Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 2927

Moment of inertia as a simple shape descriptor for diffusion-based shape-constrained molecular generation

Denis Sapegin, *[ab] Fedor Bakharev, [b] Dmitriy Krupenya, [b] Azamat Gafurov, [b] Konstantin Pildish [b] and Joseph C. Bear *[a]

The article introduces *MLConformerGenerator*, a machine-learning framework for shape-constrained molecular generation that combines an Equivariant Diffusion Model (EDM), guided by a compact shape descriptor based on the principal components of the moment of inertia tensor, and a Graph Convolutional Network (GCN) model for bond prediction. The compact yet informative descriptor provides concise representation of molecular shape, enabling scalable learning from large datasets and synthetic conformers generated from 2D molecular inputs. The use of a GCN for bond prediction is evaluated in comparison to deterministic methods. The suggested approach provides an ability to fine-tune the model to generate datasets with chemical-feature distributions closely matching those of target datasets of real conformers. The proposed model supports generation conditioned on both explicit conformers and arbitrary shapes, offering flexibility for applications such as dataset augmentation and structure-based molecule design. Trained on over 1.6 million molecules, the model demonstrates the ability to generate chemically valid, structurally diverse molecules that conform to target shape constraints. It achieves an average shape similarity of 0.53 to a reference conformer, with peak similarity exceeding 0.9 – a performance comparable to that of analogous models relying on more complex descriptors. The results show that integrating physically grounded descriptors with modern generative architectures provides a robust and effective strategy for shape-constrained molecular design.

Received 18th July 2025
Accepted 14th August 2025

DOI: 10.1039/d5dd00318k

rsc.li/digitaldiscovery

1 Introduction

Machine-learning-based generative methods offer powerful tools for the automatic creation of a wide range of objects.[1] Within cheminformatics, these approaches are especially valuable for tasks of conditional molecular generation, as they allow balancing the many interdependent factors which need to be considered for successful construction of a molecule. Numerous strategies have been reported,[2–7] and can broadly be grouped into string-based methods – relying on textual representations such as SMILES[2,3] and graph-based methods, which aim at explicit construction of molecular graphs.[4–7] While the relative simplification of the molecule in a string-based representation is advantageous in many applications, where the generation of chemical structure alone is the main point of interest, for some tasks the expressionability of these formats is not sufficient.

One of such tasks is shape-constrained generation, which can be formulated as generation of a molecular conformer capable of fitting into or replicating a shape of interest. Beyond generating a valid chemical structure, this task requires ensuring that the molecule can adopt a geometry consistent with the given constraint. Such challenges are abundant in many areas of chemistry, including design of host–guest molecular systems,[8] structural based drug-design,[9] organometallic chemistry, especially catalysis,[10,11] and are typically addressed manually through expert-based design approaches. Because molecular conformer design is inherently complex and demands substantial expertise, the application of machine-learning-based generative methods seems extremely attractive for its facilitation.

String-based molecular representations usually do not adequately represent the spatial geometry of the molecule to an extent necessary to be applied to conformer design. In contrast, graph-based representations provide a more information-rich alternative by directly encoding the topology, relationships and features of atoms. Models based on graph neural networks have demonstrated remarkable success in generative tasks. An illustrative example is the equivariant diffusion model (EDM) introduced by Hoogeboom *et al.*,[5] which relies on graph networks capable of handling both discrete (categorical) and continuous features to perform conditional generation of sensible three-dimensional molecular geometries.

[a]*Department of Chemical and Pharmaceutical Sciences, Kingston University, Penrhyn Rd, Kingston upon Thames KT1 2EE, UK. E-mail: denis.sapegin@quantori.com*
[b]*Quantori, 625 Massachusetts Ave, Cambridge, MA 02139, USA*

A key consideration in application of EDMs to shape-constrained generation is the choice of the shape descriptor. Many existing approaches rely on autoencoders to capture shape from point clouds representing molecular surfaces, as demonstrated by Chen *et al.*[6] and Adams *et al.*[7] Although autoencoders can effectively produce latent embeddings that encode a molecule's geometry, this strategy requires additional training of the encoder. In contrast, physical property-based shape descriptors require no extra training step, making them more straightforward to implement. A particularly simple yet powerful descriptor is the set of principal components of the moment of inertia (MOI) tensor, which is inherently O(3)-invariant in a principal axis frame with the origin at the center of mass. As shown by Cheng and Lo in (KREED)[12] and (Stiefel Flow Matching)[13] for the case of molecular structure elucidation from rotational spectroscopy data,[14] three floating-point values of the principal moments of inertia can robustly capture a molecule's overall geometry. This general, physically grounded descriptor can be applied for generation either guided by a specific reference molecule or an entirely arbitrary shape constraint. The MOI tensor can be defined for any object with a specific shape and density. It inherently acts as a dimensionality reduction operator for shapes by efficiently representing the mass distribution in 3D space using a 3×3 symmetric tensor. Through diagonalization, the tensor yields three principal moments of inertia, which serve as compact and rotation-invariant descriptors of the object's geometry. Such representation not only captures essential geometric characteristics but also enables generalization of model predictions to arbitrary shapes, even when the model is initially trained on molecular structures. As long as the principal moments of inertia of the arbitrary shape are similar to those in the training dataset, the model can effectively generalize across diverse geometries.

As noted by Vignac *et al.*,[15] most 3D molecule generators focus on predicting atom positions and types only, while depending on semi-empirical methods[16] for restoration of the bonds within the generated molecules.[5–7,17] Although these techniques can achieve reasonable accuracy, their flexibility is often limited. Furthermore semi-empirical algorithmic tools for bond evaluation do not consider the target distribution of the chemical features, therefore may negatively impact the quality of the generated molecular sets.

This study aims to demonstrate how a simple physically grounded descriptor can facilitate efficient, geometry-aware molecular design within an EDM-based framework. We introduce an equivariant diffusion-based model augmented with a graph convolution network (GCN) module for atom adjacency restoration – MLConformerGenerator. The model utilises the principal components of the moment of inertia as a simple shape descriptor for conditional molecule generation. To address challenges inflicted by semi-empirical algorithmic prediction of molecular graph connectivity, we consider Structure Seer, which infers atom adjacency from general atom descriptors,[18] as a potential alternative for adjacency restoration. Given 3D coordinates and initial connectivity, a modified Structure Seer model can reconstruct bonds in a trainable manner. The suggested approach enables shape-constrained generation from either a reference molecule or an arbitrary shape.

Our central claim is that by training the model to match the chosen shape descriptor, it effectively learns both reference-specific and arbitrary shapes when trained on automatically generated 3D conformers. This helps to significantly augment the training dataset and achieve promising results even for generation of relatively large molecules, containing up to 39 heavy atoms.

2 Methods

2.1 MLConformerGenerator architecture

The MLConformerGenerator framework consists of three primary components: an EDM block for the initial generation of atom coordinates and types, a Graph Convolutional Network (GCN) block for bond classification and a deterministic structure standardisation pipeline. The EDM block generates atoms based on the requested number of heavy atoms in the molecule, while the GCN block utilizes interatomic pairwise distance information to predict bond types.

The EDM block follows the conditional generation framework described by Hoogeboom *et al.*,[5] with several modifications. Atomic charges are not considered during the generation process. The denoising model is structured as an Equivariant Graph Neural Network composed of nine equivariant blocks, each containing two Graph Convolutional Layers (GCL) with 420 hidden features and one equivariant update layer with 420 hidden features. The model operates on eight atom types: carbon (C), nitrogen (N), oxygen (O), fluorine (F), phosphorus (P), sulfur (S), chlorine (Cl), and bromine (Br), while hydrogen atoms are not explicitly considered during generation. To condition the generation, the model uses the principal components of the MOI tensor as a context, represented with a floating-point vector of size three. The context was normalised using mean-MAD (Mean absolute Deviation) normalisation based on the distribution of context values within the training dataset. The normalisation was aimed at reducing the impact of scale differences and enhancing model stability during training.

For bond prediction, the GCN block (termed AdjMatSeer) has been redesigned from the original model described in ref. 18 to better address the adjacency prediction task within the proposed pipeline. The GCN encoder generates adjacency matrices using embedded atom types with embedding dimension of 64 and pairwise interatomic distances obtained from the EDM output. An initial distance matrix is utilized for preliminary embedding generation, while a Boolean adjacency matrix, derived from the distance matrix by applying a threshold, is used for final bond classification. Bond types are classified into five categories: no bond (0), single bond (1), double bond (2), triple bond (3), and aromatic bonds (4). To simplify the architecture while maintaining predictive performance, the Transformer decoder layer from the original model was omitted. The revised architecture consists of three layers dedicated to embedding generation from the distance matrix, followed by

four additional layers that operate on the Boolean adjacency matrix for final bond classification based on the embeddings. Each layer contains 2048 hidden features. The models were implemented using the PyTorch library.[19]

2.2 Datasets (preparation)

To construct a suitable dataset for training, the ChEMBL database[20] was selected as the primary source of molecular structures. The ChEMBL database was considered well-suited for this purpose because it contains manually curated chemical compounds represented in a unified and standardised manner.

For this study, the small-molecule subset of ChEMBL was examined, focusing on molecules containing 15 to 39 heavy atoms. Key features of the training dataset are presented in Table 1. The heavy atom count range was chosen because it encompasses 85.9% of the small-molecule subset, making it representative for modeling. The distribution of molecules within this atom count range (Table 1) was suitable and representative of the broader chemical space, as molecule frequencies across different heavy atom counts are comparable within the selected subset. The balanced distribution ensures that the dataset adequately captures the diversity of chemical structures necessary for model training.

For training the EDM block, it is crucial to examine the distribution of the principal components of the MOI tensor values, within the dataset. Calculating their mean and mean absolute deviation is essential for normalisation. Since the structures in the ChEMBL database are generally represented as 2D molecular graphs without explicit information on their 3D conformations, we opted to generate random conformers using the Distance Geometry Embedding Algorithm[21] implemented in the RDKit library.[22]

To assess how random conformer generation affects the mean and mean absolute deviation of the context values within the dataset, they were calculated independently three times for the entire dataset. The results are provided in Table 2.

Our experiments demonstrated that the mean and MAD values of the principal MOI tensor components remain relatively stable across different runs when conformers are generated algorithmically (Table 2). This consistency indicates that:

- The synthetic dataset adequately represents the general molecular shape, ensuring that the generated data is representative of the shape of the structures.
- The algorithmic conformer generation approach may be considered viable and reliable for the training process.

Since the EDM block learns to produce molecular structures that match the principal MOI components, we argue that the geometry optimization of molecules during training is not strictly necessary. Even if random conformers are used, a subset of these generated structures will inevitably be close to the optimized or experimentally determined geometries of the target compounds. This justifies the use of randomly generated conformers for training without compromising the model's ability to generalize to real molecular structures. When considering this approach further, it should be noted that the rationale for relying on generated conformers without application of any energy-based filtering is twofold. First, we aim to expose the models to a broader and more diverse conformational space. In many real-world scenarios, a molecule's actual conformation can vary significantly depending on its environment and may not correspond to the minimum-energy geometry under specific conditions. Less energetically favorable conformers may still be valid in particular contexts—such as binding to a protein, interacting with other molecules, or existing within a crystal lattice. By training on a wide range of conformers, the model learns to associate the global shape descriptor (MOI) with plausible molecular structures, based on realistic bond lengths and angles, rather than being biased toward a narrow set of energy-minimized geometries that are only valid under certain assumptions. This encourages the EDM

Table 1 Key parameters of the ChEMBL subset used for training

Total number of molecules	1 641 644		
Permitted types of heavy atoms	C, N, O, F, P, S, Cl, Br		
Principal components of MOI tensor	I_{xx}	I_{yy}	I_{zz}
Mean value	104.79 ± 0.01	472.96 ± 0.07	537.33 ± 0.07
Mean absolute deviation within dataset	52.03 ± 0.01	219.79 ± 0.08	232.97 ± 0.07

Distribution of molecules by number of heavy atoms

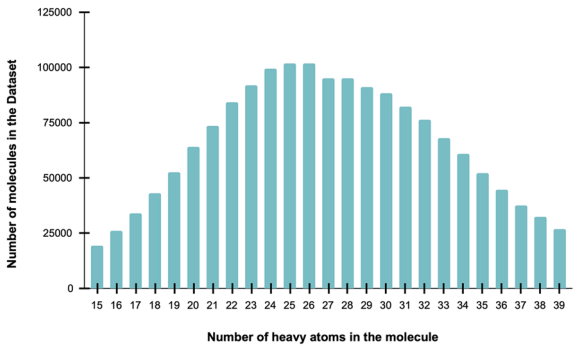


Table 2 Mean and MAD values of principal MOI tensor components (random generation of conformers using Distance Geometry Embedding Algorithm[21])

Run number		I_{xx}	I_{yy}	I_{zz}
1	Mean	104.79	473.03	537.40
	Mean absolute deviation within the dataset	52.03	219.88	233.04
2	Mean	104.81	472.96	537.32
	Mean absolute deviation within the dataset	52.04	219.76	232.97
3	Mean	104.79	472.89	537.26
	Mean absolute deviation within the dataset	52.02	219.74	232.89
Values, averaged over 3 runs	Mean	104.79 ± 0.01	472.96 ± 0.07	537.33 ± 0.07
	Mean absolute deviation within the dataset	52.03 ± 0.01	219.79 ± 0.08	232.97 ± 0.07

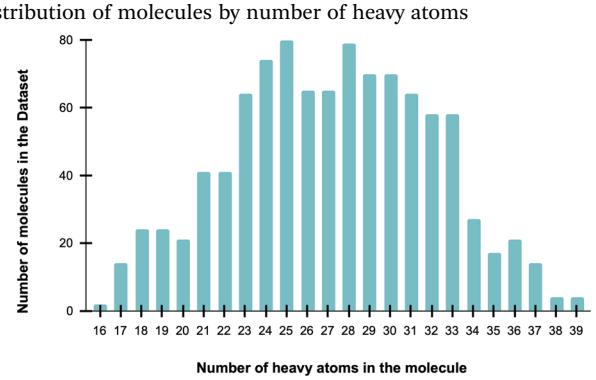
block, which is responsible for generating initial atom positions, to produce a wide variety of structurally sound molecules. At the same time, the GCN block, which predicts the molecular adjacency matrix, benefits from exposure to a broader distribution of correlations between interatomic distances and bond patterns. Second, this approach allows for independent control over the structural validity of conformers through a deterministic standardization step. To address concerns about the realism of randomly generated geometries, we perform geometry refinement using molecular dynamics after the generation (see Section 2.5). The overall architecture is intentionally modular: instead of embedding a rigid definition of "optimal" geometry into the training process, we allow users to pair the EDM block with their own conformer optimization and filtering pipeline. This design enhances adaptability and makes the model extensible across a range of cheminformatics applications and workflows.

For validation of the model's ability to generate structures similar to real molecule's geometry the Cambridge Crystallographic Data Centre (CCDC) virtual screening set (Table 3)[23] was used as a source of reference molecules for generation. A thousand real molecules, which satisfied constraints on heavy atom account and elemental composition with annotated geometries were selected to test the generation performance of the model. The mean and MAD values as well as distribution of the examples by atom count (Table 3) correlates well with the training dataset.

Table 3 Key parameters of CCDC virtual screening subset used for model evaluation

Total number of molecules	1000		
Permitted types of heavy atoms	C, N, O, F, P, S, Cl, Br		
Principal components of MOI tensor	I_{xx}	I_{yy}	I_{zz}
Mean value	117.70	443.07	517.63
Mean absolute deviation within dataset	49.15	201.20	211.42

Distribution of molecules by number of heavy atoms



2.3 Training of the EDM block

The training procedure for the EDM block was adapted from ref. 5. The initial dataset, consisting of SMILES strings, was processed using RDKit to generate random conformers based on the Distance Geometry Embedding Algorithm.[21] All conformers were stripped of hydrogen atoms, retaining only heavy atoms. It is important to note that during training, a new random conformer was generated for each sample every time it was included in a batch, ensuring that the model was exposed to diverse conformations throughout the training process.

The generated conformer coordinates were then used to center each molecule at the apparent center of mass, with the mass of all atoms assumed to be equal to one. After centering, a MOI tensor was calculated using the same assumption of equal mass for all atoms. To orient the molecule into a principal frame, a rotation matrix was computed to eliminate all non-diagonal components of the MOI tensor. The three non-zero principal components were concatenated into a floating-point vector of size three and used as a context.

Calculated context values, one-hot encoded atom types, and coordinates, rotated into a principal frame, were then subjected to a forward diffusion process, introducing noise according to a polynomial noise schedule of the form $1 - x^2$, with a noise precision of 10^{-5} and 1000 noising steps. Noised representations were passed to the model, with the optimization objective of

$$Lt = \left[\frac{1}{2}(|\varepsilon - \hat{\varepsilon}|^2\right] \quad (1)$$

where ε is a noise vector sampled from a standard multivariate normal distribution with mean 0 and identity covariance matrix, $\hat{\delta}$ – neural net prediction.

The model was initially trained on a random train/validation/test split of 60/20/20 for 1000 epochs, ensuring stable and smooth convergence. This was followed by training for an additional 500 epochs on the entire dataset to maximize performance. Training was performed on a single virtual machine (VM) equipped with 8 Nvidia H200 GPUs and 60 CPUs, with a batch-size of 2048 using the AdamW optimizer with

a weight decay of 10–12 and a learning rate of 10–4, along with the AMSGrad variant.[24] Gradient clipping was applied to limit gradients to a maximum of 150% plus 2 standard deviations from the mean of recent gradient history. The average training time per epoch was 730 seconds, and the complete training process took approximately 13 days.

2.4 Training of the AdjMatSeer block

Random conformers for the case of AdjMatSeer training were prepared similarly to the EDM block training case. To ensure the predictive performance of the model on the noised input, the coordinates of each conformer were intentionally disturbed by displacing each atom from its original position within a ball of radius R. The radius R for each atom was independently selected from the range [0, R_{max}], where R_{max} was chosen independently for each conformer from the range [0, 0.4]. The introduced disturbance allowed for controlled variability while preserving the overall molecular structure.

The model was trained to predict the adjacency matrix using three types of input data derived from the disturbed random conformers: the complete pairwise distance matrix representing atom-to-atom distances, the Boolean adjacency matrix indicating bond connectivity between atoms, and the atom types encoded to guide adjacency predictions. The loss function, as described in ref. 18, was defined as a cross-entropy loss between the predicted and expected bond types, effectively formulating the adjacency matrix prediction as a multi-class classification problem.

The model was initially trained on a random 60/20/20 train/validation/test split of the ChEMBL dataset for 200 epochs using AdamW optimiser with a learning rate of 10^{-4} and a batch size of 2048. This was followed by an additional 140 epochs at a reduced learning rate of 10^{-5}. Finally, the model underwent fine-tuning for 20 additional epochs on the entire dataset at a learning rate of 10^{-5}, achieving a 98.57% correct bond rate on the test set. The average epoch time was approximately 330 seconds, and the entire training process took around 2 days when conducted on a system equipped with 60 CPUs and one NVIDIA H100 GPU.

2.5 Structure standardisation pipeline

To ensure the quality and validity of the molecular structures generated by the model, a deterministic standardisation pipeline was introduced as an integral block of the module. The pipeline steps were implemented using RDKit[22] and followed a well-defined sequence to maximize the number of valid and chemically sensible molecules accessible to the user. The standardisation process consists of the following steps, executed in the given order:

(1) Selection of the largest fragment: if the generated molecule is not fully connected, only the largest fragment is retained. This step increases the likelihood of obtaining a valid molecule and ensures the usability of the generated structures.

(2) Kekulization: the molecule's aromatic systems are explicitly represented in their Kekulé form. This process improves chemical accuracy and ensures compatibility with downstream applications.

(3) Sanitization: the molecule is checked for chemical correctness and structural integrity. This step involves standardising atom valences, checking for aromaticity, and ensuring that the molecular graph is valid.

(4) Position-constrained MMFF94 geometry optimization: to improve the quality of the resulting geometry while preserving the overall molecular conformation suggested by the model, a position-constrained geometry optimization was performed using the MMFF94 force field.[26] This step refines atomic positions while maintaining the original geometry as much as possible.

By following these standardised procedures, the pipeline ensures that the generated molecular structures are chemically sound and geometrically consistent, significantly enhancing the reliability and interpretability of the model outputs. Since standardised molecules contain correct atom valence information, the positions and connectivity of hydrogen atoms can be straightforwardly calculated using conventional methods if required.

2.6 Shape similarity

2.6.1 Similarity to a reference molecule. The shape similarity between two molecular structures was defined as a Tanimoto-like metric based on their molecular volume overlap:

$$S_{\text{mol A/mol B}} = \frac{V_A \cap V_B}{V_A + V_B - V_A \cap V_B} \quad (2)$$

Here, V_i represents the molecular volume of structure i, and $V_A \cap V_B$ denotes the volume intersection between molecules mol A and mol B.

To calculate the molecular volumes and their intersections, the Gaussian method proposed by Grant *et al.*[25] was applied. This approach provides an accurate estimation of molecular volume through Gaussian-based integration. For the chosen metric to effectively describe overall molecular shape similarity, it is essential to align the molecules in a way that maximizes their volume intersection. To achieve satisfactory alignment while maintaining reasonable computational efficiency, the structures were aligned based on shape-multipole approach.[25] First, the center of coordinates was moved to nullify the first moment of the volume density function. Subsequently, the second moment – a symmetric 3×3 tensor referred to as the shape quadrupole – was employed to rotate the molecule into its "shape – principal" frame. This was accomplished by calculating the rotation matrix that diagonalizes the shape quadrupole. Once positioned in the principal frame, the molecule is assumed to be aligned with the axes according to its molecular volume distribution, allowing for the comparison of molecular shapes using the defined similarity metric.

2.6.2 Similarity to a reference arbitrary shape. A Tanimoto-type score was defined to assess the similarity between an arbitrary structure represented by an STL file, which models a binding protein pocket, and a set of molecular conformers. Both the STL mesh and the molecular structures were aligned to their principal axes using Principal Component Analysis (PCA). This alignment ensures that the comparison is orientation-independent, allowing for a fair assessment of similarity.

Once aligned, the molecules are voxelized, meaning they are represented as a grid of points in 3D space, with each point indicating the presence of a part of the molecule within a certain radius. The voxel size is determined by the grid_spacing parameter, set to 0.5 Angstroms, which means each voxel represents a cube with sides of 0.5 Angstroms. This size strikes a balance between capturing sufficient detail for accurate Tanimoto score calculation and maintaining computational efficiency.

The van der Waals (VdW) radii of the atoms are used to determine the extent of each atom's influence in the voxel grid. The arbitrary shape Tanimoto coefficient is then calculated by comparing the voxelized representations of the STL mesh and the molecules. It is defined as the ratio of the intersection of the voxel grids (common occupied voxels) to the union of the voxel grids (total occupied voxels).

$$S_{\text{shape/mol}} = \frac{\text{VG}_{\text{shape}} \cap \text{VG}_{\text{mol}}}{\text{VG}_{\text{shape}} \cap \text{VG}_{\text{mol}}} \quad (3)$$

Here, VG_{shape} and VG_{mol} represent the voxel grids of the arbitrary shape and a molecule correspondingly.

Different scale factors were applied to the VdW radii to observe how the defined Tanimoto score changes. As the scale factor increases, the effective size of the atoms increases, which initially leads to a higher overlap and thus a higher Tanimoto score. However, beyond a certain point, further increasing the scale factor causes excessive overlap, reducing the score. In practice, a scale factor of 1.7–1.8 was discovered to yield the maximum Tanimoto score for most molecules, indicating an optimal balance between overlap and separation. This behavior reflects the sensitivity of the Tanimoto score to the spatial configuration and size of the molecules relative to the binding pocket.

2.7 Generation based on a reference molecule

The generative performance of the model, in the context of generating conformers similar to a given reference conformer, was evaluated by requesting 100 samples for each molecule from a subset of the CCDC virtual screening set, consisting of one thousand real conformers. The generation process involved calculating the context for a reference conformer and randomly selecting the number of heavy atoms for the molecule to be generated. The chosen number of heavy atoms for each sample was within the range

$$[N_{\text{ref}} + \text{variance}, N_{\text{ref}} - \text{variance}],$$

where N_{ref} is the number of heavy atoms in the reference molecule and variance is an integer parameter. The generative performance of the model was assessed at two denoising step settings: 100 and 1000 steps. To compare the performance of the proposed GCN-based bond prediction approach with a deterministic bond-prediction method, OpenBabel[16] was used as a benchmark. To comprehensively evaluate the generative performance in terms of both efficiency and generation quality, the following set of metrics was considered appropriate:

2.7.1 Generation speed. Measured as the number of valid molecules generated per second on a NVidia H100 GPU. This metric quantifies the model's efficiency in producing chemically sound structures.

$$\text{Generation speed} = \frac{N_{\text{valid molecules}}}{t_{\text{generation}}} \quad (4)$$

where $N_{\text{valid molecules}}$ is the total number of valid molecules generated by the model; $t_{\text{generation}}$ – total generation time, sec.

2.7.2 Total number of valid molecules. The ratio between the number of valid molecules generated and the total number of requested molecules. This metric reflects the reliability of the generation process.

$$\text{FR}_{\text{valid}} = \frac{N_{\text{valid}}}{N_{\text{requested}}} \quad (5)$$

where N_{valid} is the total number of valid molecules generated by the model and $N_{\text{requested}}$ is the total number of molecules requested to be generated.

2.7.3 Average shape Tanimoto similarity. Calculated as described in Section 2.5 shape similarity for each generated molecule and a corresponding reference. The values were averaged across all generated molecules and separately for those with a specific number of heavy atoms.

$$\text{Average similarity}_{\text{shape}} = \frac{\sum_{\{i=1\}}^{\{N\}} S_{i/\text{ref } i(\text{shape})}}{N} \quad (6)$$

where $S_{i/\text{ref } i(\text{shape})}$ is shape similarity of the i-th molecule to a corresponding reference and N – number of molecules in the set of interest.

2.7.4 Maximal shape Tanimoto similarity. The highest shape similarity observed for molecules with a given number of heavy atoms.

$$\text{Maximal similarity}_{\text{shape}} = \max_{0<i<N} \left(S_{i/\text{ref } i(\text{shape})}\right) \quad (7)$$

where $S_{i/\text{ref } i\ (\text{shape})}$ is shape similarity of the i-th molecule to a corresponding reference.

2.7.5 Average chemical Tanimoto similarity. The chemical similarity was defined as the Tanimoto coefficient between bit-Morgan fingerprints (2048 bits, 2 hops) of a generated molecule and a corresponding reference conformer. The values were averaged over all generated molecules and separately within a subset of molecules with a specified number of heavy atoms. This metric assesses how chemically similar the generated molecules are to the references.

$$\text{Average similarity}_{\text{chemical}} = \frac{\sum_{\{i=1\}}^{\{N\}} S_{i/\text{ref } i(\text{chemical})}}{N} \quad (8)$$

where $S_{i/\text{ref } i\ (\text{chemical})}$ is chemical similarity of the i-th molecule to a corresponding reference and N – number of molecules in the set of interest.

2.7.6 Number of chemically unique molecules. Chemical uniqueness of the molecules was assessed by comparing InChI

strings generated using RDKit. The assessment of chemical uniqueness was performed in two contexts:

Compared to training dataset – the count of unique molecules generated that are not present in the training data. This metric evaluates the model's capacity to generate novel structures.

$$N_{\text{unique}} = |\{\text{Gen}\}\backslash\{\text{Train}\}| \quad (9)$$

where {Gen} is a set of generated molecules, {Train} is a set of molecules used for training the model.

Within the generated set – the number of unique molecules within the entire set of generated samples, reflecting diversity within the generated outputs.

$$N_{\text{unique}} = |\{x \in \text{Gen}\}| \quad (10)$$

$$\text{FR}_{\text{unique}} = \frac{N_{\text{unique}}}{N_{\text{generated}}} \quad (11)$$

where N_{unique} is the number of unique molecules and $N_{\text{generated}}$ is the total number of generated molecules.

2.7.7 Frechet fingerprint distance (FFD). The Fréchet Fingerprint Distance (FFD) employs the Fréchet distance formula applied to molecular fingerprint distributions, combining both mean and covariance information. The calculation begins by computing Morgan fingerprints for each molecule using RDKit's Morgan fingerprint generator with radius 2 and 2048 bits. For two sets of molecular fingerprints, the FFD is calculated as

$$\text{FFD} = \|\mu_1 - \mu_2\|^2 + \text{Tr}(\Sigma_1) + \text{Tr}(\Sigma_2) - 2\,\text{Tr}\left(\sqrt{\Sigma_1\Sigma_2}\right) \quad (12)$$

where μ_1 and μ_2 represent the mean fingerprint vectors of the two sets, Σ_1 and Σ_2 are their respective covariance matrices, and Tr() denotes the matrix trace operation.

The covariance matrices are regularized with a small epsilon (10^{-6}) to ensure numerical stability and positive definiteness. The square root of the matrix product $\Sigma_1\Sigma_2$ is computed using the matrix square root function, with additional numerical safeguards to handle potential complex components. This formulation captures both the difference in central tendency (mean term) and the structural diversity (covariance term) between the two molecular fingerprint distributions, providing a comprehensive measure of molecular set similarity that accounts for both positional and distributional differences in the fingerprint space. FFD was computed to compare generated molecules against three established compound databases: ChEMBL,[20] PubChem,[27] and ZINC 250k dataset[28] as described in ref. 29. A random set of 100 000 molecules was selected from each database for FFD calculation. This metric quantifies how similar the distribution of chemical features of generated molecules is to the distributions found in known molecule databases.

2.8 Generation based on an arbitrary shape

The generation of molecules constrained with an arbitrary shape was studied on the example of mimicking a shape of a protein pocket. As a model system, the CLK1 protein ref. 30 was chosen, specifically its complex with *N*2-(3-(morpholin-1-yl) propyl)pyrido[3,4-*g*]quinazoline 2,10-diamine (6q8k).[31] The binding pocket was defined based on the bound ligand using the PyVol[32] extension of the PyMol viewer, the image of the selected pocket is provided in Fig. 1. The pocket shape was extracted as an.stl file, aligned to the principal frame, and the principal components of the moment of inertia (MOI) tensor were calculated and supplied to the model as input for generation. The number of heavy atoms was set to be randomly chosen for the range between 37 and 39.

In addition to evaluating the similarity of the generated molecules to the target pocket shape, molecular docking of the generated molecules was performed using the target CLK1 binding pocket. Although docking scores obtained from unspecified protocols are not highly predictive of experimental binding affinities, they were used as an additional sanity check to assess the quality of the generated compounds.

2.9 Docking

The generated ligands were docked to CLK1 (ref. 30) using the united-atom algorithm of AutoDock Vina (version 1.2.7)[33] with the exhaustiveness of 32 and energy range of 6. Ligands

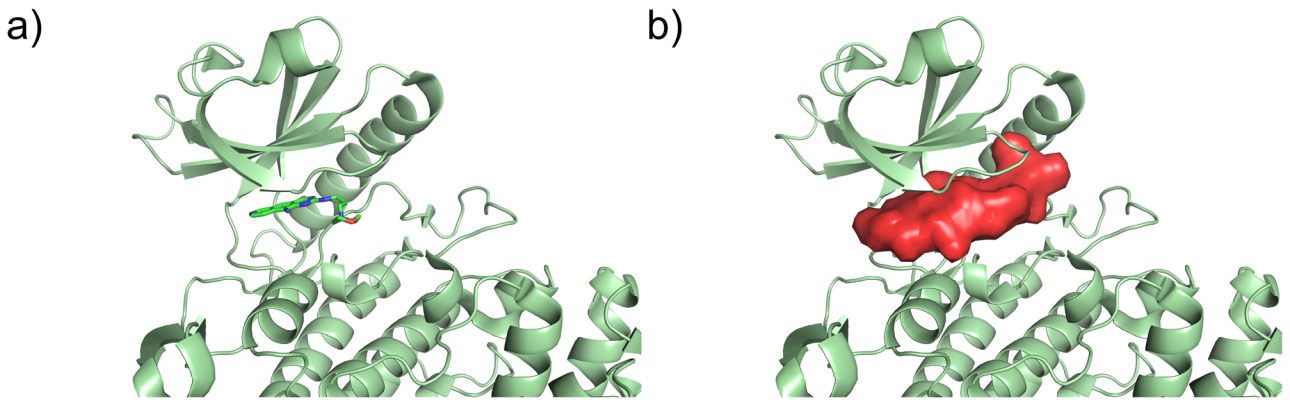


Fig. 1 The image of the pocket selected for generation based on an arbitrary shape. (a) The view of the binding site with a reference ligand, (b) the surface of the selected pocket.

were treated as flexible and the receptor as rigid. The grid box was positioned around the coordinate centre of the active site and encompassed all the active site residues. The binding energy was converted to affinity using the temperature of 310.15 K.

The 6q8k.pdb file was prepared in AutoDockTools (part of MGLTools version 1.5.7) and Swiss-PdbViewer (version 4.1.1). All ligands, the expression tag and water molecules were removed. Missing residues and missing atoms were repaired. Histidine hydrogens were assigned to the $\tau(\varepsilon)$ nitrogen, missing hydrogens were added to the protein. Kollman charges were calculated and distributed across the protein. The protein coordinates were rotated to minimise the gridbox size, and the structure was converted to .pdbqt format.

Ligands were converted to .pdbqt format with OpenBabel (supplied along with MGLTools version 1.5.7)[16] with Gasteiger charges and polar hydrogens added per pH 7.2.

3 Results and discussion

3.1 Generative performance – reference molecules

To evaluate the generative performance of the *MLConformerGenerator*, 100 molecules for each molecular shape present in CCDC virtual screening set were generated using 1000 and 100 denoising steps. The general metrics are presented in Table 4. As expected, the time required for generation of samples per reference increases linearly from 11.48 s to 96.01 s with the increase in the number of denoising steps. The small difference

Table 4 General generative performance of *MLConformerGenerator*, evaluated at 100 and 1000 denoising steps

Parameter	100 denoising steps	1000 denoising steps	Parameter	100 denoising steps	1000 denoising steps
Total generation time, sec	11 476	96 009	Average shape Tanimoto similarity	0.5332	0.5338
Averaged time for generation (per single reference context), s	11.48	96.01	Average chemical Tanimoto similarity	0.1087	0.1086
Total valid molecules (% from requested)	47.94%	48.60%	FFD PubChem	2.64	2.57
Generation speed (valid molecules per s)	4.18	0.51	FFD ChEMBL	4.14	3.98
Chemically unique molecules (not found in training dataset)	99.84%	99.81%	FFD ZINC 250k	4.95	4.84
Chemically unique molecules (within the generated set)	99.94%	99.94%			

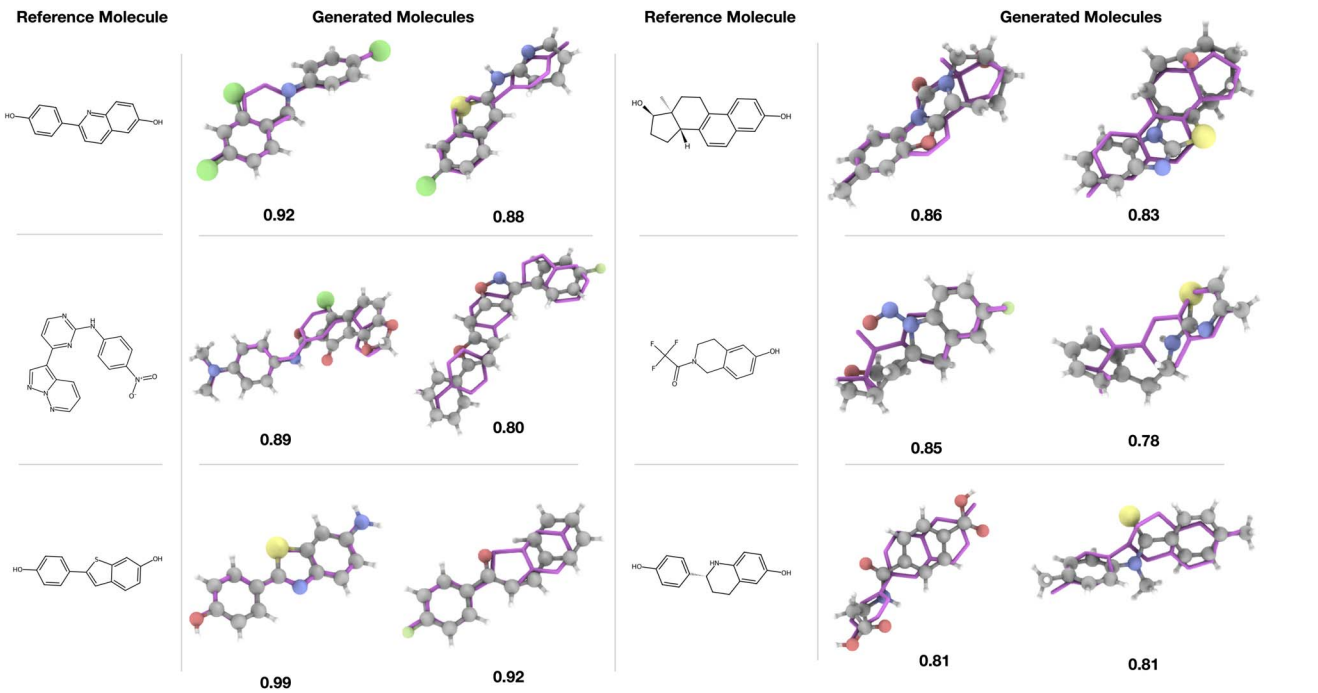


Fig. 2 Examples of molecules generated by MLConformerGenerator alongside their corresponding reference structures and shape Tanimoto similarity scores. In the 3D visualizations, reference molecules are shown in stick representation and highlighted in magenta for visual comparison. Molecules were generated using 100 denoising steps.

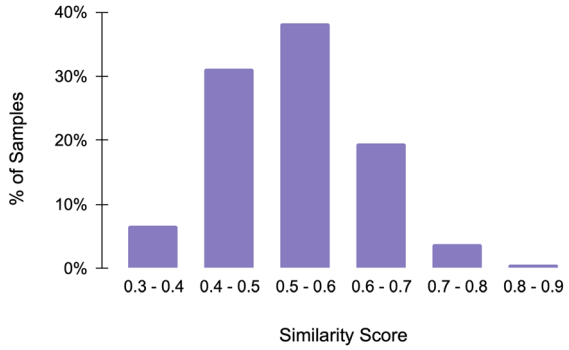


Fig. 3 The distribution of shape Tanimoto similarity scores for the dataset generated from CCDC virtual screening subset.

in the values of average shape similarity values indicates that generation with 100 denoising steps still produces molecules with resemblance to the target shapes comparable to that for 1000. Both runs produced molecules with low chemical similarity (Tanimoto coefficient lower than 0.11) to the reference molecule, indicating good chemical variability of generated molecules. FFD values calculated for generated molecule sets and common chemical datasets of real molecules – ChEMBL, PubChem and ZINC 250k are relatively low (<5), suggesting that the distribution of chemical features in the generated molecules is reasonably similar to that of real-world compounds.

To visually illustrate the generative performance and allow for an overall qualitative assessment, representative examples of the generated molecules, along with their corresponding reference structures, are presented in Fig. 2. Furthermore, to quantitatively evaluate the model's performance, the distribution of shape Tanimoto similarity scores between generated molecules and their respective references is presented in Fig. 3.

The distribution of shape Tanimoto similarity values among the generated compounds reveals that over 62% of the molecules exhibit a similarity score greater than 0.5 relative to the reference shape. The relatively narrow distribution, with a median located above 0.5 threshold, indicates that the model demonstrates a strong ability to capture and reproduce the target geometry.

A more detailed assessment of the model's generative ability was conducted by analysing the average shape Tanimoto similarity, maximum shape Tanimoto similarity, and average chemical similarity across subsets of generated molecules grouped by their number of heavy atoms. The corresponding results are presented in Fig. 4.

Reduction in the number of denoising steps from 1000 to 100 does not lead to a significant decline in sample quality, as evidenced by both average and maximum Tanimoto similarity scores shown in Fig. 4a and b. The evaluated metrics remain relatively constant with a decrease in the number of denoising steps, indicating that the generation quality is preserved while inference time is reduced significantly. This indicates that alike in the case of diffusion models for image generation,[34] a reduction in denoising steps can still yield high-quality outputs.

Another notable, though expected, dependency, is the decrease in both average and maximum shape Tanimoto

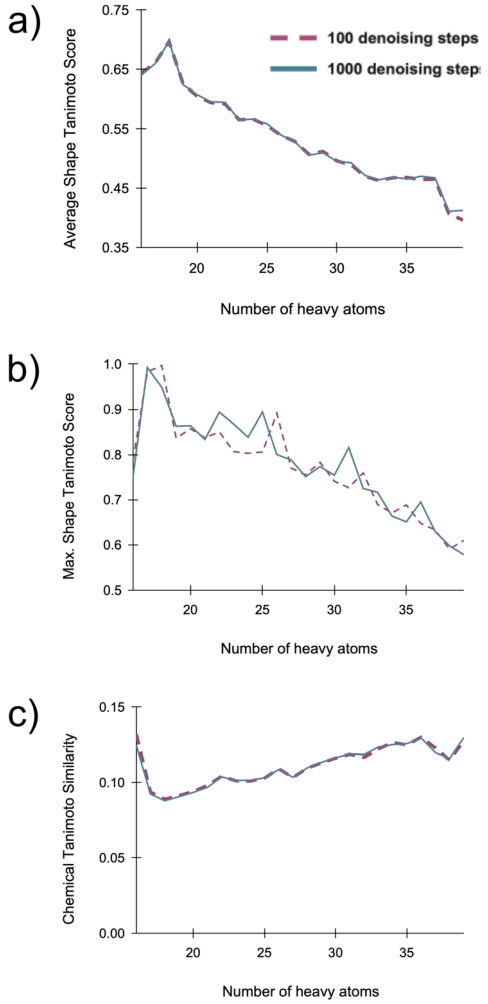


Fig. 4 Generation performance at 100 and 1000 denoising steps. (a) Average shape Tanimoto similarity, (b) maximal shape Tanimoto similarity and (c) chemical Tanimoto similarity as functions from the number of heavy atoms in the reference molecule, with variance parameter set to 2.

similarity with increasing heavy atom count in the reference molecules. While the maximum shape similarity ranges from 0.80 to 0.99 for molecules with 15 to 27 heavy atoms, it drops to 0.6 to 0.8 for larger structures. This is likely due to the increased structural complexity of larger molecules and suggests that the model would benefit from greater exposure to such examples during training. In contrast, the chemical similarity remains relatively stable, fluctuating within the range of 0.08 to 0.13 as can be seen from Fig. 4c. The observed low values of chemical similarity between generated molecules and corresponding references indicate that the model tends to produce chemically diverse structures. A slight upward trend in chemical Tanimoto coefficient can be observed with an increase in the number of heavy atoms in the molecule. This may be attributed to the fact that, as molecule size increases, and given that the model has learned a limited set of chemical features, the probability of reusing known fragments rises – leading to a modest increase in similarity. Despite the minimal change in similarity metrics,

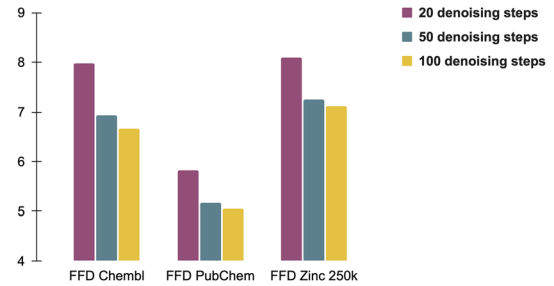


Fig. 5 FFD values to different datasets for different numbers of denoising steps – 20, 50 and 100.

a slight decline in generation quality is observed, as indicated by an increase in FFD (Table 4) values when the number of denoising steps is reduced.

To further assess the impact of the number of denoising steps on FFD values, a small subset of 100 molecules from the CCDC virtual screening set was selected. For each molecule, 100 samples were requested while varying the number of diffusion steps. The resulting FFD values – calculated with respect to ChEMBL, PubChem, and ZINC250k as a function of the number of steps are shown in Fig. 5.

It can be observed that the difference in FFD values between 50 and 100 denoising steps is relatively small, while the average generation time per reference structure decreases from 11.69 seconds to 6.92 seconds. However, reducing the number of denoising steps further to 20 results in a noticeable decline in generation quality, as indicated by a substantial increase in FFD values across all reference datasets, with a linear decrease in averaged generation time per reference to 4.01 seconds.

3.2 Generative performance – arbitrary shape

Since the principal components of the moment of inertia (MOI) tensor can be computed for any object with a defined geometry and mass distribution, virtually any shape can be used to condition the generation process – provided that its principal MOI components are similar to those observed in the training dataset and its overall size is comparable to molecules with the targeted number of heavy atoms.

The performance of *MLConformerGenerator* in application to generation of molecules conditioned on an arbitrary shape was evaluated on a total of 360 molecules generated using the shape of a selected CLK1 binding pocket (Fig. 1) as a reference. The average shape similarity between the generated molecules and the reference shape, computed using eqn (3) (with a scaling factor of 1.8), was 0.436, with a maximum similarity reaching 0.534. The lower average shape similarity observed when generating molecules from an arbitrary target protein pocket shape is primarily due to how the similarity metric is defined (see Section 2.6.2). Since the binding pocket typically occupies a much larger volume than any individual ligand, the resulting shape similarity scores tend to be lower than those characteristic for the case of generating from a reference molecular conformer, even when the generated structures are reasonably well-aligned with the pocket.

Six molecules with the highest similarity score to the pocket are illustrated along with the distribution of shape similarity scores for the generated samples to help assess the overall quality of generation are illustrated in Fig. 6a and b correspondingly.

Visual inspection of the results shown in Fig. 6a suggests that *MLConformerGenerator* effectively captures the overall pattern of the target arbitrary shape by attempting to fill the reference volume with the specified number of atoms. While some examples reveal that, after alignment, a few atoms fall outside the boundaries of the reference shape, the visualizations nonetheless demonstrate that the model is capable of generating molecules that approximate a given shape. These observations, along with the values of shape similarity metric and its narrow distribution (Fig. 6b), support the applicability of the model for arbitrary-shape-constrained molecular design.

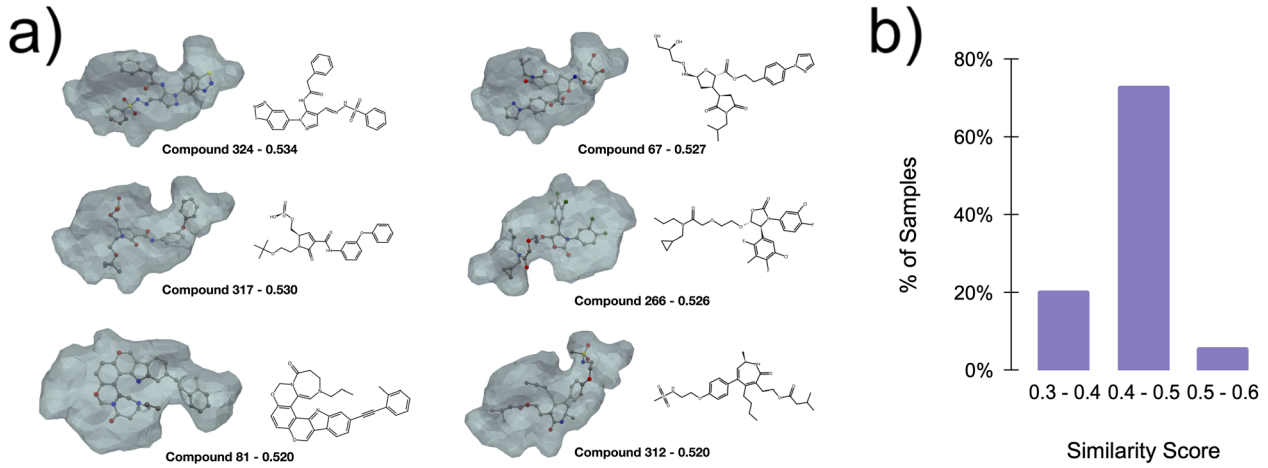


Fig. 6 (a) Illustrative examples of molecules generated based on the shape of the selected CLK1 binding pocket. Each example is aligned with the reference pocket shape and annotated with the corresponding similarity value. (b) Distribution of generated samples by shape similarity to a reference pocket.

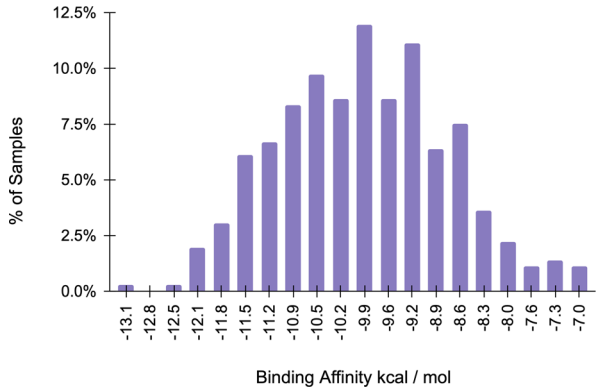


Fig. 7 Distribution of binding affinities (docking scores) for molecules generated with CLK1 binding site shape constraint.

Docking experiments were performed with molecules generated based on the shape of the selected CLK1 binding pocket to evaluate whether the generated compounds can fit the intended site. This served as an additional validation of the viability of the shape-constrained generation process. The distribution of docking scores for the 360 generated molecules along with illustration of the best poses of top scoring compounds is presented in Fig. 7 and 8 correspondingly. The generated ligands exhibited a distribution of affinities close to normal. The top candidate dissociation constant reached K_d = 438 pM indicating exceptional affinity to the target pocket. The performance of the model conditioned on the arbitrary shape of the pocket showcased that created molecules successfully fitted into a target binding site and show reasonable affinities which attests applicability of *MLConformerGenerator* to the task of molecule generation based on the extracted pocket shape, even though trained only on generated molecular conformers.

It should be noted that little to no correlation was observed between the arbitrary shape similarity and the binding affinity of the corresponding structures. This is expected, as protein–ligand interactions are not solely governed by the compound's ability to fit within the binding site. The arbitrary shape similarity metric is introduced to assess the model's ability to reproduce a given shape constraint, rather than to predict its binding affinity. While the docking experiments demonstrate that *MLConformerGenerator* is capable of generating chemically valid structures with reasonable docking scores, the overall success of the approach in generating high-affinity ligands using the proposed model will depend on the careful selection, definition and configuration of the target pocket shape, as well as the use of a specialized docking protocol tailored to the system of interest.

3.3 Comparative analysis of deterministic and GCN-based bond prediction strategies

Deterministic methods for bond prediction are commonly utilised in molecular generation pipelines due to simplicity of usage and effectiveness in determining atomic adjacency from the atom types and spatial coordinates produced by EDMs.[5–7] However, these methods present several limitations. Such methods do not account for the underlying distribution of bond types or chemical features present in the training data. Additionally, they may require extra installation steps or runtime dependencies, which can complicate their integration into production environments. To address these challenges, we explored the use of a GCN – based model as an alternative to traditional deterministic approaches, specifically comparing its performance against OpenBabel[16] in the scope of the EDM-based molecule generation pipeline. Such an approach, while also relying on interatomic distances, offers the advantage of explicitly learning correlations that reflect the chemical feature distribution of the target dataset. The comparative performance of the generation results using both bond prediction

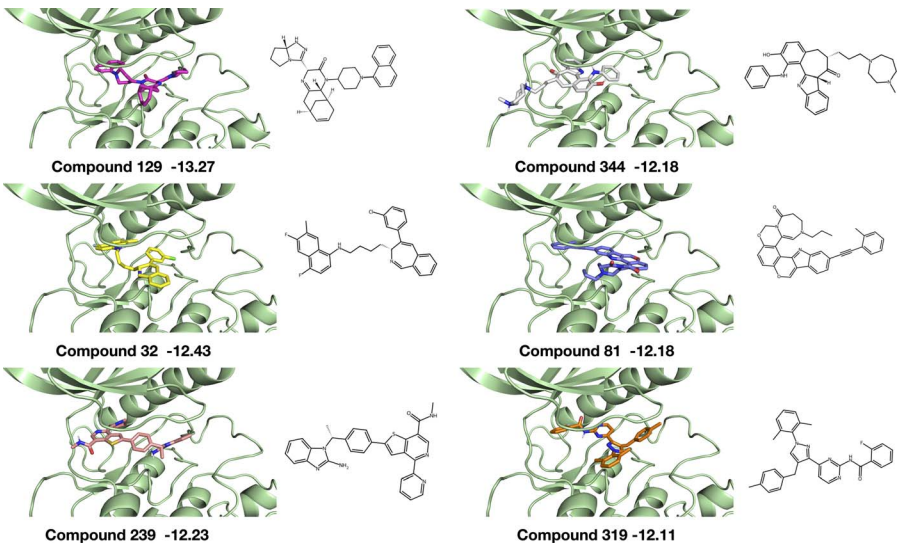


Fig. 8 Visual representation of the top six compounds with the highest predicted affinities, annotated with their corresponding docking scores in kcal mol^{-1}.

Table 5 Performance of AdjMatSeer (GCN) and OpenBabel bond prediction with EDM at 1000 denoising steps

Parameter	AdjMatSeer (GCN)	OpenBabel	Parameter	AdjMatSeer (GCN)	OpenBabel
Total valid molecules (% from requested)	48.16%	93.56%	Average chemical Tanimoto similarity	0.1086	0.1056
Chemically unique molecules (not found in training dataset)	99.81%	99.93%	FFD PubChem	2.57	2.89
Chemically unique molecules (within the generated set)	99.94%	99.87%	FFD ChEMBL	3.98	4.63
Average shape Tanimoto similarity	0.5338	0.5336	FFD ZINC 250k	4.84	5.38

approaches at 1000 denoising steps is presented in Table 5 and Fig. 9a and b.

As shown in Fig. 9, both the average and maximum values of shape Tanimoto similarity are nearly identical across the two bond prediction approaches. This trend also holds for the overall average shape and chemical similarity to reference across the entire generated dataset, as summarized in Table 5. However, notable differences emerge in terms of the fraction of valid molecules generated and the FFD values. While OpenBabel yields a significantly higher fraction of valid structures after standardisation (93.47% *vs.* 48.60%), it also exhibits consistently higher FFD values (by approximately 11–18%) when compared to real datasets – suggesting that the proposed GCN-based method may produce molecules that are closer in distribution of features to real chemical structures. These observations suggest that, while deterministic methods may outperform the proposed GCN-based bond prediction approach in terms of the absolute number of valid structures generated, the GCN method offers greater flexibility in tuning the model to generate structures that more closely resemble a target distribution of chemical features.

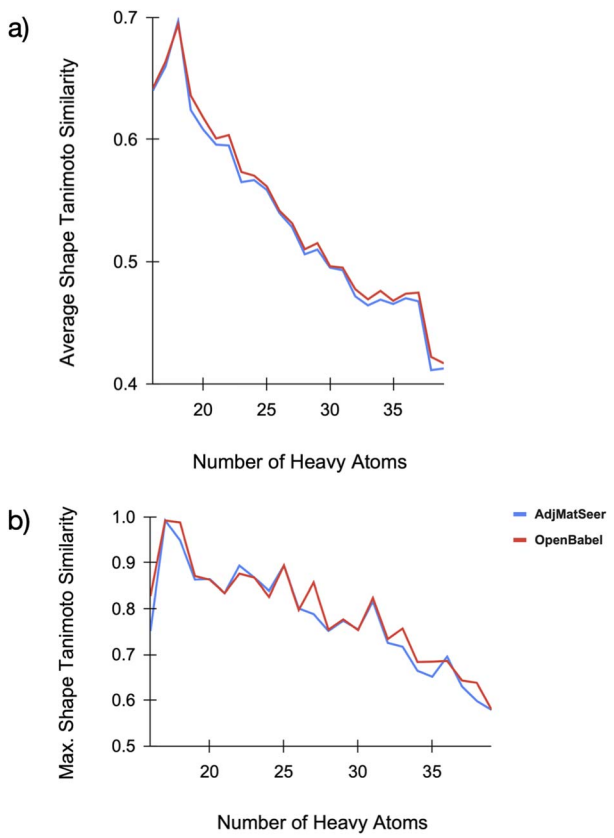


Fig. 9 Generation performance with AdjMatSeer and OpenBabel bond prediction. (a) Average shape Tanimoto similarity, (b) maximal shape Tanimoto similarity.

A key factor contributing to the lower rate of valid structures was suggested to be the occasional lack of precision in the 3D coordinates generated by the EDM block, which can hinder accurate bond prediction by the GCN block. This sensitivity is further influenced by the fact that the GCN was trained on moderately-noised structures, limiting the level of precision it can reliably leverage during inference. To address this, we explored an inference-time resampling strategy, as described in ref. 35, which introduces iterative refinement of intermediate states during the denoising process. Specifically, at each denoising step i, we apply a predefined number of intermediate denoising updates in the direction of step $i-1$. This additional refinement helps harmonize structural intermediates and smooth out potential outliers, thereby improving geometric stability and enhancing the reliability of subsequent adjacency predictions. The impact of resampling on generation quality is summarized in Table 6.

Applying resampling during inference results in a noticeable – but relatively modest—improvement in both molecular validity and average shape Tanimoto similarity, increasing from approximately ∼48% to ∼52% and from ∼0.53 to ∼0.54, respectively, while maintaining a generation throughput of over one molecule per second. Nonetheless, the GCN block may still benefit from further architectural refinements and improved training strategies to enhance molecular validity, while retaining its capacity to generate structures with chemical feature distributions closely aligned with those of the target dataset.

3.4 Performance assessment with respect to existing models

A careful evaluation of the MLConformerGenerator, incorporating both deterministic and GCN-based bond prediction strategies, demonstrated the validity of using principal MOIs as a concise and computationally efficient physical descriptor for shape-constrained molecular generation. A comparative performance analysis of the proposed framework against other approaches for similar tasks is presented in Table 7.

The performance values for the considered models were obtained from their respective original publications.[4,6,7] Although

Table 6 Performance of AdjMatSeer (GCN) bond prediction with EDM at 100 denoising steps with 1 and 4 resampling steps

Parameter	1 resampling step	4 resampling steps	Parameter	1 resampling step	4 resampling steps
Generation speed (valid molecules per s)	2.53	1.10	Chemically unique molecules (within the generated set)	99.79%	99.18%
Total valid molecules (% from requested)	51.57%	53.05%	Average shape Tanimoto similarity	0.5409	0.5452
Chemically unique molecules (not found in training dataset)	99.47%	98.99%	Average chemical Tanimoto similarity	0.1146	0.1163

Table 7 Comparative performance of the *MLConformerGenerator* and other state-of-the-art models for shape-constrained molecular generation.[4,6,7]

	Valid molecules (% from output samples)	Average shape similarity	Maximal shape similarity	Reference
MLConformerGenerator (deterministic bond prediction)	93.6%	0.536	>0.99	This study
MLConformerGenerator (GCN bond prediction)	48.2%	0.533	>0.99	
MLConformerGenerator (GCN bond prediction, 4 resampling steps)	53.05%	0.545	>0.99	
ShapeMol + g	98.7%	0.746	0.852	6
ShapeMol	98.8%	0.689	0.803	
Shepherd	73.7–96.2%	0.799	—	7
SQUID ($\lambda = 0.3$)	100.0%	0.717	0.904	4
SQUID ($\lambda = 1.0$)	100.0%	0.670	0.842	

the generative performance was assessed on different datasets, these results are included to provide a general overview of relative model capabilities. For ShapeMol[6] and SQUID,[4] the percentage of connected molecules reported by the authors was interpreted as the percentage of valid structures, in accordance with our own validity criteria, to enable a consistent comparison.

The *MLConformerGenerator*, which uses a simple float vector of the size of three as a shape-capturing context, was trained on artificially generated conformers derived from SMILES representations. Despite the simplicity of this approach, the model, when paired with a deterministic bond prediction module, achieves competitive performance in generating valid molecular structures and even surpasses other models in terms of maximum achieved shape Tanimoto similarity. The combination of lower average and higher maximum similarity suggests that the structures generated with *MLConformerGenerator* exhibit a broader variance in shape similarity values. It should be noted that while the proposed model achieves a notably high maximum shape Tanimoto similarity (up to 0.99), its average similarity (∼0.53) is lower than that reported for models employing more expressive shape descriptors.[6,7] We attribute this discrepancy primarily to the limited representational capacity of the MOI tensor. While the MOI provides a computationally efficient means of capturing overall molecular shape, it offers only a coarse approximation and lacks the resolution to grasp finer 3D features. As a result, even when the model accurately learns to reproduce the target MOI, the generated conformers may still diverge in fine-grained geometry, leading to lower average shape similarity across diverse molecules. Nonetheless, we consider this trade-off to be a deliberate and acceptable design decision. Compared to more detailed descriptors, such as voxel grids or surface-based representations, the utilization of MOI tensor as a shape descriptor significantly reduces computational overhead and model complexity. This makes it particularly well-suited for scalable, shape-aware generation in early-stage molecular design tasks, where speed and simplicity are often prioritized over precision. Additionally, the physical properties of the MOI tensor – such as additivity and translational invariance- in theory enable its use in generating molecules of arbitrary size. This can be achieved by splitting the initial shape constraint into smaller fragments, generating the corresponding molecular substructures independently, and subsequently merging them into a complete molecule. This approach is currently under investigation as part of our future research.

The generic nature of the chosen shape descriptor enables generation based on arbitrary input shapes without the need for additional retraining or fine-tuning – a capability not reported for other models evaluated. While the proposed GCN-based bond prediction approach (AdjMatSeer) may result in a lower percentage of valid structures, it enables finer control over the distribution of chemical features for the within the sets of generated molecules, as evidenced by lower FFD values. However, due to the absence of detailed information on the distribution of chemical features within the datasets generated by competitor models in the original publications, this metric was excluded from the analysis.

4 Conclusions

The applicability of the principal components of the moment of inertia tensor as a compact and efficient shape descriptor for

the task of shape-constrained molecular generation was illustrated, using the proposed EDM-based model – *MLConformerGenerator*. The simplicity of the considered descriptor enables effective training on synthetically generated conformers derived from 2D molecular representations. We demonstrate the gain in efficiency by training our model on a subset of the ChEMBL database containing over 1.6 million molecules. The versatility of the chosen descriptor supports flexible generation scenarios, allowing the model to operate using either a reference molecular conformer or an arbitrary target shape. This capability was validated through experiments on two distinct shape-conditioning tasks: generation based on a set of target conformers, and generation based on the shape of a binding pocket of CLK1.

Generative performance evaluation showed that *MLConformerGenerator* produces molecules with chemical feature distributions closely aligned with real datasets, such as ChEMBL, PubChem, and ZINC 250k, as evidenced by FFD values consistently below 5. Despite relying on a compact yet expressive descriptor, the model achieves competitive performance relative to approaches using more complex shape representations. With deterministic bond prediction, the model achieved a 93.6% validity rate for generated molecules. Switching to a GCN-based bond prediction module (AdjMatSeer) reduced validity to 48.2–53.0%, while in turn resulted in lower FFD values – indicating a closer match to chemical feature distributions in the datasets containing real molecules. This trade-off suggests that the GCN-based bond prediction is better suited for applications focused on generating chemically realistic datasets, even at the cost of lower validity per attempt. When conditioned on a reference conformer, the generated molecules showed moderate to high shape similarity to the target, ranging from 0.3 to 0.99, with an average of 0.53–0.54. At the same time, the chemical similarity to the reference molecules remained low (<0.2), confirming the model's capacity to produce chemically diverse outputs within a given shape constraint.

Finally, the model's practical utility in shape-constrained molecular design was demonstrated through an end-to-end experimental pipeline: Extract Target Protein Pocket Shape → Generate Candidate Molecules → Dock to Protein. This highlights the potential of the suggested approach in generative structure-based molecular design workflows.

Conflicts of interest

There are no conflicts to declare.

Data availability

The subset of the CCDC virtual screening set used for evaluating generative performance, along with a smaller subset of 100 compounds used to analyze the impact of the number of diffusion steps, is provided in the SI. See DOI: **https://doi.org/10.1039/d5dd00318k**.

This section also includes datasets generated by the trained model using both AdjMatSeer (48 167 compounds) and OpenBabel (93 560 compounds) bond prediction approaches, as well as a set of 360 compounds generated based on a target arbitrary shape, accompanied by the corresponding .stl file used as a context for generation. The inference code of MLConformerGenerator is available under the Apache 2. License on GitHub,[36] and the initial release is published on Zenodo (**https://doi.org/10.5281/zenodo.15243143**) and can be installed as a Python package *via* PyPI.[37] The trained model weights are hosted on Hugging Face under the CC-BY-NC-ND 4.0 License.[38]

Acknowledgements

We gratefully acknowledge Nebius Ltd for supporting this research by providing access to cloud-based NVIDIA H100 and H200 GPUs, enabling efficient large-scale computation and model development.

Notes and references

1 N. Polson and V. Sokolov, *arXiv*, 2025, preprint, arXiv:2501.05458, DOI: **10.48550/arXiv.2501.05458**.
2 H. H. Loeffler, J. He, A. Tibo, J. P. Janet, A. Voronov, L. H. Mervin and O. Engkvist, *ChemInform*, 2024, **16**, 20, DOI: **10.1186/s13321-024-00812-5**.
3 D. Reidenbach, M. Livne, R. K. Ilango, M. Gill and J. Israeli, *arXiv*, 2022, preprint, arXiv:2208.09016, DOI: **10.48550/arXiv.2208.09016**.
4 K. Adams and C. W. Coley, *arXiv*, 2022, preprint, arXiv:2210.04893, DOI: **10.48550/arXiv.2210.04893**.
5 E. Hoogeboom, V. G. Satorras, C. Vignac and M. Welling, *arXiv*, 2022, preprint, arXiv:2203.17003, DOI: **10.48550/arXiv.2203.17003**.
6 Z. Chen, B. Peng, S. Parthasarathy and X. Ning, *arXiv*, 2023, preprint, arXiv:2308.11890, DOI: **10.48550/arXiv.2308.11890**.
7 K. Adams, K. Abeywardane, J. Fromer and C. W. Coley, *arXiv*, 2024, preprint, arXiv:2411.04130, DOI: **10.48550/arXiv.2411.04130**.
8 S. Murkli, J. N. McNeill and L. Isaacs, *Supramol. Chem.*, 2018, **31**, 150–158. DOI: **10.1080/10610278.2018.1516885**.
9 A. C. Anderson, *Chem. Biol.*, 2003, **10**, 787–797, DOI: **10.1016/j.chembiol.2003.09.002**.
10 D. J. Durand and N. Fey, *Chem. Rev.*, 2019, **119**, 6561–6594, DOI: **10.1021/acs.chemrev.8b00588**.
11 Z. Kaya, E. Bentouhami, K. Pelzer and D. Armspach, *Coord. Chem. Rev.*, 2021, **445**, 214066, DOI: **10.1016/j.ccr.2021.214066**.
12 A. H. Cheng, A. Lo, S. Miret, B. H. Pate and A. Aspuru-Guzik, *J. Chem. Phys.*, 2024, **160**, 124115, DOI: **10.1063/5.0196620**.
13 A. Cheng, A. Lo, K. L. K. Lee, S. Miret and A. Aspuru-Guzik, *arXiv*, 2024, preprint, arXiv:2412.12540, DOI: **10.48550/arXiv.2412.12540**.
14 A. Bauder, *Handbook of High-resolution Spectroscopy*, 2011, DOI: **10.1002/9780470749593.hrs002**.
15 C. Vignac, N. Osman, L. Toni and P. Frossard, *arXiv*, 2023, preprint, arXiv:2302.09048, DOI: **10.48550/arXiv.2302.09048**.

16 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *ChemInform*, 2011, **3**, 33, DOI: **10.1186/1758-2946-3-33**.
17 J. Guan, W. W. Qian, X. Peng, Y. Su, J. Peng and J. Ma, *arXiv*, 2023, preprint, arXiv:2303.03543, DOI: **10.48550/arXiv.2303.03543**.
18 D. A. Sapegin and J. C. Bear, *Digit. Discov.*, 2024, **3**, 186–200, DOI: **10.1039/D3DD00178D**.
19 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, *arXiv*, 2019, preprint, arXiv:1912.01703, DOI: **10.48550/arXiv.1912.01703**.
20 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2011, **40**, D1100–D1107, DOI: **10.1093/nar/gkr777**.
21 J. M. Blaney and J. S. Dixon, *Reviews in Computational Chemistry*, 1994, pp. 299–335, DOI: **10.1002/9780470125823.ch6**.
22 G. Landrum, P. Tosco, B. Kelley, Ric, D. Cosgrove, Sriniker, R. Vianello, Gedeck, N. Schneider, G. Jones, E. Kawashima, N. Dan, A. Dalke, B. Cole, M. Swain, S. Turk, A. Savelev, A. Vaucher, M. Wójcikowski, I. Take, V. F. Scalfani, D. Probst, K. Ujihara, G. Godin, A. Pahl, R. Walker, J. Lehtivarjo and F. Berenger, *strets123 and jasondbiggs, rdkit/rdkit: Release_2023.09.5*, Zenodo, 2024, DOI: **10.5281/zenodo.10633624**.
23 CCDC Virtual Screening Set, accessed on 19.04.2025, **https://www.ccdc.cam.ac.uk/support-and-resources/downloads/**.
24 S. J. Reddi, S. Kale and S. Kumar, *arXiv*, 2019, preprint, arXiv:1904.09237, DOI: **10.48550/arXiv.1904.09237**.
25 J. A. Grant and B. T. Pickup, *Computer Simulation of Biomolecular Systems*, 1997, pp. 150–176, DOI: **10.1007/978-94-017-1120-3_5**.
26 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519, DOI: **10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P**.
27 S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang and S. H. Bryant, *Nucleic Acids Res.*, 2015, **44**, D1202–D1213, DOI: **10.1093/nar/gkv951**.
28 T. Akhmetshin, A. I. Lin, D. Mazitov, E. Ziaikin, T. Madzhidov and A. Varnek, *ChemRxiv*, 2021, preprint, DOI: **10.26434/chemrxiv-2021-18x0d**.
29 K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter and G. Klambauer, *J. Chem. Inf. Model.*, 2018, **58**, 1736–1741, DOI: **10.1021/acs.jcim.8b00234**.
30 M. Song, L. Pang, M. Zhang, Y. Qu, K. V. Laster and Z. Dong, *Signal Transduct. Target. Ther.*, 2023, **8**, 148, DOI: **10.1038/s41392-023-01409-4**.
31 H. Tazarki, W. Zeinyeh, Y. J. Esvan, S. Knapp, D. Chatterjee, M. Schröder, A. C. Joerger, J. Khiari, B. Josselin, B. Baratte, S. Bach, S. Ruchaud, F. Anizon, F. Giraud and P. Moreau, *Eur. J. Med. Chem.*, 2019, **166**, 304–317, DOI: **10.1016/j.ejmech.2019.01.052**.
32 R. H. B. Smith, A. C. Dar and A. Schlessinger, *bioRxiv*, 2019, preprint, DOI: **10.1101/816702**.
33 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2009, **31**, 455–461, DOI: **10.1002/jcc.21334**.
34 A. Nichol and P. Dhariwal, *arXiv*, 2021, preprint, arXiv:2102.09672, DOI: **10.48550/arXiv.2102.09672**.
35 A. Schneuing, *et al*, *arXiv*, 2022, preprint, arXiv:2210.13695, DOI: **10.48550/arXiv.2210.13695**.
36 D. Sapegin, A. Gafurov and F. Bakharev, *Membrizard/ml_conformer_generator: Initial Release*, Zenodo, 2025, DOI: **10.5281/zenodo.15243143**, **https://github.com/Membrizard/ml_conformer_generator**.
37 D. Sapegin, A. Gafurov and F. Bakharev, mlconfgen v0.0.1, PyPi. **https://pypi.org/project/mlconfgen/**.
38 Denis Sapegin, 2025, DOI: **10.57967/hf/5165**.