molecular Design

Previous research has encountered challenges in facilitating effective collaboration between humans and machines in the realm of molecular design. Addressing this issue, our study explores using LLMs to narrow this gap through natural language. We first assessed the capability of LLMs to comprehend and

manipulate chemical structures. To this end, we asked GPT-4—a leading-edge LLM created by OpenAI—to generate new OSDAs from the basic OSDA, tetraethylammonium (TMA) via natural language and SMILES notation[35] (see Table S1 for the exact conversation). GPT-4 responded by generating five SMILES strings representing TMA derivatives, each adorned with additional functional groups, thereby demonstrating its potential as a tool for molecular generation. To further explore how we can guide the LLM to the desired direction of the molecular design, we instructed GPT-4 to incorporate a carbon ring into its outputs. In response, GPT-4 provided five different candidates with carbon rings. This confirmed that GPT-4 can not only generate molecular structures but also refine them based on guidance provided via natural-language dialogue, showcasing its value in human-machine collaborative efforts in molecular design.

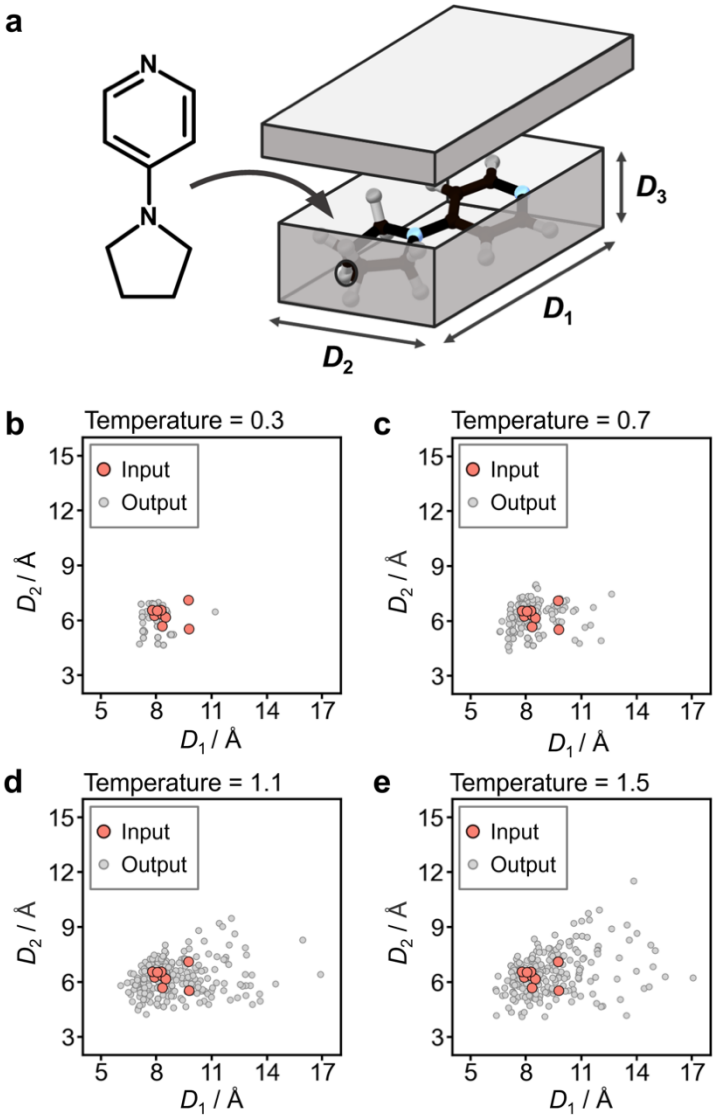

Figure 1. Shoebox parameters of input OSDAs and candidates generated by GPT-4 with different temperatures. **a**. Shoebox parameters can be calculated by putting an OSDA to a cuboid so that the longest distance of the OSDA becomes D_1. D_2 is the longer distance orthogonal to D_1. D_3 is orthogonal to D_1 and D_2. **b**–**e**. Input OSDAs (orange) were supplied to GPT-4 to generate new OSDAs (grey) at temperatures of 0.3 (b), 0.7 (c), 1.1 (d), and 1.5 (e).

GPT-4 determines the likelihood of the subsequent token based on the prompt or the preceding token it generated during text production[36]. If it only opts for tokens with high probability, the result is deterministic and not stochastic, which means that the generated texts become precise but not very far away from the given texts. Conversely, if it opts for tokens with low probability, the result becomes more stochastic[37,38]. More "creative" text is generated in this case, while information and even grammar become far away from the given texts. This deterministic-stochastic tradeoff is regulated by temperature, which is a common parameter in related text generation models. Selecting the right temperature is thus crucial for controlling the operation of GPT-4 as desired.

To examine how GPT-4 behaves differently with temperature when dealing with molecular data, we provided GPT-4 with a few OSDAs using a notation called SMILES and asked it to generate new molecules. This was repeated to obtain a total of 100 molecules. The exact prompt is shown in Figure S1. Shape and/or size descriptors on OSDAs are useful tools to predict the structure-directing ability[39]. We applied shoebox algorithm[40] (Figure 1a) to the input OSDAs and output OSDAs. Figure 1b–e visualizes the input and output OSDAs using them being D_1, the maximum distance in each

molecule, and D_2, the longer distance diagonal to the first axis. When the temperature was set to 0.3, GPT-4 generated molecules slightly different from input molecules, as shown in Figure 1b and Figure S2. A noticeable number of duplicated molecules were generated in this case. When the temperature was increased to 0.7, more diverse molecules were generated from the identical inputs, as apparent from the spread data points in Figure 1c. As the temperature increased further to 1.1 and 1.5, more and more diverse outputs were obtained (Figure 1d, e and Figure S2). This observation is consistent with the fact that the higher temperature leads to more stochastic responses[42].

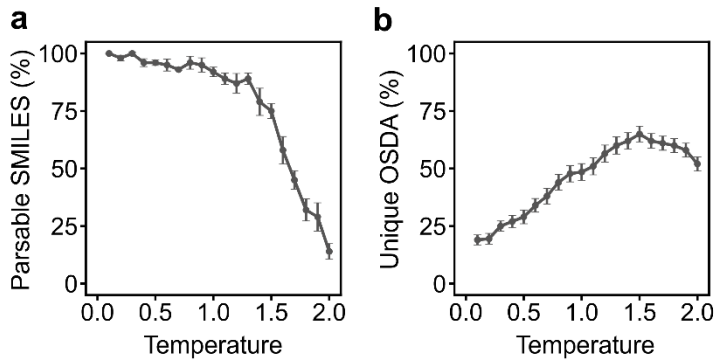

Figure 2. a. The fraction of parsable SMILES generated by GPT-4 with different temperature values. **b**. The fraction of unique OSDAs generated by GPT-4 with different temperature values. The error bars represent the standard errors among three independent runs.

Upon the above test, we observed that some SMILES strings produced by GPT-4 could not be parsed, particularly at elevated temperatures. A similar phenomenon, so-called hallucination, has been documented in text generation by GPT-4: as the temperature escalates, the text becomes excessively nonsensical and approaches a random character sequence[43]. To quantitatively examine how temperature affects the correctness of the syntax of SMILES, we measured the fraction of parsable SMILES in 100 texts generated by GPT-4 under different temperatures, across three independent trials. We further calculated the average and standard error of each temperature. As shown in Figure 2a, more than 80% of generated strings obey the syntax of SMILES at temperatures up to 1.3. As the temperature increased further, the fraction of parsable SMILES dramatically decreased. At the temperature of 2.0, we frequently observed nonsensical texts. These results suggest that low temperatures are required to obtain reliable results.

Another observation in the preliminary experiment in Figure 1 was that too low temperature led to a significant number of SMILES duplicating with inputs or outputs, which can lead to inefficient exploration of the chemical space. To examine how temperature affects the uniqueness of the generated molecules, we counted the number of unique molecules from the three runs. As shown in Figure 2b, only 20% of the generated molecules were unique at temperatures less than 0.3. As the temperature increased, the fraction of unique entries increased (Figure 2b). The peak of the unique rate was the temperature of 1.5. Above that, the number of unique entries started to decrease, due to the hallucination effect observed in Figure 2b.

Collectively, we chose the temperature range from 0.7 to 1.1 for developing our de novo molecular design workflow, considering the tradeoff between the correctness and the novelty of the generated texts shown in Figure 1 and Figure 2, which resembles the exploration-exploitation trade-off[44].

De novo molecular design workflow

Figure 3 provides the overview of our de novo molecular design workflow, which incorporates elements such as GPT-4, prescreening filters, atomistic simulation, and a database. The data stored within the database encompasses the SMILES notation for the molecule and the corresponding stabilization energy for a specific zeolite. The algorithm retrieves the top 10 entries among 100 recent records from the

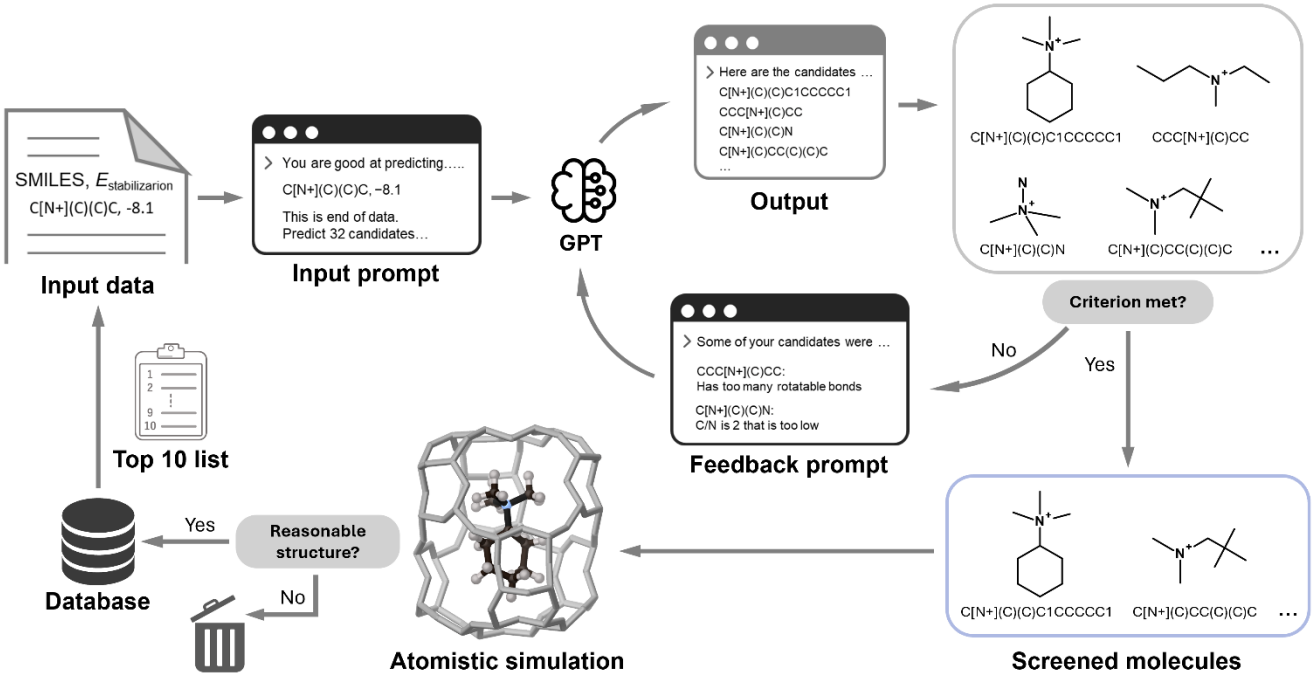

Figure 3. Schematic illustration of the computational workflow. TMA (C[N+](C)(C)C) is chosen as the first input data. GPT-4 generates new OSDAs based on the input OSDAs and corresponding stabilization energies. Some of the generated OSDAs are rejected by the screening, and their reasons are fed to GPT-4. Screened molecules are docked in the inner space of a specific zeolite for calculation of their affinity. If the structure of the zeolite–OSDA complex is reasonable, the OSDA and the calculated stabilization energy are stored in a database. 10 molecules and corresponding stabilization energies out of 100 recent records in a database are used to update the input data.

database and formulates a prompting text to instruct GPT-4 to generate potential molecules. The precise prompt is depicted in Figure S3. The formulated text is then submitted to GPT-4. To initiate the workflow, tetramethylammonium (TMA), one of the simplest OSDAs, and its stabilization energy were prepared as the first input data.

In response, GPT-4 generates candidate SMILES-like strings as requested. These strings are parsed and subjected to a series of empirical filtering criteria to eliminate candidates unsuitable as OSDAs for zeolites. The criteria check rigidity/flexibility, stability, and hydrophobicity/hydrophilicity. It has been found that the rigidity/flexibility influences the structure-directing ability. In particular, OSDAs with an excessive number of degrees of freedom may adopt multiple configurations during the hydrothermal synthesis of zeolites, thereby reducing the selectivity[45]. To avoid them, we precluded the molecules containing many rotatable bonds and too large rings. Zeolites are typically synthesized under hydrothermal conditions[46–48] so that OSDA needs sufficient stability. Molecules containing a N-ring with triple bonds may be unsuitable for hydrothermal synthesis. If the compounds contain rings with fewer than five members, the steep bond angle could potentially break the ring. We exploited these properties to assess chemical stability. It is established that a balance of hydrophobicity and hydrophilicity is vital for OSDA to promote the crystallization of zeolites in aqueous media. This is because OSDA requires sufficient hydrophilicity to be dissolved in water, and they also need to interact with the relatively hydrophobic aluminosilicate species that ultimately form zeolites[49]. C/N ratio is an experimentally established metric to evaluate the hydrophobicity and hydrophilicity of OSDA[50]. We limit the C/N ratio from 4 to 20. We restrict the

permissible elements in OSDA candidates to those typically observed—hydrogen, carbon, nitrogen, and oxygen. This is done to ensure the effectiveness of the C/N ratio in evaluating hydrophobicity/hydrophilicity (see details in Table S2).

The rejected entries by the above screening criteria along with the reasons for the rejection are then sent back to GPT-4. This process mimics the human-machine interaction in molecular design. GPT-4 is then asked to generate new candidates based on the feedback. This step is repeated up to five times according to the preliminary testing (Figure S4).

Molecules that meet the prescreening criteria undergo atomistic simulation to assess their affinity against the targeted zeolite. We use stabilization energy as the metric to evaluate this affinity, a common indicator for assessing potential OSDAs for targeted zeolites[51–53].

$$E_S = E_{complex} - E_{zeolite} - nE_{OSDA}$$

$E_{complex}$ represents the energy of complex of the zeolite with the OSDA inside of its cages, $E_{zeolite}$ and E_{OSDA} are the energy of the vacant zeolite and the isolated OSDA respectively, and n is number of OSDAs in the unit cell of the zeolite. We set $n = 3$ in this study. The "frozen pose" method[54], as recommended by a previous study[55], is utilized, where the structure of relaxed zeolite–OSDA complex was directly used to calculate $E_{complex}$, $E_{zeolite}$, and E_{OSDA} to assess the stabilization energy. This method, however, occasionally distorts the zeolite frameworks unrealistically, leading to significant under- or over-estimation of the stabilization energy. To avoid this, we dismiss any zeolite–OSDA complexes with excessive atomic displacement (see the Methods section for details). Candidates that passed this criterion were stored in the database for subsequent runs.

Designing OSDAs for AEI, CHA, and ITE

The trajectory of the optimization

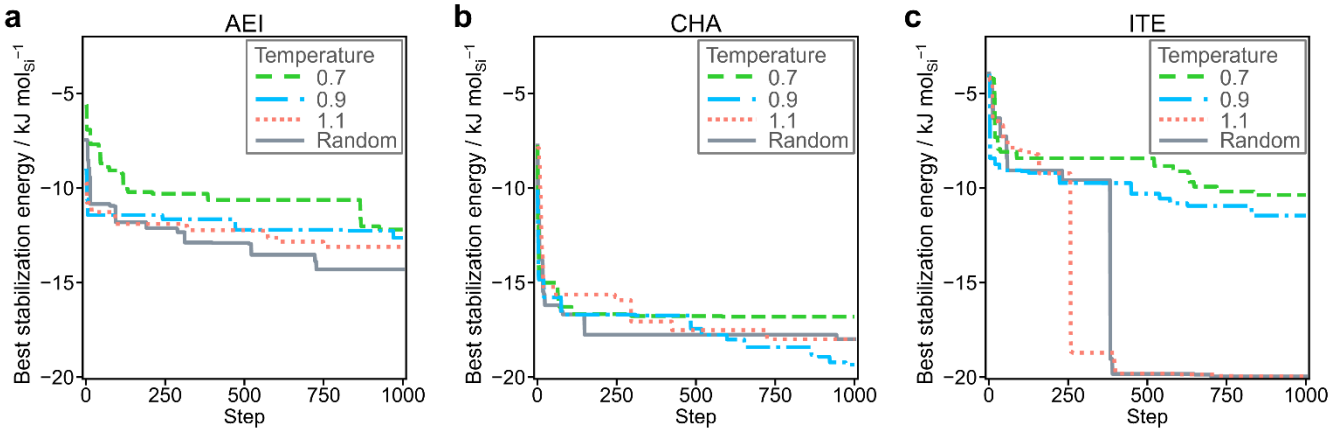

Figure 4. The representative trajectories of best stabilization energy from the runs at fixed temperatures of 0.7, 0.9, 1,1 and random temperature settings. **a**. **AEI** zeolite. **b**. **CHA** zeolite. **c**. **ITE** zeolite.

Our de novo molecular design methodology was applied to three cage-type zeolites: **CHA**, **AEI**, and **ITE** (Figure S6), which are frequently used structures in demonstrating de novo molecular design for OSDAs[31,32,56–58].

Temperatures play a crucial role in balancing the deterministic and stochastic behavior of GPT-4, as discussed earlier. We evaluated our workflow at three temperatures: 0.7, 0.9, and 1.1. Figure 4a shows the lowest stabilization energy at each step of the runs for **AEI**. A temperature of 0.7 yielded a slower improvement in stabilization energy, likely due to the deterministic nature of lower temperatures. Increasing the temperature to 0.9 resulted in a steep initial improvement in stabilization energy, likely due to the increased stochastic nature of GPT-4. The most negative stabilization energy for the entire run at a temperature of 0.9 was more negative than that of 0.7, indicating that the higher temperature improved the results. At the higher temperature of 1.1, a steep initial improvement was observed akin to the temperature of 0.9. This temperature resulted in the OSDA with more negative stabilization energy than the other temperatures.

Higher temperatures do not necessarily yield OSDAs with higher affinity. For **CHA**, all three temperature runs exhibited similar initial profiles (Figure 4b), with the best OSDA achieved at 0.9, neither the highest nor the lowest temperature. The sensitivity towards temperatures of GPT-4 was more evident with ITE (Figure 4c). At temperatures of 0.7 and 0.9, the trajectories were similar, achieving stabilization energies around −10 kJ mol_{Si}^{-1}. Increasing the temperature to 1.1 generated OSDAs with more negative stabilization energies of approximately −20 kJ mol_{Si}^{-1}, indicating that temperature effects are highly system-dependent.

While the optimal temperatures vary among targeted zeolites, the necessity for parameter tuning to achieve optimal results can be disadvantageous. To address this, we randomly selected a temperature from 0.7, 0.9, and 1.1 each time we asked queries to GPT-4. The grey solid lines in Figure 4 show the trajectory of the best stabilization energies. This stochastic scheme, as shown in the figure, demonstrated comparable or superior performance to the other conditions, effectively balancing exploration, and exploitation by GPT-4 at different temperatures.

Suggested OSDA candidates

Figure 5 shows some OSDAs predicted by our de novo molecular design workflow under varying temperature conditions. GPT-4 successfully generated several OSDAs that are identical or closely resembling experimentally validated OSDAs. OSDAs **1**, **8**, and **3** are identical to those known from experimental studies of corresponding zeolites[58–60]. Predicted candidates **2**, **3**, **4**, **9**, **10**, **11**, **15**, and **16** demonstrate structural similarities to known OSDAs **20**, **21**, **22**, **23**, **24**, **25**, **26**, and **27**, respectively[50,56,58] (see Figure S5). For example, OSDAs **2** and **9** possess an extra methyl group compared to **20** and **23** (Figure S5); **3** and **11** have fewer methyl group than **21** and **25** with similar chemical structures (Figure S5).

Some generated candidates exhibited superior stabilization energies compared to experimentally proven OSDAs. **5**, **6**, and **7** showed sufficiently negative stabilization energies for **AEI** compared with experimentally proven OSDA and candidates having resembling structures to them (**1**, **2**, **3**, and **4** in Figure 5). For **CHA**, **12**, **13**, and **14** displayed stabilization energies better than known ones, likely due to their slightly larger size. While the runs of **ITE** successfully predicted piperidine-based OSDAs including experimentally verified ones such as **1** and **3**, it also identified promising candidates exhibiting more negative stabilization energies. Particularly, it expanded its exploration to include OSDAs containing benzene rings such as **17**, **18**, and **19**. Despite the lack of benzene rings in the experimentally validated OSDAs, the relatively large cavity in **ITE** seems to have great compatibility with OSDAs with benzene rings. This compatibility merits future experimental investigation. We presume this contributed to the

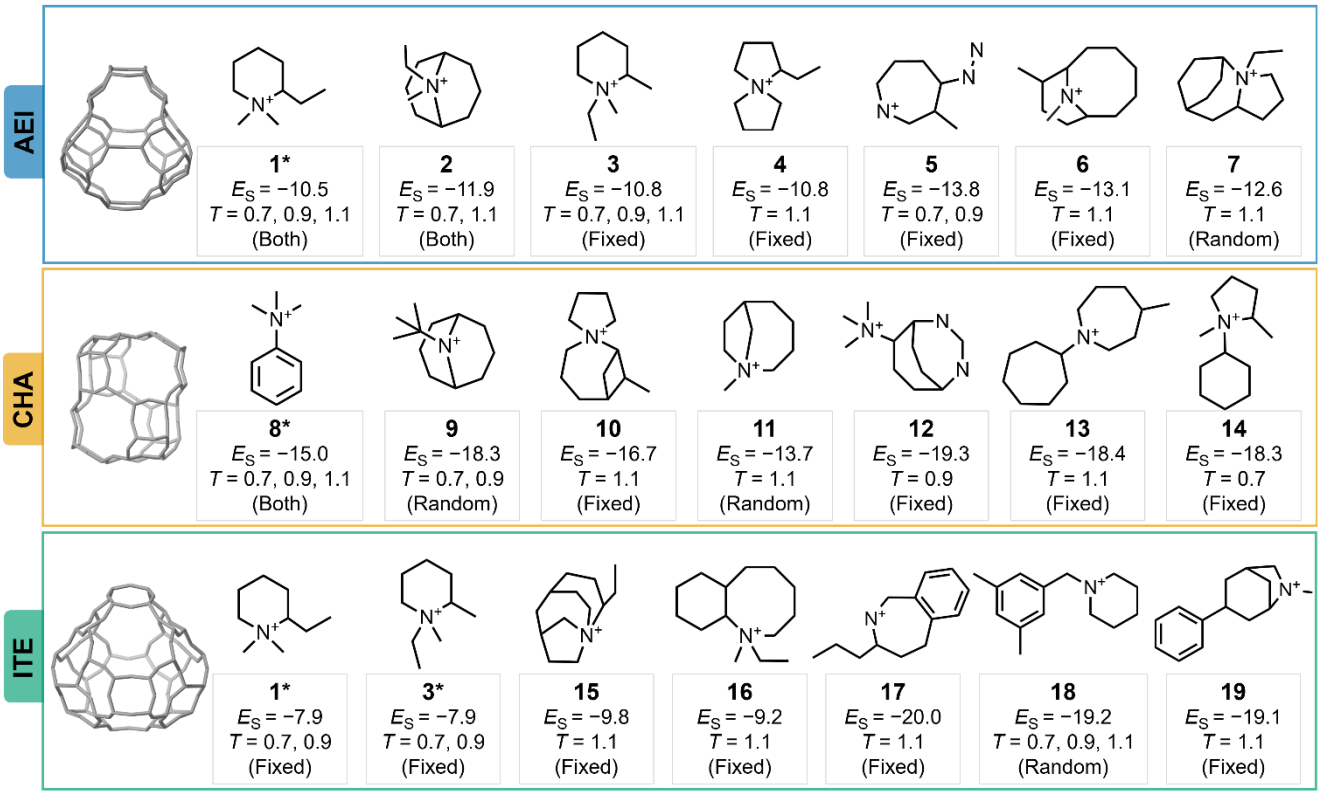

Figure 5. Computationally predicted OSDAs for **CHA**, **AEI**, and **ITE** zeolites, along with their associated stabilization energy (E_S), temperature (T), and the methods employed for temperature settings (random, fixed, or both). Experimentally validated OSDAs for each zeolite type are marked with an asterisk. The first three on the left for each zeolite type correspond to known or highly similar OSDAs, while the remaining candidates exhibit favorable stabilization energies.

difference in the trajectory of best stabilization energies (**Figure 4**). While under lower temperatures, the probability of generating OSDAs with benzene rings seems to be low, at a higher temperature of 1.1, such OSDAs can easily be generated to achieve more negative stabilization energy.

A limitation of the current prediction regarding the experimental validation is the potential synthesizability of OSDA candidates because prescreening criteria did not consider it, and GPT-4 was not asked to take it into account. The difficulty in computer-aided synthesis planning is well-established within the scientific literature, even when utilizing specialized algorithms[61,62]. To ascertain if LLMs could aid in computer-aided retrosynthesis, we asked GPT-4 to estimate the synthesis pathways for experimentally validated OSDAs, as depicted in Table S3. While GPT-4 correctly identified a synthesis pathway of the simplest OSDA, TMA, it did not successfully estimate the pathway for **1**. This may be attributed to the dearth of information on the organic synthesis pathway in the training data of GPT-4. It would be interesting to study LLMs for data-driven retrosynthesis through the specialized LLMs or by more general-purpose models with post-training.

Dimensionality reduction

To better understand the nature of explored OSDAs, we performed principal component analysis (PCA) for descriptors of OSDAs for all of the suggested molecules as well as experimentally known OSDAs[39] that pass our prescreening criteria, as shown in Figure 6. Considering that the shape, size, and

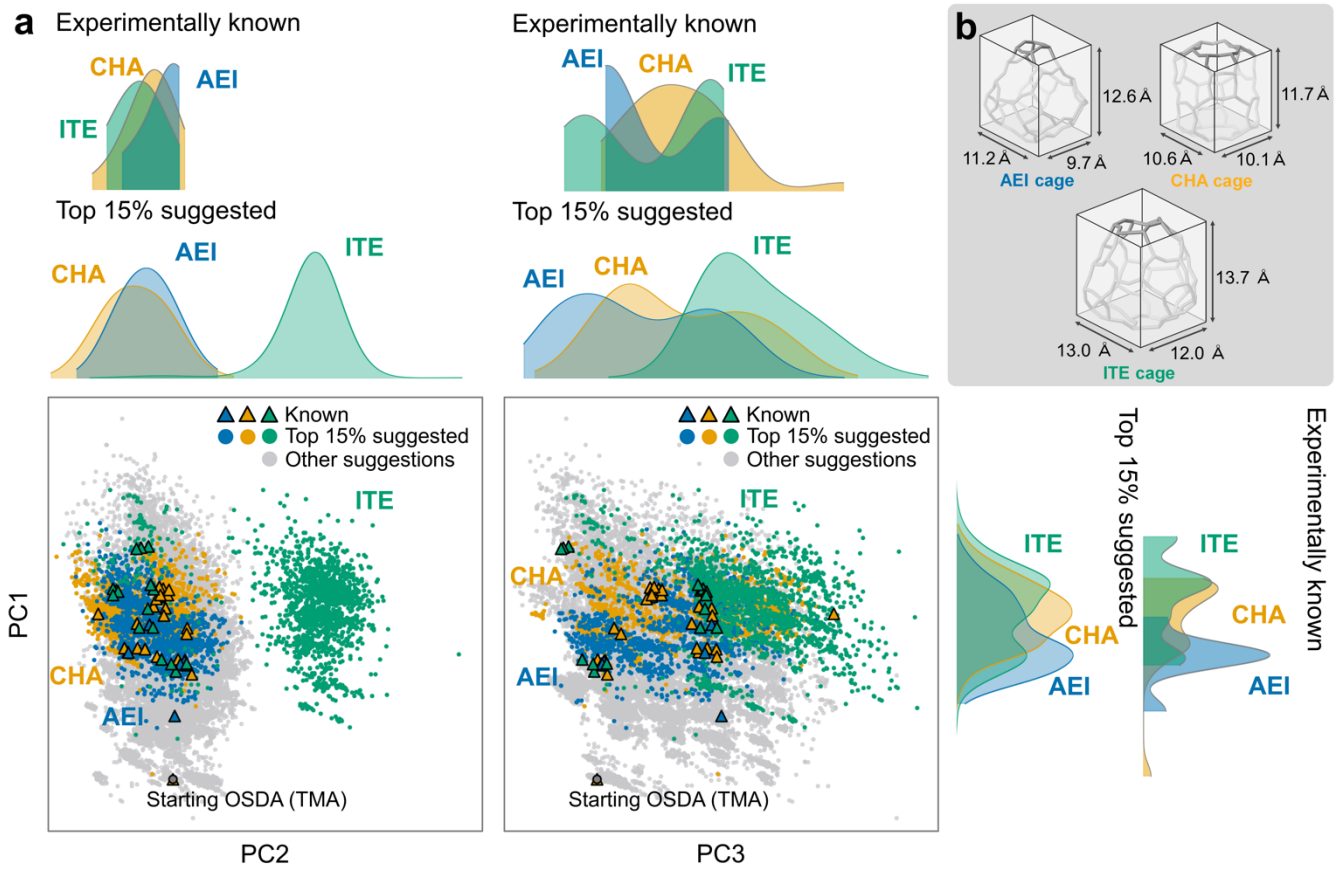

Figure 6 a. PCA for suggested and experimentally proven OSDAs for **CHA**, **AEI**, and **ITE** zeolites together with Gaussian kernel density estimates of PCs. All suggested OSDA candidates are scattered in grey color on the PCA maps. TMA, the starting OSDA in our workflow is also plotted. The top 15% of suggested candidates based on stabilization energy are denoted as circles with respective colors. Experimentally proven OSDAs are represented as triangles, with each color corresponding to a specific zeolite. Two kinds of gaussian kernel density estimates of distributions are depicted. One describes the top 15% of explored OSDA candidates, and the other shows experimentally proven OSDAs. **b.** Cages in **ITE**, **CHA**, and **AEI**, with boxes representing their size and shape based on interatomic distances.

rigidity/flexibility of OSDAs significantly affect zeolite crystallization[23,63], we calculated molecular descriptors including volume, number of atoms (N), number of rotatable bonds, number of rings, number of benzene rings, number of atoms creating branches, and geometrical parameters (D_1, D_2, D_3, and ($D_1 \times D_2 \times D_3$) $\times N$) derived from the shoebox algorithm (Figure 1a)[40]. The explained variance values of the principal components (PCs) confirm that PC1, PC2, and PC3 adequately captured most of the variance in data (Figure S6a). After carefully examining the elements of the PCs as detailed in Figure S6b, we found that PC1 describes the overall size of OSDAs; OSDAs with larger PC2 show high aspect ratios; OSDAs with larger value of PC3 is closely related to rigid property.

As shown in Figure 6a, our de novo molecular design workflow using GPT-4 explored a vast chemical space covering the plots of experimentally proven OSDAs for **AEI**, **CHA**, and **ITE**. We extracted top 15% of the generated OSDAs in terms of the stabilization energy for each zeolite. The distributions of PCs for these OSDAs with high affinity is highly consistent with the experimentally proven OSDAs, suggesting the validity of our approach. The capability of the LLM pretrained against a large number of

internet texts and provided feedback informed by empirical screening criteria and stabilization energies effectively enable the workflow to produce reasonable candidates using only the structure of the simplest OSDA as a starting point. The adaptive exploration and exploitation process can be visualized using the PCA map as shown in Figure S7. At first, GPT-4 predominantly sampled points close to the initial OSDA, TMA. As the process advanced, GPT-4 increasingly explored more diverse regions of the chemical space, as depicted in Figure S7.

A noticeable point in Figure 6a is that PC2 is divided into two clusters. The figure shows that the cluster with high PC2 was explored to generate effective OSDAs for **ITE**. The thorough investigation of the constituting OSDAs reveals that this cluster is for benzene-containing OSDAs characterized by high aspect ratios, such as **17**, **18**, and **19** in Figure 6. These points are sampled at the later stages of the runs for **ITE**, as depicted in Figure S7 and Figure 4. Furthermore, promising OSDAs for **ITE** tend to have larger PC1 and PC3 (Figure 6a), indicating that bulkier and more rigid OSDAs are effective for **ITE**. To fill and stabilize relatively large cages in **ITE**, depicted in Figure 6b, OSDAs need to be bulkier and more rigid, as suggested earlier[58,64].

OSDAs effective for **CHA** and **AEI** overlap in the PCA map (Figure 6a), presumably due to the similar size of cages in the two zeolites[65] (Figure 6b). For both the top 15% of the generated OSDA candidates and experimentally proven OSDAs, OSDAs for **CHA** exhibit marginally elevated values in PC1 and PC3, alongside reduced values in PC2, compared to **AEI** counterparts (Figure 6a). This suggests that the cage for **AEI** favors smaller, flatter, and more flexible OSDAs, while **CHA** favors bulkier, more spherical, and more rigid ones. To effectively fill the avocado-like uneven cages in **AEI** (Figure 6b), OSDAs with more rotational freedom can be advantageous[56].

These results confirm that the general-purpose LLM successfully sampled the chemical space covering existing OSDAs with similar distribution, recommending the candidates with properties and affinity tailored to each zeolite.

Conclusion

Prior studies documented de novo molecular design algorithms tailored for diverse applications. Despite the advanced development of algorithms and the vast amount of data available, the effective integration of computationally generated data by chemists remains a significant challenge. Consequently, molecular design continues to rely heavily on experimental trial and error. In this study, we focused on general-purpose LLMs to demonstrate the potential human-machine collaboration through natural language and chemical notation to design novel molecules, specifically OSDAs for synthesizing zeolites.

Our de novo molecular design framework using an LLM can efficiently explore the chemical space for OSDAs under appropriate temperature parameters. The feedback provided in natural language, stemming from empirical knowledge, along with affinity scores derived from atomistic simulations, effectively steers the LLM toward identifying more viable candidates. This highlights the strong capability of LLM to design novel molecules. Owing to the extensive pretraining of the general-purpose LLM, it can propose to add various functional groups not in the starting OSDA. This feature will benefit from transferring chemical knowledge across various domains documented in diverse texts used during the pretraining of LLMs.

Although the current study is limited to designing OSDAs for synthesizing zeolites, this approach can be applied to drug design and other molecular design problems. This paves the way for future collaborations between humans and machines in molecular design, utilizing natural language as the communication interface.

Methods

Dataset

Structures of pure silica zeolites were obtained from the International Zeolite Association[66]. The experimentally verified OSDAs for the **AEI**, **CHA**, and **ITE** zeolites were acquired from the OSDB database[39].

LLM

We utilized GPT-4 via an API developed by OpenAI. The queries were performed between June 26th 2023 and March 11th 2023. The exact prompts are shown in Figure S1, S3, and S4 and Table S2

Prescreening

GPT-4 sometimes produces candidates that are not as viable as OSDAs. We rejected such candidates by applying empirical screening criteria. The properties of OSDAs were calculated by RDKit software[67]. The complete list of screening criteria is presented in Table S2. To avoid too flexible candidates, the number of rotatable bonds was restricted to less than 5, and bridge-free rings larger than the nine-membered ring were rejected. To avoid candidates unstable under hydrothermal conditions, we rejected the candidates with rings smaller than the five-membered ring, NH^+ group, NH_3^+ group, and rings having nitrogen and triple bonds. Furthermore, if rings with N possessed bonds in conjugated systems (one-and-a-half bonds) and/or double bonds, it would not be included if it met the empirically determined equation[32].

$$3n_2 + 4.5n_3 \geq n_1$$

Where n_1, n_2, and n_3 represent the count of single bonds, one-and-a-half bonds, and double bonds in the ring, respectively. We also rejected candidates with unconventional atoms and restricted C/N ratio as described above.

Atomistic modeling

Proposed OSDA candidates were optimized by means of the universal force field (UFF) as implemented in RDKit software[67]. The location and rotation of OSDAs in a zeolite were determined by Bayesian optimization with the following objective function.

$$\theta = n_{1.1} \times 10^{17} + n_{1.3} \times 10^{14} + n_{1.5} \times 10^{12} + n_{1.7} \times 10^{10}$$

$n_{1.1}$, $n_{1.3}$, $n_{1.5}$, and $n_{1.7}$ represent the number of interatomic distances between zeolite and OSDAs under specific distance thresholds of 1.1, 1.3, 1.5, and 1.7 Å, respectively. The resulting zeolite–OSDA complex underwent structure optimization with the GULP program[68] using the charge-less DREIDING force field[69], which is used and verified in various zeolite–OSDA systems[32,55,56].

We dismissed zeolite–OSDA complexes exhibiting stabilization energy smaller than −20 kJ molSi−1. Zeolite frameworks in zeolite–OSDA complexes after structure optimization were evaluated by Structure Matcher implemented in pymatgen[70]. If the algorithm is unable to determine that a zeolite in an optimized complex is equivalent to its corresponding zeolite before optimization, it is subsequently rejected.

Cheminformatics

Three-dimensional coordinates of a candidate OSDA, as obtained by UFF, were subjected the shoebox algorithm[40] to calculate D_1, D_2, D_3, and the size descriptor, $(D_1 + D_2 + D_3) \times N$, where N is the number of atoms. First, maximum interatomic distances in a candidate OSDA were found and defined as distance D_1. The candidate was rotated to define a new plane perpendicular to the axis along with D_1. On the new plane, the two atoms of maximum distance were obtained to define the distance D_2. From these two axes along with D_1 and D_2, we created a new axis, which we defined as D_3, denoting the maximum interatomic distance along with it. RDKit calculated the other molecular descriptors. The molecular descriptors underwent standard scaling and PCA as implemented in the scikit-learn package.

Data availability

The data of this study are available from the corresponding author upon request.

References

1. Gurung, A. B., Ali, M. A., Lee, J., Farah, M. A. & Al-Anazi, K. M. An Updated Review of Computer-Aided Drug Design and Its Application to COVID-19. *BIOMED Res. Int.* **2021**, 8853056 (2021).
2. Austin, N. D., Sahinidis, N. V. & Trahan, D. W. Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques. *Chem. Eng. Res. Des.* **116**, 2–26 (2016).
3. Churi, N. & Achenie, L. E. K. Novel Mathematical Programming Model for Computer Aided Molecular Design. *Ind. Eng. Chem. Res.* **35**, 3788–3794 (1996).
4. Li, Z., Jiang, M., Wang, S. & Zhang, S. Deep learning methods for molecular representation and property prediction. *Drug Discov. Today* **27**, 103373 (2022).
5. Odele, O. & Macchietto, S. Computer aided molecular design: a novel method for optimal solvent selection. *Fluid Phase Equilibria* **82**, 47–54 (1993).
6. Venkatasubramanian, V., Chan, K. & Caruthers, J. M. Computer-aided molecular design using genetic algorithms. *Comput. Chem. Eng.* **18**, 833–844 (1994).
7. Meyers, J., Fabian, B. & Brown, N. De novo molecular design and generative models. *Drug Discov. Today* **26**, 2707–2715 (2021).
8. Sabe, V. T. *et al.* Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *Eur. J. Med. Chem.* **224**, 113705 (2021).
9. Schneider, G. & Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.* **4**, 649–663 (2005).
10. Winter, R. *et al.* Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **10**, 8016–8024 (2019).
11. Sundin, I. *et al.* Human-in-the-loop assisted de novo molecular design. *J. Cheminformatics* **14**, 86 (2022).

12. Liu, Y. *et al.* Generative artificial intelligence and its applications in materials science: Current situation and future perspectives. *J. Materiomics* **9**, 798–816 (2023).
13. Maik Jablonka, K. *et al.* 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digit. Discov.* **2**, 1233–1250 (2023).
14. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
15. OpenAI *et al.* GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).
16. Gemini Team *et al.* Gemini: A Family of Highly Capable Multimodal Models. Preprint at https://doi.org/10.48550/arXiv.2312.11805 (2023).
17. Ali, R. *et al.* Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery* **93**, 1090 (2023).
18. Eloundou, T., Manning, S., Mishkin, P. & Rock, D. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2303.10130 (2023).
19. Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T. & Yaghi, O. M. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *J. Am. Chem. Soc.* **145**, 18048–18062 (2023).
20. Zheng, Z. *et al.* A GPT-4 Reticular Chemist for Guiding MOF Discovery. *Angew. Chem. Int. Ed.* **62**, e202311983 (2023).
21. Chen, H. & Bajorath, J. Designing highly potent compounds using a chemical language model. *Sci. Rep.* **13**, (2023).
22. Bagal, V., Aggarwal, R., Vinod, P. K. & Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J. Chem. Inf. Model.* (2021) doi:10.1021/acs.jcim.1c00600.
23. Lobo, R. F., Zones, S. I. & Davis, M. E. Structure-direction in zeolite synthesis. *J. Incl. Phenom. Mol. Recognit. Chem.* **21**, 47–78 (1995).
24. Li, R. *et al.* Diverse Physical States of Amorphous Precursors in Zeolite Synthesis. *Ind. Eng. Chem. Res.* **57**, 8460–8471 (2018).
25. Burkett, S. L. & Davis, M. E. Mechanism of Structure Direction in the Synthesis of Si-ZSM-5: An Investigation by Intermolecular 1H-29Si CP MAS NMR. *J. Phys. Chem.* **98**, 4647–4653 (2002).
26. Li, J., Corma, A. & Yu, J. Synthesis of new zeolite structures. *Chem. Soc. Rev.* **44**, 7112–7127 (2015).
27. Jo, D. & Hong, S. B. Targeted Synthesis of a Zeolite with Pre-established Framework Topology. *Angew. Chem. Int. Ed.* **58**, 13845–13848 (2019).
28. Schwalbe-Koda, D. *et al.* Repurposing Templates for Zeolite Synthesis from Simulations and Data Mining. *Chem. Mater.* **34**, 5366–5376 (2022).
29. Zones, S. I. Translating new materials discoveries in zeolite research to commercial manufacture. *Microporous Mesoporous Mater.* **144**, 1–8 (2011).
30. Lewis, D. W., Willock, D. J., Catlow, C. R. A., Thomas, J. M. & Hutchings, G. J. De novo design of structure-directing agents for the synthesis of microporous solids. *Nature* **382**, 604–606 (1996).
31. Pophale, R., Daeyaert, F. & W. Deem, M. Computational prediction of chemically synthesizable organic structure directing agents for zeolites. *J. Mater. Chem. A* **1**, 6750–6760 (2013).
32. Muraoka, K., Chaikittisilp, W. & Okubo, T. Multi-objective: De novo molecular design of organic structure-directing agents for zeolites using nature-inspired ant colony optimization. *Chem. Sci.* **11**, 8214–8223 (2020).
33. Xu, L., Peng, X., Xi, Z., Yuan, Z. & Zhong, W. Predicting organic structures directing agents for zeolites with conditional deep learning generative model. *Chem. Eng. Sci.* **282**, 119188 (2023).

34. Sastre, G. & Daeyaert, F. *AI-Guided Design and Property Prediction for Zeolites and Nanoporous Materials*. (John Wiley & Sons, 2023).
35. Gong, H., Liu, Q., Wu, S. & Wang, L. Text-Guided Molecule Generation with Diffusion Language Model. Preprint at https://doi.org/10.48550/arXiv.2402.13040 (2024).
36. Ghojogh, B. & Ghodsi, A. Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey. Preprint at https://doi.org/10.31219/osf.io/m6gcn (2020).
37. Ouyang, S., Zhang, J. M., Harman, M. & Wang, M. LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation. Preprint at https://doi.org/10.48550/arXiv.2308.02828 (2023).
38. Momennejad, I. *et al.* Evaluating Cognitive Maps and Planning in Large Language Models with CogEval. *Adv. Neural Inf. Process. Syst.* **36**, (2023).
39. Schwalbe-Koda, D. *et al.* A priori control of zeolite phase competition and intergrowth with high-throughput simulations. *Science* **374**, 308–315 (2021).
40. León, S. & Sastre, G. Computational Screening of Structure-Directing Agents for the Synthesis of Pure Silica ITE Zeolite. *J. Phys. Chem. Lett.* **11**, 6164–6167 (2020).
41. Boyett, R. E., Stevens, A. P., Ford, M. G. & Cox, P. A. A quantitative shape analysis of organic templates employed in zeolite synthesis. *Zeolites* **17**, 508–512 (1996).
42. Rosoł, M., Gąsior, J. S., Łaba, J., Korzeniewski, K. & Młyńczak, M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Sci. Rep.* **13**, 20512 (2023).
43. Beutel, G., Geerits, E. & Kielstein, J. T. Artificial hallucination: GPT on LSD? *Crit. Care* **27**, 148 (2023).
44. Chang, C.-C., Reitter, D., Aksitov, R. & Sung, Y.-H. KL-Divergence Guided Temperature Sampling. Preprint at (2023).
45. Gies, H. & Marker, B. The structure-controlling role of organic templates for the synthesis of porosils in the systems SiO2/template/H2O. *Zeolites* **12**, 42–49 (1992).
46. Kim, E. *et al.* On the synthesis and characterization of all-silica CHA zeolite particles. *Microporous Mesoporous Mater.* **184**, 47–54 (2014).
47. Guo, Y. *et al.* Cost-effective synthesis of CHA zeolites with controllable morphology and size. *Chem. Eng. J.* **358**, 331–339 (2019).
48. Tsunoji, N., Shimono, D., Tsuchiya, K., Sadakane, M. & Sano, T. Formation Pathway of AEI Zeolites as a Basis for a Streamlined Synthesis. *Chem. Mater.* **32**, 60–74 (2020).
49. Burkett, S. L. & Davis, M. E. Mechanism of Structure Direction in the Synthesis of Pure-Silica Zeolites. 2. Hydrophobic Hydration and Structural Specificity. *Chem. Mater.* **7**, 1453–1463 (1995).
50. Kubota, Y., Helmkamp, M. M., Zones, S. I. & Davis, M. E. Properties of organic cations that lead to the structure-direction of high-silica molecular sieves. *Microporous Mater.* **6**, 213–229 (1996).
51. Schmidt, J. E., Deem, M. W. & Davis, M. E. Synthesis of a Specified, Silica Molecular Sieve by Using Computationally Predicted Organic Structure-Directing Agents. *Angew. Chem. Int. Ed.* **53**, 8372–8374 (2014).
52. Daeyaert, F., Ye, F. & Deem, M. W. Machine-learning approach to the design of OSDAs for zeolite beta. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 3413–3418 (2019).
53. Boal, B. W. *et al.* Facile Synthesis and Catalysis of Pure-Silica and Heteroatom LTA. *Chem. Mater.* **27**, 7774–7779 (2015).
54. Zones, S. I., Nakagawa, Y., Yuen, L. T. & Harris, T. V. Guest/Host Interactions in High Silica Zeolite Synthesis: [5.2.1.$0^{2,6}$]Tricyclodecanes as Template Molecule. *J. Am. Chem. Soc.* **118**, 7558–7567 (1996).

55. Schwalbe-Koda, D. & Gómez-Bombarelli, R. Benchmarking binding energy calculations for organic structure-directing agents in pure-silica zeolites. *J. Chem. Phys.* **154**, 174109 (2021).
56. Schmidt, J. E., Deem, M. W., Lew, C. & Davis, T. M. Computationally-Guided Synthesis of the 8-Ring Zeolite AEI. *Top. Catal.* **58**, 410–415 (2015).
57. Schwalbe-Koda, D. & Gómez-Bombarelli, R. Supramolecular Recognition in Crystalline Nanocavities through Monte Carlo and Voronoi Network Algorithms. *J. Phys. Chem. C* **125**, 3009–3017 (2021).
58. Wagner, P. *et al.* Guest/Host Relationships in the Synthesis of the Novel Cage-Based Zeolites SSZ-35, SSZ-36, and SSZ-39. *J. Am. Chem. Soc.* **122**, 263–273 (2000).
59. Liang, J. *et al.* CHA-type zeolites with high boron content: Synthesis, structure and selective adsorption properties. *Microporous Mesoporous Mater.* **194**, 97–105 (2014).
60. Zones, S. I., Burton, A. W., Lee, G. S. & Olmstead, M. M. A Study of Piperidinium Structure-Directing Agents in the Synthesis of Silica Molecular Sieves under Fluoride-Based Conditions. *J. Am. Chem. Soc.* **129**, 9066–9079 (2007).
61. Gao, W., Raghavan, P. & Coley, C. W. Autonomous platforms for data-driven organic synthesis. *Nat. Commun.* **13**, 1075 (2022).
62. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
63. Moliner, M. Design of Zeolites with Specific Architectures Using Self-Assembled Aromatic Organic Structure Directing Agents. *Top. Catal.* **58**, 502–512 (2015).
64. Cantín, A., Corma, A., Diaz-Cabanas, M. J., Jordá, J. L. & Moliner, M. Rational Design and HT Techniques Allow the Synthesis of New IWR Zeolite Polymorphs. *J. Am. Chem. Soc.* **128**, 4216–4217 (2006).
65. Jensen, Z. *et al.* Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks. *ACS Cent. Sci.* **7**, 858–867 (2021).
66. Baerlocher, C. *et al.* Database of Zeolite Structures. https://www.iza-structure.org/databases/.
67. RDKit. Open-Source Cheminformatics Software. https://www.rdkit.org/.
68. Gale, J. D. & Rohl, A. L. The General Utility Lattice Program (GULP). *Mol. Simul.* **29**, 291–341 (2003).
69. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909 (1990).
70. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).

Acknowledgments

This work is supported by JSPS KAKENHI (22K14751) and JST PRESTO (JPMJPR2378).

Author Contribution

K.M. and A.N. directed the project. K.M. conceived the project. S.I. and K.M. developed the de novo molecular design workflow. S.I. and K.M. discussed, validated, and analyzed the data. S.I. and K.M. wrote the manuscript with input from all authors. All authors reviewed and commented on the manuscript.

Competing financial interests

The authors declare no competing financial interests.