Unveiling the Potential of AI for Nanomaterial Morphology Prediction

Anonymous Authors[1]

Abstract

Creation of nanomaterials with specific morphology remains a complex experimental process, even though there is a growing demand for these materials in various industry sectors. This study explores the potential of AI to predict the morphology of nanoparticles within the data availability constraints. For that, we first generated a new multi-modal dataset that is double the size of analogous studies. Then, we systematically evaluated performance of classical machine learning and large language models in prediction of nanomaterial shapes and sizes. Finally, we prototyped a text-to-image system, discussed the obtained empirical results, as well as the limitations and promises of existing approaches.

1. Introduction

Nowadays, nanomaterials are spread across many fields of science and industry (Zebarjadi et al., 2011; Liu & Lal, 2015; Kairdolf et al., 2017; Shifrina et al., 2020; Gao et al., 2021; Takechi-Haraya et al., 2022). In each of those fields, for a nanomaterial to be fit for purpose, its size, shape, and other morphological parameters must be precisely controlled, as they directly influence toxicity, catalytic activity and other properties of nanomaterials crucial for applications. Altering these parameters also allows to improve efficiency of drug delivery systems (Sen Gupta, 2016), catalysts (Shifrina et al., 2020), energy storage systems (Pomerantseva et al., 2019), etc.

Typically, creating a nanomaterial with a specific set of properties requires a significant number of experiments ranging from a few repetitive syntheses to a dozen of substantially different synthesis procedures (Vaidyanathan & Sendhilnathan, 2008; Sun et al., 2021). Each synthesis is followed by a specific method of analysis to confirm the experimental outcome. One of the most prominent methods for analyzing nanomaterials is the scanning electron microscopy (SEM) (Smith & Oatley, 1955). With SEM, it is possible to obtain information about the size and shape of nanoparticles (NPs), as well as the structure of the surface, surface flaws and contaminants. Currently, the SEM method is deemed irreplaceable despite being costly and time-consuming (Singh, 2016). On average, one analysis with SEM can cost up to a few hundred US dollars, leading to vast amounts of resources required to run any large-scale study. Because of the huge number of interdependent synthesis parameters affecting the final result, it is currently impossible to theoretically predict the outcome of a particular synthesis. Therefore, there is a high demand (AbdelHamid et al., 2022) for predictive models capable of characterizing the properties of nanomaterials bypassing the need of costly experimental work.

Artificial intelligence (AI) offers the most promising set of tools to meet this demand. In fact, classical machine learning (ML) models including artificial neural networks have already been successfully applied to many tasks related to nanomaterial science (Serov & Vinogradov, 2022; Chen et al., 2023; Banaye Yazdipour et al., 2023). With recent astonishing advances in deep learning (Jumper et al., 2021; Rombach et al., 2021; Ramesh et al., 2022; OpenAI, 2023; Touvron et al., 2023; Jiang et al., 2023; Merchant et al., 2023), the potential of AI in the design of nanomaterials seems truly immense. However, one has to possess large volumes of carefully curated data to fully exploit the power of AI. As discussed earlier, accumulating the data appropriate for the prediction of nanomaterial morphology has been a major challenge for decades. Within realistic data constraints, the boundaries of AI in the design of nanomaterials are underexplored.

In this study, we aim to unveil the capabilities and limitations of AI in predicting morphology of nanomaterials. For that, we first conduct 215 experimental syntheses of calcium carbonate-based nanomaterials of different shapes and sizes. We carefully document the synthesis procedures with the parameters of experimental conditions, take SEM-images of the resulting nanoparticles, segment and manually

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

annotate them with expert knowledge. We investigate the statistical associations in this multimodal dataset and identify features informative of nanoparticle morphology. We further use these findings to train classical ML models to predict sizes and shapes of nanoparticles and achieve 0.77 and 0.80 average accuracy, respectively. For the first time in the field of nanomaterial synthesis, we explore the potential of LLMs for prediction tasks. Using few-shot methods, we utilize state-of-the-art models, such as GPT-4, to predict the shapes of nanomaterials and achieve an impressive 0.81 average accuracy. Finally, we augment the available data to prototype a text-to-image system aimed at generating an image of a nanoparticle based on the description of its synthesis procedure. In conclusion, we review the obtained empirical results and discuss the future of AI in the field of nanomaterial design.

2. Related work

Over the past 10 years, there have been several works predicting morphological properties of nanoparticles. However, the majority of them focused on size prediction considering a single experimental system, where the resulting particles conform to the same shape and their sizes can be easily standardized. Some particular examples include size prediction for silver nanoparticles (Chen et al., 2016; Shafaei & Khayati, 2020), carbon nanotubes (Iakovlev et al., 2019), agar nanospheres (Zaki et al., 2015), chitosan nanoparticles (Baharifar & Amani, 2017), polymeric nanoparticles (Shahsavari et al., 2013; Soliman et al., 2014; Youshia et al., 2017), TiO_2 nanoparticles (Pellegrino et al., 2020) and different methacrylates (Kimmig et al., 2021). In our work, there is no attachment to nanoparticles of a certain shape. Instead, we generate a dataset containing multiple different shapes, which greatly expands the generalizability of our approach and enables future transfer learning applications. In addition, unlike many previous studies, we provide the data for benchmarking and the code for reproducibility.

A few published works specialize in predicting the shapes of nanoparticles (Timoshenko et al., 2017; Chen et al., 2020; Yao et al., 2022), but they too have certain shortcomings. For example, Timoshenko et al. created a model that takes experimental X-ray absorption near-edge structure (XANES) spectroscopy data as input to predict the 3D structure of metallic nanoparticles (Timoshenko et al., 2017). Although circumventing the need for SEM analysis, this approach still requires actual synthesis and experimental evaluation of other properties to predict the shape of the nanomaterial. This narrows down the list of possible applications significantly. In contrast, our work explores data-driven approaches that only use features of the past syntheses to predict morphology of potentially new nanomaterials.

More advanced deep learning algorithms have also found applications in the creation of new nanomaterials (Roccapriore et al., 2021; Xu et al., 2023). In the paper by Kim, Han, and Han, a model based on convolutional neural networks was proposed capable of determining the morphology of nanomaterials based on the SEM images (Kim et al., 2020). Such efforts help to better understand morphological properties of nanomaterials and simplify data labeling for the future predictive approaches. However, they do not avoid tedious experimental work preparing the datasets of SEM images, by design. Ultimately, our work stands out by predicting SEM images of nanoparticles of different morphologies based on the properties of the corresponding syntheses, which is an inverse problem formulation.

Recent advances in natural language processing (OpenAI et al., 2023; Jiang et al., 2023; Touvron et al., 2023) have also been reflected in some areas of chemistry. Recently, there have been studies that describe the use of LLMs, in particular using the few-shot method, to predict the characteristics of various chemical objects (Zheng et al., 2023) and even to generate new chemical structures (Jablonka et al., 2022). However, the potential application of LLMs to predict the morphology of nanomaterials has not yet been investigated.

Various multimodal systems have been proposed recently in application to nanomaterial science (Kononova et al., 2019; Lee et al., 2020; Hiszpanski et al., 2020). Since the emergence of Stable Diffusion (Rombach et al., 2021) and DALL-E (Ramesh et al., 2022), image generation models have attracted particularly much public attention. A recent work in nanofabrication presented an image-to-image system capable of predicting the postfabrication appearance of structures manufactured by focused ion beam milling (Buchnev et al., 2022). Although a very specialized application, it demonstrates how the field of nanotechnology already benefits from generative AI. In this work, we prototyped a text-to-image solution predicting morphologies of the previously unseen nanomaterials.

3. Dataset preparation

To obtain the most reliable and standardized dataset, we performed 215 syntheses of calcium carbonate-based nanomaterials. As mentioned above, usually up to several dozen experiments are conducted to perform optimization of nanomaterial properties. When using machine learning, more samples are usually needed to build predictive models, but due to the resource-intensive and time-consuming nature of synthesizing and analyzing nanomaterials, most of the morphology prediction works described above are limited to about a hundred syntheses. In this work, we generated a dataset that is double that size.

We considered a single chemical system of calcium car-

bonate, because of its rich variety of nanoparticle shapes and sizes. By making this study design choice, we were hoping to achieve better generalization of our work to other nanoparticles, since most of the known shapes are already represented in our dataset.

For each synthesis, we documented all variable parameters, such as names of reagents, solvents, etc., their concentrations, temperature and reaction time, as well as other synthesis parameters. Additionally, for each synthesis, one most representative SEM-image was taken, which clearly shows nanoparticles with distinguishable sizes and shapes.

We thoroughly analyzed shapes and sizes of the resulting nanoparticles and identified five different shapes: cubic, spherical, stick-shaped, flat, and amorphous. For each shape, except flat and amorphous, we distinguished small-, medium- and large-sized nanoparticles applying an empirical threshold. In the case of amorphous and flat particles, the number of samples was too small to consider differentiation. Altogether, we used 5 different shape categories and 9 different categories combining shapes and sizes to label the dataset.

To train the variational autoencoder in the text-to-image setup described later, each image from the original dataset with 215 syntheses was segmented to extract multiple images of individual nanoparticles using ImageJ (Rueden et al., 2017). The resulting dataset was further augmented to increase the size of the dataset and decrease the probability of overfitting. For that, we generated new images by applying random rotations, different blurring and brightness settings. In total, the training dataset contained 46,800 images of individual nanoparticles.

4. Feature selection

Each synthesis in our dataset was described by 10 continuous and 3 categorical variables that might be influencing the shapes of nanomaterials in different ways. This section describes statistical evaluation of those features to determine whether they are indeed informative of the geometry of nanomaterials, which served as a basis for downstream AI applications.

4.1. Analysis of continuous variables

Let $(X_1^1, X_2^1, \ldots, X_n^1)$ denote real values of a parameter of a synthesis which produces cubic nanoparticles. Let $(X_1^2, X_2^2, \ldots, X_m^2)$ denote the real values of the same parameter of any synthesis which always results in nanoparticles of different shapes. We wondered whether the two samples came from the same population or not. If so, each value of the first sample would have had an equal chance of being larger than each value of the second sample. Therefore, the null hypothesis can be formulated as follows:

$$H_0 : p(X_i^1 > X_j^2) = \frac{1}{2}$$

In fact, this formulation represents the Mann-Whitney U test (Nachar, 2008). We applied it for each of the real-valued parameters of synthesis and each type of the nanomaterial shapes. We found that formation of stick-shaped nanoparticles was dependent on the reaction temperature, synthesis time, and polymer mass and/or concentration. Cubic shapes of nanoparticles were also associated with certain temperatures and polymer concentrations, as well as the molar mass of the polymer. We used the Kruskal-Wallis H test (Kruskal & Wallis, 1952), which is analogous to Mann-Whitney U test but applicable to three and more sample groups, Kolmogorov-Smirnov test (Smirnov, 1939) and ANOVA (Marsal, 1987) to corroborate these findings. Herewith, we used the significance level = 0.05 and the Bonferroni correction method to account for multiple hypothesis testing.

4.2. Analysis of categorical variables

To establish relationships between categorical parameters of synthesis procedures and the corresponding shapes of nanomaterials, we composed contingency tables as shown in Table 1.

Table 1. Example contingency table for testing categorical variables of synthesis procedures.

	Compound in synthesis	Compound not in synthesis
NPs of a given shape	a	b
NPs of other shapes	c	d

According to Fisher, $a \sim Hypergeometric(N, K, n)$, where $N = a+b+c+d$ is the population size, $K = a+b$ is the number of successes and $n = a+c$ is the number of draws (Fisher, 1922). Therefore, the probability of this outcome is given by:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \\ = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

We computed these probabilities for each combination of nanoparticle shape and polymer/surfactant/solvent involved in the synthesis. Using the same significance level and the correction for multiple hypothesis testing as before, we

observed several strong associations: stick-shaped nanoparticles with polyethylene glycol (PEG) and polyethylenimine (PEI) polymers; flat nanoparticles with presence of PEDOT:PSS and polyvinylpyrrolidone (PVP); cubic nanoparticles with presence of polyacrylic acid (PAA) and PEDOT:PSS. We also found strong dependencies of nanoparticles' shapes on the following surfactants: Myristyltrimethylammonium bromide and Sodium dodecylsulfate. In the case of amorphous nanoparticles, the presence of Propylene glycol and tert-Butanol solvents was also found significant. Finally, we applied the Chi-squared test (Magnello, 2005) to confirm the aforementioned findings. For more information on how the statistical tests and the most significant associations between particular synthesis parameters and nanomaterial shapes, see Appendix A.1.

Notably, many of the parameters of syntheses had no effect on the shapes of nanomaterials, e.g., stirring speed, concentrations of Ca and CO_3 ions, presence of Hexadecyltrimethylammonium bromide and Triton X-100 surfactants, and 1-Hexanol and Methyl alcohol solvents. For the downstream machine learning applications, we excluded those features from the data.

5. Shape and size prediction

Statistical tests proved certain associations between the parameters of syntheses and the morphologies of the resulting nanomaterials. Therefore, we attempted to exploit them in predicting shapes and sizes of nanomaterials using classical machine learning algorithms.

In some cases, several nanoparticles of different shapes and sizes were present on the same image, so initial 215 syntheses produced 314 training examples of nanoparticles of different types. Following the logic of the statistical evaluation, we formulated a set of binary classification tasks, one for each type of shape or a combination of shape and size. In this formulation, we first trained a separate model to distinguish nanoparticles of each particular shape. Then, we ran multiple predictions for each sample during the inference to establish what shapes of nanoparticles were present on the corresponding image. The same logic applied to combinations of shapes and sizes. Notably, some syntheses consistently result in nanoparticles of several different shapes. Our approach allows dealing with such ambiguities without the need to determine the prevailing nanomaterial shape or size.

5.1. Classical machine learning

5.1.1. Tree-based ensemble models

We trained the tree-based models, namely Random Forest (RF) and Gradient Boosted Trees (XGB), to predict 9 categories representing combinations of shapes and sizes and 5 categories representing shapes only. Therein, we followed all the good practices in data preprocessing and model selection. A thorough description of the process of development, optimization and evaluation of classical machine learning models is presented in the Appendix A.2.

5.1.2. Results

The accuracy and the F1 scores of the best models evaluated on the test dataset are presented in Table 2 and Table 3. Each experiment was performed 5 times at different random states, and the mean value and standard deviation were calculated.

Table 2. Prediction of shapes. Top average accuracy and F1 scores achieved by the Random Forest classifiers on the test set.

Shape	# samples	Accuracy	F1 score
Cube	140	0.76 ± 0.02	0.73 ± 0.03
Stick	84	0.78 ± 0.01	0.77 ± 0.01
Sphere	40	0.82 ± 0.06	0.67 ± 0.08
Flat	16	0.82 ± 0.11	0.52 ± 0.09
Amorphous	34	0.80 ± 0.02	0.62 ± 0.04
Average		0.80 ± 0.04	0.66 ± 0.05

Based on our results, the shapes of nanoparticles can be predicted reasonably well (Table 2). For every nanomaterial shape, RF performed better than XGB, so only RF metrics are displayed. The average accuracy and F1 score were 0.80 and 0.66, respectively. Unsurprisingly, the samples of the least represented categories (namely, flat and amorphous shapes) produced lower F1 scores, which decreased the overall metrics.

Extending the number of categories to include the sizes of nanoparticles as well resulted in superior performance of XGB in most cases (Table 3). The overall average accuracy for the task was 0.77, and the average F1 score – 0.53 . This drop in performance was expected, as the number of samples per category became smaller, increasing the risk of overfitting. Underrepresentation becomes even more apparent as well for some classes. Apart from evaluating the models on the test set, which had never been used during training, we also explored feature importances as an additional validation step. In most cases, we observed that the top 5 most important parameters were well in agreement with the statistical tests described in the previous section and presented in Table 6 of the Appendix A.1. An example of feature importance analysis for the Random Forest model predicting whether a nanoparticle belongs to a stick shape is shown on Figure 4 of the Appendix A.2.

Thus, we demonstrated the possibility of predicting shapes and sizes of NPs with machine learning models, confirmed by average test accuracy of 0.80 and by feature importance analysis coherent with the statistical evaluation. The trained

Table 3. Prediction of shapes and sizes with tree-based ensemble models. Average accuracy, and F1 scores for Random Forest (RF) and Gradient Boosting (XGB) classifiers on the test set.

Shape & size	# samples	Accuracy		F1 score	
		XGB	RF	XGB	RF
Cube_small	25	**0.85 ± 0.01**	0.82 ± 0.04	**0.58 ± 0.07**	0.57 ± 0.08
Cube_medium	49	0.64 ± 0.05	**0.64 ± 0.03**	0.48 ± 0.05	**0.52 ± 0.06**
Cube_large	66	0.67 ± 0.04	**0.70 ± 0.03**	0.61 ± 0.02	**0.64 ± 0.03**
Stick_small	30	**0.83 ± 0.03**	0.82 ± 0.03	0.52 ± 0.04	**0.54 ± 0.04**
Stick_medium	28	**0.83 ± 0.06**	0.81 ± 0.07	**0.61 ± 0.07**	0.59 ± 0.09
Stick_large	26	**0.79 ± 0.04**	0.79 ± 0.04	**0.64 ± 0.05**	0.63 ± 0.05
Sphere_small	11	**0.70 ± 0.36**	0.68 ± 0.34	0.37 ± 0.19	**0.37 ± 0.18**
Sphere_medium	19	**0.86 ± 0.04**	0.84 ± 0.07	**0.55 ± 0.06**	0.55 ± 0.10
Sphere_large	10	**0.72 ± 0.27**	0.61 ± 0.28	**0.44 ± 0.16**	0.40 ± 0.18
Average		**0.77 ± 0.10**	0.75 ± 0.10	**0.53 ± 0.08**	0.53 ± 0.09

models can already be used to predict morphological properties of new nanomaterials based on their synthesis procedures. However, with recent advances in large language models, we wondered whether similar prediction performance can be achieved with state-of-the-art LLMs in a few-shot scenario. That would allow material scientists to use natural language for prediction tasks, bypassing the need to develop and optimize complex machine learning pipelines. In the following section, we describe applications of LLMs to nanomaterial shape and size prediction.

5.2. Large language models

5.2.1. Texts of synthesis procedures

A dataset of texts describing synthesis procedures was prepared for morphology predictions with LLMs and to train the text-to-image model. For that, we created a dozen of semantic templates with gaps for particular values of synthesis parameters. We leveraged the publicly available GPT-3.5 (Liu et al., 2023) model to generate such templates based on a few examples taken from the scientific articles. Thanks to GPT's strong ability to paraphrase while maintaining the writing style, we managed to collect texts of synthesis procedures sufficiently different from each other in semantics but identical in contents (i.e., the sequence of actions and the list of relevant parameters). We provide two examples of the generated templates in the Appendix A.3.

5.2.2. Few-shot classification

It is now known that LLMs can achieve quite high performance in domain-specific regression and classification tasks, often on par with the other widely accepted methods (Jablonka et al., 2022). In this study, we investigated applications of LLMs to nanomaterial morphology prediction.

For this purpose, we used a few-shot method, in which we show the model only a few samples from our training set and then prompt it to make a prediction for a test sample. In all experiments, we used a special prompt describing the task that the LLM was given. It starts as follows:

You are an expert in the synthesis of nanomaterials. You analyze the conditions for obtaining a nanomaterial and predict what particle shapes will be present in the synthesized material. There are five particle shapes: 'Cube', 'Stick', 'Sphere', 'Flat' and 'Amorphous'. A nanomaterial can contain particles of different shapes. If you cannot say exactly what it is, list the forms that have the highest probability in those conditions.

We then appended several random examples from the training set with the corresponding true labels and a single example from the test subset to the prompt. While doing so, we varied the number of random examples N, the sampling method and the data format. We used from $N = 2$ to $N = 10$ training examples in the prompt. We experimented with two sampling strategies: *i)* at least one training example belongs to the same target class as the test sample, *ii)* all training examples belong to the same class as the test sample. Finally, we used either of the two formats: textual (described in subsection 5.2.1) or tabular. In the tabular format, features of the training examples were concatenated to a string along with their values separated by colon, e.g., "Ca ion, mM: 44; CO3 ion, mM: 159...". Finally, the LLM was instructed to produce the list of nanoparticle shapes corresponding to the test synthesis as an answer. A more detailed description of prompts is presented in the Appendix A.3.

Using the above prompt structure, we applied 6 state-of-the-art LLMs, including GPT-4-turbo (`gpt-4-0125-preview`), GPT-4 (`gpt-4-0613`) and GPT-3.5-turbo (`gpt-3.5-turbo-1106`) from OpenAI (OpenAI, 2023), as well as the latest versions of Mistral Medium, Small and Tiny from Mistral AI (Jiang

et al., 2023), to the same classification tasks described earlier. To systematically evaluate performance, we repeated each computational experiment 5 times and calculated mean and standard deviation for the standard classification metrics.

5.2.3. Results

Table 4 shows top performance of LLMs predicting shapes of nanomaterials. Strikingly, GPT-4 achieved an even higher average accuracy than tree-based ensemble models. Among the other LLMs, it also demonstrated the smallest standard deviation, which speaks for better consistency. Interestingly, the second best model was Mistral-small. Given that its inference time and pricing are much lower than GPT-4, this model could be a pragmatic choice for practitioners as a balanced cost-quality trade-off. A detailed comparison of the pricing, inference time and rate limits is summarized in the Table 9 of Appendix A.4. In addition, we observed some mysterious drops in performance when predicting spherical shape. More specifically, Mistral-medium and GPT-4-turbo produced the accuracy of 0.38 and 0.44, respectively, which dramatically decreased their average scores, while the other models under identical experimental conditions coped with the problem reasonably well.

Analyzing the impact of sampling methods and data formats (Table 5 shows results for one of the GPT-4 experiments), we came to the following conclusions. First, including more examples from the training set belonging to the same class as the test sample definitely benefits the prediction. We observed improvements in accuracy in all related cases. Second, textual and tabular data formats performed similarly. However, textual format consistently resulted in a 4% increase in average accuracy, which was expected due to the nature of LLMs.

Finally, the number of training samples in the prompt also correlated with the performance metrics (Figure 1). For all shapes except the cube, we observed an increase in accuracy as more examples from the training set were included for prediction. However, longer prompts are also known to trigger hallucination. On top of that, there is a hard limit on the maximum prompt size for many models. Therefore, for any particular application, one has to seek another trade-off between the number of training samples and the total prompt size. In our case, the performance seemed to reach a plateau with 8 samples (see Appendix A.4 for more details). The same configuration demonstrated the overall top performance (Table 4).

Achieving state-of-the-art performance for nanomaterial morphology prediction with LLMs is very exciting for several reasons. First, it makes it possible for domain experts and experimentalists to avoid implementing complex data engineering pipelines and optimizing machine learning models, and use natural language to obtain the predictions instead. Second, it is obvious from our empirical results (Table 4) that an ensemble of LLMs would by far outperform the best classical ensemble models. Third, based on our empirical results, LLMs look especially advantageous in classification of underrepreresented classes, or in small data scenarios. In particular, GPT-4 demonstrated a significant increase in accuracy when predicting less represented spherical, flat and amorphous nanoparticles (Table 2). Altogether, our results look very promising for the broader adoption of LLMs in the nanomaterial science.

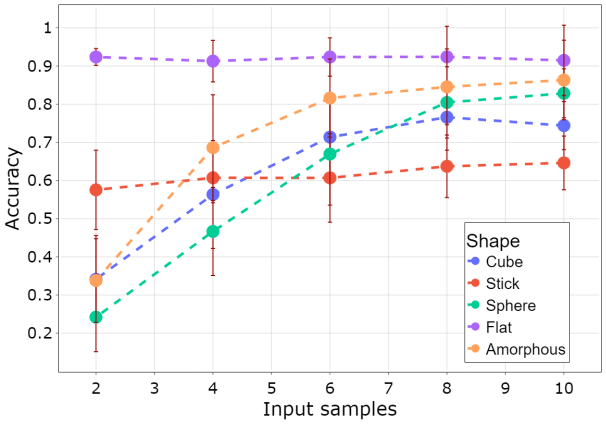

Figure 1. Average accuracy of GPT-4 for different number of samples in prompt taken from the training set. Sampling method: only target classes in prompt. Syntheses presented in the textual format. Colors correspond to different shapes of nanoparticles.

5.3. Text-to-image system

Prediction of a nanoparticle shape as a categorical variable based on the selected set of properties describing the synthesis procedure is inherently subject to information loss. Intuitively, images are much better representations of shapes than any handcrafted categories, and the full text of a nanoparticle synthesis carries more information compared to a set of numerical features extracted from it. Therefore, a text-to-image paradigm previously explored in general-purpose applications (Rombach et al., 2021) and other domains (Khwaja et al., 2022) looks appealing in the context of our problem. In the following, we attempt to prototype such a system to explore its potential despite the hard constraints on the sample size.

We break down the text-to-image system into three main components. The first one is the natural language processing model converting the text of a synthesis procedure to a vector of numerical features. The second component is the generative model with an encoder-decoder architecture

Table 4. Top performance achieved by the LLMs in prediction of nanomaterial shapes. Average accuracy corresponds to the following prompting strategy: only target classes in prompt, syntheses presented in the textual format, number of training examples $N = 8$.

	Mistral-medium	Mistral-small	Mistral-tiny	GPT-3.5-turbo	GPT-4	GPT-4-turbo
Cube	0.70±0.11	**0.76±0.08**	0.76±0.19	0.69±0.18	0.71±0.05	0.60±0.15
Stick	**0.71±0.04**	0.67±0.11	0.71±0.10	0.62±0.16	0.68±0.05	0.61±0.13
Sphere	0.38±0.12	0.77±0.18	0.62±0.24	0.63±0.15	**0.88±0.05**	0.44±0.12
Flat	0.89±0.08	**0.92±0.07**	0.81±0.17	0.90±0.06	0.90±0.10	0.91±0.06
Amorphous	0.70±0.15	0.88±0.08	0.53±0.16	0.80±0.13	0.87±0.12	**0.88±0.08**
Average accuracy	0.68±0.10	0.80±0.10	0.69±0.17	0.73±0.13	**0.81±0.07**	0.69±0.11

Table 5. Average accuracy for different prompting strategies of GPT-4 with $N = 4$ training examples: combinations of sampling methods and data formats.

Sampling method	At least one target class in prompt		Only target classes in prompt	
Data format	Textual	Tabular	Textual	Tabular
Cube	0.52±0.10	0.49±0.13	0.56±0.16	**0.67±0.11**
Stick	0.54±0.03	0.59±0.05	**0.61±0.10**	**0.61±0.05**
Sphere	0.43±0.16	0.43±0.14	**0.54±0.08**	0.32±0.10
Flat	0.88±0.11	0.86±0.08	0.92±0.05	**0.94±0.02**
Amorphous	0.57±0.20	0.37±0.15	**0.68±0.07**	0.59±0.15
Average	0.59±0.12	0.55±0.11	**0.66±0.09**	0.62±0.09

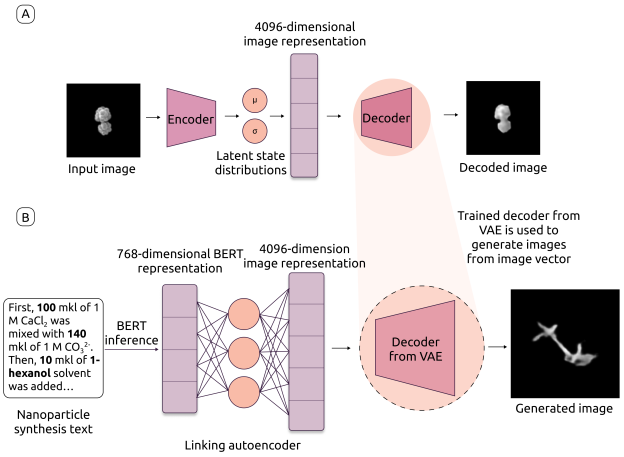

Figure 2. A schematic of the text-to-image system prototype. A) VAE training. The images of nanoparticles are used to train a variational autoencoder (VAE). B) Final model inference. The corresponding synthesis procedures are converted into vector representations with a pretrained BERT (bottom left). The "linking" autoencoder is trained to map text and image representations (bottom center). Finally, the decoder of the VAE is used to generate new images of nanomaterials based on the descriptions of syntheses (bottom right).

designed to learn representations of images of nanoparticles. Finally, the third component is the "linking" model translating the text representations into the image representations. When combined, the three components make a generative system capable of drawing the morphology of a nanomaterial based on the description of its synthesis (Figure 2).

5.3.1. Natural language processing model

The main requirement for the NLP model used for feature extraction was the ability to retain information about the qualitative and the quantitative features of a synthesis. In order to select the NLP model, we formulated several classification and regression tasks related to the key features of a synthesis procedure. We used the linear evaluation setup with standard metrics (Kolesnikov et al., 2019) to compare several pretrained transformer-based models. We found that the classic BERT model (Devlin et al., 2018) achieved perfect scores in most tasks and, therefore, used BERT as the feature extractor in the text-to-image setup (Figure 2). It also met the requirement of being relatively lightweight, easy to start up and use.

5.3.2. Autoencoder-based generative model

The most widely spread deep learning model architectures capable of generating images are generative adversarial networks (GANs) (Goodfellow et al., 2014), variational autoencoders (VAEs) (Kingma & Welling, 2013), and diffusion models (Rombach et al., 2021; Ramesh et al., 2022). We opted for a variational autoencoder as a more stable and a more suitable solution for small datasets, given the limited amount of data available for training.

The central idea of autoencoders is to learn a compressed representation of the input data while solving a data reconstruction problem. Variational autoencoders also imply a certain probabilistic distribution in the input data, which

allows it to generate meaningful outputs by sampling the latent representation after the training is complete (Kingma & Welling, 2013). In order to plug the VAE into the text-to-image system, we first trained it on the set of SEM images and then froze the decoder part (Figure 2). Refer to Appendix A.5 and A.6 for tested VAE architectures and evaluation metrics used.

We validated the final VAE model by monitoring training losses and evaluation metrics (Figure 3), analyzing individual examples of reconstructed images and visualizing the space of learned representations allowing to distinguish different clusters of nanoparticle shapes (Appendix A.9).

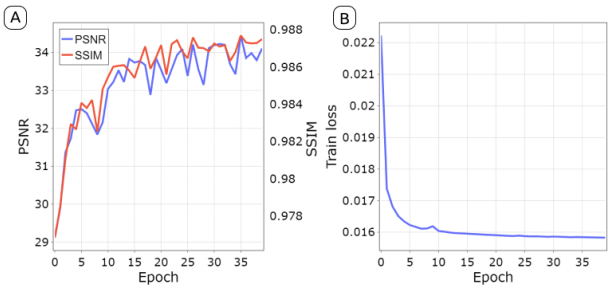

Figure 3. A) PSNR and SSIM metrics of the selected VAE by epoch. B) Training loss of the selected VAE by epoch.

5.3.3. "Linking" autoencoder model

The last component of the proposed text-to-image system is the "linking" neural network learning to map representations of the two modalities. Considering the data limitations and the empirical results described earlier, we refrained from using complex model architectures for this task. Instead, we developed another set of shallow autoencoder networks having from 3 to 8 linear layers. Like in the case of VAE, we optimized hyperparameters for each network, including the dimensionality of the latent space, to achieve the lowest reconstruction MSE (see Appendix A.7 for details on training and generation phases). The best architecture for the "linking" autoencoder is given in Appendix A.8.

5.3.4. Results

We observed that our prototype of the text-to-image system copes best with the generation of cubic nanoparticles, which was expected since the cubic shape was the most represented in the training data. For syntheses of this type of nanomaterials, the generated images were often distinct and well-shaped. It was also easier to grasp the size of cubic nanoparticles compared to other types. In general, however, the size of the dataset was insufficient to generate high quality images directly from text. Several examples of generated images are shown on Figure 5 of the Appendix A.8.

Despite the limited applicability of this prototype, we realized that repeated image generation based on the same synthesis parameters can provide insights into the polydispersity of NPs. Polydispersity is normally defined as $PdI = (\frac{\sigma}{2a})^2$, where σ is the standard deviation of the particle diameter, and a is the mean hydrodynamic radius. We performed 50 generations of amorphous NPs with the same synthesis parameters and observed maximal diameters ranging from 30 to 80 pixels. As polydispersity characterization is critical for many applications (Clayton et al., 2016), a generative model, such as the proposed prototype, could be instrumental in fast *in silico* screening of NPs by estimating PdI based on the predicted images.

6. Discussion and conclusion

In this work, we explored the potential of AI in predicting morphological properties of nanomaterials using the newly generated multimodal dataset of calcium carbonate nanoparticles. First, we investigated statistical associations between synthesis procedures and the resulting morphologies. Then, we trained and optimized tree-based ensemble models to predict multiple categories of nanomaterial shapes and sizes. After that, we systematically evaluated capabilities of the state-of-the-art LLMs in the same prediction tasks. Finally, we prototyped a text-to-image system to predict images of nanoparticles directly from the descriptions of syntheses.

Notably, this work stands out by creating a new dataset of multiple types of nanoparticle shapes, which can later be used for benchmarking. Also, to our knowledge, we are the first to train machine learning models to distinguish between nanomaterial shapes based on the synthesis parameters. Despite these achievements, there are still several unresolved issues in the field and certain limitations to the proposed models. A more detailed discussion and a comparison with previous works are offered in Appendix A.11 and Appendix A.10, respectively.

While text-to-image applications remain largely infeasible due to the limited data availability, we identified a huge potential for future LLM applications. Not only did we observe on par performance with the classical ensemble models, we also managed to collect evidence for the superior performance of LLMs, especially in the small data scenarios. Ensemble methods for LLMs now look as a promising research direction, as less computationally expensive models like Mistral-small approach the market leader's performance in the domain-specific tasks.

7. Data and code availability

All datasets, scripts and results described in this work are available in the [ANONYMIZED] repository for reproducibility and possible transfer learning applications.

References

AbdelHamid, A., Mendoza-Garcia, A., and Ying, J. Advances in and prospects of nanomaterials' morphological control for lithium rechargeable batteries. *Nano Energy*, 93, March 2022. ISSN 2211-2855. doi: 10.1016/j.nanoen.2021.106860. Publisher Copyright: © 2021 The Authors.

Baharifar, H. and Amani, A. Size, Loading Efficiency, and Cytotoxicity of Albumin-Loaded Chitosan Nanoparticles: An Artificial Neural Networks Study. *Journal of Pharmaceutical Sciences*, 106(1):411–417, January 2017. ISSN 00223549. doi: 10.1016/j.xphs.2016.10.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0022354916417922.

Banaye Yazdipour, A., Masoorian, H., Ahmadi, M., Mohammadzadeh, N., and Ayyoubzadeh, S. M. Predicting the toxicity of nanoparticles using artificial intelligence tools: a systematic review. *Nanotoxicology*, 17(1):62–77, January 2023. ISSN 1743-5390, 1743-5404. doi: 10.1080/17435390.2023.2186279. URL https://www.tandfonline.com/doi/full/10.1080/17435390.2023.2186279.

Behrmann, J., Grathwohl, W., Chen, R. T. Q., Duvenaud, D., and Jacobsen, J.-H. Invertible Residual Networks. 2018. doi: 10.48550/ARXIV.1811.00995. URL https://arxiv.org/abs/1811.00995. Publisher: arXiv Version Number: 3.

Buchnev, O., Grant-Jacob, J. A., Eason, R. W., Zheludev, N. I., Mills, B., and MacDonald, K. F. Deep-Learning-Assisted Focused Ion Beam Nanofabrication. *Nano Letters*, 22(7):2734–2739, April 2022. ISSN 1530-6984, 1530-6992. doi: 10.1021/acs.nanolett.1c04604. URL https://pubs.acs.org/doi/10.1021/acs.nanolett.1c04604.

Chen, F., Li, X., Zhou, Y., and Yang, D. Prediction of Size of Silver Nanoparticles Using Support Vector Machine and Artificial Neural Networks. *Journal of Computational and Theoretical Nanoscience*, 13(11):8666–8673, November 2016. ISSN 1546-1955. doi: 10.1166/jctn.2016.6028. URL http://www.ingentaconnect.com/content/10.1166/jctn.2016.6028.

Chen, H., Zheng, Y., Li, J., Li, L., and Wang, X. AI for Nanomaterials Development in Clean Energy and Carbon Capture, Utilization and Storage (CCUS). *ACS Nano*, 17(11):9763–9792, June 2023. ISSN 1936-0851, 1936-086X. doi: 10.1021/acsnano.3c01062. URL https://pubs.acs.org/doi/10.1021/acsnano.3c01062.

Chen, P., Tang, Z., Zeng, Z., Hu, X., Xiao, L., Liu, Y., Qian, X., Deng, C., Huang, R., Zhang, J., Bi, Y., Lin, R., Zhou, Y., Liao, H., Zhou, D., Wang, C., and Lin, W. Machine-Learning-Guided Morphology Engineering of Nanoscale Metal-Organic Frameworks. *Matter*, 2(6):1651–1666, June 2020. ISSN 25902385. doi: 10.1016/j.matt.2020.04.021. URL https://linkinghub.elsevier.com/retrieve/pii/S2590238520301922.

Clayton, K. N., Salameh, J. W., Wereley, S. T., and Kinzer-Ursem, T. L. Physical characterization of nanoparticle size and surface modification using particle scattering diffusometry. *Biomicrofluidics*, 10(5):054107, September 2016. ISSN 1932-1058. doi: 10.1063/1.4962992. URL https://doi.org/10.1063/1.4962992.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. doi: 10.48550/ARXIV.1810.04805. URL https://arxiv.org/abs/1810.04805. Publisher: arXiv Version Number: 2.

Fardo, F. A., Conforto, V. H., de Oliveira, F. C., and Rodrigues, P. S. A Formal Evaluation of PSNR as Quality Measurement Parameter for Image Segmentation Algorithms. 2016. doi: 10.48550/ARXIV.1605.07116. URL https://arxiv.org/abs/1605.07116. Publisher: arXiv Version Number: 1.

Fisher, R. A. On the Interpretation of 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87, January 1922. ISSN 09528385. doi: 10.2307/2340521. URL https://www.jstor.org/stable/2340521?origin=crossref.

Gao, C., Lyu, F., and Yin, Y. Encapsulated Metal Nanoparticles for Catalysis. *Chemical Reviews*, 121(2):834–881, January 2021. ISSN 0009-2665, 1520-6890. doi: 10.1021/acs.chemrev.0c00237. URL https://pubs.acs.org/doi/10.1021/acs.chemrev.0c00237.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Networks. 2014. doi: 10.48550/ARXIV.1406.2661. URL https://arxiv.org/abs/1406.2661. Publisher: arXiv Version Number: 1.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. 2015. doi: 10.48550/ARXIV.1512.03385. URL https://arxiv.org/abs/1512.03385. Publisher: arXiv Version Number: 1.

Hiszpanski, A. M., Gallagher, B., Chellappan, K., Li, P., Liu, S., Kim, H., Han, J., Kailkhura, B., Buttler,

D. J., and Han, T. Y.-J. Nanomaterial Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge. *Journal of Chemical Information and Modeling*, 60(6): 2876–2887, June 2020. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.0c00199. URL https://pubs.acs.org/doi/10.1021/acs.jcim.0c00199.

Iakovlev, V. Y., Krasnikov, D. V., Khabushev, E. M., Kolodiazhnaia, J. V., and Nasibulin, A. G. Artificial neural network for predictive synthesis of single-walled carbon nanotubes by aerosol CVD method. *Carbon*, 153:100–103, November 2019. ISSN 00086223. doi: 10.1016/j.carbon.2019.07.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0008622319307006.

Jablonka, K. M., Schwaller, P., and Smit, B. Is GPT-3 all you need for machine learning for chemistry? In *AI for Accelerated Materials Design NeurIPS 2022 Workshop*, 2022. URL https://openreview.net/forum?id=dgpgTEZ6G__.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2.

Kairdolf, B. A., Qian, X., and Nie, S. Bioconjugated Nanoparticles for Biosensing, in Vivo Imaging, and Medical Diagnostics. *Analytical Chemistry*, 89(2):1015–1031, January 2017. ISSN 0003-2700, 1520-6882. doi: 10.1021/acs.analchem.6b04873. URL https://pubs.acs.org/doi/10.1021/acs.analchem.6b04873.

Khwaja, E., Song, Y. S., and Huang, B. CELL-E: Biological Zero-Shot Text-to-Image Synthesis for Protein Localization Prediction. preprint, Bioinformatics, May 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.05.27.493774.

Kim, H., Han, J., and Han, T. Y.-J. Machine vision-driven automatic recognition of particle size and morphology in SEM images. *Nanoscale*, 12(37):19461–19469, 2020. ISSN 2040-3364, 2040-3372. doi: 10.1039/D0NR04140H. URL http://xlink.rsc.org/?DOI=D0NR04140H.

Kimmig, J., Schuett, T., Vollrath, A., Zechel, S., and Schubert, U. S. Prediction of Nanoparticle Sizes for Arbitrary Methacrylates Using Artificial Neuronal Networks. *Advanced Science*, 8(23):2102429, December 2021. ISSN 2198-3844, 2198-3844. doi: 10.1002/advs.202102429. URL https://onlinelibrary.wiley.com/doi/10.1002/advs.202102429.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. 2013. doi: 10.48550/ARXIV.1312.6114. URL https://arxiv.org/abs/1312.6114. Publisher: arXiv Version Number: 11.

Kolesnikov, A., Zhai, X., and Beyer, L. Revisiting Self-Supervised Visual Representation Learning. 2019. doi: 10.48550/ARXIV.1901.09005. URL https://arxiv.org/abs/1901.09005. Publisher: arXiv Version Number: 1.

Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V., and Ceder, G. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6(1):203, October 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0224-1. URL https://doi.org/10.1038/s41597-019-0224-1.

Kruskal, W. H. and Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621, December 1952. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1952.10483441. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441.

Lee, B., Yoon, S., Lee, J. W., Kim, Y., Chang, J., Yun, J., Ro, J. C., Lee, J.-S., and Lee, J. H. Statistical Characterization of the Morphologies of Nanoparticles through Machine Learning Based Electron Microscopy Image Analysis. *ACS Nano*, 14(12):17125–17133, December 2020. ISSN 1936-0851, 1936-086X. doi: 10.1021/acsnano.0c06809. URL https://pubs.acs.org/doi/10.1021/acsnano.0c06809.

Liu, R. and Lal, R. Potentials of engineered nanoparticles as fertilizers for increasing agronomic productions. *Science of The Total Environment*, 514:131–139, May 2015. ISSN 00489697. doi: 10.1016/j.scitotenv.2015.01.104. URL https://linkinghub.elsevier.com/retrieve/pii/S0048969715001266.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., and Ge, B. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. 2023. doi: 10.48550/ARXIV.2304.01852. URL https://arxiv.org/abs/2304.01852. Publisher: arXiv Version Number: 3.

Magnello, M. Karl Pearson, paper on the chi square goodness of fit test (1900). In *Landmark Writings in Western Mathematics 1640-1940*, pp. 724–731. Elsevier, 2005. ISBN 978-0-444-50871-3. doi: 10.1016/B978-044450871-3/50137-6. URL https://linkinghub.elsevier.com/retrieve/pii/B9780444508713501376.

Marsal, D. Introduction to the Analysis of Variance (ANOVA). In *Statistics for Geoscientists*, pp. 141–143. Elsevier, 1987. ISBN 978-0-08-026260-4. doi: 10.1016/B978-0-08-026260-4.50025-9. URL https://linkinghub.elsevier.com/retrieve/pii/B9780080262604500259.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 2023. doi: 10.1038/s41586-023-06735-9.

Nachar, N. The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology*, 4(1):13–20, March 2008. ISSN 1913-4126. doi: 10.20982/tqmp.04.1.p013. URL http://www.tqmp.org/RegularArticles/vol04-1/p013.

OpenAI. GPT-4 Technical Report. 2023. doi: 10.48550/ARXIV.2303.08774. URL https://arxiv.org/abs/2303.08774. Publisher: arXiv Version Number: 3.

OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2023.

Pedregosa, F., 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., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pellegrino, F., Isopescu, R., Pellutiè, L., Sordello, F., Rossi,

A. M., Ortel, E., Martra, G., Hodoroaba, V.-D., and Maurino, V. Machine learning approach for elucidating and predicting the role of synthesis parameters on the shape and size of TiO2 nanoparticles. *Scientific Reports*, 10(1): 18910, November 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-75967-w. URL https://www.nature.com/articles/s41598-020-75967-w.

Pomerantseva, E., Bonaccorso, F., Feng, X., Cui, Y., and Gogotsi, Y. Energy storage: The future enabled by nanomaterials. *Science*, 366(6468):eaan8285, November 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aan8285. URL https://www.science.org/doi/10.1126/science.aan8285.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. 2022. doi: 10.48550/ARXIV.2204.06125. URL https://arxiv.org/abs/2204.06125. Publisher: arXiv Version Number: 1.

Roccapriore, K. M., Ziatdinov, M., Cho, S. H., Hachtel, J. A., and Kalinin, S. V. Predictability of Localized Plasmonic Responses in Nanoparticle Assemblies. *Small*, 17(21):2100181, May 2021. ISSN 1613-6810, 1613-6829. doi: 10.1002/smll.202100181. URL https://onlinelibrary.wiley.com/doi/10.1002/smll.202100181.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. 2021. doi: 10.48550/ARXIV.2112.10752. URL https://arxiv.org/abs/2112.10752. Publisher: arXiv Version Number: 2.

Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., and Eliceiri, K. W. ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics*, 18 (1):529, December 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1934-z. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1934-z.

Sen Gupta, A. Role of particle size, shape, and stiffness in design of intravascular drug delivery systems: insights from computations, experiments, and nature. *WIREs Nanomedicine and Nanobiotechnology*, 8(2):255–270, March 2016. ISSN 1939-5116, 1939-0041. doi: 10.1002/wnan.1362. URL https://onlinelibrary.wiley.com/doi/10.1002/wnan.1362.

Serov, N. and Vinogradov, V. Artificial intelligence to bring nanomedicine to life. *Advanced Drug Delivery Reviews*, 184:114194, May 2022. ISSN 0169409X. doi: 10.1016/j.addr.2022.114194. URL https://linkinghub.elsevier.com/retrieve/pii/S0169409X22000849.

Shafaei, A. and Khayati, G. R. A predictive model on size of silver nanoparticles prepared by green synthesis method using hybrid artificial neural network-particle swarm optimization algorithm. *Measurement*, 151:107199, February 2020. ISSN 02632241. doi: 10.1016/j.measurement.2019.107199. URL https://linkinghub.elsevier.com/retrieve/pii/S0263224119310656.

Shahsavari, S., Bagheri, G., Mahjub, R., Bagheri, R., Radmehr, M., Rafiee-Tehrani, M., and Dorkoosh, F. Application of Artificial Neural Networks for Optimization of Preparation of Insulin Nanoparticles Composed of Quaternized Aromatic Derivatives of Chitosan. *Drug Research*, 64(03):151–158, September 2013. ISSN 2194-9379, 2194-9387. doi: 10.1055/s-0033-1354372. URL http://www.thieme-connect.de/DOI/DOI?10.1055/s-0033-1354372.

Shifrina, Z. B., Matveeva, V. G., and Bronstein, L. M. Role of Polymer Structures in Catalysis by Transition Metal and Metal Oxide Nanoparticle Composites. *Chemical Reviews*, 120(2):1350–1396, January 2020. ISSN 0009-2665, 1520-6890. doi: 10.1021/acs.chemrev.9b00137. URL https://pubs.acs.org/doi/10.1021/acs.chemrev.9b00137.

Singh, A. K. Experimental Methodologies for the Characterization of Nanoparticles. In *Engineered Nanoparticles*, pp. 125–170. Elsevier, 2016. ISBN 978-0-12-801406-6. doi: 10.1016/B978-0-12-801406-6.00004-2. URL https://linkinghub.elsevier.com/retrieve/pii/B9780128014066000042.

Smirnov, N. Estimate of deviation between empirical distribution functions in two independent samples. 2(2)(6.1, 6.2):3–16, 1939.

Smith, K. C. A. and Oatley, C. W. The scanning electron microscope and its fields of application. *British Journal of Applied Physics*, 6(11):391–399, November 1955. ISSN 0508-3443. doi: 10.1088/0508-3443/6/11/304. URL https://iopscience.iop.org/article/10.1088/0508-3443/6/11/304.

Soliman, M., Shalaby, K., Casettari, L., Bonacucina, G., Cespi, M., Palmieri, G. F., El Shamy, A., and Sammour, O. Determination of factors controlling the particle size and entrapment efficiency of noscapine in PEG/PLA nanoparticles using artificial neural networks. *International Journal of Nanomedicine*, pp. 4953, October 2014. ISSN 1178-2013. doi: 10.2147/IJN.S68737. URL https://doi.org/10.2147/IJN.S68737.

Sun, M., Fang, Q., Li, Z., Cai, C., Li, H., Cao, B., Shen, W., Liu, T. X., and Fu, Y. Co-precipitation synthesis of CuCo2O4 nanoparticles for supercapacitor electrodes with large specific capacity and high rate capability. *Electrochimica Acta*, 397:139306, November 2021. ISSN 00134686. doi: 10.1016/j.electacta.2021.139306. URL https://linkinghub.elsevier.com/retrieve/pii/S0013468621015966.

Takechi-Haraya, Y., Ohgita, T., Demizu, Y., Saito, H., Izutsu, K.-i., and Sakai-Kato, K. Current Status and Challenges of Analytical Methods for Evaluation of Size and Surface Modification of Nanoparticle-Based Drug Formulations. *AAPS PharmSciTech*, 23(5):150, July 2022. ISSN 1530-9932. doi: 10.1208/s12249-022-02303-y. URL https://link.springer.com/10.1208/s12249-022-02303-y.

Timoshenko, J., Lu, D., Lin, Y., and Frenkel, A. I. Supervised Machine-Learning-Based Determination of Three-Dimensional Structure of Metallic Nanoparticles. *The Journal of Physical Chemistry Letters*, 8(20):5091–5098, October 2017. ISSN 1948-7185, 1948-7185. doi: 10.1021/acs.jpclett.7b02364. URL https://pubs.acs.org/doi/10.1021/acs.jpclett.7b02364.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023. doi: 10.48550/ARXIV.2307.09288. URL https://arxiv.org/abs/2307.09288. Publisher: arXiv Version Number: 2.

Vaidyanathan, G. and Sendhilnathan, S. Characterization of Co1xZnxFe2O4 nanoparticles synthesized by co-precipitation method. *Physica B: Condensed Matter*, 403(13-16):2157–2167, July 2008. ISSN 09214526. doi: 10.1016/j.physb.2007.08.219. URL https://linkinghub.elsevier.com/retrieve/pii/S0921452607008216.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13 (4):600–612, April 2004. ISSN 1057-7149. doi: 10.1109/TIP.2003.819861. URL http://ieeexplore.ieee.org/document/1284395/.

Xu, Y., Xu, D., Yu, N., Liang, B., Yang, Z., Asif, M. S., Yan, R., and Liu, M. Machine Learning Enhanced Optical Microscopy for the Rapid Morphology Characterization of Silver Nanoparticles. *ACS Applied Materials & Interfaces*, 15(14):18244–18251, April 2023. ISSN 1944-8244, 1944-8252. doi: 10.1021/acsami.3c02448. URL https://pubs.acs.org/doi/10.1021/acsami.3c02448.

Yao, L., An, H., Zhou, S., Kim, A., Luijten, E., and Chen, Q. Seeking regularity from irregularity: unveiling the synthesis–nanomorphology relationships of heterogeneous nanomaterials using unsupervised machine learning. *Nanoscale*, 14(44):16479–16489, 2022. ISSN 2040-3364, 2040-3372. doi: 10.1039/D2NR03712B. URL http://xlink.rsc.org/?DOI=D2NR03712B.

Youshia, J., Ali, M. E., and Lamprecht, A. Artificial neural network based particle size prediction of polymeric nanoparticles. *European Journal of Pharmaceutics and Biopharmaceutics*, 119:333–342, October 2017. ISSN 09396411. doi: 10.1016/j.ejpb.2017.06.030. URL https://linkinghub.elsevier.com/retrieve/pii/S0939641117303570.

Zaki, M. R., Varshosaz, J., and Fathi, M. Preparation of agar nanospheres: Comparison of response surface and artificial neural network modeling by a genetic algorithm approach. *Carbohydrate Polymers*, 122:314–320, May 2015. ISSN 01448617. doi: 10.1016/j.carbpol.2014.12.031. URL https://linkinghub.elsevier.com/retrieve/pii/S0144861714012272.

Zebarjadi, M., Esfarjani, K., Bian, Z., and Shakouri, A. Low-Temperature Thermoelectric Power Factor Enhancement by Controlling Nanoparticle Size Distribution. *Nano Letters*, 11(1):225–230, January 2011. ISSN 1530-6984, 1530-6992. doi: 10.1021/nl103581z. URL https://pubs.acs.org/doi/10.1021/nl103581z.

Zheng, Z., Rong, Z., Rampal, N., Borgs, C., Chayes, J. T., and Yaghi, O. M. A gpt-4 reticular chemist for guiding mof discovery. *Angewandte Chemie International Edition*, 62(46):e202311983, 2023.

A. Appendix

A.1. Details on statistical testing

The Kolmogorov-Smirnov test (Smirnov, 1939). If the null hypothesis is rejected, this test indicates that the two samples are not drawn from the same distribution, that is, the two samples of a synthesis parameter differ in the presence or absence of a particular class of nanoparticles. Let $(X_1^1, X_2^1, \ldots, X_m^1)$ be independent, identically distributed real values of a parameter of a synthesis that produces cubic nanoparticles with the common cumulative distribution function $F_{1,n}$. Let $(X_1^2, X_2^2, \ldots, X_m^2)$ be independent, identically distributed real values of the same parameter of a synthesis which always results in nanoparticles of different shapes with the common cumulative distribution function $F_{2,m}$. The Kolmogorov–Smirnov statistic in this case is: $D_{n,m} = sup_x|F_{1,n}(x^1) - F_{2,m}(x^2)|$, where sup is the supremum function.

The null hypothesis is that the two samples are from the same continuous distribution. The null hypothesis is rejected at level $\alpha = 0.05$ if $D_{n,m} > \sqrt{-ln(\alpha/2)(1 + m/n)/2m}$. We applied this test for each of the real-valued parameters of synthesis and each type of nanomaterial shape and used the Bonferroni correction method similarly to the previous tests. The results of this test were similar to those of the previous two, except that in the case of stick-shaped nanoparticles, the dependence was observed on the parameter characterizing the mass of the polymer rather than its concentration, which is not surprising given the similar nature of these two parameters.

ANOVA (Marsal, 1987) was used to compare distributions of continuous parameters corresponding to different shapes of nanomaterials. Let $(X_1^i, X_2^i, \ldots, X_n^i)$ be independent, identically distributed real values of a parameter of a synthesis that produces nanoparticles of a specific shape with the common cumulative distribution function $F_{i,n}$with the mean $\overline{X}^i$. The formula for the one-way ANOVA F-test statistic is: $F = \frac{\sum_{i=1}^{K} n_i(\overline{X}^i - \overline{X})^2/(K-1)}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_j^i - \overline{X}^i)^2/(N-K)}$, where $\overline{X}^i$ denotes the sample mean in the i-th group, n_i is the number of observations in the i-th group, $\overline{X}$ denotes the overall mean of the population, and K denotes the number of groups, where X_j^i is the j^{th} observation in the i^{th} out of K groups and N is the overall sample size. The null hypothesis can be formulated as follows: $\overline{X}^i = \overline{X}^j$, for each two groups i and j. If F-statistic is greater than critical p-value (at the significance level $\alpha = 0.05$), then the null hypothesis is rejected and distributions of this synthesis parameter in the case of at least two different shapes are different. We applied this test for each of the real-valued parameters of synthesis and each type of nanomaterial shape and used the Bonferroni correction method similarly to the previous tests. The results of this test were consistent with the results of the first two tests, except that it failed to confirm the relationship between the shape of nanoparticle and polymer mass, although polymer concentration was still a significant parameter. All major associations between features of the synthesis and the corresponding shapes of nanoparticles are presented in Table 6.

Table 6. Significant associations between features of the synthesis and the corresponding shapes of nanoparticles. The table shows the parameters that turned out to be determinant in the synthesis of nanomaterials of one or another shape. For continuous features, the following tests were used: Mann–Whitney U test, Kruskal–Wallis H test, Kolmogorov–Smirnov test, ANOVA. Fisher exact and Chi-squared tests was used for categorical features.

Shape	Stick-shaped	Spherical	Flat	Cubic	Amorphous
Continuous features	Temperature, C Synthesis time Polymer, % wt. Polymer Mwt, kDa	Solvent, % vol.	-	Polymer, % wt. Temperature, C Polymer Mwt, kDa	-
Categorical features	Sodium dodecylsulfate PEG PEI	Myristyltrimethylammonium bromide	PAA PSS PVP	Sodium dodecylsulfate PAA PSS Polymer absence	Sodium dodecylsulfate Isopropyl alcohol tert-Butanol Propylene glycol

A.2. Tree-based ensemble models

In order to achieve the best results for each of the Random Forest and Gradient Boosted Trees models, we optimized the hyperparameters for each of them, and built the models with different splits of the original dataset. The final metrics for each model were calculated by predicting at 5 different random states, after which the mean value as well as the standard deviation were calculated. Most of the functions used to prepare the dataset and use the models were implemented using the

scikit-learn library (Pedregosa et al., 2011).

In case of the Random Forest, optimization of the following parameters was performed: `n_estimators`, `max_features`, `max_depth`, `min_samples_leaf`, `max_leaf_nodes`. In case of Gradient Boosted Trees, the optimized parameters were: `gamma`, `colsample_bytree`, `max_depth`, `n_estimators`, `learning_rate`. Hyperparameter optimization was performed using 5-fold cross-validated grid-search. Given that some target classes were underrepresented, we prepared three test sets in advance for a more thorough assessment of performance. The test sets contained 33%, 20% or 15% of the total number of samples. A summary Table 7 provides motivation for testing several data splits. In our case, the lowest mean standard deviation was observed in 33% test split for both, accuracy and F1 score, among all the experiments. Also, for each model, the optimal threshold was found to solve the problem of class imbalance. This was achieved by balancing precision and recall metrics.

Table 7. Comparison of different data splits. A representative test set was obtained with 33% of total number of samples.

Validation subset size of dataset, %	Average accuracy	Average standard deviation of accuracy	Average F1 score	Average standard deviation of F1 score
33	**0.74**	**0.11**	**0.53**	**0.09**
20	0.69	0.13	0.51	0.11
15	0.65	0.18	0.49	0.15

For the best models, we also performed feature importance analysis by constructing SHAP diagrams showing the most important features in model performance. Figure 4 below shows the 10 most important features for the Random Forest model with optimal parameters predicting the stick shaped nanomaterials. Among these features, statistical relationship with the given shape of nanomaterials was confirmed for the following features: 'Temperature, C', 'Synthesis time', 'Polymer, % wt.', 'Polymer Mwt, kDa', 'Sodium dodecylsulfate', 'PEI'. This is an additional validation of our models, as the results of the analysis of feature importance almost completely correspond to the previously discovered statistical patterns that were presented in Table 6.

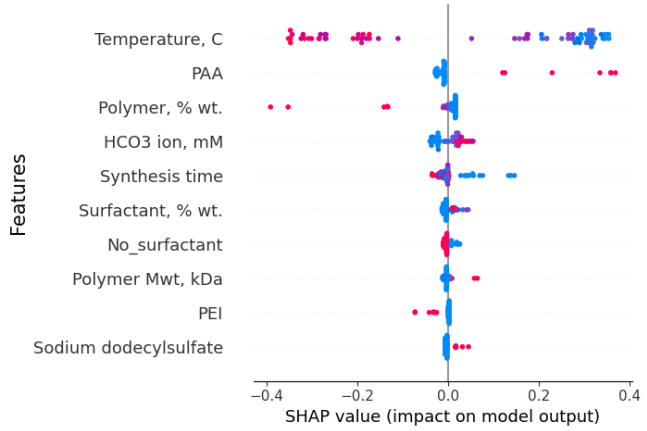

Figure 4. Results of feature importance analysis in the form of SHAP values for the top 10 features for the best Random Forest model for predicting stick shaped nanoparticles.

A.3. Texts of synthesis procedures and prompts

In order to make predictions of the morphology of nanomaterials based on their synthesis text using LLMs, special templates were created, which were then used to be filled with parameters for a particular synthesis. From these, the final textual prompt for LLM was compiled. These templates were similarly used in the development of a generative text-to-image system. Two examples of such templates are given below.

Template example 1:

"Synthesis was carried out using the co-precipitation technique. Initially, ca_conc mkl of 1 M CaCl2 was combined with

pol_vol mkl of pol_conc % wt. polymer polymer having a molecular weight of pol_mass kDa. Subsequently, solvent_volume mkl of solvent was introduced, and the volume adjusted to 500 mkl using distilled water. Following that, co3_conc mkl of 0.1 M Na2CO3 was mixed with hco3_conc mkl of 0.1 M NaHCO3, along with surf_vol mkl of surf_conc % wt. surfactant serving as the surfactant. Another solvent_volume mkl of solvent was added, and the volume adjusted to 500 mkl using distilled water. Two resulting solutions, both heated to r_temp C prior to the reaction, were combined under continuous stirring at stir_ratio rpm while maintaining the temperature. The reaction proceeded for r_time min, followed by centrifugation."

Template example 2:

"All materials were synthesized via the co-precipitation technique. In the first step, ca_conc mkl of 1 M CaCl2 was combined with pol_vol mkl of pol_conc % wt. polymer polymer, characterized by a molecular weight of pol_mass kDa. This was followed by the addition of solvent_volume mkl of solvent, and the volume was adjusted to 500 mkl using distilled water. In the subsequent step, co3_conc mkl of 0.1 M Na2CO3, hco3_conc mkl of 0.1 M NaHCO3, and surf_vol mkl of surf_conc % wt. surfactant surfactant were combined. Once more, solvent_volume mkl of solvent was added, and the volume was adjusted to 500 mkl using distilled water. Finally, two solutions, both heated to r_temp C before the reaction, were mixed under stirring at stir_ratio rpm while maintaining the temperature. The reaction proceeded for r_time min, followed by centrifugation."

Below is also one of the text-based prompts that were given to the model before predicting the morphology of nanomaterials on the test subset.

"You are an expert in the synthesis of nanomaterials. You analyze the conditions for obtaining a nanomaterial and predict what particle shapes will be present in the synthesized material. There are five particle shapes: 'Cube', 'Stick', 'Sphere', 'Flat' and 'Amorphous'. A nanomaterial can contain particles of different shapes. If you cannot say exactly what it is, list the forms that have the highest probability in those conditions.

CaCO3 nanoparticles were synthesized by the co-precipitation approach according to the following manner. In separate burettes two solutions were made, 57 mkl of 1 M CaCl2 and 20 mkl of 0.155 % wt. PEI with molecular weight of 25.0 kDa were mixed in 200.0 mkl of 1-Hexanol before dilution with distilled water up to 500 mkl. Similarly, 140 mkl of 0.1 M Na2CO3 and 200 mkl of 0.1 M of NaHCO3 were combined with 20 mkl of 0.43 % wt. Myristyltrimethylammonium bromide and 200.0 mkl of 1-Hexanol. Then, the solution was also diluted in 500 mkl of water. Both solutions were heated up to 68 C right before mixing under stirring at 1000 rpm for 8 min 0 sec min following centrifugation.

Answer: 'Cube, Stick'"

An example is also given for the case of prompts that used tabular data. In this case, only the way the synthesis was presented differed, but the overall structure of the prompt remained the same.

"You are an expert in the synthesis of nanomaterials. You analyze the conditions for obtaining a nanomaterial and predict what particle shapes will be present in the synthesized material. There are five particle shapes: 'Cube', 'Stick', 'Sphere', 'Flat' and 'Amorphous'. A nanomaterial can contain particles of different shapes. If you cannot say exactly what it is, list the forms that have the highest probability in those conditions.

Ca ion, mM: 148; CO3 ion, mM: 0; HCO3 ion, mM: 100; Polymer Mwt, kDa: 0.0; Polymer, % wt.: 0.0; Surfactant, % wt.: 0.0; Solvent, % vol.: 0.0; Stirring, rpm: 0; Temperature, C: 31; Synthesis time: 129; Hexadecyltrimethylammonium bromide: 0; Myristyltrimethylammonium bromide: 0; No_surfactant: 1; Sodium dodecylsulfate: 0; Triton X-100: 0; 1-Hexanol: 0; Dimethylformamide: 0; Ethylene glycol: 0; Isopropyl alcohol: 0; Methyl alcohol: 0; No_solvent: 1; Propylene glycol: 0; tert-Butanol: 0; No_polymer: 1; PAA: 0; PEG: 0; PEI: 0; PSS: 0; PVP: 0

Answer: 'Flat'"

A.4. Few-shot classification

To optimize the number of input examples and the proportion of the test subset, experiments were conducted with the GPT-4 model for the text subset. The table below summarizes these results (Table 8). The low impact of the proportion of the test subset is obvious, but the number of input examples has a significant impact on the metrics.

A.5. VAE: implementation details

We have experimented with several ResNet architectures but also developed a few custom architectures for the VAE. ResNet is the classical convolutional neural network originally proposed for the classification tasks (He et al., 2015). It consists

Table 8. Average accuracy of GPT-4 for different number of input samples in prompt N taken from the training set. Sampling method: only target classes in prompt. Syntheses presented in textual format.

Input samples	Test subset size	Shape					Average accuracy
		Cube	Stick	Sphere	Flat	Amorphous	
2	0.15	0.29±0.14	0.58±0.15	0.27±0.10	0.92±0.02	0.33±0.12	0.48±0.10
	0.2	0.40±0.09	0.56±0.12	0.20±0.10	0.92±0.03	0.38±0.15	0.49±0.10
	0.33	0.33±0.12	0.59±0.05	0.25±0.07	0.93±0.02	0.30±0.07	0.48±0.07
4	0.15	0.54±0.15	0.64±0.05	0.38±0.14	0.93±0.01	0.64±0.23	0.63±0.12
	0.2	0.59±0.13	0.56±0.06	0.49±0.12	0.89±0.09	0.73±0.12	0.65±0.10
	0.33	0.56±0.15	0.62±0.09	0.54±0.09	0.92±0.05	0.68±0.07	0.66±0.09
6	0.15	0.76±0.08	0.65±0.11	0.63±0.14	0.94±0.00	0.79±0.14	0.75±0.09
	0.2	0.67±0.08	0.56±0.13	0.68±0.17	0.90±0.08	**0.88±0.03**	0.74±0.10
	0.33	0.71±0.17	0.61±0.11	0.70±0.10	0.93±0.07	0.78±0.14	0.74±0.12
8	0.15	**0.84±0.07**	0.62±0.11	0.81±0.10	**0.96±0.03**	0.84±0.08	**0.81±0.08**
	0.2	0.72±0.08	0.67±0.05	0.80±0.10	0.91±0.08	0.87±0.10	0.80±0.08
	0.33	0.74±0.11	0.63±0.08	0.80±0.08	0.90±0.13	0.83±0.12	0.78±0.10
10	0.15	0.78±0.06	0.59±0.11	0.84±0.07	0.94±0.03	**0.88±0.07**	**0.81±0.07**
	0.2	0.71±0.05	**0.68±0.05**	**0.88±0.05**	0.90±0.10	0.87±0.12	**0.81±0.07**
	0.33	0.74±0.07	0.66±0.06	0.77±0.07	0.90±0.15	0.84±0.12	0.78±0.09

Table 9. Comparison of computational resources of LLMs: time per one complete experiment (in case of text dataset, an average prompt was around 3000 tokens), price per 1M tokens in USD, limits for requests per minute and 1000 tokens per minute.

	Mistral-medium	Mistral-small	Mistral-tiny	GPT-3.5-turbo	GPT-4	GPT-4-turbo
Time per experiment, s	46	21	19	44	63	53
Input price per 1M tokens, USD	2.7	0.7	0.1526	0.5	30	10
Tokens limit, 1000/min	2000	2000	2000	60	10	150
Requests limit, 1/min	120	120	120	500	500	500

of several blocks of convolutional, batch normalization and ReLU layers, and several depth options are available. Jens Behrmann et al. showed that invertible ResNets can also be used as generative models (Behrmann et al., 2018) and, therefore, we used the reversed ResNet from PyTorch Lightning Bolts[1] as a decoder for the VAE. Additionally, we developed several shallow networks varying the number of convolutional blocks and the dimensionality of the bottleneck layer as custom VAE architectures.

We trained all the architectures in the grid search setup optimizing several hyperparameters, such as batch size, learning rate, Kullback-Leibler (KL) divergence coefficient and image size, to achieve the lowest BCE loss.

Based on the training losses and the evaluation metrics described in Appendix A.6, we selected one of the custom architectures as the best. We failed to achieve on-par performance with the ResNet backbones, likely due to insufficient number of training examples. The top-performant VAE architecture had only 4 convolutional blocks for the encoder and 4 upsampling blocks for the decoder with 4096 dimensions in the latent space. The optimal set of hyperparameters was 128×128 for the image size, 64 for the batch size, 0.001 for the learning rate, and 0.01 for the KL divergence coefficient. The corresponding training curves are depicted on Figure 3.

[1]https://lightning-bolts.readthedocs.io/en/latest/

A.6. VAE: metrics

For each combination of hyperparameters, the trained VAEs were evaluated on the test set. Two metrics reflecting the similarity of the original and the decoded images were used to compare architectures: structural similarity index measure (SSIM) (Wang et al., 2004) and peak signal-to-noise ratio (PSNR) (Fardo et al., 2016). SSIM is a standardized measure of the difference between the compressed and the original image, ranging from -1 to 1. It is defined by the following formula:

$$SSIM(X,Y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where X, Y are the images, x, y are the mean pixel values of the images, x, y are the variances of the pixel values, xy is the covariance, and c_1, c_2 are the coefficients stabilizing the division. PSNR is a simpler metric showing the ratio of the contribution of the maximum value of the original image to the contribution of noise in the compressed image. This metric is calculated using the following formula:

$$PSNR(X,Y) = 20log_{10}(\frac{max(X)}{\sqrt{e(X,Y)}})$$

where X is the original image of size m×n, Y is the compressed image of the same size, and $e(X,Y) = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ji} - y_{ji})^2$ is the mean squared deviation between the pixels of two images.

A.7. "Linking" VAE: training and generation phases

The training process was organized into the following key steps. For each training example:

1. Choose a text template of a synthesis procedure randomly and fill in the corresponding experimental parameters.
2. Obtain text representations with a pretrained BERT.
3. Obtain image representations with a pretrained VAE.
4. Perform a forward pass to convert text representations into image representations.
5. Calculate the loss and backpropagate the error.

After the training is done, morphology of a new nanomaterial described in a synthesis procedure can be predicted as follows:

1. Obtain representations of the synthesis procedure text with a pretrained BERT.
2. Apply the "linking" autoencoder to predict the corresponding image representations.
3. Apply the decoder part of the VAE to predict the image of the nanomaterial.

A.8. "Linking" VAE: best architecture

The final architecture of the "linking" VAE is shown on Figure 5A. It consists of 4 linear layers and has 768-dimensional latent space. The optimal hyperparameters were 8 for the batch size, 0.00001 for the learning rate. Figure 5B shows individual examples of nanoparticles reconstructed and generated from texts. Three different shapes are given.

A.9. Data visualization

We visualized the space of learned representations of the VAE to validate the model and gain additional insights into various dependencies between the features and the target classes. For that, we used UMAP to compress the bottleneck 4096 dimensions to 2D (Figure 6). Each dot represents a single representative nanoparticle from one of the 215 syntheses. We observed five clusters having 2-3 particular shapes as the most prominent. Based on the literature and the statistical evaluation, we expected to see drastic differences in temperatures for different clusters. However, we could observe a single bottom-right cluster having lower synthesis temperatures on average.

A.10. Comparison with previous works

A comparison with the most relevant previous works is given in Table 10.

Table 10. Comparison with other works. *a prototype of the text-to-image system

Prediction task	Data points	Number of shapes	Best metric	Availability	Generative design	Reference
Size	103	1	MAPE = 0.70%	Only dataset	No	(Shafaei & Khayati, 2020)
Size	98	1	MAPE = 4%	-	No	(Iakovlev et al., 2019)
Size	26	1	MAPE = 9.10%	-	No	(Pellegrino et al., 2020)
Size and shape	215	5	Accuracy = 0.93	Code and dataset	Yes*	Our work

A.11. Discussion on limitations

Our ML models were trained to predict 5 types of nanomaterial shapes, some of which were underrepresented (Table 2). This limitation can be mitigated by either adjusting the prediction threshold, or oversampling techniques. Ultimately, this issue can only be resolved by expanding the dataset for underrepresented classes. In the context of the text-to-image system, we always refer to a prototype acknowledging its limitations, such as low diversity of generated images and their quality, caused by the limited training examples available. Therefore, most of the barriers to training a more universal and accurate model for prediction of nanomaterial morphology are related to insufficient quality and number of existing datasets. There is currently no unified database with syntheses and properties of different nanoparticles that is well documented and publicly available. Therefore, applied AI researchers have to resort to small single study datasets or larger datasets of a single experimental system extensively studied in the past (Table 10). Both approaches impose severe limitations on machine learning and, even more so, on deep learning applications that typically require a lot more training data. Thus, a collective effort towards assembling a curated database of nanomaterials with deep characterization of their properties is long overdue.

Additional challenges arise from the data preprocessing steps dealing with SEM. Many syntheses result in numerous overlaying NPs on a single SEM image, such that it is difficult even for a human eye to distinguish between individual NP units. Since image segmentation methods have already reached quite an advanced level, we anticipate major breakthroughs rather on the experimental and the imaging technology side.

A.12. Computing infrastructure

Table 11. Computing infrastructure used for study experiments.

CPU	AMD Ryzen 7 3700X 3.60 GHz 8-Core Processor
GPU	NVIDIA GeForce RTX 3090 24 GB of GPU memory
RAM	32.0 GB
Operating system	Windows 11 Pro N
Python	3.9

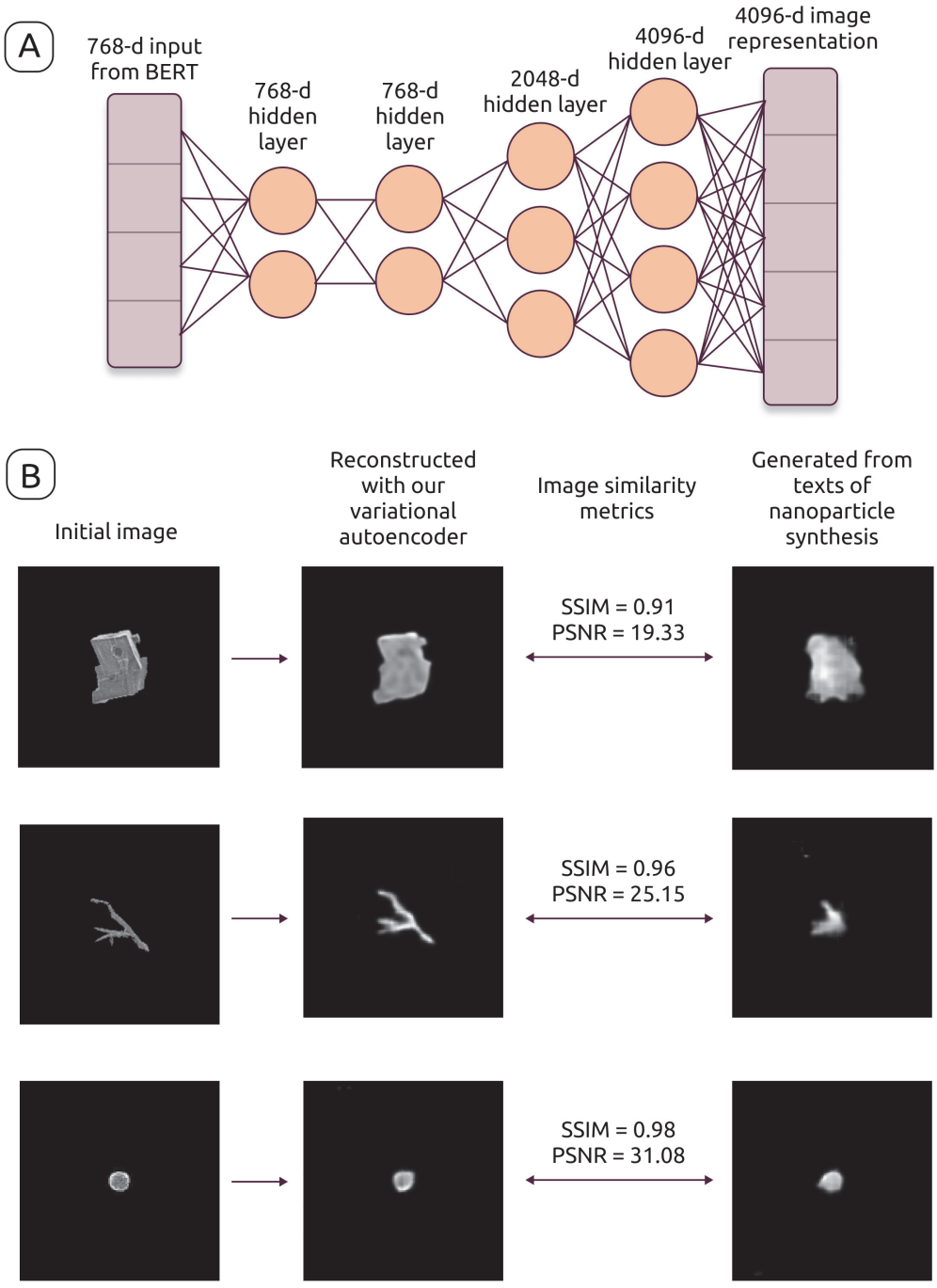

Figure 5. A) The encoder-decoder architecture of the linking neural network. B) Comparison of real images to their VAE reconstructions and the images generated from the corresponding synthesis texts.

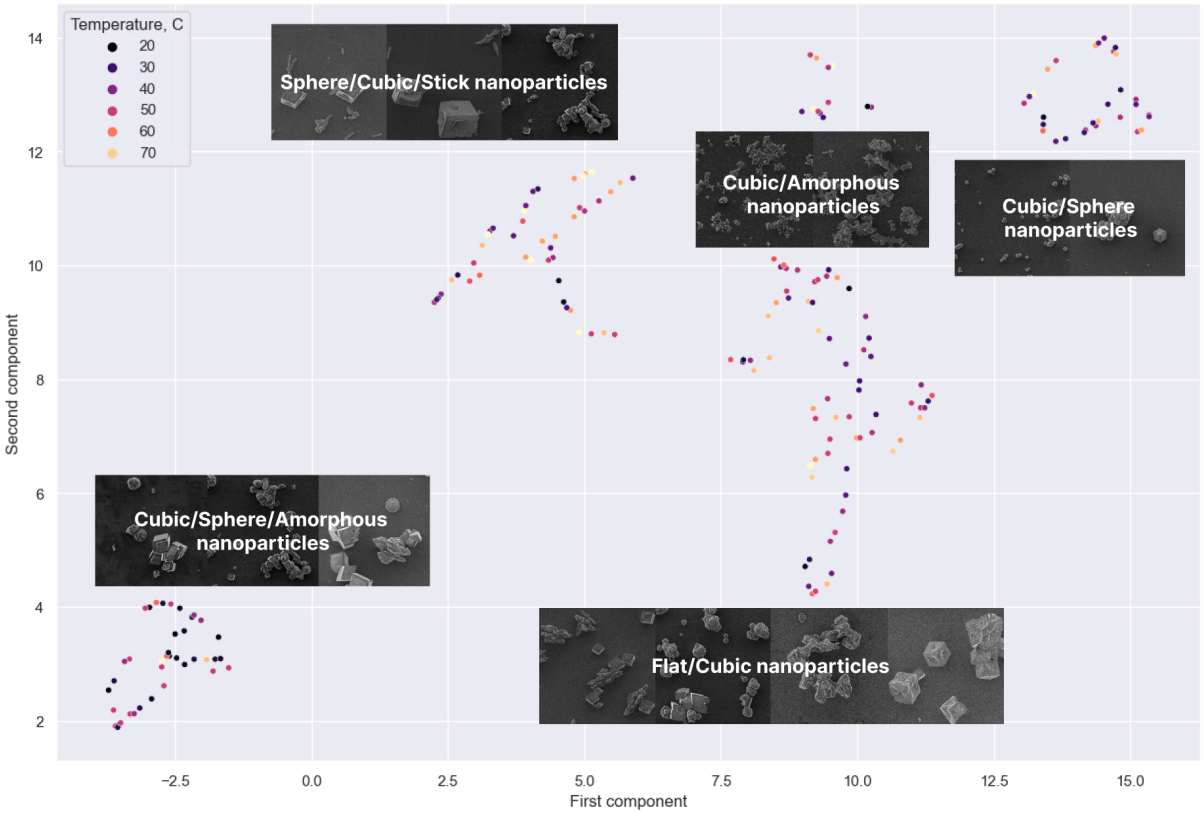

Figure 6. Representation of the latent space of the variational autoencoder trained on our image dataset. Colors indicate the shapes of the nanomaterials, and the axes are the UMAP components after dimensionality reduction.