# Inverse design of 3d molecular structures with conditional generative neural networks

* Abstract: - Generative neural networks have emerged as a powerful approach to sample novel molecules from a learned distribution.
* In this paper researchers proposed a conditional generative neural network for 3d molecular structures with specified chemical and structural properties.
* The researchers demonstrate the utility of their method for inverse design by generating molecules with specified motifs or composition, discovering particularly stable molecules, and jointly targeting multiple electronic properties beyond the training regime.
* Introduction: - As an exhaustive exploration of the vast chemical compound space is infeasible, progress in these areas can benefit substantially from inverse design methods.
* ML become more prominent in the sector of exploration of chemical compound space. Here, the number of reference calculations required from training ML models depends on the size of the domain to be explored.
* Therefore, the naïve exploration schemes may still require a prohibitive number of electronic structure calculations.
* It gives rise to the idea of inverse molecular design, where the structure-property relationship is reversed. The challenge with this is that to directly construct molecular structures corresponding to a given set of properties.
* Generative ML models have recently gained traction as a powerful, data-driven approach to inverse design as they enable sampling from a learned distribution of molecular configurations.
* These methods typically represent molecules as graphs or SMILES strings, which lack information about the three-dimensional structures with desirable characteristics for further evaluation.
* Connectivity-based representations are problematic in chemical systems where bonding is ambiguous, e.g., in transition metal complexes, conjugated systems or metals.
* Recently, generative models that enable sampling of 3d molecular configurations have been proposed.
* The other models aim at sampling directly from distributions of 3d molecules with arbitrary composition, making them suitable for general inverse design settings.
* These models need to be biased towards structures with properties of interest, e.g., using reinforcement learning, fine-tuning on a biased dataset, or other heuristics.
* Some of the researchers previously proposed G-SchNet, the G-SchNet has been biased by fine-tuning on a fraction of the training dataset containing all molecules with a small HOMO-LUMO gap. For this, a sufficient amount of training examples in the target space is required.
* In this paper the researchers proposed conditional G-SchNet (cG-SchNet), a conditional generative neural network for the inverse design of molecules.
* Building on G-SchNet, the model learns conditional distributions depending on structural or chemical properties allowing us to sample corresponding 3d molecular structures.
* The researchers’ architecture is designed to generate molecules of arbitrary size and does not require the specification of a target composition.
* It also learns the relationship between the composition of molecules and their physical properties in order to sample candidates exhibiting given target properties, e.g., preferring smaller structures when targeting small polarizabilities.
* The researchers conditional approach permits searching for molecules with any desired set of target property values after training is completed.
* The model is able to jointly target multiple properties without the need to retain or otherwise indirectly constrain the sampling process.
* They demonstrate that cG-SchNet enables the exploration of sparsely populated regions that are hardly accessible with unconditional models.
* They conduct extensive experiments with diverse conditioning targets including chemical properties, atomic compositions and molecular fingerprints.
* Through this way the researchers generate novel molecules with predefined structural motifs, isomers of a given composition that exhibit specific chemical properties, and novel configurations that jointly optimize HOMO-LUMO gap and energy.
* Results: - target 3d molecule generation with cG-SchNet – the researchers’ molecules as tuple of atom positions with and corresponding atom types with .
* The cG-SchNet assembles these structures from sequences of atoms that are placed step by step in order to build the molecule in an autoregressive manner, where the placement of the next atom depends on the preceding atoms.
* The cG-SchNet, learns as unconditional distribution over molecules, cG-SchNet samples from target-dependent conditional probability distributions of 3d molecular structures.
* A tuple of *k* conditions cG-SchNet learns a factorization of the conditional distribution of molecules, i.e., the joint distribution of atom positions and atom types conditioned on the target properties:



* The researchers can split up the joint probability of the next type and the next position into the probability of the next type and the probability of the next given the associated next type:



* This allows predicting the next type before the next position. The researchers approximate the distribution over the absolute position from distributions over distances to already placed atoms



* It guarantees that it is equivariant with respect to translation and rotation of the input. In this is the normalization constant and is the distance between the new atom *i* and a previously placed atom *j*.
* The conditions are each embedded into a latent space and concatenated, followed by a fully connected layer.
* According to the principle, any combination of properties can be used as conditions with researcher architecture with a suitable embedding network.
* The researcher used three scalar-valued electronic properties such as isotropic polarizability, vector-valued molecular fingerprints, and the atomic composition of molecules.
* Vector-valued properties are directly processed by the network while scalar-valued targets are first expanded on a Gaussian basis.
* To localize the atom placement and stabilize the generation procedure, cG-SchNet makes use of the same two auxiliary tokens as in the unconditional setting, namely the origin and the focus token.
* Auxiliary tokens are treated like regular atoms by the model, i.e., they possess positions and token types, which are contained in the tuples of atom positions and atom types serving as input at each step.
* The original token marks the center of the mass of molecules and allows the architecture to steer the growth from inside to outside.
* The focus token localizes the prediction of the next position in order to assure scalability and allows to break symmetries of partial structures.
* At each step, the focus token is randomly assigned to a previously placed atom. This way, they can use a small grid localized on the focus that does not grow with the number of atoms when predicting the distribution of the next position.
* They train the cG-SchNet on a set of molecular structures, where the values of properties used as conditions are known for each molecule. Given the conditions and the partial molecular structure at each step, cG-SchNet predicts a discrete distribution for the type of next atom.
* After the sampling a type, cG-SchNet predicts distributions for the distance between the atom to be placed and each preceding atom and auxiliary token.
* During the training the researchers minimize the cross-entropy loss between the predicted distributions and the ground-truth distribution known from the reference calculations.
* Generating molecules with specified motifs. – many applications, it is advantageous for molecules to process specific functional groups or structural motifs. It can be correlated with desirable chemical properties, e.g., polar groups that increases solubility, or with improved synthetic accessibility.
* The researchers condition cG-SchNet on a path-based, 1024 bits long fingerprint that checks molecular graphs for all linear segments of the seven atoms.
* The researchers condition the sampling on fingerprints of unseen molecules, i.e., structures not used during training.
* They observed that the generated molecules have a higher similarity with the target similarity are also sampled with higher probability, as scan be seen from the increased similarity score of generated duplicates.
* In conclusion we can say that the researchers find that the conditional sampling with cG-SchNet is sensitive to the target fingerprint and allows for the generation of molecules with desired structural motifs.
* Although there are no molecules with the same fingerprint in the training data for three of the four fingerprint targets, the ML model successfully generates perfectly matching molecules, demonstrating its ability to generalize and explore unseen regions of chemical compound space.
* Generalization of condition-structure relationship across compositions – for inverse design tasks, integrating information gained from different structures and properties is vital to obtain previously unknown candidates with desired properties.
* The model has to learn from other compositions how molecules with particularly high or low HOMO-LUMO gaps are structured, and transfer this knowledge to the target composition.
* The researchers restrict the training data consisting of 55k molecules from QM9 to contain no C7N1O1H11 isomers with HOMO-LUMO gap values outside the intermediate range. Therefor the model can only learn to generate molecules with gaps outside this range from compositions other than C7N1O1H11.
* The majority of generated isomers exhibit gap values close to the respective target (1 eV), i.e., outside of the range observed for these isomers by the model during training.
* This demonstrates that cG-SchNet is able to transfer knowledge about the relationship between structural patterns and HOMO-LUMO gaps learned from molecules of other compositions to generate unseen C7N1O1H11 isomers with outlying gap values upon request.
* Discovery of low-energy conformations – a generative model needs to fill the sparsely sampled regions of the space, effectively enhancing the available data with novel structures that show property value of interest.
* The identification of low-energy conformations is desirable in many practical applications, since they tend to be more stable.
* Because the researchers are only interested in the energy contribution of the spatial arrangement sampled by the model, they require a normalized energy, which indicates whether the internal energy per atom is relatively high or low compared to other molecules of the same composition in the dataset.
* Negative values indicate comparatively low energy, and thus higher stability than the average structure of this composition.
* Using the relative atomic energy allows cG-SchNet to learn the influence of the spatial arrangement of atoms on the energy and transfer this knowledge to the unseen target composition.
* The researchers’ sample 100k molecules with the trained cG-SchNet conditioned on the composition C7O2H10 and a relative atomic energy value of -0.1 eV, i.e., close to the lowest energies occurring for these isomers in QM9.
* 169 of the 200 isomers with the lowest relative atomic energy in the test set have been recovered by the model as well as 67% of the 1k isomers with relative atomic energy lower than -0.05 eV.
* The researchers found that 32% more unique C7O2H10 isomers with relative atomic energy lower than -0.05 eV with their model than already contained in QM9.
* Examples illustrate that cG-SchNet samples molecules that are close to equilibrium configuration and thus require only a few steps of relaxation with DFT or a neural network potential.
* The model has discovered several atomic energies than those in QM9. As cG-SchNet generalizes beyond the chemical diversity of QM9, this demonstrates that it can be employed to systematically enhance a database of molecular structures.
* Targeting multiple properties: Discovery of low-energy structures with small HOMO-LUMO gap – a method for exploration needs to allow for the specification of several conditions at the same time.
* The researchers demonstrate the above ability by targeting HOMO-LUMO gap as well as relative atomic energy, i.e., two complex electronic properties as the same time.
* Previously the researchers biased an unconditional G-SchNet in order to sample molecules with small HOMO-LUMO gap.
* In this they demonstrate that improved results can be achieved with the cG-SchNet architecture while using fewer training samples from the target region.
* They even further conditions the sampling to particularly stable, low-energy conformations. In a fine-tuning approach, this would limit the training data to only a few molecules that are both stable and exhibit small gaps. In contrast, the conditional model is able to learn also from reference, calculations where only one of the desired properties is present.
* Because the energy range has not been restricted for the biased G-SchNet, it samples structures that capture the whole space spanned by the training data, i.e., also fewer stable molecules with higher relative atomic energy.
* The molecules generated with cG-SchNet, in contrast, are mostly structures with low relative atomic energy.
* The researchers observed that cG-SchNet samples a significantly larger number of structures from the low-energy domain than the biased G-SchNet.
* The model has learned to build molecules close to the target conditions that contain more than nine heavy atoms, i.e., larger than the structures from the training data.
* The molecules found with cG-SchNet contain more double bonds and a larger number of rings, mainly consisting of five or six atoms.
* It indicates a prevalence of aromatic rings and conjugated systems with alternating double and single bonds, which are important motifs in organic semiconductors.
* The same patterns can be found for molecules from biased G-SchNet, however, there is an increased number of nitrogen and oxygen atoms stemming from less stable motifs such as rings dominated by nitrogen.
* The molecules of biased G-SchNet tend to contain highly strained small cycles of three or four atoms. cG-SchNet successfully averts these undesirable motifs when sampling molecules with a low relative atomic energy target.
* The researchers conclude that cG-SchNet has learned to build stable molecules with a low HOMO-LUMO gap even though it has seen less than half of the structures that the biased model was finetuned on.
* The model is able to leverage information even from structures where one of the properties is outside the targeted range. Consequently, it is able to sample a significantly higher number of unseen molecules from the target domain than there is structure in the training data that fulfill both targets.
* Through this way multiple properties can be targeted at once in order to efficiently explore chemical compound space.
* The efficiency of cG-SchNet in finding molecular structures close to the target conditions is particularly evident compared to an exhaustive enumeration of graphs with subsequent relaxation using DFT.
* In both cases the relaxation required to obtain equilibrium coordinates and the physical properties is the computational bottleneck and takes more than 15 min per structure for the molecules generated in this experiment.
* If we calculation the internal energy at zero Kelvin (U0) we need additional 40 min per molecule.
* In contrast, the generation with cG-SchNet takes only 9 ms per structure on the GPU Nvidia A100 when sampling in batches of 1250.
* The efficiency is determined by the number of molecules that need to be relaxed for each method. The QM9 dataset was assembled by relaxing structures from the GDB enumeration of graphs for small organic compounds.
* The model has unveiled valid molecules close to the target that are not contained in the dataset. The researchers obtain more than two times number of molecules close to the target property values with cG-SchNet than with the exhaustive enumeration method while requiring less than 10% of the computation time.
* The conditional model is not restricted to the space of low-energy/low gap molecules, but can also sample low-energy/high gap structures or any other combination of interest.
* The efficiency of the generative model becomes even more pronounced when there are multiple sets of desirable target values.
* In this the same model is employed to sampled molecules for five different target values. Again, cG-SchNet is able to generalize to isotropic polarizabilities beyond the values present in the training data.
* Discussion: - cG-SchNet enables the targeted discovery of 3d molecular structures conditional on arbitrary combinations of multiple structural and chemical properties.
* The neural network captures global and local symmetries of molecular structures by design, enabling it to learn complex relationships between chemical properties and 3d structures.
* In previous approaches the model does not require target-specific biasing procedures. Instead, the explicit conditioning enables cG-SchNet to learn efficiently from all available reference calculations.
* Through this way, cG-SchNet generates novel 3d candidate molecules that exhibit the target properties with high probability and thus are perfectly suited for further filtering and evaluation using ML force fields.
* Further work is required to apply the cG-SchNet architecture to the exploration of significantly larger systems and a more diverse set of atom types.
* Adjustments are necessary to make the scalability to materials.
* In current implementation, the researchers employ all preceding atoms to predict the type and reconstruct the positional distribution of the next atom.
* Another direction for future work is the extended comparison of cG-SchNet to established methods in different fields, e.g., for the discovery of drugs or materials, to identify promising applications and possible short-comings.
* In cases where not all targeted properties can be fulfilled simultaneously, finding suitable molecules becomes harder, if not impossible.
* The researchers applied cG-SchNet to sample particular stable, low-enrgy C7O2H10 isomers. During this process they find that molecules and motifs that are absent from the QM9 database, such as isomers with carboxylic acid groups.
* Their model considerably accelerates the process by providing reasonable candidate structures. cG-SchNet also enables the data-efficient, systematic improvement of chemical databases, which is particularly valuable considering the computational cost and unfavorable scaling of electronic structure calculations.
* Methods: - Training data – for each training run, 55k reference structures are randomly sampled from the QM9 dataset, collection of 133,885 molecules with up to nine heavy atoms from carbon, nitrogen, oxygen, and fluorine.
* They removed 915 molecules from the training pool which are deemed invalid by their validation procedure that checks the valency and connectedness of generated structures.
* They train the neural network using 50k randomly sampled molecules and employ the remaining 5k for validation.
* Details on the neural network architecture – they used the shifted softplus non-linearity throughout the architecture.



* Successive linear neural network layers with intermediate shifted softplus activation are written as



* With input x , weights , , and biases , . It shows the succession of two linear layers.
* The inputs to cG-SchNet when placing atom *I* is a partial molecule consisting of *i –* 1 atoms including two auxiliary tokens and *k* target properties .
* The atoms and tokens are given as tuples of positions with and types and .
* The first two entries correspond to the auxiliary tokens, which are treated like ordinary atoms by the neural network. Therefore, whenever the researchers refer to atoms in the following, this also encompasses the tokens.
* The token doesn’t influence the sampling probability of a molecule in Eq. 1, since they are placed with probability *p*=1.
* Each target property is first mapped into vector space using an individual embedding network that depends on the form of the specific property.
* In this work, the researchers employ different embedding networks for scalar-valued properties, vector-valued properties, and atomic composition.
* Scalar-valued properties are processed by an MLP after applying a Gaussian radial basis function expansion



* Here the minimum and maximum property values and the grid spacing are hyper-parameters chosen per target property.
* Vector-valued properties such as molecular fingerprints are directly processed by an MLP:



* The researchers used two embedding blocks for atomic composition. The researchers map atom types to learnable embeddings .
* These vectors are weighted by the fraction of the corresponding atom type in the target atomic composition, concatenated, and processed by an MLP.
* The atomic composition of hydrocarbons would be encoded as:



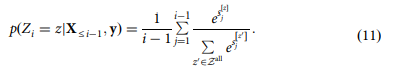
* Here “” is the concatenation of two vectors and and is the fraction of hydrogen and carbon atoms in the target atomic composition, respectively.
* At last, property feature vectors are aggregated by an MLP to obtain the combined conditional features y.



* The cG-SchNet architecture predicts distributions for the type of the next atom and its pairwise distances to all preceding atoms with two output networks.
* The type prediction network first computers atom-wise ||-sized vectors containing a scalar score for each atom type.



* The probability for the next atom being of type *z* is obtained by taking the software over all types and averaging the atom-wise predictions:



* The distance distributions are discretized on a grid with *L* bins, each covering a span of . The bin of a distance *d* is given by *b*: ,



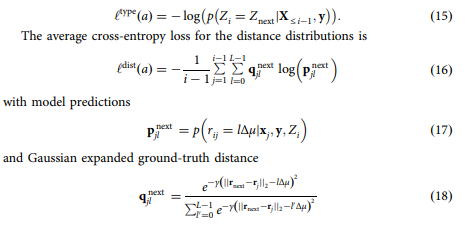
* given the type of the next atom, the distance prediction network computers scores for each preceding atom and distance bin



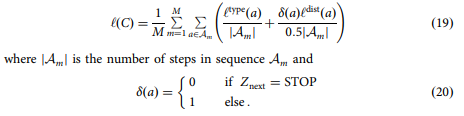
* Here “” is the Hadamard product and is a learnable atom type embedding. The probability of any distance between the new atom and a preceding atom is obtained by applying a softmax over all bins



* Where is the score of bin *b*(*d*) predicted for preceding atom *j*.
* Sampling atom placement sequences for training – the number of sequences in which a molecule can be built by placing *n* atoms grows factorially with *n*. The researchers focus and origin tokens to constrain how molecules are built by cG-SchNet and thus significantly reduce the number of possible sequences.
* The model approach ensures that the molecules grow outwards starting from the center of mass and that each new atom is placed close to one of the already placed atoms.
* The researchers set the position of the focus and origin tokens to the center of mass of the training molecule and choose the atom closest to it as the first atom to be placed. If multiple atoms are equally close, one of them is randomly chosen as the first atom.
* One of already token is set to the position of the chosen atom. Then, closest neighbors of focus atom are chosen. If there are no neighbors of the focus among the unplaced atoms, then the researchers insert a step where the type prediction network shall predict the stop marker type.
* Through this way the model’s focus atom is marked as finished before randomly choosing a new focus and proceeding with the next atom placement step.
* Neural network training – Mini batches *M* contains one atom placement sequence per molecule, randomly sampled in each epoch.
* Each step of the atom placement sequence consists of types and positions of already placed atoms and the two auxiliary tokens, of the values of molecule *m* for the target properties of the model, and of the type Znext and position **r**next of the next atom.
* For each replacement of the atom, the researchers minimize the cross-entropy between the distributions predicted by model given , and and the distributions obtained from the ground-truth next type Znext and position **r**next.
* The cross-entropy loss for the type distributions is



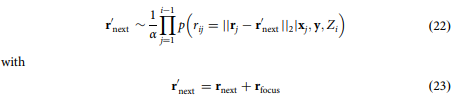
* Here *L* is the number of bins of the distance probability grid with spacing . The width of the Gaussian expansion can be tuned with , which the researchers set to in their experiments.
* The loss for a mini-batch *C* is the average type and distance loss of all atom placement steps of all *M* molecules in the mini-batch:



* The indicator function is zero for steps where the type to predict is the stop marker.
* The NN were trained with stochastic gradient descent using the ADAM optimizer.
* Conditional generation of molecules – the generation is an iterative process where the type and position of each atom are sampled sequentially using the distributions predicted by cG-SchNet.
* A molecule with *n* atoms takes 2*n* steps, as each atom needs to be placed and furthermore marked as finished in order to terminate the generation process.
* At each step, the researchers want to sample the type and position of the next atom given the types and positions of already placed atoms and the conditions.
* In this is the set of all atom types in the training data including an additional stop marker type and **G** is a grid of candidate position in 3d space.
* They predict the distribution of the type of the next atom with the model to sample the next type



* Because cG-SchNet is trained to placed atoms in close proximity to the focused atom, the researchers align the local grid of candidate positions with the focus at each step regardless of the number of atoms in the unfinished molecule.
* The position of the next atom is drawn accordingly



* Here is the normalization constant and is the position of the focus token.
* The generation process terminates when all regular atoms have been marked as finished. The researchers limit the model to maximum number of 35 atoms. If the model attempts to place more atoms, the generation terminates and the molecule is marked as invalid.
* Checking validity and uniqueness of generated molecules – the researcher use Open Babel, Open Babel assigns bonds and bond orders between atoms to translate the generated 3d representation of atom positions and types into a molecular graph.
* The generated structure is considered invalid if it consists of multiple disconnected graphs.
* The researchers found that the Open Babel may struggle to assign correct bond orders even for training molecules if they contain aromatic sub-structure made of nitrogen and carbon.
* The uniqueness of generated molecules is checked using their canonical SMILES string representation obtained from the molecular graph with Open Babel.
* The researchers check the canonical SMILES string of mirror images of generated structures, which means that mirror-image stereoisomers are considered to be the same molecule in researchers’ statistics.
* Molecules from the training and test data are matched with generated structures in the same way, using their canonical SMILES representations obtained with Open Babel and the custom heuristic for bond order assignment.
* The researchers use isomeric SMILES strings that encode information about the stereochemistry of 3d structures.
* Prediction of property values of generated molecules – the researchers use pretrained SchNet models from SchNetPack to predict the HOMO-LUMO gap, isotropic polarizability, and internal energy at zero Kelvin of generated molecules.
* The reported mean absolute error (MAE) of these models is 0.074 eV, 0.124 Bohr, and 0.012 eV, respectively.
* They relax generated molecules for every experiment in order to assess how close the equilibrium configurations and to calculate the MAE between predictions for generated, unrelaxed structures and the computed ground-truth property value of the relaxed structure.

Niklas W. A. Gebauer, Michael Gastegger, Stefaan S. P. Hessmann, Klaus-Robert Müller, Kristof T. Schütt