MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™

PAPER • OPEN ACCESS

LoGAN: local generative adversarial network for novel structure prediction

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 035079 https://doi.org/10.1088/2632-2153/ad7a4d

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
8 April 2024

REVISED
2 July 2024

ACCEPTED FOR PUBLICATION
12 September 2024

PUBLISHED
23 September 2024

PAPER

LoGAN: local generative adversarial network for novel structure prediction

Péter Kovács, Esther Heid, Jasper De Landsheere and Georg K H Madsen*

Institute of Materials Chemistry, Technical University of Vienna, Getreidemarkt 9/165-TC, A-1060 Vienna, Austria
* Author to whom any correspondence should be addressed.

E-mail: georg.madsen@tuwien.ac.at

Keywords: WGAN, structure prediction, local descriptors

Supplementary material for this article is available online

Abstract

The efficient generation and filtering of candidate structures for new materials is becoming increasingly important as starting points for computational studies. In this work, we introduce an approach to Wasserstein generative adversarial networks for predicting unique crystal and molecular structures. Leveraging translation- and rotation-invariant atom-centered local descriptors addresses some of the major challenges faced by similar methods. Our models require only small sets of known structures as training data. Furthermore, the approach is able to generate both non-periodic and periodic structures based on local coordination. We showcase the data efficiency and versatility of the approach by recovering all stable $C_5H_{12}O$ isomers using only 39 $C_4H_{10}O$ and $C_6H_{14}O$ training examples, as well as a few randomly selected known low-energy SiO_2 crystal structures utilizing only 167 training examples of other SiO_2 crystal structures. We also introduce a filtration technique to reduce the computational cost of subsequent characterization steps by selecting samples from unique basins on the potential energy surface, which allows to minimize the number of geometry relaxations needed after structure generation. The present method thus represents a new, versatile approach to generative modeling of crystal and molecular structures in the low-data regime, and is available as open-source.

1. Introduction

Identifying new compounds or materials with a desired set of properties is a complex challenge. The computational task primarily involves two stages: the generation of physically sound candidate structures, and their subsequent characterization using computational methods. For the primary characterization step of assessing structural stability, density functional theory (DFT) offers a reasonable trade-off between accuracy and computation cost. Recently, advances in accurate machine-learned force fields (MLFFs) have substantially accelerated the stability evaluation at only a small cost in accuracy [1, 2]. However, the vast search space, i.e. all the plausible conformations of molecules or crystals at a given stoichiometry, renders an exhaustive exploration unfeasible for all but the most trivial cases. Hence, the generation and selection of candidate structures is an equally important challenge to address. Historically, several approaches, including substitution into known structures based on data mined rules [3], simulated annealing [4, 5], and various methods based on evolutionary algorithms [6–8], have been employed to address this challenge. While these methods are able to cover a wide range of structures, they are often limited by their underlying assumptions and require a frequent evaluation of the potential energy (PE). Furthermore, the vast knowledge about stable structures is at best only included as heuristic rules to filter the number of candidates to be considered.

The success of generative models for images, text, and videos has sparked interest for their application for the direct generation of three-dimensional candidate structures for molecules and materials [9–18]. Despite some success, the field is still in its infancy. Most early approaches adopt a global representation of the full molecule or unit cell [10–14, 18], requiring a large amount of training data. Only recently, models based on

local environments have begun to appear [15–17]. These are however limited to molecular, non-periodic structures and it is unclear how a step-wise atom placement approaches can be generalized to periodic structures.

One conceptually attractive approach to generative models are generative adversarial networks (GANs), [19] which circumvent the need of explicitly evaluating a likelihood by adversarial training of two competing neural networks. GANs have been successfully applied to the generation of periodic structures [10–13]. However, existing approaches have been based on global representations which means that the critic/decoder has to understand whole structures, of which there are many. As a result, the GANs generating periodic structures have required rather large amounts of training data on the order of 10 k–1 M structures [10–13].

Focusing on learning the comparatively fewer physically significant local environments for evaluation could mitigate this data requirement. Recognizing viable compounds based on transferable local coordination or functional groups is fundamental to chemistry. The main challenge in producing stable and chemically accurate 3D structures using GANs lies in selecting the computational representation. The representation should continuously encode molecules and crystals while preserving periodicity, as well as rotational and translational symmetry, regardless of the number of atoms. In this paper we propose a 'local GAN' (LoGAN), a GAN-based architecture which functions in a new, data-efficient manner. The architecture amalgamates the success of local structural descriptors used for MLFFs with the ability of the generator network to predict the whole structure in one step without requiring an evaluation of the PE or even local descriptors. Importantly, a prediction of the full structure in a single step allows our model to be equally successful for periodic and non-periodic systems. During training, our critic network utilizes the effectiveness of local descriptors, thus requiring a low amount of training data, and enabling the generation of structures at different stoichiometries from the training set. The rotational- and translational symmetries are taken care of by our choice of descriptors, while the permutational symmetry of similar atoms is handled by the critic model. We also include a programmable post-processing step which can be used to enforce various symmetries or boundary conditions of the goal systems. The LoGAN approach is tested on both on $C_5H_{12}O$ isomers using only 39 similar molecules as training data and SiO_2 crystal structures using only 167 training samples, demonstrating its capability in a low-data regime.

2. Methodology

Our structure generation pipeline consists of two main steps. Initially we train multiple instances of LoGAN, a modified Wasserstein-GAN [20] (WGAN) model, which are subsequently used to generate candidate structures. Given the finite number of physically plausible structures, this approach inevitably results in multiple candidates residing within the same energy basins. To avoid unnecessary computations we subsequently apply a post-generation filter step to select representatives of these basins to be evaluated, which is based on structural descriptors and thus does not require an energy evaluation. The training data consists solely of a collection of structures, and also does not require their energies to be known. We note, however, that the trained model will sample from a distribution similar to the training set. Therefore, to search for low energy structures, the training set should primarily consist of low energy structures. In the following the various parts of our pipeline are discussed in detail.

2.1. Data

We selected a non-periodic and periodic system each as benchmark systems, as described in the following.

For the non-periodic system, we chose subsets of the QM9 database [21]. QM9 reports equilibrium structures and energies for 134 k molecules. Here, we selected all $C_4H_{10}O$ and $C_6H_{14}O$ isomers as training examples (resulting in 39 datapoints of equilibrium structures). As target structures, $C_5H_{12}O$ isomers were chosen. We thus generate structures at different stoichiometries as the training data, namely generating the coordinates for five carbon, twelve hydrogen and one oxygen atoms, each. Importantly, we only use the structural information in QM9, and not the energies or other target properties. In QM9, 14 stable $C_5H_{12}O$ isomers are recorded, which we aim to recover using LoGAN by comparing the SMILES strings of the true structures to the generated structures after geometry relaxation.

For the periodic system, we downloaded all SiO_2 bulk structures from the MaterialsProject [22] (MP) database with < 0.02 eV atom^{-1} energy above the convex hull, amounting to 168 structures with varying unit cell size. We then randomly picked six SiO_2 bulk structures with up to 18 atoms per unit cell as target structures, and used the remaining 167 structures as training data for each target. For each target, atom coordinates for the given atoms in the unit cell were then predicted. The energies of the subsequently relaxed

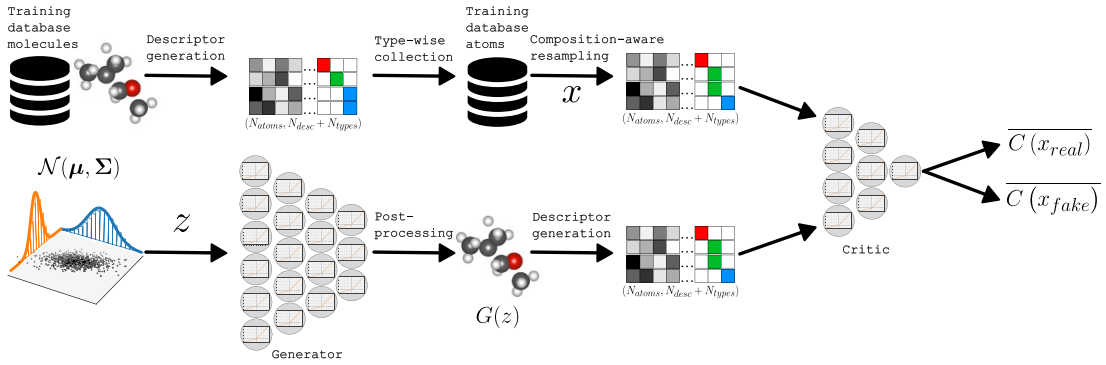

Figure 1. Simplified architecture of our proposed WGAN model. The top and bottom halves depict the flow of information from the training (real) and the generated (fake) samples respectively.

structures were then compared to the energies of the target structures to identify whether a structure was found by LoGAN. A full list of all SiO_2 structures is given in the Supporting Information.

2.2. WGAN model architecture

The backbone of our model follows the general WGAN architecture which consists of two competing neural networks, the generator and the critic [20]. The generator aims to produce samples which are similar to the ones in the training data given an atom composition chosen by the user, while the critic tries to assign low scores to the training and large scores to the generated samples. Both the generator and critic models are fully connected neural networks with core widths of $512:256:128:64:32$ and $256:128:64:32$ respectively. The generator takes 20 latent variables of random noise (see Supporting Information for results using 10 and 30 latent variables) as input and outputs $3N_{\text{atoms}}$ coordinates ordered according to the list of requested atom types. The hyperparameters reported above are the result of a manual hyperparameter search and may be changed by the user. We note that the output of the generator does not satisfy permutation invariance, which is only handled later within the critic. We use leaky ReLU [23] activation functions for both the generator and the critic, with the generator having an additional sigmoid activation applied to its output before being passed to the postprocessing step.

To handle the specific requirements of the present study, we augmented the general WGAN architechture with extra steps as shown in figure 1. For our intermediate representation to feed to the critic model, we encode the local atomic environments using element-specific spherical-Bessel descriptors [24, 25]. These begin with atom-centered density distributions:

$$\rho_{iA}(\mathbf{R}) = \sum_{a \in A; a \neq i} \delta(\mathbf{R} - \mathbf{R}_{i,a}), \tag{1}$$

where a runs through all atoms with type A and $\mathbf{R}_{i,a}$ is the vector pointing from atom i to atom a. The density distribution is limited by a chosen cut-off radius $|\mathbf{R}_{i,a}| < R_{\text{cut}}$. This density is then projected on an orthonormal basis:

$$c_{nlm;iA} = \int g_{n-l,l}(|\mathbf{R}|)\, Y_l^{m*}(\hat{\mathbf{R}})\, \rho_{iA}(\mathbf{R})\, \mathrm{d}\mathbf{R} \tag{2}$$

where Y_l^m are the spherical harmonics and $g_{n-l,l}$ the spherical Bessel radial functions [24]. The resolution of the descriptor is controlled by $n_{\max}$ which sets the number of basis functions to $(n_{\max}+1)(n_{\max}+2)/2$. Subsequently, these descriptor fragments are made rotationally invariant [26] by contracting the angular parts:

$$P_{iAB;nl} = \sum_{m=-l}^{l} c_{nlm;iA} c^*_{nlm;iB}. \tag{3}$$

The final descriptor is created by concatenating all fragments based on A, B, n, l and appending the one-hot encoded atom type of the center atom [25].

In addition to the above-mentioned parameters, R_{cut} and n_{max}, the descriptor is controlled by the parameter $N_{\text{neighbour}}$, which sets the maximum number of neighbour atoms included in the descriptor. Usually, $N_{\text{neighbour}}$ is set to large enough values so that all atoms inside the sphere defined by R_{cut} are considered. However, in our tests we found that we require relatively large R_{cut} values to avoid fragmentation of our predictions in non-periodic cases where atom clusters can be separated by more than R_{cut} distance. The constraint on the $N_{\text{neighbour}}$ value could, on the other hand, be relaxed, thereby in effect setting an adaptive cutoff radius. These descriptors are more short-sighted in dense environments, while in sparser regions they include information up to the maximum R_{cut} distance, with constant low computational cost.

The critic thus receives N_{atoms} atomic Bessel descriptors of the generated structure as input, ensuring rotational, translational, and permutation invariance. Each descriptor is of dimensionality $N_{\text{desc}} + N_{\text{types}}$ (determined by the choice of n_{max} and the number of atom types). In the Supporting Information, we report manual hyperparameter searches for R_{cut}, n_{max}, $N_{\text{neighbour}}$, and the total number of models.

The whole model, including the descriptor generation, is implemented in JAX [27], which allows us to backpropagate through the whole process for gradient descent, and utilize various accelerators such as GPUs and TPUs for training and inference.

2.3. Composition-aware resampling

One of the advantages of our model is its relatively low training data requirement. We achieve this by not using the training structures as a whole, but instead we extract every atom-centered local environment as 'real environments' for our critic model as illustrated on the top part of figure 1. Since the atomic composition of the training structures do not necessarily match the composition of the predicted structure, during preprocessing we extract the atom centered local descriptors for every atom in the structure and order them in lists based on the central atom type. For the critic to evaluate a 'real' structure, descriptors are randomly drawn from these lists according to the goal composition. For example, if the training database consists of structures with the composition $C_4H_{10}O$ and $C_6H_{14}O$, but the aim is to generate isomers with the $C_5H_{12}O$ composition, then the atomic descriptors of the generated structure, as well as 5 C, 12 H and 1 O atomic descriptors sampled randomly from the training structures are input to the critic, figure 1. This is necessary to prevent the critic from distinguishing between compositions (flagging all $C_5H_{12}O$ as fake, and all $C_4H_{10}O$ and $C_6H_{14}O$ as real), rather than discerning between physically valid and invalid environments. It furthermore enables the model to predict on entirely unseen stoichiometries without a single data entry, which is in contrast to some other current models requiring retraining on new stoichiometries [15]. Also, the resampling allows our model to operate in low-data regimes, as there is a much larger set of possible resampled descriptors than actual training structures.

2.4. Training

For training we used the RMSProp [28] optimizer as suggested by Arjovsky *et al* [20]. in the original formulation of the WGAN. Our models are trained for a fixed number of iterations, where each iteration consist of three steps. First, the generator loss is minimized w.r.t. the generator weights:

$$\mathscr{L}_{\text{gen}} = \mathbb{E}_z\left[C(G(z))\right] \tag{4}$$

where C and G stand for critic and generator model respectively, while z is a latent vector sampled from a multivariate normal distribution, see figure 1. This update is responsible for guiding the generator to predict structures, for which the local environments are similar to the ones seen in the training examples. Then the critic model is trained to distinguish between the real and fake descriptors by minimizing:

$$\mathscr{L}_{\text{crit}} = \mathbb{E}_{x,z}\left[C(x) - C(G(z))\right] \tag{5}$$

where x stands for the composition-aware resampled descriptors, figure 1. This step is applied five times per generator update to ensure a good quality of the critic predictions. Finally, the gradient penalty [29] is enforced on the critic by minimizing:

$$\mathscr{L}_{\text{grad}} = \mathbb{E}_y\left(||\nabla C(y)|| - 1\right)^2 \tag{6}$$

where y are linearly interpolated descriptors of real and fake environments of atoms of the same type. This step is applied ten times per generator update, to ensure a smooth surface of the critic predictions.

The significant advantage of the WGAN compared to the original GAN is its more stable training due to the WGAN's reliance on the Wasserstein-distance as opposed to the GAN's Kullback–Liebler divergence. For a WGAN to have stable training the critic model has to obey the Lipschitz contraint

$$|C(P_i) - C\left(P_j\right)| > K|P_i - P_j| \tag{7}$$

where P_i, P_j are two local descriptors, K is an arbitrary positive constant and $C(P_i)$ corresponds to the prediction of the critic on P_i.

In most WGANs the output of the generator model is directly fed into the critic, thus the linearly interpolated points where the gradient penalty is enforced lie in the actually accessible regions of the intermediate representation space. In our case this is not true, since the generator outputs are post-processed and turned into local descriptors before criticism, and the interpolated descriptors might not relate to geometries accessible to the generator. This potentially hinders the training of our generator, since the loss surface experienced by it might still be ragged. As a partial resolution we tested replacing the spherical Bessel radial function, g in equation (2), by Gaussian radial basis functions similar to those used in the original SOAP descriptors [26]. While these descriptors do not satisfy the same conditions for optimal completeness as their spherical Bessel function-based counterparts, they depend monotonically on $|\mathbf{R}|$. This means that if we compare the descriptors of three structures where the distances measured from the central atom are scaled with c_1, c_2 and c_3, with $c_1 < c_2 < c_3$ and the directions are unchanged, then all descriptor values of the c_2 scaled structure will lie between the descriptors of the c_1 and c_3 scaled structures. While we wanted to note this theoretical drawback of our model, we did not find significant differences between the two kind of descriptors in our tests. In the following we use both, since larger variety in the trained models can lead to larger variety in predicted structures.

2.5. Post-generation selection and relaxation

GAN models are susceptible to 'mode collapse' [30] which means that they generate samples only from a subset of the expected categories. This also occurs for the models described in this study, where a single model generates only one or a few of the possible relevant configurations of the system. We therefore always train multiple models with different starting weights and hyperparameters, and incorporate predictions from all of these. Even if our models experience mode collapse, they can collapse into different basins, thus the use of several models results in a much better sampling overall.

Subsequently we apply a filtering step to the union of generated structures. While the structure generation itself is fast, relaxing these structures using DFT can be a computationally heavy task, so that a filtering process based on structural features to avoid relaxing similar structures is beneficial. This involves calculating the same local atom centered descriptors, equation (3), as we used in the training, but aggregating them over each molecule or structure. Namely, the atomic descriptors are then averaged for each atom type and concatenated, yielding a permutation-invariant descriptor that encapsulates the entire structure:

$$\text{Structural descriptor} = \left(\frac{1}{N_A}\sum_{i\in A} P_i\right) \oplus \left(\frac{1}{N_B}\sum_{i\in B} P_i\right) \oplus \ldots \tag{8}$$

where N_A is the number of A atoms. These structural descriptors are then projected onto a lower dimensional space using principal component analysis (PCA) and clustered using k-means clustering. Finally the structures closest to the k-means centroids are chosen as candidate structures. Thus, we use a global descriptor that encompasses the full structure for clustering, while for training and structure generation, we leverage the power of atomic, local descriptors.

The selected structures were then relaxed using GPAW [31] with the LDA functional using a 400 eV plane wave cutoff. For the periodic structures a $2 \times 2 \times 2$ k-point mesh was used. For the relaxations only these coarse settings were used to reduce the computational cost, since we just required qualitative comparison with the goal structures, but no accurate energies, or even ordering of energy minima. For non-periodic systems, we furthermore confirmed that the same isomers were obtained when using the HF/STO-3G level of theory compared to DFT with the LDA functional (data not shown).

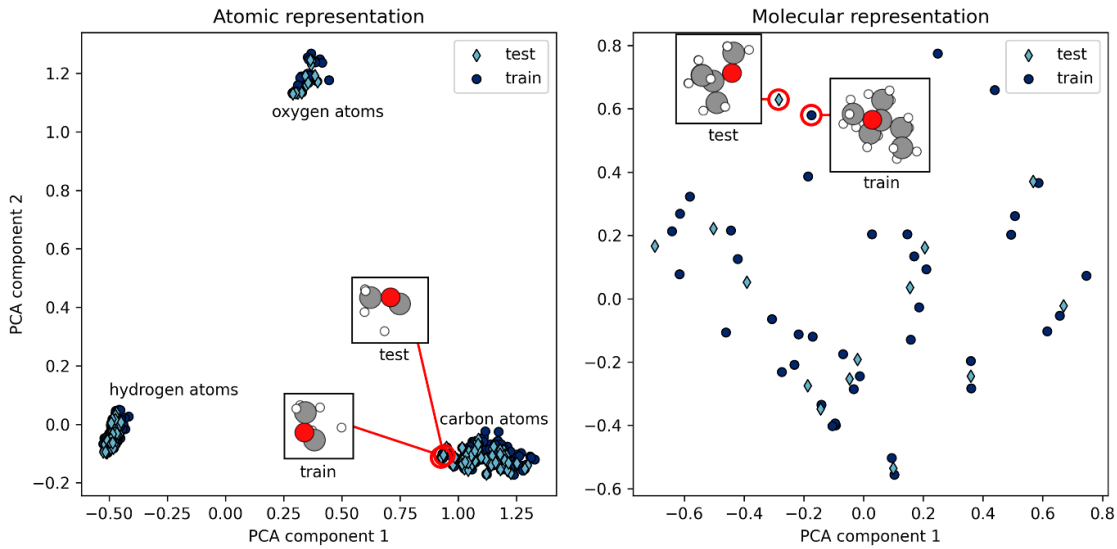

Figure 2. PCA of atomic representations of all atom types (left) and global representations of the full molecules (right). In the insets, atomic local environments are only shown up to $R_{\text{cut}} = 3$ for clarity. Bessel descriptors were computed with $R_{\text{cut}} = 5$, $N_{\text{max}} = 4$, and $N_{\text{neighbour}} = 20$.

3. Results

3.1. $C_5H_{12}O$ isomers

To benchmark the performance of LoGAN for non-periodic systems at different stoichiometries from the training set, we aimed to reproduce the 14 $C_5H_{12}O$ isomers found in the QM9 database. As training data all instances of $C_4H_{10}O$ and $C_6H_{14}O$ were used, which resulted in overall only 39 training examples.

Figure 2 depicts a PCA of the atomic and molecular representations based on Bessel descriptors encountered in the training ($C_4H_{10}O$ and $C_6H_{14}O$ isomers) and test ($C_5H_{12}O$ isomers) sets. The molecular structural descriptors are obtained via equation (8), while the atomic descriptors are the raw P_i. Thus, the atomic descriptors reflect the information that the critic receives during training, namely the local environments of which there are only few relevant ones. In contrast, the molecular descriptors reflect the conformation and diversity of the resulting predicted structures, as well as the information that the filtering procedure receives. The atom representations (figure 2 left) cluster strongly according to their atom-type. Within one atom type, the local environments of the training and test sets are very similar and follow a similar distribution. In contrast, the molecular representations (figure 2 right) of the training and test set are less similar, underlining our assumption that there are many physically valid structures, while there are only few important local environments. Therefore, a discriminator between real and fake structures can be trained more easily and efficiently on atomic environments.

We trained 12 LoGAN models with $N_{\text{neighbour}} = 12$ using the Bessel descriptors, three with each combinations of $R_{\text{cut}} = 5$ Å, $R_{\text{cut}} = 6$ Å and $n_{\text{max}} = 3$, $n_{\text{max}} = 4$. For evaluation we generated 2000 examples with each model. Figure 3(left) depicts a PCA of the predictions of two out of the twelve models, which showcases how different models experience mode collapse. Namely, both models find configurations that are close to the ground truth $C_5H_{12}O$ isomers, but neither model can find all 14 isomers. Since each model learns to focus on different structures, we can use this mode collapse to our advantage and generate structures with different models.

To break down the 24 000 predicted molecules into a more manageable number, we then applied the previously described filtration method to find 20 representative structures to relax per model. Namely, all generated structures are k-means clustered as described in the methodology and only the structure closest to each cluster center is relaxed. After relaxation using DFT we used RDKit [32] to parse the molecules to SMILES and discarded all stereoisomer information, these SMILES were then used to compare our results with the ground truth. Figure 4(left) depicts the structures each model was able to identify, where again we find that none of the models alone was able to find all possible $C_5H_{12}O$ isomers, but the union of the relaxed structures contained all target molecules. Figure 4(left) furthermore shows that both different hyperparameters and different model initializations with the same hyperparameters can help the model to

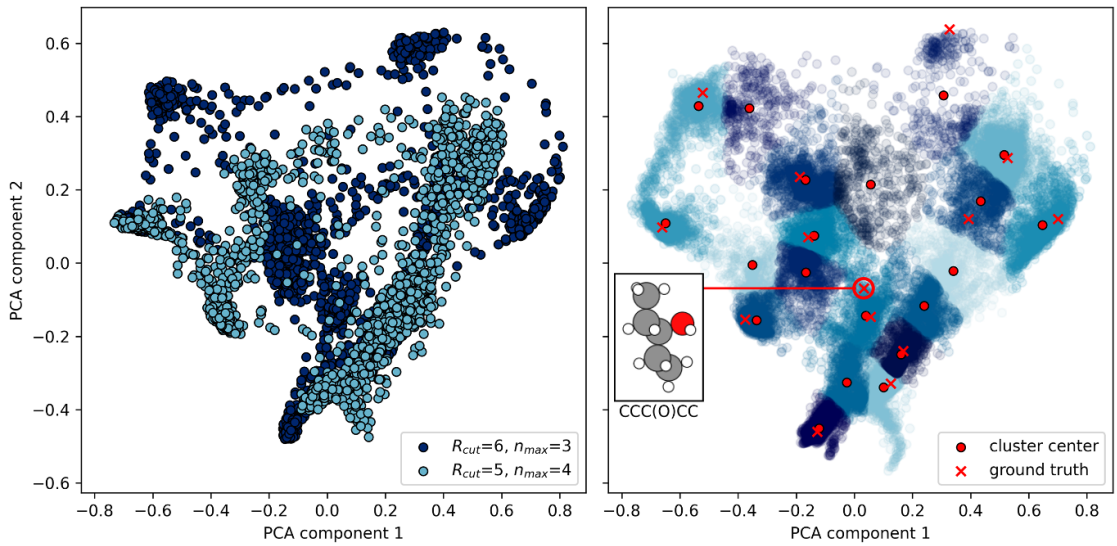

Figure 3. Left: PCA of the molecular representation of two out of twelve WGAN models. Right: Clustering of the PCA of all generated structures. Red crosses correspond to the ground truth $C_5H_{12}O$ isomers. The two ground truth structures highlighted correspond to the ones that are hardest to rediscover. Bessel descriptors were computed with $R_{\text{cut}} = 5$, $N_{\text{max}} = 4$, and $N_{\text{neighbour}} = 20$ for plotting, irrespective of the descriptors used by the individual models.

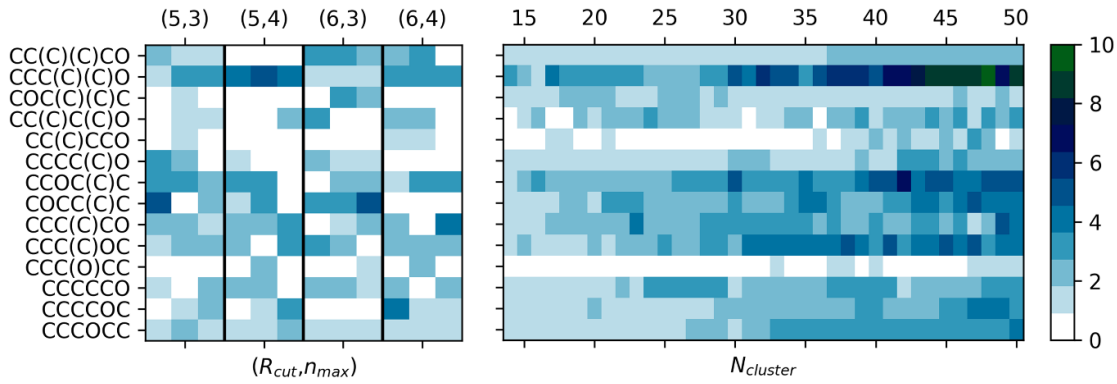

Figure 4. Left: $C_5H_{12}O$ structures found after relaxing 20 filtered candidate structures from the 12 models. While none of the individual models found all 14 isomers, every isomer occurred at least in one model. Right: $C_5H_{12}O$ structures found after relaxing N_{cluster} filtered candidate structures obtained from each cluster center, for all generated structures of all models.

collapse into different basins. For example, the first model with $R_{\text{cut}} = 5$ and $n_{\text{max}} = 3$ fails to recover five structures (white boxes in the leftmost column), which are all found by the second model with $R_{\text{cut}} = 5$ and $n_{\text{max}} = 3$ except for a single structure. Likewise, the structure most often discovered in the first model is not discovered even once by the second model.

To further reduce the number of DFT computations, we applied our filtration method on the union of all 24 000 generated structures instead of the individual models outputs to again select only a few molecules to relax. Figure 3(right) shows the clustered data for 20 clusters, where we again selected one structure per cluster to relax. Figure 4(left) depicts the discovered structures as a function of the number of clusters, ranging from 14 to 50. We find that as few as 15 relaxations can already discover 12 isomers, and all 14 isomers are discovered with at least 38 relaxations. Interestingly, two specific structures are rarely found, namely `CC(C)CCO` and `CCC(O)CC`. The structure `CCC(O)CC` is highlighted in figure 3(right), and shares a cluster with another target structure. Similarly, `CC(C)CCO` often shares a cluster with another target structure depending on the number of clusters. Since only one structure per cluster is relaxed, this makes them difficult to discover and necessitates a large N_{cluster}. Nevertheless, by relaxation of only 38 structures we are able to rediscover all possible $C_5H_{12}O$ isomers, showcasing the good performance of both our generator model and filtering method, as well as highlighting that the developed architecture is capable to easily predict systems at different stoichiometries as the training data. This is in contrast to other generative approaches where the whole structure is evaluated by the critic or decoder, corresponding to a molecular descriptor similar to the left panel in figure 2.

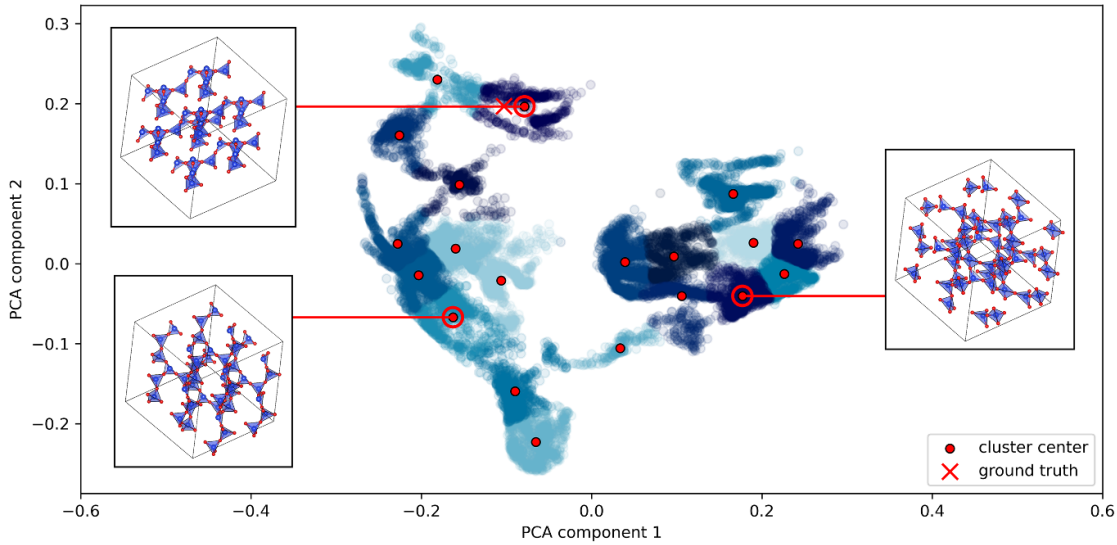

Figure 5. Illustration of our filtering step using all structures generated by all models for mp-559 550. Bessel descriptors were computed with $R_{\text{cut}} = 5$, $N_{\text{max}} = 4$, and $N_{\text{neighbour}} = 20$ for plotting, irrespective of the descriptors used by the individual models.

In the Supporting Information, we furthermore report other benchmark experiments including a comparison to GSchNet [9] for $C_5H_{12}O$, $C_6H_{14}O$, and $C_7H_{16}O$, as well as hyperparameter scans. We find that LoGAN is especially suited to extrapolate to new stoichiometries, which again is a direct result of the critic architecture being based on local instead of global environment descriptors.

3.2. SiO_2 structures

To benchmark the developed model on a periodic system, we picked six SiO_2 bulk structures with up to 18 atoms per unit cell from the materials project [22] (MP) database with low energy above the convex hull. In the following, we refer to these structures as the target structures, which we aim to recover with the LoGAN. As training data we used all MP structures with energies within 0.02 eV atom^{-1} of the convex hull. Excluding the six target structures, this led to a training data set with 167 examples in each case. As input we also required the number and types of atoms, along with the lattice vectors of the unit cell.

We trained 20 LoGAN models for each selected target structure (5 models per hyperparameter setting), with $R_{\text{cut}} = 4$ Å and $N_{\text{max}} = 4$ while using all four combinations of Gaussian or Bessel radial functions with either 12 or 25 maximum neighbour atoms. For each combination, we generated 5000 structures using the five models (1000 structures per individual model), leading to 20 000 structures overall. We then applied the previously described filtering algorithm to identify 20 structures for relaxation for each of the four hyperparameter sets. Figure 5 depicts the clustered structures obtained from a PCA of the per-structure Bessel descriptors of all 20 000 generated structures. The spread of the points and the clear clusters show that the 2D projection of the structural descriptors is able to distinguish between different structures, and that the k-means clustering algorithm is able to pick unique samples for relaxation. The insets in figure 5 furthermore depict a few structures (prior to DFT relaxation). Each structure is mainly made up of Si-O tetrahedra, which are then arranged differently inside the unit cell. This showcases again that the local structure is similar (silicon prefers to be surrounded by four oxygen atoms in a tetrahedral arrangement), while the global structure differs vastly, making the LoGAN approach both viable and efficient.

For each hyperparameter setting, each of the 20 structures identified by clustering was then geometry-optimized using DFT. Figure 6 depicts the energy distribution of the relaxed structures for each target. As a baseline, Figure 6 furthermore depicts DFT relaxations started from 80 randomly generated structures, where we enforced that atoms cannot be placed closer than $0.9d_{\text{min}}$, where d_{min} is the smallest distance between atoms in the goal structure. For each of the six target structures, the developed WGAN outperforms random initialization by a large margin. Namely, the WGAN yields lower energy structures (figure 6) and required less relaxation steps to reach convergence (not shown). Most importantly, the union of all models was able to find 100% of the target structures, while the random search only yielded a success rate of 50%. With the LoGAN approach, we are therefore able to also identify new crystal structures based on only a small set of training data without any expert knowledge in a fast and data-efficient fashion.

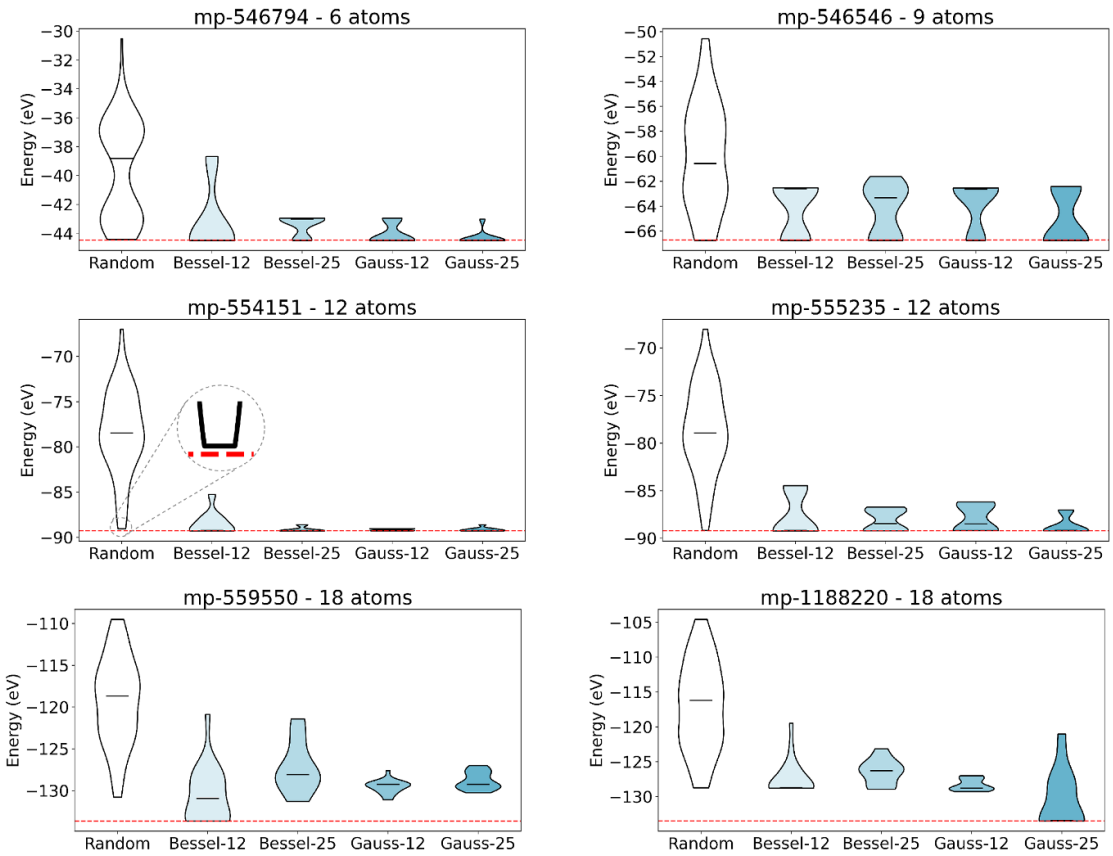

Figure 6. Energy distribution of the relaxed structures with the random search (80 structures, white) and the four different combinations of radial basis functions and $N_{\text{neighbour}}$ values (20 structures per combination, blue).

4. Conclusion

We presented LoGAN, a proof-of-concept implementation of a novel Wasserstein-GAN based architecture for atomic structure prediction of various systems. We combine ideas from previous GAN implementations with approaches seen in neural-network force fields to optimize our model for the efficient use of training data. We also introduce a selection step in our pipeline to find the unique structures predicted by our models, thus avoiding wasting large amount of computational power to evaluate the properties of similar candidates.

The idea of functional groups and local environments determining atomic structure is prevalent in chemistry, making a local approach to generate structures both attractive and plausible. We demonstrated the ability of the approach on both periodic SiO_2 target structures and on aperiodic $C_5H_{12}O$ isomers. Our model was able to find all six target SiO_2 crystal structures while the random search only succeeded in three cases. Our predicted candidates also resulted in lower average energy and required less relaxation steps until convergence during the quantum-mechanical geometry optimization. In the search for $C_5H_{12}O$ isomers, our model required only 39 training samples of similar molecules at different stoichiometries. We were able to recover all 14 possible isomers with only relaxing 38 selected structures from all 24 000 generated ones, which also highlights the efficiency of our filtering approach.

LoGAN only requires a training set of plausible structures, possibly at different stoichiometries than the target structures, without any notion of energy. We note that although LoGAN does not directly allow for a conditional structure generation with respect to a target property such as energy, the generated structures can be steered in a specific direction by changing the distribution of the training data, or by adapting the post-processing filtering steps if a model for property prediction is available. Due to its modular code, the adaption of the filtering step is straightforward, e.g. to only include low-energy conformers.

We have thus presented, for the first time, an efficient and versatile generative approach for the low-data and variable stoichiometry regime, which can generate both periodic and non-periodic structures. Due to its ease of use, fast training and inference, and clever filtration of the generated structures, we envision LoGAN to be applicable to diverse fields of research relying on the generation of plausible candidate structures.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://zenodo.org/doi/10.5281/zenodo.10944145.

Acknowledgment

This research was funded in part by the Austrian Science Fund (FWF) through SFB-TACO [10.55776/F81]. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

Data and code availability

The full datasets and code to reproduce this study, as well as utility scripts to train LoGAN on custom datasets is freely available at github.com/Madsen-s-research-group/logan and https://zenodo.org/doi/10.5281/zenodo.10944145.

ORCID iDs

Esther Heid https://orcid.org/0000-0002-8404-6596
Jasper De Landsheere https://orcid.org/0009-0003-5958-117X
Georg K H Madsen https://orcid.org/0000-0001-9844-9145

References

[1] Behler J 2015 *Int. J. Quantum Chem.* **115** 1032–50
[2] Watanabe S, Li W, Jeong W, Lee D, Shimizu K, Mimanitani E, Ando Y and Han S 2020 *J. Phys. Energy* **3** 012003
[3] Hautier G, Fischer C C, Jain A, Mueller T and Ceder G 2010 *Chem. Mater.* **22** 3762–7
[4] Schön J and Jansen M 1999 Predicting structures of compounds in the solid state by the global optimisation approach *Theoretical and Computational Chemistry* (Elsevier) pp 103–27
[5] Pannetier J, Bassas-Alsina J, Rodriguez-Carvajal J and Caignaert V 1990 *Nature* **346** 343–5
[6] Glass C W, Oganov A R and Hansen N 2006 *Comput. Phys. Commun.* **175** 713–20
[7] Wang Y, Lv J, Zhu L and Ma Y 2012 *Comput. Phys. Commun.* **183** 2063–70
[8] Arrigoni M and Madsen G K H 2021 *npj Comput. Mater.* **7** 71
[9] Gebauer N W A, Gastegger M and Schütt K T 2019 Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules (arXiv:1906.00957)
[10] Court C J, Yildirim B, Jain A and Cole J M 2020 *J. Chem. Inf. Mod.* **60** 4518–35
[11] Kim S, Noh J, Gu G H, Aspuru-Guzik A and Jung Y 2020 *ACS Cent. Sci.* **6** 1412–20
[12] Zhao Y, Al-Fahdi M, Hu M, Siriwardane E M D, Song Y, Nasiri A and Hu J 2021 *Adv. Sci.* **8** 2100566
[13] Long T, Fortunato N M, Opahle I, Zhang Y, Samathrakis I, Shen C, Gutfleisch O and Zhang H 2021 *npj Comput. Mater.* **7** 66
[14] Xie T, Fu X, Ganea O E, Barzilay R and Jaakkola T 2021 Crystal diffusion variational autoencoder for periodic material generation (arXiv:2110.06197)
[15] Meldgaard S A, Köhler J, Mortensen H L, Christiansen M P V, Noé F and Hammer B 2021 *Mach. Learn.: Sci. Technol.* **3** 015008
[16] Gebauer N W A, Gastegger M, Hessmann S S P, Müller K R and Schütt K T 2022 *Nat. Commun.* **13** 973
[17] Daigavane A, Kim S, Geiger M and Smidt T 2023 Symphony: symmetry-equivariant point-centered spherical harmonics for molecule generation (arXiv:2311.16199)
[18] Zeni C, Pinsler R, Zügner D, Fowler A, Horton M, Fu X, Shysheya S, Crabbé J, Sun L, Smith J, Nguyen B, Schulz H, Lewis S, Huang C W, Lu Z, Zhou Y, Yang H, Hao H, Li J, Tomioka R and Xie T 2024 Mattergen: a generative model for inorganic materials design (arXiv:2312.03687)
[19] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial networks (arXiv:1406.2661)
[20] Arjovsky M, Chintala S and Bottou L 2017 Wasserstein GAN (arXiv:1701.07875)
[21] Ramakrishnan R, Dral P O, Rupp M and Von Lilienfeld O A 2014 *Sci. Data* **1** 1–7
[22] Jain A *et al* 2013 *APL Mater.* **1** 011002
[23] Maas A L, Hannun A Y and Ng A 2013 Rectifier nonlinearities improve neural network acoustic models *Proc. 30th Int. Conf. on Machine Learning* vol 28 p 3 (available at: https://ai.stanford.edu/ amaas/papers/relu_hybrid_icml2013_final.pdf)
[24] Kocer E, Mason J K and Erturk H 2020 *AIP Adv.* **10**
[25] Montes-Campos H, Carrete J, Bichelmaier S, Varela L M and Madsen G K H 2022 *J. Chem. Inf. Mod.* **62** 88–101
[26] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev.* B **87** 184115
[27] Bradbury J, Frostig R, Hawkins P, Johnson M J, Leary C, Maclaurin D, Necula G, Paszke A, VanderPlas J, Wanderman-Milne S and Zhang Q 2018 JAX: composable transformations of Python+NumPy programs (available at: http://github.com/google/jax)
[28] Tieleman T and Hinton G 2012 *COURSERA: Neural Netw. Machine Learn.* **4** 26–31
[29] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A 2017 Improved training of wasserstein gans (arXiv:1701.07875)
[30] Lala S 2018 Evaluation of mode collapse in generative adversarial networks (available at: https://api.semanticscholar.org/CorpusID:214622026)
[31] Enkovaara J *et al* 2010 *J. Phys.: Condens. Matter* **22** 253202
[32] Landrum G 2006 Rdkit: Open-source cheminformatics (available at: www.rdkit.org/)