SuperDiff: Diffusion Models for Conditional Generation of Hypothetical New Families of Superconductors

Samuel Yuan[*]

Homestead High School, Cupertino, CA 95014, USA

S.V. Dordevic[†]

Department of Physics, The University of Akron, Akron, OH 44325, USA

(Dated: February 6, 2024)

Effective computational search holds great potential for aiding the discovery of High-Temperature Superconductors (HTSs), especially given the lack of systematic methods for their discovery. Recent progress has been made in this area with machine learning, especially with deep generative models, which have been able to outperform traditional manual searches at predicting new superconductors within existing superconductor families but have yet to be able to generate completely new families of superconductors. We address this limitation by implementing conditioning—a method to control the generation process—for our generative model and develop SuperDiff, a Denoising Diffusion Probabilistic Model (DDPM) with Iterative Latent Variable Refinement (ILVR) conditioning for HTS discovery—the first deep generative model for superconductor discovery with conditioning on reference compounds. With SuperDiff, by being able to control the generation process, we were able to computationally generate completely new families of hypothetical superconductors for the very first time. Given that SuperDiff also has relatively fast training and inference times, it has the potential to be a very powerful tool for accelerating the discovery of new superconductors and enhancing our understanding of them.

I. INTRODUCTION

Superconductors exhibit zero resistivity and perfect diamagnetism. These traits lend them useful for various important technologies, including Maglev trains, MRI magnets, power transmission lines, and quantum computers. However, a major current limitation is that the superconducting transition temperatures (T_c) of all known superconductors at ambient pressures are well below room temperature, restricting their broader practical application. Consequently, the search for superconductors with higher T_c is a very active field, as they have significant potential to considerably improve the efficiency of current technologies while also enabling new ones.

Currently, however, superconductivity in high T_c superconductors is not very well understood. As a result, there exists no systematic method for searching for new high T_c superconductors [1], and the most common method for searches for new high T_c superconductors is essentially trial-and-error. For instance, the study in Ref. [2] surveyed approximately 1000 compounds over four years, of which they found only about 3% to be superconducting. That study is a testament to the extreme inefficiency of finding new high T_c superconductors through pure manual search.

Understanding this, more recently, computational techniques have been applied to assist researchers in the search for new high T_c superconductors. Specifically, a number of works have applied machine learning to this search for superconductors. Although serving as very valuable tools in many respects, most of these attempts [3–5], have been limited to classification and regression models, which only search through existing databases and are not able to generate any new compounds. Only recently, with deep generative models applied to superconductor discovery, have new hypothetical superconductors not found in most popular compound datasets been generated [6–8]. In Ref. [6], a Generative Adversarial Network (GAN) [9] was applied for unconditional high T_c superconductor generation, and in Ref. [7], a Crystal Diffusion Variational Autoencoder (CDVAE) [10] was also applied for unconditional superconductor generation so that crystal structure could be accounted for. However, Ref. [7] was applied on a different dataset and focused on a different task of generating stoichiometric Bardeen–Cooper–Schrieffer (BCS) conventional superconductors [11] and so did not generate any superconductors with $T_c \gtrsim 20$ K.

New attempts at high T_c superconductor discovery with generative models are not without limitations, however. Most notably, although past models have been able to successfully generate new superconductors within existing superconductor families, they have not been able to generate completely *new families* of superconductors, which would be particularly desirable. This is because they are only unconditional models, which learn only the training dataset distribution. As unconditional models, the generation process of these models cannot be controlled. In other words, past models lack conditioning functionality—a method for controlling the generation process, that, in this context, means giving an example superconductor, the reference compound, and having the model generate similar superconductors, ideally by inter-

---
[*] sdkyuan@gmail.com
[†] dsasa@uakron.edu

polating between the example and what the model has learned from the training dataset. With conditioning, the possibility of generating new families of superconductors can be opened, and researchers can be given control over the generation process. This can be especially useful for researchers looking to find only specific types of superconductors or expand on their own new discoveries. Parallel to our work, Ref. [8] also applied a diffusion model for high T_c superconductor discovery; however, their model was, like previous GANs, greatly limited by its lack of support for conditional generation with reference compounds—which is our main focus. Thus, their diffusion model shared with previous models the major limitation of being unable to generate any *new families* of superconductors—essentially, their work was only recreating the performance of the GAN in Ref. [6] but with a diffusion model instead and added T_c label control only. Once again, we note that, in this work, we consider "conditioning" to mean conditioning the model on reference compounds only, as only this allows for the controlled generation of known and new families of superconductors. Moreover, Ref. [6] also struggled at generating unique (distinct from others in the given generated set) pnictides because of the small number of pnictides in SuperCon, the training dataset.

To resolve these limitations, in this work, we implement a Denoising Diffusion Probabilistic Model (DDPM) [12, 13] for superconductor generation as our unconditional model and further implement conditioning with the Iterative Latent Variable Refinement (ILVR) [14] extension to DDPM, which allows for one-shot generation without additional training. With conditioning, we hope to be able to generate *new families* of superconductors for the first time, as identified by the clustering analysis proposed in Ref. [15], by experimenting with feeding the model different reference superconductors—this would mark a leap in the capabilities of computational searches for superconductors.

Diffusion models are a class of deep generative models that are inspired by nonequilibrium thermodynamics [13] and have recently shown superior performance and outperformed GANs in image synthesis [16] and materials discovery [17]. Diffusion Models are also at the heart of popular new image generation software, such as DALL·E 2 [18] and Stable Diffusion [19]. More recently, these models have also been implemented and shown considerable promise for a variety of scientific applications, such as for drug discovery [20].

We coin this first approach to conditionally generating new superconductors with reference compounds as "SuperDiff". With SuperDiff, we aim to resolve the issues found in past works as a result of the small pnictide training dataset with the unconditional DDPM and, as our main focus, explore how the conditional DDPM can adapt to new information to generate completely new families of superconductors for the first time. This paper is structured as follows: in Section II, we give an overview of the underlying models behind SuperDiff and detail how
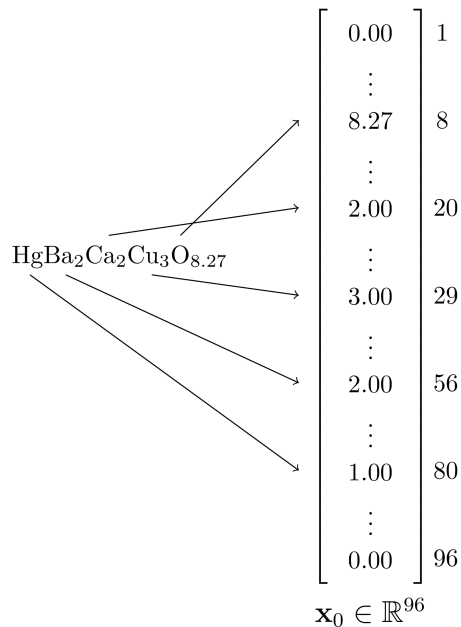

FIG. 1. The column vector encoding method used. The figure shows the chemical composition of $HgBa_2Ca_2Cu_3O_{8.27}$ being encoded as a vector in $\mathbb{R}^{96}$ which is fed to the diffusion model as $\mathbf{x}_0$.

they were applied; in Section III, we present the main results of the study, most notably a selection of new families of superconductors generated by SuperDiff—a first for the computational search for superconductors; and in Section IV, we summarize our contributions and most important findings.

II. METHODOLOGY

As stated in the introduction, we leverage the capabilities of Denoising Diffusion Probabilistic Models and Iterative Latent Variable Refinement to propose a method for conditionally generating new hypothetical superconductors. Here, we discuss the details of the creation of SuperDiff. In Section II A, we discuss the sourcing and processing of superconductor data. We then provide a brief overview of the details of the underlying DDPM in Section II B and the Iterative Latent Variable Refinement method in Section II C. Lastly, we discuss the techniques we use to evaluate the quality of SuperDiff outputs in Section II D and II E.

A. Data Processing

All data for the model was sourced from SuperCon[21], which is the largest database for superconducting materials. The dataset was processed by the steps in Ref. [6] and, like in previous studies [4, 6, 15], only the chemi-
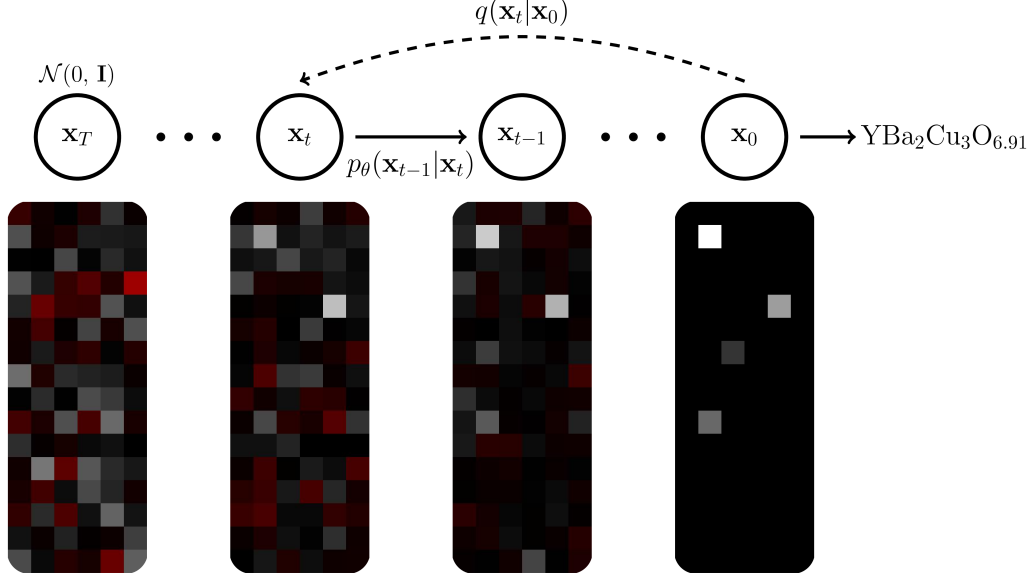
FIG. 2. Overview of the unconditional DDPM used. Compounds are encoded as vectors in $\mathbb{R}^{96}$; however, for illustration purposes, the vectors are represented as 16×6 pixel images in this figure, where each pixel in the image represents an element of the vector, starting from the top-left corner and proceeding horizontally row by row. Whiter pixels represent more positive values (all values are divided by the maximum element of $\mathbf{x}_0$), and redder pixels represent more negative values (black is zero). Starting from noise $\mathbf{x}_T$, the model generates a compound $\mathbf{x}_0$ by denoising $\mathbf{x}_t$ iteratively. Note that $YBa_2Cu_3O_{6.91}$ was picked from SuperCon for illustration purposes only, and is not a compound generated by SuperDiff.

cal composition data was used. Every compound from SuperCon was represented as a column vector for input into the model. As shown in Fig. 1, each compound was encoded as a 96×1 column vector as 96 is the maximum atomic number present in the dataset.

B. Denoising Diffusion Probabilistic Model

Denoising Diffusion Probabilistic Models (DDPMs) [12, 13] function by learning a Markov chain to progressively transform an isotropic Gaussian into a data distribution. The general structure of the DDPM used is shown in Fig. 2. The DDPM consists of two parts: a forwards "diffusion" process that adds noise to data, and a generative reverse process that learns the reverse of the forwards process—"denoising" the forwards process. The forward process is a fixed Markov chain that gradually adds Gaussian noise to data. Each step in the forward process is defined as

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}; \beta_t\mathbf{I}), \qquad (1)$$

where $\beta_1, ..., \beta_T$ is the variance schedule, $\mathbf{I}$ is the identity matrix, and $\mathbf{x}_0$ is dimensionally equivalent to latent variables $\mathbf{x}_1, ..., \mathbf{x}_T$ (all vectors in $\mathbb{R}^{96}$). In this work, we adopt the cosine variance schedule proposed in Ref. [22].

A notable property of the forwards process is that given clean data $\mathbf{x}_0$, noised data at any time-step $\mathbf{x}_t$ can be sampled in closed-form:

$$q(\mathbf{x}_t|\mathbf{x}_0) := \mathcal{N}(\mathbf{x}_t; \sqrt{\overline{\alpha}_t}\mathbf{x}_0; (1-\overline{\alpha}_t)\mathbf{I}), \qquad (2)$$

where $\alpha_t := 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. This can be reparametrized [23] as:

$$\mathbf{x}_t = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, \qquad (3)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and is dimensionally equivalent to $\mathbf{x}_0$.

The reverse process is then defined to be

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t); \sigma_t^2\mathbf{I}). \qquad (4)$$

In this work, we fix $\sigma_t^2 = \beta_t$. Then, as shown in Ref. [12], by rewriting $\boldsymbol{\mu}_\theta$ as a linear combination of $\mathbf{x}_t$ and $\boldsymbol{\epsilon}_\theta$, a neural network that predicts $\boldsymbol{\epsilon}$ from $\mathbf{x}_t$ with input and output dimensions equal to that of the noise it predicts, the reverse process may be rewritten as:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}, \qquad (5)$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$.

To train the DDPM, noise is added to $\mathbf{x}_0$ using the forward process $q(\mathbf{x}_t|\mathbf{x}_0)$ for a randomly sampled $t \sim$ Uniform($\{1, ..., T\}$), which the neural network then learns to remove through the reverse process.

Four different versions of the DDPM were trained on SuperCon: one for cuprates, one for pnictides, one for others, and one for all classes ("everything"). The training datasets for each version of the DDPM were randomly split into training and validation sets in an approximately $95\% - 5\%$ proportion. Training curves for all versions of the DDPM were able to converge and stabilize after around 50 epochs, and each version of the DDPM was
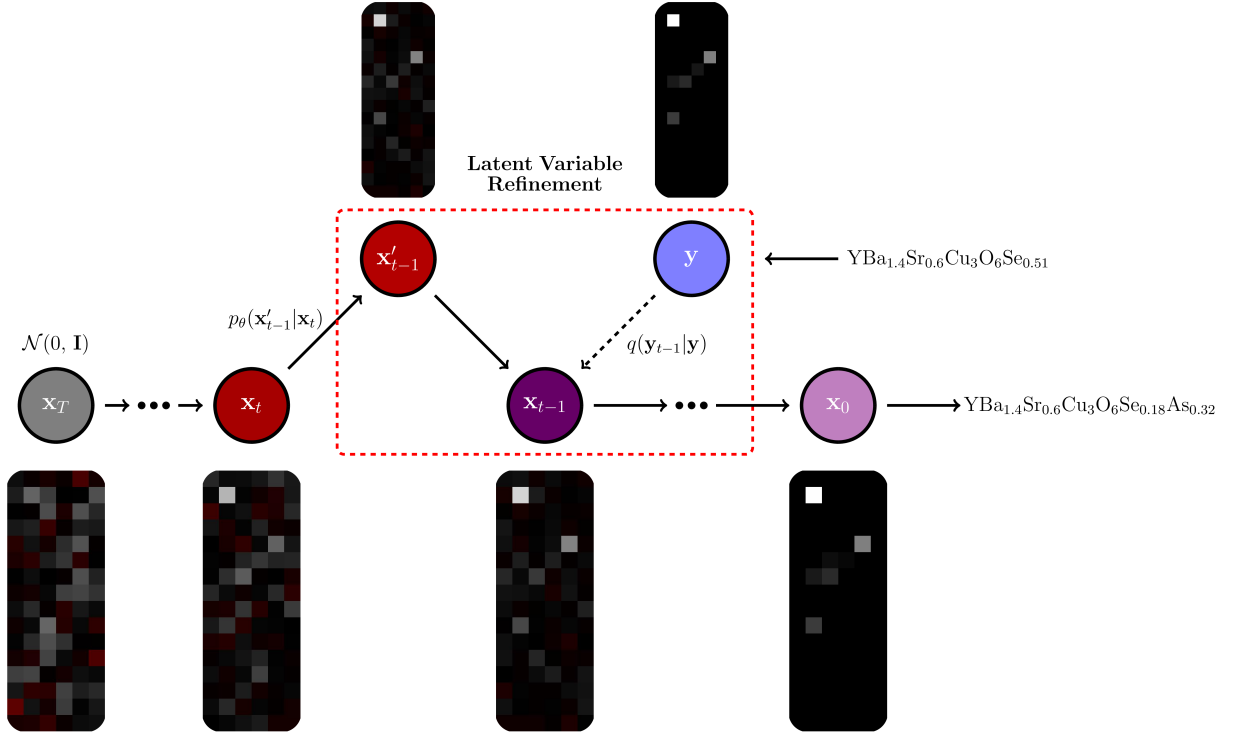
FIG. 3. Overview of the Iterative Latent Variable Refinement [14] method used. The vector image representation is the same as explained in Figure 2. $YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.51}$ [24] is an example of a reference superconductors and $YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.18}As_{0.32}$ is an example of a generated output.

trained for between 50 and 100 epochs, depending on the approximate lowest validation loss. For all versions of the DDPM, NAdam [25] was chosen as the optimizer, and provided satisfactory results. Moreover, like in Ref. [12], T was set to 1000 and the U-Net [26] neural network architecture was used for ϵ_θ (for this work, a 1D U-Net was used as opposed to the 2D U-Net used for images).

C. Conditioning

Iterative Latent Variable Refinement (ILVR) [14] was used to condition the DDPM. Because ILVR is training-free, the same four trained unconditional DDPMs could be relatively easily modified for conditioning.

ILVR is a slight modification to the reverse diffusion process, and the general structure of ILVR used is shown in Figure. 3. At each step of the reverse "denoising" process, instead of sampling $\mathbf{x}_{t-1}$ directly from $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ like in unconditional DDPM, $\mathbf{x}_{t-1}$ instead becomes

$$\mathbf{x}_{t-1} = \phi_N(\mathbf{y}_{t-1}) + \mathbf{x}'_{t-1} - \phi_N(\mathbf{x}'_{t-1}), \qquad (6)$$

where $\mathbf{x}'_{t-1} \sim p_\theta(\mathbf{x}'_{t-1}|\mathbf{x}_t)$ is the original unconditional proposal, $\mathbf{y}_{t-1} \sim q(\mathbf{y}_{t-1}|\mathbf{y})$ is the condition encoding by the noising process in Equation (2), and ϕ_N is a linear low-pass filtering operation that maintains the dimensionality of the input.

The goal of ILVR conditioning is to have $\phi_N(\mathbf{x}_0) = \phi_N(\mathbf{y})$, thereby allowing the generated output $\mathbf{x}_0$ to share high-level features with reference $\mathbf{y}$. In this case, the generated superconductor should have similar chemical composition as the reference superconductor.

In Ref. [14], it was stated that the scale factor N could be changed to control the amount of information brought from the reference to the generated output, where lower N results in greater similarity between generated output and reference and higher N results in only coarse information from the reference being brought by the model to the generated output. In our work, we found that $N > 4$ resulted in large numbers of invalid compounds with negative amounts of elements. As a result, we used $N = 2$ up to $N = 4$, but we still found the conclusions made about changing N in Ref. [14] applicable.

D. Sampling

As mentioned previously, we trained four versions of the unconditional model, each of which was then copied and modified with ILVR conditioning to also create four versions of the conditional model. We thus have four versions of the unconditional DDPM (without ILVR), which we call "unconditional SuperDiff", and four versions of the conditional DDPM (with ILVR), which we call "conditional SuperDiff". On a single consumer Nvidia RTX 3060 Ti GPU, each version of SuperDiff was trained in under 2 hours, and we sampled 500,000 compounds from

SuperDiff Version	# Generated	Inference Time (s)
Entirety of SuperCon	500,000	35,567
Cuprate	500,000	35,781
Pnictides	500,000	35,720
Others	500,000	35,679

TABLE I. Inference metrics for the various versions of unconditional SuperDiff. These results were all achieved on a single consumer Nvidia RTX 3060 Ti GPU.

each of the four unconditional SuperDiff versions. Reasonably fast inference times were achieved, making SuperDiff a system that can be trained and used using resources at most universities and even consumers. Detailed inference metrics for unconditional SuperDiff are in Table I. For conditional SuperDiff, we sampled varying amounts of compounds for different reference superconductors, and we discuss those results in Section III E.

All sampled compounds were initially screened through various quality checks to ensure that all generated compounds were reasonably realistic. First, we obviously eliminated all generated compounds with negative amounts of elements. Note that we round all amounts of elements to two decimal places beforehand. Next, we eliminate compounds with either too few (only 1) or too many elements—for Cuprates, we limit outputs to compounds with a maximum of 7 elements, and for Pnictides and Others, we limit outputs to compounds with a maximum of 5 elements. After these basic checks, we removed duplicates and further evaluated compound validity with the charge neutrality and electronegativity checks from the `SMACT` package [27]. Finally, we ran formation energy prediction with ElemNet [28, 29]. We will discuss the performance of model generations against these checks later in Section III.

E. Clustering

To identify if SuperDiff could generate new superconductor families, clustering analysis was performed. Clustering, which is an unsupervised machine-learning method used to find hidden patterns within data, was applied to the SuperCon database in Ref. [15], which established that these methods, when applied to superconductors, could exceed human performance in identifying different "families" of superconductors, which are represented as clusters. In this work, we use the clustering method for superconductors from Ref. [15] to evaluate generated outputs for new families. In Ref. [15], it was also found that, for superconductors, to visualize clustering results, the t-SNE method worked best. t-SNE is a non-linear dimensionality reduction technique that allows higher dimensional data (96-dimensional superconductor data points in this case) to be represented in 2D or 3D [30] (which do not have any physical meaning).

As discussed in the introduction, a major objective

of this work was to generate new families of superconductors, as identified by the clustering model—that is, to generate new clusters of superconductors. This was something not accomplished by previous works, including the GAN in Ref. [6]. In order to achieve this goal, we experimented with the conditional model's ability to interpolate between the reference compound and the training dataset. This idea of experimenting with a conditional DDPM's ability to interpolate between the reference set and training set was proposed in Ref. [31] to attempt to achieve few-shot generation on image classes never seen during training. We attempt to do this with superconductors in this work. For instance, we experiment with conditioning the cuprate version of conditional SuperDiff on new, different reference cuprates not in the families of cuprate superconductors in the training dataset. We examine the model's ability to generate new clusters or families of superconductors using information from the reference compound with this technique. We report our clustering results in Section III D.

III. RESULTS

In this section we report the performance of SuperDiff on various checks and discuss some interesting new findings. We first evaluate the performance of unconditional SuperDiff with the 500,000 compounds we generated for each of the four classes by performing various computational tests, which included some general compound checks in Sections III A and III B as well as checks for superconductivity in Section III C. We use the same computational tests for unconditional SuperDiff as used for the GAN in Ref. [6] and are thus able to directly compare unconditional performance. Afterward, as our most notable results, we evaluate the performance of both the unconditional and conditional versions of SuperDiff on clustering in Section III D and manually identify and present some promising new families of superconductors generated by the conditional SuperDiff in Section III E.

A. Duplicates and Validity

For the 500,000 compounds generated by each version of unconditional SuperDiff, we first screened for duplicates between the generated set and the training set (the portion of the SuperCon database of the same class) and for duplicates within the generated set itself. After this, we ran the charge neutrality and electronegativity checks on the generated compounds with the `SMACT` package [27]. We present the results of these general tests in Table II, and then we remove all duplicates from the generated sets.

We notice that the novelty % and uniqueness % of generated results are all very high, which means that unconditional SuperDiff is able to successfully generate both diverse and novel compounds. Unconditional SuperDiff,
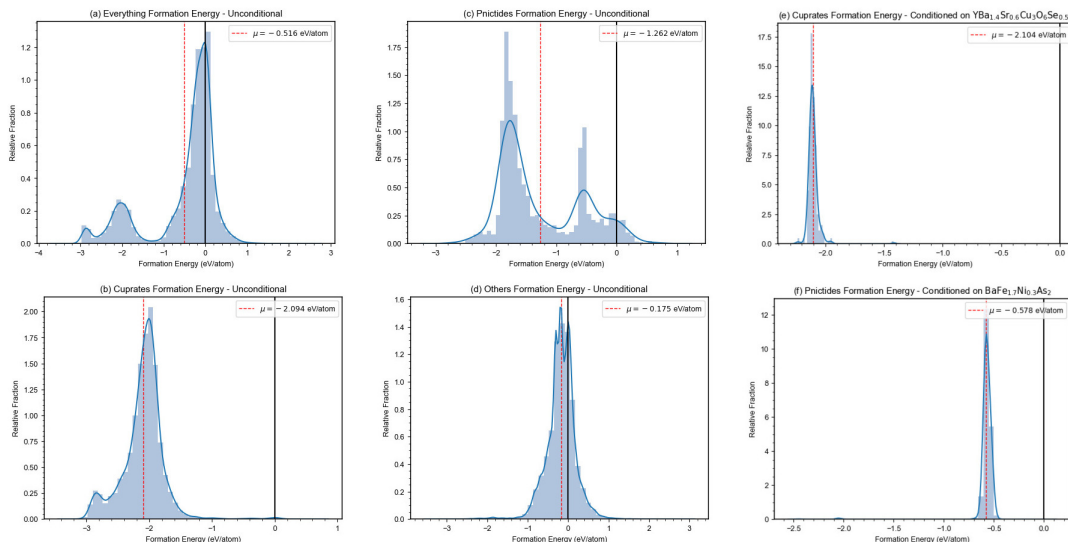
FIG. 4. Distribution of ElemNet [28, 29] predicted formation energies of the generated compounds from the four versions of unconditional SuperDiff—(a) Everything, (b) Cuprates, (c) Pnictides, and (d) Others—as well as (e) Cuprates version of conditional SuperDiff conditioned on $YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.51}$ [24] and (f) Pnictides version of conditional SuperDiff conditioned on $BaFe_{1.7}Ni_{0.3}As_2$ [32]. Also shown are the average formation energy for each distribution.

SuperDiff Version	Novel %	Unique %	# Valid
Entirety of SuperCon	100.00%	99.32%	79,828
Cuprate	100.00%	99.92%	10,971
Pnictides	100.00%	99.98%	2,184
Others	100.00%	99.39%	172,739

TABLE II. The percentage of generated compounds that were novel (not in the training set) and unique (distinct from others in the given generated set) before `SMACT` [27] filters, and the number of generated compounds that were valid (passed `SMACT` filters) for the 500,000 compounds generated by each of the four versions of unconditional SuperDiff.

here, outperforms the GAN in Ref. [6] in all metrics regarding generation novelty and uniqueness. Notably, unconditional SuperDiff maintains a very high uniqueness % for pnictides despite the small training set, something not accomplished by the Wasserstein GAN in Ref. [6]. This corroborates the observation of the superior ability of DDPMs to generate diverse results when compared to a GAN in other disciplines [16]. Lastly, although the `SMACT` check [27] results varied greatly between classes and the proportion of valid compounds for some classes was fairly low, the fast inference time justifies that SuperDiff is still able to generate valid compounds reasonably well for all classes.

Overall, these results indicate that all versions of unconditional SuperDiff are able to generate both novel and unique compounds—overcoming the past issues faced by Ref. [6]—as well as valid compounds. As conditional SuperDiff maintains much of the same components as the unconditional model, it was unsurprising that—in most

cases—conditional SuperDiff was also able to generate novel, unique, and valid compounds; however, for conditional SuperDiff, these qualities were very much dependent on the reference compound—we still run these checks on all compounds generated by conditional SuperDiff and filter out invalid compounds.

B. Formation Energy

We further validated the performance of SuperDiff on generating synthesizable compounds by predicting the formation energies of the generated compounds with ElemNet [28, 29], which is a deep neural network model for predicting material properties from only elemental compositions. We chose ElemNet for our formation energy prediction because of its ability to use only chemical composition, as we do not consider crystal structure in our generation process. Because ElemNet does not take in compounds as column vectors in $\mathbb{R}^{96}$, as SuperDiff does, but instead takes them in as column vectors in $\mathbb{R}^{86}$ with certain elements removed, we ran the ElemNet formation energy prediction on only the compounds generated by SuperDiff that ElemNet would directly support—this did constitute the great majority of generated compounds. We display the distributions for the predicted formation energies of generated compounds in Fig. 4.

As shown in the figure, unconditional SuperDiff generated a majority of compounds with negative formation energy for all classes of superconductors, with the mean formation energy for all classes predicted to be negative as well. In Ref. [28], it was stated that negative formation energy values for compounds are a good

SuperDiff Version	Raw % Superconducting	True % Superconducting	Mean T_c	Standard Dev.	Max T_c
Entirety of SuperCon	67.81%	62.22%	8.51 K	8.79 K	97.0 K
Cuprates	64.53%	58.22%	64.68 K	14.61 K	110.5 K
Pnictides	95.74%	96.39%	22.00 K	2.43 K	29.5 K
Others	55.82%	47.56%	6.33 K	2.50 K	30.5 K

TABLE III. The percentages of generated compounds determined to be superconducting by the classification model for each of the unconditional SuperDiff versions, along with the estimated true percentages according to Eq. 7 and summary T_c statistics from the predictions by the regression model.
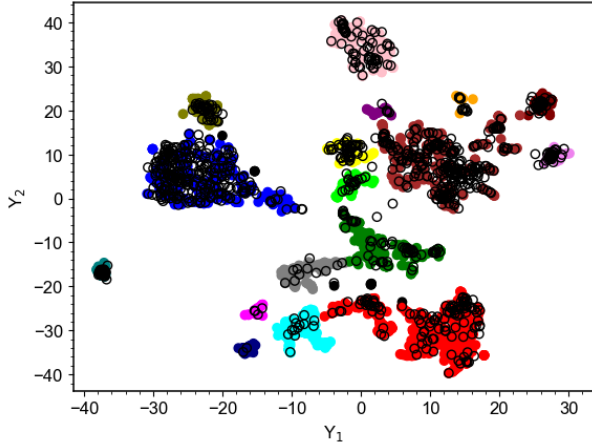

FIG. 5. Clustering of the valid generated compounds from the Cuprates version of unconditional SuperDiff. Full circles represent Cuprate data points from SuperCon, and black open circles are generated compounds. We notice that unconditional SuperDiff did not generate any new families of superconductors, as all generated compounds fall within the existing clusters of superconductors from SuperCon.
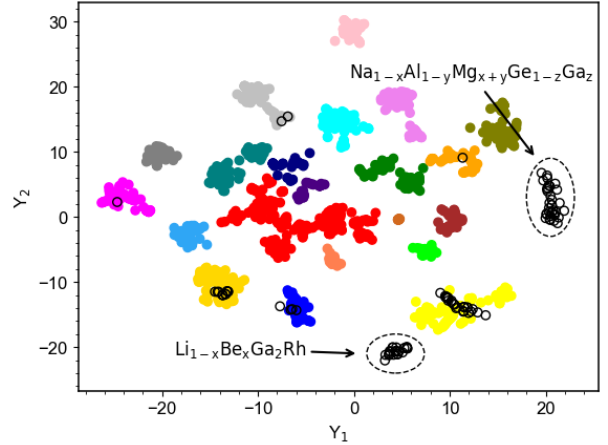

FIG. 6. Clustering of the valid generated compounds from the Others version of conditional SuperDiff conditioned on various different compounds. Full circles represent Others data points from SuperCon, and black open circles are generated compounds. Although some generated superconductors fall within the existing SuperCon clusters, we were able to identify two new clusters consisting of only generated superconductors (marked with arrows). These two new clusters correspond to two new families of superconductors generated by SuperDiff: $Li_{1-x}Be_xGa_2Rh$ and $Na_{1-x}Al_{1-y}Mg_{x+y}Ge_{1-z}Ga_z$.

indicator of their stability and synthesizability; therefore, although these predictions are not definitive proof—experimentation validation would be necessary—these predictions provide an indication that most of the compounds generated by unconditional SuperDiff are plausibly stable and synthesizable.

For conditional SuperDiff, the distribution of formation energies for generated compounds is heavily dependent on the reference. However, given a reasonable reference compound—that is, a valid reference compound that belongs to the class of superconductor that the version of SuperDiff was trained on—we demonstrate that conditional SuperDiff is able to generate compounds predicted to be stable by ElemNet. Specifically, as shown in Fig. 4, for the cuprates version of conditional SuperDiff conditioned on $YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.51}$ [24] and the pnictides version of conditional SuperDiff conditioned on $BaFe_{1.7}Ni_{0.3}As_2$ [32]—some of the compounds we conditioned conditional SuperDiff on to find new families of superconductors in Section III E—the predicted distribution of formation energies for generated compounds show all generated compounds to have negative formation energy. These results indicate that, given reasonable

reference compounds, conditional SuperDiff can generate plausibly stable and synthesizable compounds, which is not surprising given the fundamental architecture similarities between conditional and unconditional SuperDiff.

C. Superconductivity

After those general checks, we performed some computational checks for superconductivity in order to verify that unconditional SuperDiff is indeed able to generate probable superconductors. We ran the compounds generated by unconditional SuperDiff through the K-Nearest Neighbors (KNN) classification model and regression model from Ref. [4] for predicting superconductivity and critical temperature, respectively, based on elemental composition.

For the predicted proportion of generated compounds that were superconducting, we accounted for the inherent probabilistic error of the classification model by us-

Reference Compound	SuperDiff Version	Output Examples	Predicted T_c	General Formula
$YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.51}$ [24]	Cuprates	$YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.35}As_{0.12}$	55 K	
		$YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.28}As_{0.21}$	41 K	$YBa_{1.4}Sr_{0.6}Cu_3O_6Se_xAs_y$
		$YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.18}As_{0.32}$	46 K	
$YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.51}$ [24]	Cuprates	$YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.18}Br_{0.19}$	54 K	
		$YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.11}Br_{0.25}$	33 K	$YBa_{1.4}Sr_{0.6}Cu_3O_6Se_xBr_y$
		$YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.17}Br_{0.25}$	41 K	
$SrCu_2O_3$ [33]	Cuprates	$SrCu_{1.77}Ni_{0.08}Zn_{0.15}O_3$	31 K	
		$SrCu_{1.58}Ni_{0.11}Zn_{0.31}O_3$	10 K	$SrCu_{2-x-y}Zn_xNi_yO_3$
		$SrCu_{1.85}Ni_{0.08}Zn_{0.07}O_3$	28 K	
$Ba_2CuO_{3.25}$ [34]	Cuprates	$Ba_{1.88}Cs_{0.12}CuO_{3.28}$	30 K	
		$Ba_{1.91}Cs_{0.09}CuO_{3.28}$	28 K	$Ba_{2-x}Cs_xCuO_{3.3}$
		$Ba_{1.77}Cs_{0.23}CuO_{3.3}$	13 K	
$LiCu_2O_2$ [35]	Cuprates	$Li_{0.67}Be_{0.34}Cu_2O_2$	33 K	
		$Li_{0.89}Be_{0.11}Cu_2O_2$	22 K	$Li_{1-x}Be_xCu_2O_2$
		$Li_{0.72}Be_{0.28}Cu_2O_2$	34 K	
$LiGa_2Rh$ [36]	Others	$Li_{0.67}Be_{0.34}Ga_2Rh$	9 K	
		$Li_{0.87}Be_{0.13}Ga_2Rh$	33 K	$Li_{1-x}Be_xGa_2Rh$
		$Li_{0.71}Be_{0.29}Ga_2Rh$	27 K	
$NaAlGe$ [37]	Others	$Na_{0.8}Al_{0.92}Mg_{0.28}Ge_{0.84}Ga_{0.16}$	10 K	
		$Na_{0.36}Al_{0.63}Mg_{0.99}Ge_{0.88}Ga_{0.12}$	12 K	$Na_{1-x}Al_{1-y}Mg_{x+y}Ge_{1-z}Ga_z$
		$Na_{0.79}Al_{0.75}Mg_{0.46}Ge_{0.68}Ga_{0.32}$	8 K	
$BaFe_{1.7}Ni_{0.3}As_2$ [32]	Pnictides	$BaFe_{1.72}Co_{0.13}Ni_{0.15}As_2$	24 K	
		$BaFe_{1.74}Co_{0.08}Ni_{0.08}As_2$	16 K	$BaFe_{2-x-y}Co_xNi_yAs_2$
		$BaFe_{1.7}Co_{0.12}Ni_{0.11}As_2$	30 K	
$BaFe_{1.7}Ni_{0.3}As_2$ [32]	Pnictides	$BaFe_{1.84}Co_{0.16}As_{1.8}Ge_{0.2}$	17 K	
		$BaFe_{1.77}Co_{0.23}As_{1.81}Ge_{0.19}$	24 K	$BaFe_{2-x}Co_xAs_{2-y}Ge_y$
		$BaFe_{1.82}Co_{0.18}As_{1.79}Ge_{0.21}$	27 K	

TABLE IV. Promising new families of superconductors generated by conditional SuperDiff. Shown are the reference compound used to condition the SuperDiff, the version of conditional SuperDiff used, a few output examples from the family and their predicted critical temperatures [4], and the general formula for the new family.

ing Bayesian statistics to estimate the true proportion of superconducting generated compounds given the classification model's predicted proportion p_{sc} and the true positive tp and false positive rates fp of the classification model. The true proportion of generated compounds that are superconductors ρ_{sc} may be estimated as [6]

$$\rho_{sc} \approx \frac{p_{sc} - fp}{tp - fp}, \qquad (7)$$

where $tp = 98.69\%$ and $fp = 16.94\%$ are reported by Ref. [4].

For the generated compounds that were predicted to be superconducting, we used the regression model in Ref. [4] to predict their critical temperatures. Like all other tests done so far, this computational prediction is only an approximation. We tabulated the results of the classification and regression predictions on the compounds generated by unconditional SuperDiff in Table III. We will discuss the predicted superconductivity of compounds generated by conditional SuperDiff later in Section III E.

As seen in the table, all versions of unconditional SuperDiff were able to generate predicted superconductors at a rate comparable to the GAN in Ref. [6] and much higher than the 3% achieved by manual search in Ref. [2]—notably, unconditional SuperDiff seems to perform much better on pnictides despite the small training set. This is further indication of the effectiveness of computational search for superconductors when compared to manual searches. Moreover, unconditional SuperDiff seems to capture the critical temperature distribution of the SuperCon training dataset much better than the GAN in Ref. [6].

Although actual synthesis and testing in a lab are required to confirm superconductivity, these checks, combined with the clustering analysis results that we will discuss in Section III D, provide a general indication that unconditional SuperDiff is able to generate highly plausible superconductors.

D. Clustering Results

We ran the clustering analysis described in Section II E on both unconditional and conditional SuperDiff. We display the clustering results for the cuprates version of unconditional SuperDiff in Fig. 5. Superconductors from the SuperCon database are shown with full circles of different colors, whereas our predictions are shown with open black circles. Although unconditional SuperDiff generated compounds in all known clusters or families of superconductors, no new families of supercon-

ductors were generated by unconditional SuperDiff—this was true for the other versions of unconditional SuperDiff as well. This was the expected result for unconditional SuperDiff as the underlying DDPM's goal is to just find a mapping from Gaussian noise to the training data distribution, not some other new distribution. However, superconductor discovery has a particular interest in the generation of new families of superconductors, so a method to control the generation process to change the generated data distribution is desirable. With conditional SuperDiff, we are able to control the generation process to computationally generate new families of superconductors for the first time.

In Fig. 6, we display a sample clustering result from the "others" version of conditional SuperDiff conditioned on various compounds. As seen in the plot, we identified two new clusters: $Li_{1-x}Be_xGa_2Rh$, which was generated by conditioning SuperDiff on $LiGa_2Rh$ [36], and $Na_{1-x}Al_{1-y}Mg_{x+y}Ge_{1-z}Ga_z$, which was generated by conditioning SuperDiff on NaAlGe [37]. Those and other predicted families will be discussed in more detail below.

These clustering results show that, with this ability to control generation, and by conditioning SuperDiff on compounds not in the SuperCon training set, SuperDiff is able to use information from various reference compounds to generate completely new families of superconductors. As expected, we note that for these generated new families, the reference compound does belong to the new cluster; however, the contribution presented in this work is that we are able to extrapolate a new family of superconductors from the single reference compound. We performed this clustering analysis on all versions of conditional SuperDiff conditioned on a variety of different reference compounds, and we discuss the promising new families of superconductors generated by conditional SuperDiff in more detail and verify their superconductivity in Section III E.

E. Promising Generated New Families

After running clustering analysis for the different versions of conditional SuperDiff conditioned on a variety of reference compounds, we manually identified the most promising new families of superconductors generated by conditional SuperDiff. Beyond the novelty, uniqueness, and SMACT checks, we further checked for the novelty of these newly generated families by searching on the internet and through other databases—these newly generated families could not be found anywhere else. We tabulated these most promising new families of superconductors generated by conditional SuperDiff in Table IV. There, we identified the reference compound used as well as a few output examples and their respective predicted T_c using the regression model in Ref. [4], and determined the general formula for the new family. We notice that most compounds generated with conditional SuperDiff are predicted to be superconducting, with predicted T_c
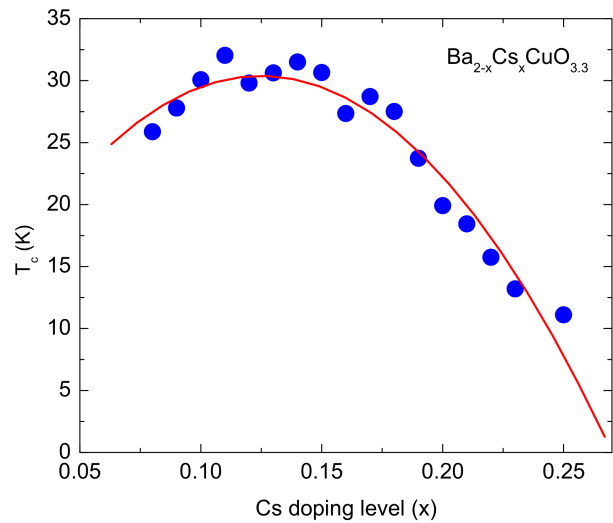

FIG. 7. Plot of T_c predicted by the regression model in Ref. [4] versus Cesium content (x) for $Ba_{2-x}Cs_xCuO_{3.3}$ family generated by conditional SuperDiff (see Table IV). We notice a characteristic parabolic dependence of T_c versus doping, observed previously in other cuprate families [38].

being reasonable for each class. Additionally, a particularly interesting result to note was that our model seemed to generate some new families of superconductors with double or, in one case, even triple doping. This is an interesting new avenue for superconductor discovery that has not been extensively studied, which our model suggests should be explored in more detail by material scientists.

We further demonstrate that conditional SuperDiff is able to generate realistic new families of superconductors by plotting the predicted T_c using the regression model in Ref. [4] against the Cesium doping content for the newly generated $Ba_{2-x}Cs_xCuO_{3.3}$ family in Fig. 7. We notice that the generated $Ba_{2-x}Cs_xCuO_{3.3}$ family is predicted to exhibit the expected parabolic T_c doping dependence relationship for this type of cuprate superconductor, which was observed previously in other cuprate families [38].

These findings again show that SuperDiff is not only able to generate new superconductors within known families but is also able to overcome the limitations of previous generative models to generate completely new families of superconductors that are also realistic—although we note that for some reference compounds, SuperDiff was also unable to generate new families of superconductors.

IV. CONCLUSION

With the lack of a systematic approach, the discovery of new high T_c superconductors has long depended on material scientists' serendipity. Recently, machine learning has been applied to this field to help assist scientists,

but past works still lacked many key capabilities, for instance, the ability to computationally find new families of superconductors. Moreover, recent generative model approaches applied to this field also lacked methods of controlling the generation process by incorporating information from reference compounds [6–8].

In this paper, we have introduced a novel method for superconductor discovery using diffusion models with conditioning functionality that has addressed these major issues. Like previous works applying generative models to superconductor discovery, we were able to generate novel, realistic, and highly plausible superconductors that lie outside of existing databases—leveraging this "inverse design" approach to significantly outperform manual search and previous classification model approaches. With our unconditional model, we were also able to address the low generated compound uniqueness issues that plagued previous works due to the small training data set for pnictides. Most importantly, however, beyond the unconditional performance improvements the diffusion model brought, our contribution of implementing conditioning with ILVR for superconductor discovery to allow the generation process to be controlled enabled the creation of a tool for computationally generating completely new families of superconductors. We verified the

generation of new families of superconductors with our clustering analysis, and we presented several of these promising new families of generated superconductors for several different classes of superconductors in Table IV. Once again, we point out that no previous computational model for superconductor discovery would have been capable of generating these new families of superconductors, as they attempt to produce only samples that match the training data.

The application of deep generative models for superconductor discovery continues to be a very promising and exciting approach. Future studies can benefit from possible improvements that can be made to SuperDiff, including implementing a physics-informed diffusion model and creating and utilizing a better, more comprehensive training dataset of superconductors. Nevertheless, SuperDiff in its current form is still very powerful as a tool for superconductor discovery, and researchers can currently benefit from it in a myriad of ways, such as by using its novel generations as inspiration—starting with the new families introduced here, using it to expand on their own new discoveries, or by simply experimenting with many more reference compounds to continue using it to generate completely new families of superconductors.

[1] J. Hirsch, M. Maple, and F. Marsiglio, Superconducting materials classes: Introduction and overview, Physica C: Superconductivity and its Applications **514**, 1 (2015), superconducting Materials: Conventional, Unconventional and Undetermined.

[2] H. Hosono, K. Tanabe, E. Takayama-Muromachi, H. Kageyama, S. Yamanaka, H. Kumakura, M. Nohara, H. Hiramatsu, and S. Fujitsu, Exploration of new superconductors and functional materials, and fabrication of superconducting tapes and wires of iron pnictides, Science and Technology of Advanced Materials **16**, 033503 (2015).

[3] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, Machine learning modeling of superconducting critical temperature, npj Computational Materials **4**, 29 (2018).

[4] B. Roter and S. Dordevic, Predicting new superconductors and their critical temperatures using machine learning, Physica C: Superconductivity and its Applications **575**, 1353689 (2020).

[5] T. Konno, H. Kurokawa, F. Nabeshima, Y. Sakishita, R. Ogawa, I. Hosako, and A. Maeda, Deep learning model for finding new superconductors, Phys. Rev. B **103**, 014509 (2021).

[6] E. Kim and S. V. Dordevic, Scgan: A generative adversarial network to predict hypothetical superconductors, Journal of Physics: Condensed Matter **36**, 025702 (2023).

[7] D. Wines, T. Xie, and K. Choudhary, Inverse design of next-generation superconductors using data-driven deep generative models, The Journal of Physical Chemistry Letters **14**, 6630 (2023), pMID: 37462366, https://doi.org/10.1021/acs.jpclett.3c01260.

[8] C. Zhong, J. Zhang, Y. Wang, Y. Long, P. Zhu, J. Liu, K. Hu, J. Chen, and X. Lin, High-performance diffusion model for inverse design of high t_c superconductors with effective doping and accurate stoichiometry, InfoMat (2024).

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14 (MIT Press, Cambridge, MA, USA, 2014) p. 2672–2680.

[10] T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, and T. Jaakkola, Crystal diffusion variational autoencoder for periodic material generation, arXiv preprint arXiv:2110.06197 (2021).

[11] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Microscopic theory of superconductivity, Phys. Rev. **106**, 162 (1957).

[12] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, arXiv preprint arxiv:2006.11239 (2020).

[13] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 37, edited by F. Bach and D. Blei (PMLR, Lille, France, 2015) pp. 2256–2265.

[14] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, Ilvr: Conditioning method for denoising diffusion probabilistic models (2021), arXiv:2108.02938 [cs.CV].

[15] B. Roter, N. Ninkovic, and S. Dordevic, Clustering superconductors using unsupervised machine learning, Physica C: Superconductivity and its Applications **598**, 1354078

(2022).

[16] P. Dhariwal and A. Nichol, Diffusion models beat gans on image synthesis (2021), arXiv:2105.05233 [cs.LG].

[17] M. Alverson, S. Baird, R. Murdock, and T. Sparks, Generative adversarial networks and diffusion models in material discovery (2022).

[18] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, Hierarchical text-conditional image generation with clip latents (2022), arXiv:2204.06125 [cs.CV].

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models (2021), arXiv:2112.10752 [cs.CV].

[20] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. S. Jaakkola, Diffdock: Diffusion steps, twists, and turns for molecular docking, in *The Eleventh International Conference on Learning Representations* (2023).

[21] N. I. for Materials Science, Supercon (2020).

[22] A. Nichol and P. Dhariwal, Improved denoising diffusion probabilistic models (2021), arXiv:2102.09672 [cs.LG].

[23] D. P. Kingma and M. Welling, Auto-encoding variational bayes (2022), arXiv:1312.6114 [stat.ML].

[24] V. Grinenko, A. Dudka, S. Nozaki, J. Kilcrease, A. Muto, J. Clarke, T. Hogan, V. Nikoghosyan, I. de Paiva, R. Dulal, S. Teknowijoyo, S. Chahid, and A. Gulian, Extraordinary physical properties of superconducting $YBa_{1.4}Sr_{0.6}Cu_3O_6Se_{0.51}$ in a multiphase ceramic material (2023), arXiv:2309.16814 [cond-mat.supr-con].

[25] T. Dozat, Incorporating Nesterov Momentum into Adam, in *Proceedings of the 4th International Conference on Learning Representations*, pp. 1–4.

[26] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi (Springer International Publishing, Cham, 2015) pp. 234–241.

[27] D. W. Davies, K. T. Butler, A. J. Jackson, A. Morris, J. M. Frost, J. M. Skelton, and A. Walsh, Computational screening of all stoichiometric inorganic materials, Chem **1**, 617 (2016).

[28] D. Jha, L. Ward, A. Paul, W.-k. Liao, A. Choudhary, C. Wolverton, and A. Agrawal, Elemnet: Deep learning the chemistry of materials from only elemental composition, Scientific Reports **8**, 17593 (2018).

[29] D. Jha, K. Choudhary, F. Tavazza, W.-k. Liao, A. Choudhary, C. Campbell, and A. Agrawal, Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning, Nature Communications **10** (2019).

[30] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research **9**, 2579 (2008).

[31] G. Giannone, D. Nielsen, and O. Winther, Few-shot diffusion models (2022), arXiv:2205.15463 [cs.CV].

[32] M. Wang, C. Zhang, X. Lu, G. Tan, H. Luo, Y. Song, M. Wang, X. Zhang, E. A. Goremychkin, T. G. Perring, T. A. Maier, Z. Yin, K. Haule, G. Kotliar, and P. Dai, Doping dependence of spin excitations and its correlations with high-temperature superconductivity in iron pnictides, Nature Communications **4**, 2874 (2013).

[33] S. Ohsugi, Y. Kitaoka, M. Azuma, Y. Fujishiro, and M. Takano, Antiferromagnetic order in the ladder compound $SrCu_2O_3$; cu-nmr/nqr measurements, Journal of Low Temperature Physics **117**, 1671 (1999).

[34] R. Fumagalli, A. Nag, S. Agrestini, M. Garcia-Fernandez, A. C. Walters, D. Betto, N. B. Brookes, L. Braicovich, K.-J. Zhou, G. Ghiringhelli, and M. Moretti Sala, Crystalline and magnetic structure of $Ba_2CuO_{3+\delta}$ investigated by x-ray absorption spectroscopy and resonant inelastic x-ray scattering, Physica C: Superconductivity and its Applications **581**, 1353810 (2021).

[35] A. A. Bush, N. Büttgen, A. A. Gippius, M. Horvatić, M. Jeong, W. Kraetschmer, V. I. Marchenko, Y. A. Sakhratov, and L. E. Svistov, Exotic phases of frustrated antiferromagnet $LiCu_2O_2$, Phys. Rev. B **97**, 054428 (2018).

[36] P. Mondal, S. Khanom, N. A. Shahed, M. K. Hossain, and F. Ahmed, An ab initio insight into the structural, physical, thermodynamic and optoelectronic properties of superconducting heusler-like $LiGa_2Rh$, Physica C: Superconductivity and its Applications **603**, 1354142 (2022).

[37] T. Ikenobe, T. Yamada, D. Hirai, H. Yamane, and Z. Hiroi, Superconductivity induced by doping holes in the nodal-line semimetal NaAlGe, Phys. Rev. Mater. **7**, 104801 (2023).

[38] J. Tallon and J. Loram, The doping dependence of t* – what is the real high-tc phase diagram?, Physica C: Superconductivity **349**, 53 (2001).