Nano Research

Open Access

Research Article

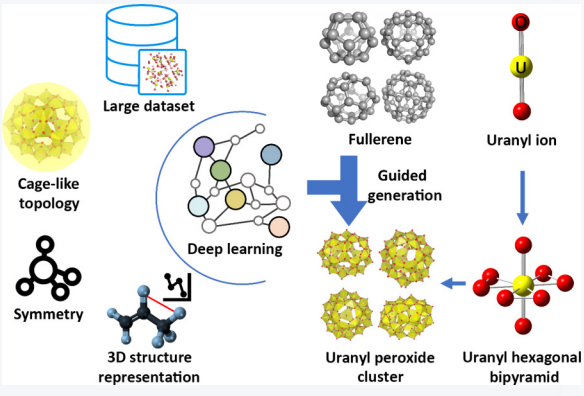
Deep learning guided generation of uranyl peroxide clusters via fullerene inspired topologies

Bin Li, Jian Ouyang, Yihan Zou, Xiaocheng Xu, and Han-Shi Hu ✉

Department of Chemistry and Engineering Research Center of Advanced Rare-Earth Materials of Ministry of Education, Tsinghua University, Beijing 100084, China

Cite this article: *Nano Research*, **2025**, *18*, 94908047. https://doi.org/10.26599/NR.2025.94908047

ABSTRACT: Uranyl peroxide clusters, formed from uranyl ions, peroxide and other ligands, exhibit complex structural diversity and have significant potential in nuclear chemistry applications. Theoretical prediction is essential for atomic-level understanding of their formation and stability, offering insights where experiments face challenges from uranium's radioactivity and complex coordination. However, their large size and varied topologies pose challenges for theoretical study. In this work, we introduce Uni-Gen, a deep learning model based on the Uni-Mol framework, which generates uranyl cluster structures by leveraging topological similarities with carbon clusters, particularly fullerene. This newly-developed Uni-Gen not only validates the stability of the U_{28} uranyl cluster consistent with previous experimental data, but also discovers a more stable U_{44} uranyl cluster than the previously reported structure, and most notably, predicts the previously unreported isomers of the U_{38} uranyl cluster. Our results demonstrate that Uni-Gen is a powerful tool for predicting diverse potentially stable uranyl cluster structures, thereby enriching our understanding of the stability landscape of uranyl clusters and offering potential for further exploration of novel uranyl materials and chemical phenomena.



KEYWORDS: uranyl cluster, fullerene, *ab initio* molecular dynamics (AIMD), deep learning

1 Introduction

The uranyl ion (UO_2^{2+}) is pivotal in actinide chemistry, governing uranium's environmental mobility, nuclear cycle behavior, and coordination-driven material design through its dual nature: thermodynamically stable yet kinetically versatile [1–3]. Among compounds formed by uranyl ion, uranyl peroxide clusters, which consist of multiple uranyl ions bound by peroxide and other ligands, exhibit extraordinary structural versatility, ranging from small aggregates [4], cup-shaped and open-ring structures [5] to large cage-like ones containing up to 124 uranyl ions [6–9]. Most cage structures are fullerene-like topologies, including U_{20} [10, 11], U_{28} [12], U_{30} [13], U_{36}, U_{44} [14], U_{50} [15], and U_{60} [16]. In these clusters, uranyl ions are typically coordinated by bridging peroxide ligands in a side-on (μ-η^2-η^2) coordination mode, while end-on (η^2) coordination is exceedingly rare and often exists in some small fragments [17]. The peroxide units are positioned at the equatorial vertices of hexagonal bipyramids, with the apexes of the bipyramids corresponding to the oxygen atoms of the uranyl dication [18–21]. These uranyl polyhedra are arranged into tetragons, pentagons and hexagons, forming close-shell clusters with fullerene topology [15]. Uranyl peroxide clusters can easily be synthesized experimentally by simply adding hydrogen peroxide and alkali metal hydroxides to uranyl solution [14]. They are thermodynamically stable, radioresistant [22], capable of encapsulating multiple uranyl ions simultaneously, and readily form crystals in the form of alkali metal salts, thus efficiently purifying uranyl ions from contaminated water without the use of toxic reagents or complex processing steps [23–27]. Moreover, their structural versatility allows control of the uranium distribution between aqueous and solid phases based on varying conditions such as solvent, pH, and the presence of counter cations, which offers a more efficient and atom-economical approach for optimizing the nuclear fuel cycle [12, 28–31]. While many of the uranyl peroxide clusters are topologically identical to the carbon fullerene clusters, with more than 10,000 structures ranging from 20 to 200 carbon atoms [32–37], comparatively

Received: June 20, 2025; **Revised:** August 29, 2025
Accepted: September 3, 2025

✉Address correspondence to hshu@mail.tsinghua.edu.cn

numerous potential stable fullerene-like structures remain undiscovered in uranyl peroxide systems. Furthermore, the structural mimicry between carbon and uranyl clusters provides a valuable heuristic for predicting more uranyl peroxide clusters [3].

Molecular conformation generation (MCG) plays a crucial role in drug discovery [38–40], material discovery [41–46], catalyst design [47, 48], protein design [49–51] and so on [52, 53], as it aims to generate low-energy spatial arrangements of atoms in a molecule or in a lattice. Traditional MCG methods rely on stochastic, rule-based or similar discrete interpolation techniques approaches to generate molecular conformations, usually followed by energy minimization using techniques like force field molecular mechanics or density functional theory (DFT) [54–56]. Examples include RDKit, TG-Min [57], ABCluster [58], etc. They can do extensive explorations of the energy landscape and accurately sample equilibrium structures, but quickly become prohibitively slow for larger molecules. Recently, deep learning generative models, such as variational autoencoders (VAEs), generative adversarial networks (GANs), graph attention networks (GATs), and diffusion models, have been applied to MCG [59]. Early models like CVGAE used VAEs to generate atomic coordinates for small molecules but struggled to maintain translation and rotation equivariance [60]. More recent models, such as GeoMol and DMCG [61, 62], refine coordinates iteratively and introduce loss functions invariant to rotational and translational transformations, advancing the design of complex molecules.

However, applying these models to uranyl peroxide clusters is challenging due to the exponential growth of possible structures as the atomic number and types increase. Uranyl peroxide clusters, typically consisting of hundreds of atoms including uranium, oxygen, and various alkali metals, involve multiple bonding types such as U=O covalent bonds and ionic interactions between alkali metals and peroxide groups. Their topologies range from simple linear chains to complex cage-like structures, greatly complicating their conformation generation.

Considering the intricate nature of uranyl peroxide clusters, we hereby developed a deep learning generative model "Uni-Gen" to generate initial structures for uranyl peroxide clusters based on a flexible three-dimensional molecular representation learning framework "Uni-Mol" [63]. The Uni-Gen model utilizes an SE(3)-equivariant encoder-decoder transformer architecture, which encodes all atomic types and coordinate information simultaneously, avoiding systematic errors that arise from neighborhood approximation in cluster molecules. We further optimize the model through a two-stage pre-training and fine-tuning process by incorporating unique structural features of uranyl peroxide clusters, such as fixed coordination environments, cage-like spherical structures, and high atomic densities. These modifications ensure the model captures complex structural information while preventing overfitting. Our model offers significant advantages over traditional structural search algorithms, such as TG-Min and ABCluster, by accelerating structural generation, thereby reducing the computational time required for exploring the conformational space of uranyl peroxide clusters. Moreover, to better capture the complex structures of uranyl peroxide clusters, our model explores topological relationships between fullerenes and uranyl peroxide clusters during the fine-tuning process. By capitalizing on the known structural diversity and stability of fullerenes, Uni-Gen is able to extrapolate these patterns to the more complex uranyl peroxide systems, resulting in stable and geometrically plausible cluster structures.

Furthermore, we employ the Uni-Gen workflow to generate structures for uranyl peroxide clusters using the structures of carbon clusters, followed by optimization using the CP2K software package [64]. Our approach successfully produced and optimized multiple clusters based on experimentally synthesized clusters, including $K_{28}[(UO_2)_{28}(O_2)_{42}]$ (U_{28}) and $K_{44}[(UO_2)_{44}(O_2)_{66}]$ (U_{44}), etc. For U_{28}, the experimentally synthesized T_d-symmetric isomer consistently exhibits lower energy than its D_2 counterpart, although small energy differences suggest that alternative configurations may also be viable. For U_{44}, we identified seven isomers, including a novel D_2-symmetric structure with lower energy than the bispherical experimental form [14], with frequency calculations confirming the reliability of these results. Moreover, the generation of 17 hitherto unreported $K_{38}[(UO_2)_{38}(O_2)_{57}]$ isomers of U_{38}, a uranyl peroxide cluster that has not been reported either theoretically or experimentally, demonstrates the potential of our method to predict novel uranyl peroxide clusters and guide future experimental synthesis.

2 Experimental

2.1 Workflow and model design of Uni-Gen

The Uni-Gen model is a deep three-dimensional generative model pre-trained on a large uranyl peroxide database and fine-tuned for generation from fullerene to uranyl peroxide clusters. Figure 1 offers a schematic overview of the whole Uni-Gen framework.

The Uni-Gen workflow is depicted in Fig. 1(a). It begins with a pre-training phase using an extensive uranyl peroxide database, where the model undergoes a three-dimensional position denoising task to learn the spatial distribution of uranyl peroxide clusters. This pre-training process enables the model to generate new isomers of known clusters. But given the complexity and diversity of atoms in uranyl peroxide clusters, which adds a high computational load, the model is further fine-tuned on a uranyl peroxide-fullerene cluster database to explore a broader conformational space. This fine-tuning step allows the model to learn structural similarities between fullerene and uranyl peroxide clusters through denoising from random coordinates. Thereafter, for each collected fullerene coordinate, Uni-Gen generates the corresponding cluster coordinates for subsequent structural optimization by leveraging the molecular formula of uranyl peroxide clusters and the fullerene's topology, and these structures are then optimized using first-principles programs like CP2K, Vienna *ab-initio* simulation package (VASP), and Gaussian. Ultimately, the newly optimized and stable uranyl peroxide structures are obtained and integrated with the initial database to form an extended one. The model will be iteratively trained on the extended database, continuously optimizing and improving as new structures are generated and integrated into it until the model is capable of capturing structural information of uranyl peroxide clusters as the training loss is small enough.

For generation of molecules, graph neural networks are frequently employed. However, their local connectivity limits their ability to effectively represent uranyl peroxide clusters. To address this, we utilize the Transformer architecture. As shown in Fig. 1(b), the Uni-Gen model adopts an encoder-decoder Transformer framework [65], comprising three key components: an encoder, a decoder, and a structural noise part. The model takes in two distinct 3D molecular structures, referred to as "prompt" and "input",

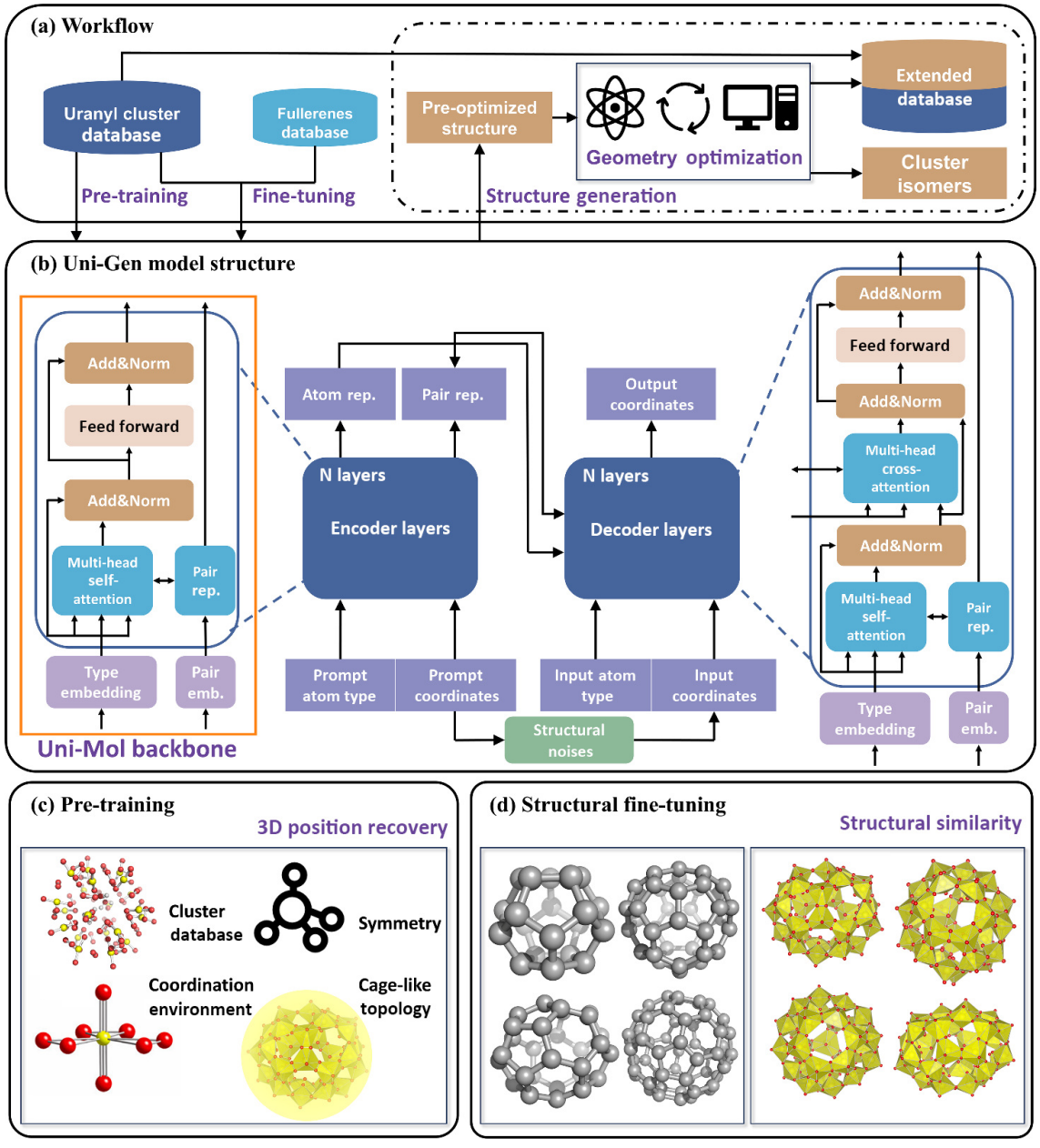


Figure 1 (a) Workflow of Uni-Gen. Through a two-stage pre-training and fine-tuning, we can generate uranyl peroxide clusters with specific fullerene topologies from a fullerene cluster structure. (b) Detailed model structure of Uni-Gen. The input and output of the model are both the three-dimensional structure of clusters. (c) Pre-training strategies. In the pre-training stage, we add random structural noise with considering the intricate properties of uranyl peroxide clusters, including high-symmetry, cage-like topology and constrained coordination environment. A three-dimensional position denoising task was carried out to learn the spatial representation of uranyl peroxide clusters. (d) Structural fine-tuning strategy. The model learns the structural similarity between fullerene and uranyl peroxide clusters through denoising from random coordinates (the yellow dots represent uranium, the gray dots represent carbon, the red dots represent oxygen, and the yellow polyhedra represent uranyl peroxide coordination polyhedra).

whose types vary across pre-training, fine-tuning, and generation stages. It then outputs the modified 3D coordinates of atoms corresponding to the input ones. In the pre-processing of the molecular 3D structures, we begin with the Cartesian coordinates and atom types of the given clusters, and then convert the coordinates into distance matrices, which are embedded into pair embedding. This approach guarantees the special Euclidean (3) (SE(3)) group symmetry of the system. Meanwhile, we embed the corresponding atom types of the coordinates into type embedding. Additionally, by encoding all atoms simultaneously through pair embedding and type embedding, we can avoid the systematic errors that typically occur due to neighborhood approximation when encoding large molecules. The initial type embedding and pair embedding will go through several layers (l) iteratively. The same initial type embeddings will go through several heads (h) with different weights in self-attention and cross-attention simultaneously.

In the encoder part, we utilized Uni-Mol as the representation learning framework to represent the molecule, which is a specialized pure three-dimensional pre-training model designed for molecules. Uni-Mol has demonstrated remarkable performance across diverse downstream tasks within the realm of drug

discovery, metal-organic frameworks (MOF) properties prediction [66] and nuclear magnetic resonance spectroscopy prediction [67]. However, the uranyl peroxide clusters have significantly different structures and three-dimensional spatial distributions compared to small-molecule drug and MOF. Moreover, their topologies are complex, and they contain a much larger number of heavy atoms and fragments. In light of these characteristics, we modified Uni-Mol. We employ Uni-Mol's self-attention with pair-to-atom communication, which is denoted as (Eq. (1)):

$$\text{Attention}\left(Q_i^{l,h}, K_j^{l,h}, V_j^{l,h}\right) = \text{softmax}\left(\frac{Q_i^{l,h}\left(K_j^{l,h}\right)^T}{\sqrt{d}} + q_{ij}^{l-1,h}\right) V_j^{l,h} \quad (1)$$

where $Q_i^{l,h}(K_j^{l,h}, V_j^{l,h})$ is the Query (Key, Value), $q_{ij}^{l,h}$ is the representation of the atom pair composed of the i-th and j-th atoms in the l-th layer and the h-th head. But for the decoder part, besides the self-attention, we perform cross-attention between the output representation of the encoder and the output of the self-attention, which is denoted as follows (Eq. (2)):

$$\text{Cross}\left(Q_i^{l,h}, K_j^{l,h}, \widetilde{V}_j^{l,h}\right) = \text{softmax}\left(\frac{Q_i^{l,h}\left(K_j^{l,h}\right)^T}{\sqrt{d}} + \widetilde{q}_{ij}^{l-1,h}\right) \widetilde{V}_j^{l,h} \quad (2)$$

where $Q_i^{l,h}\left(K_j^{l,h}\right)$ is the Query (Key) of the i-th (j-th) atom in the l-th layer and h-th head of the encoder layer. $\widetilde{V}_j^{l,h}\left(\widetilde{q}_{ij}^{l,h}\right)$ is the Value (pair representation) of the i-th (j-th) atom in the l-th layer and the h-th head of the decoder layer.

By incorporating 3D spatial positional encoding and pair representation, the model effectively captures 3D structural information. However, it lacks the capability to directly generate atomic coordinates, a crucial requirement for 3D spatial generation. To address this limitation, we employed Uni-Mol's SE(3)-equivariant head that predicts position shifts based on the SE(3)-invariant pair representation and the equivariant input $x_i - x_j$ [63], expressed as (Eq. (3)):

$$\widehat{x}_i = x_i + \sum_{j=1}^{n} \frac{(x_i - x_j)\, c_{ij}}{n}; c_{ij} = \text{ReLU}\left(\left(q_{ij}^l - q_{ij}^0\right) U\right) W \quad (3)$$

where n is the number of atoms, $\text{ReLU}(\cdot)$ is rectified linear unit [68], U, W are projection matrices of pair representations.

To further improve performance, we introduced a new kernel for geometric spatial positional encoding. Unlike drug molecules, clusters exhibit denser three-dimensional atomic arrangements, characterized by porous structures and an average of over 300 heavy atoms per molecule. While Uni-Mol employs a Gaussian kernel for spatial positional encoding, pretraining on clusters becomes unstable due to significantly larger pairwise distances and atom counts. Additionally, the Gaussian kernel exhibits excessive smoothness near zero and rapidly diminishes as d_{ij} increases, making it inadequate for capturing abrupt changes in local atomic environments and ineffective in modeling long-range interactions. To overcome these challenges, we introduced a mixed distance kernel, presented below (Eq. (4)), and incorporated it into both the encoder and decoder components of the model.

$$p_{ij} = \text{EMB}_1(t_{ij}) \cdot \phi(d_{ij}) + \text{EMB}_2(t_{ij}) \cdot \mathcal{M}_{\frac{3}{2}}(d_{ij}) \quad (4)$$

where d_{ij} is the Euclidean distance of atom pair representation between i-th and j-th atom, and t_{ij} is the edge type between i-th and j-th atom, $\text{EMB}(\cdot)$ denotes the embedding, $\phi(\cdot)$ is the conventional Gaussian kernel and $\mathcal{M}_{3/2}(d) = \left(1 + \frac{\sqrt{3}d}{\sigma}\right)\exp\left(-\frac{\sqrt{3}d}{\sigma}\right)$ is the Matérn 3/2 kernel. At this stage, the overall architecture of the Uni-Gen model has been outlined, with the encoder-decoder framework specifically tailored to facilitate fine-tuning on the uranyl peroxide-fullerene cluster dataset. In this design, attention mechanisms are primarily responsible for capturing atom-type relationships, while interatomic distances are modeled through physically informed kernels. These two sources of information are further integrated within the attention layers, allowing the model to jointly reason over both geometric and chemical features. This separation of roles enables a more interpretable and physically grounded representation, which is particularly important for modeling the complex coordination environments of uranyl peroxide clusters. And given the considerable computational demands of pretraining on large-scale datasets and the high dimensional complexity of uranyl peroxide clusters, the adoption of this efficient backbone with a total of 44 million trainable parameters is well justified. A detailed description of the model architecture is provided in Table S3 in the Electronic Supplementary Material (ESM).

2.2 Training detail

2.2.1 *Pre-training strategies*

Self-supervised learning plays a crucial role in effectively extracting knowledge from large-scale unlabeled data. For instance, bidirectional encoder representations from Transformers' (BERT's) masked token prediction task enables the model to capture contextual information [69]. Similarly, in the context of 3D data processing, Uni-Gen aims to facilitate the model's understanding of 3D structural information during pretraining. As shown in Fig. 1(c), several strategies are employed during the pre-training stage. First of all, a 3D position recovery self-supervised task is introduced to reconstruct accurate 3D positions from altered inputs on a large uranyl peroxide database. A straightforward method could be a strategy similar to BERT's token-based masking, applying a comparable principle to positions. Nevertheless, rather than implementing this position-related masking approach, the input is perturbed by adding structural noises. This is because such a position masking-like strategy might result in an over-emphasis on local information, whereas the proposed method is more effective in capturing global 3D structural relationships, enabling the model to predict the correct positions.

Following the introduction of the 3D position recovery self-supervised task in Uni-Gen, mapping from random positions to ground-truth atomic positions is inherently challenging. To optimize learning efficiency, we further implement two key strategies to minimize the discrepancy between the random and true positions. Using these input coordinates with structural noises, we use two extra network heads for position recovery. Specifically:

(1) Pair-distance prediction: The model predicts correct Euclidean distances between modified atom pairs based on pairwise representations.

(2) Coordinate prediction: It forecasts accurate atomic coordinates with an SE(3)-equivariant coordinate head. Based on these two heads, the loss of the pre-training process is as follows (Eq. (5)):

$$\mathcal{L} = \sum\nolimits_i \text{smooth}_{L_1}(\widehat{x}_i, x_i) + 2\sum\nolimits_{j,k} \text{smooth}_{L_1}(\widehat{p}_{jk}, p_{jk}) \quad (5)$$

Here, x_i, p_{jk} represent atomic coordinates and elements of the pairwise distance matrix, $\text{smooth}_{L_1}(\cdot)$ denotes the smooth L_1 loss. Ablation studies (provided in the ESM) demonstrate that incorporating pairwise distance information is crucial for efficient learning. In the pre-training stage, the "prompt" corresponds to the most stable structure of a given cluster, whereas the "input" corresponds to its structurally perturbed variants within the dataset. Specifically, in each pre-training epoch, a newly perturbed version of each training structure is generated, enhancing sample diversity and promoting the model's ability to learn robust and generalizable structural representations. Considering the intricate properties of uranyl peroxide clusters, we propose a novel random shift algorithm instead of the uniform noise to a subset of atoms. This procedure begins by calculating the molecular centroid and determining the coordination number of each atom. The calculation formula for the i-th atom's coordination number (c_i) is as follows (Eq. (6)):

$$c_i = \sum_j s_0 \frac{1}{1+\left(\frac{r_{ij}}{r_0}\right)^6} - s_0 \tag{6}$$

The default values are $s_0 = 500, r_0 = 0.5$. For heavy atoms with higher coordination numbers, no perturbation is applied, allowing the model to better capture the coordination environment. Then for the atoms that need structural noise, given the general spherical symmetry of cluster structures, perturbations are restricted to the plane normal to the vector from the centroid to the atom, preventing excessive atomic aggregation within the molecular interior and the breakage of the cage-like topology. Additionally, the number of perturbed atoms is set within a default range of 25% to 50%, which is adjustable as per the specific requirements. Ablation studies confirm that this perturbation strategy aligns well with chemical intuition and enhances training effectiveness.

2.2.2 *Structural fine-tuning*

Following the pre-training stage, the model demonstrates the ability to generate uranyl peroxide clusters with reasonable accuracy. However, due to the complex topological features and diverse assembly patterns of these clusters, the Uni-Gen model, trained solely through pre-training, struggles to effectively explore the broader conformation space. As illustrated in Fig. 1(d), numerous reported uranyl peroxide clusters share fullerene-like topological frameworks, exhibiting high degrees of topological similarity. Leveraging this chemical insight, we want to further utilize well-studied fullerene clusters to guide and enhance the structural generation of uranyl peroxide clusters. Considering that the three-dimensional structural representation of uranyl peroxide clusters has already been learned in pre-training, we focused on the similarity between fullerene clusters and uranyl peroxide clusters during the structural fine-tuning stage. For Uni-Gen, this task straightforwardly turns into a conformation optimization task: generating a new conformation from a different input one. Specifically, in the fine-tuning phase, the model undergoes self-supervised learning to map the conformations generated by Uni-Gen to the labeled ones. Importantly, the 3D structures of fullerene clusters are used as topological templates or "prompts" to guide the generation of uranyl peroxide clusters with analogous cage-like architectures, rather than to transfer bonding patterns. To ensure consistency with the pre-training phase, we employ the identical data preprocessing pipeline during the fine-tuning process and the loss function. Because fullerene clusters serve as topological references within the cross-attention mechanism, their structural representations are padded to match the dimensionality of uranyl peroxide clusters.

2.3 Dataset preparation

In the pretraining stage, we generated the dataset using *ab initio* molecular dynamics (AIMD) to ensure that the training set encompasses a diverse array of uranyl peroxide cluster structures with sufficient 3D structure coordination data. We systematically reviewed the available literature and collected 16 types of stable uranyl clusters or fragments. These structures vary in composition and topological structure significantly, incorporating elements such as U, O, H, C, Li, Na, and K. These clusters include a wide range of types, such as uranyl dimers, tetramers, pentamers, hexamers, icosamers, and hexacontamers, which exhibit fullerene topologies. The ligands of these clusters include O_2^{2-}, OH^-, $C_2O_4^{2-}$, and H_2O. For each cluster or fragment, 4000 structure coordination data points are randomly generated by AIMD for the purpose of enrichment of data and collected into the dataset. Additionally, we utilized the AIMD data from our previous studies on U_{20} and U_{60} clusters to enhance the comprehensiveness and reliability of the current dataset. Ultimately, in the pre-training phase, the total dataset comprised 64,000 structure coordination data points. To address potential redundancy in the AIMD-generated dataset, we focused on sampling only stable regions of the trajectories, as these represent chemically meaningful, low-energy configurations of uranyl peroxide clusters. Although some structural overlap is unavoidable, repeated exposure to slight variations of stable geometries helps the model learn robust coordination patterns and energetically favorable motifs. To assess structural diversity, we analyzed both the relative energy distribution and the pairwise root-mean-square deviation (RMSD) of atomic coordinates across the dataset. As shown in Figs. S1 and S2 in the ESM, the dataset spans a wide range of energies and geometries, indicating substantial diversity despites the inherent stability of the configurations.

For the structural fine-tuning, we also reviewed the literature and collected 8 reported uranyl peroxide clusters ranging from U_{20} to U_{60} and their corresponding fullerene clusters with identical topology. In order to solve the problem of insufficient data, we used AIMD to generate a large amount of 3D coordination data for these 8 clusters. Specifically, we proportionally scaled the uranium skeleton to geometrically match the topological features of fullerene frameworks. In this phase, we employed a pseudo-random number generator to randomly sample 200 structures for each of the eight clusters. This approach generated a fine-tuning dataset consisting of 1600 data points, suitable for in-depth training. The pre-training and structural fine-tuning dataset were subsequently partitioned into training, validation, and testing subsets using an 8:1:1 ratio.

3 Results

3.1 Performance

We evaluated the Uni-Gen model's performance across various aspects. The detailed results are illustrated in Fig. 2. In Fig. 2(a), the pre-training loss curve shows that the model exhibits a substantial decline in both training and validation losses during the initial epochs, indicating its effective ability to learn 3D positional patterns from the uranyl peroxide database. The training loss decreases

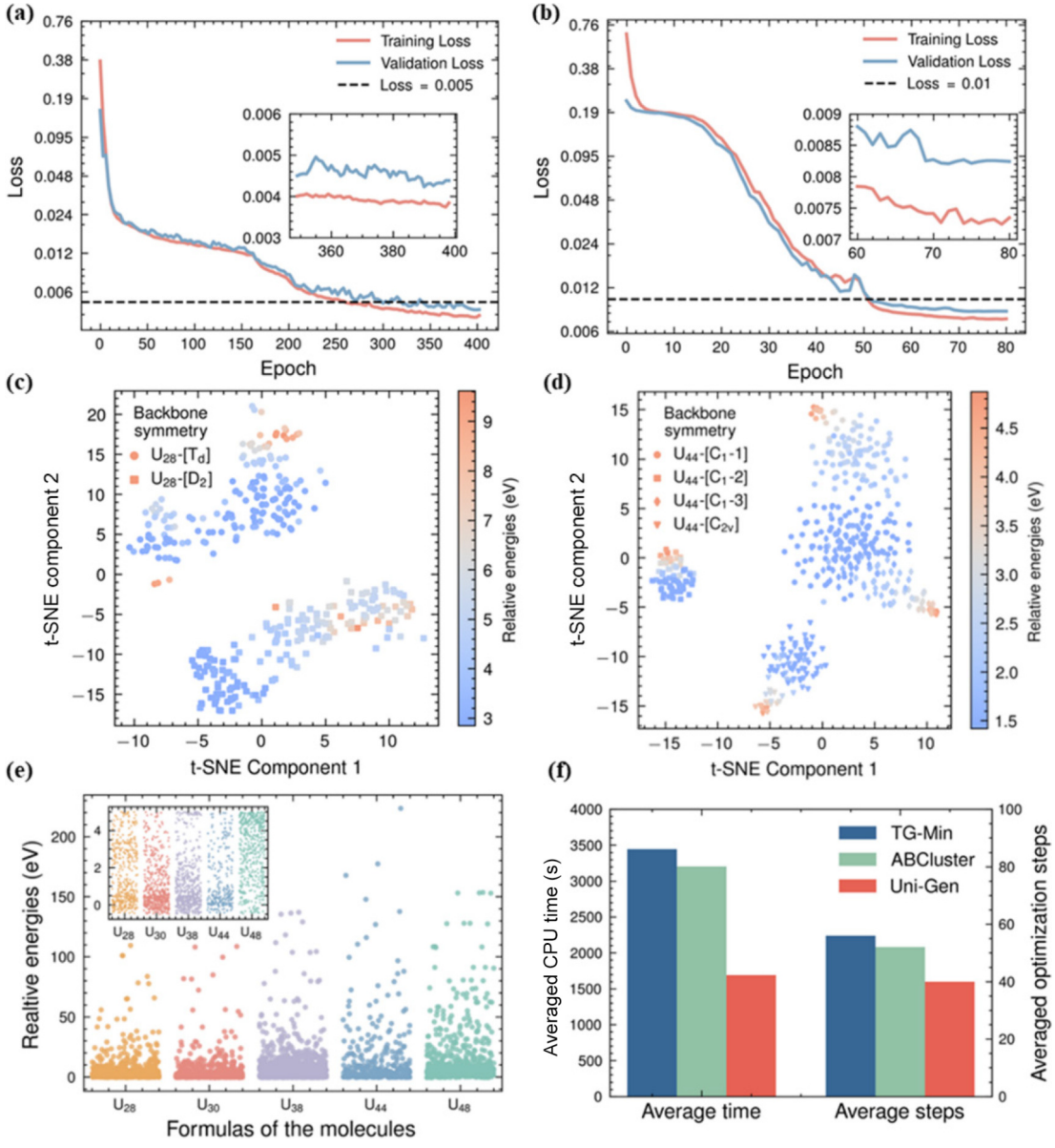


Figure 2 Performance of the Uni-Gen model. Loss curve for the (a) pre-training phase and (b) structural fine-tuning. (c) Structural representations of U_{28}-[T_d] and U_{28}-[D_2] computed by t-SNE method (U_n-[X-m] represents $K_n[(UO_2)_n(O_2)_{1.5n}]$ generated from the fullerene backbone with X symmetry, where m is used to distinguish the fullerene backbones with the same symmetry but different specific structures. (d) Structural representations of U_{44}-[C_1-1], U_{44}-[C_1-2], U_{44}-[C_1-3] and U_{44}-[C_{2v}] computed by t-SNE method. (e) Energies of generated conformations of uranyl clusters. Generated conformations are primarily concentrated near the energy minima rather than being randomly distributed in unreasonable high-energy regions. Use the energy of the most stable generated conformations as the energy zero point for each uranyl cluster. (f) Average central processing unit (CPU) time and average optimization steps for structure optimization of 35 generated uranyl cluster conformations by TG-Min, ABCluster, and Uni-Gen.

progressively in a steady manner, starting from approximately 0.38 and gradually dropping below 0.01. Eventually, after around 200 epochs, it converges precisely around 0.005, demonstrating the highly accurate mapping to the reference structure. Similarly, the validation loss follows a downward trajectory, consistent with the training loss trend. This convergence indicates that the model generalizes well during pre-training, showing minimal signs of overfitting as it performs comparably on both training and unseen data.

In Fig. 2(b), which represents the fine-tuning stage, the model undergoes additional optimization. Even though the fine-tuning commences with a relatively high loss of 0.76 compared to pre-training, the training and validation losses rapidly decrease. In just 40 epochs, they fell below 0.02 and eventually stabilized around 0.01. This stabilization not only indicates successful fine-tuning but also implies the model's enhanced generalization ability. Throughout this stage, the training and validation loss curves remain closely aligned, with negligible divergence, suggesting effective control over overfitting. The rapid loss reduction in the early epochs, followed by stabilization, further validates that the model swiftly adapts to fine-tuning, refining its understanding of similarity between fullerene and uranyl peroxide clusters while preserving its generalization ability.

To further assess whether the model effectively captures the structural characteristics of uranyl peroxide clusters and ensures diversity in the generated structures, we applied t-SNE (t-

distributed stochastic neighbor embedding) dimensionality reduction to the final layer representation on the generation of U_{28} and U_{44} [70]. In Figs. 2(c) and 2(d), the visualization of the results clearly shows that uranyl peroxide clusters with different configurations occupy distinct positions. This vividly demonstrates the model's remarkable ability to accurately differentiate between clusters featuring varying topological structures. Additionally, clusters of the same configuration exhibit a certain degree of diversity, as reflected in the regularity of their energy distribution, indicating that the model can also distinguish the energy changes corresponding to structures within the same skeleton topology.

In Fig. 2(e), we present various uranyl peroxide cluster structures containing five different uranium atom counts (U_n represent $K_n[(UO_2)_n(O_2)_{1.5n}]$) each over 400 structures and calculated their relative energy distributions using DFT. The zero point in the figure corresponds to the lowest energy of each cluster. Notably, most generated structures fall within the 0–20 eV range rather than being randomly distributed in unreasonable high-energy regions, suggesting that the Uni-Gen model primarily produces stable configurations. Meanwhile, the presence of some structures in the higher energy region indicates that the model effectively explores a broader conformational space.

Although uranyl peroxide clusters are similar to fullerene clusters, they may still exhibit structural differences compared to fullerenes with the same topology. Therefore, the diversity in energy distribution is essential for both the effective generation of uranyl peroxide clusters and for finding the global minima of specific uranyl peroxide structures. A detailed discussion of this will be provided in section 3.2.

To assess the stability of the Uni-Gen model in generating uranyl peroxide clusters, 35 different structures were produced using TG-Min, ABCluster, and Uni-Gen. Given the time-intensive nature of structural optimization tasks and the difficulty traditional methods face in converging on complex clusters, we selected a relatively simple system for this assessment. And these generated structures are subsequently evaluated through structure optimization tasks. As illustrated in Fig. 2(f), Uni-Gen achieves the shortest average optimization time, highlighting its efficiency in structural refinement. While TG-Min and ABCluster require slightly longer computation times, they still perform adequately but necessitate more optimization steps than Uni-Gen. This suggests that Uni-Gen is the most efficient approach in terms of both computational time and the number of steps needed to obtain optimized structures. Furthermore, when dealing with complex cluster molecules, TG-Min and ABCluster struggle to explore larger conformational spaces, further highlighting the powerful capabilities of the Uni-Gen model.

Furthermore, the performance of various structure generation methods was evaluated through two key metrics: Coverage (COV) and Matching (MAT). A higher COV reflects better diversity in generated structures, while a lower MAT indicates higher accuracy in matching known configurations [54]. Formally, COV and MAT are denoted as (Eqs. (7) and (8)):

$$\mathrm{COV}(S_g, S_r) = \frac{1}{|S_r|}\left\|\left\{R \in S_r \left| \mathrm{RMSD}\left(R, \widehat{R}\right) < \delta \right|, \widehat{R} \in S_g\right\}\right\| \tag{7}$$

$$\mathrm{MAT}(S_g, S_r) = \frac{1}{|S_r|}\sum_{R \in S_r} \min_{\widehat{R} \in S_g} \mathrm{RMSD}\left(R, \widehat{R}\right) \tag{8}$$

Four methods are evaluated against two uranyl molecules $Li_6[(UO_2)_2(O_2)_5]$ and $K_{28}[(UO_2)_{28}(O_2)_{42}]$, with 300 structures generated and tested each case for 10 times. The COV value was calculated with $\delta = 1.5\ AA$, and the results are shown in Table 1. Among the methods tested, RD-Kit exhibits relatively lower diversity and higher MAT, indicating a more limited range of structures with less accuracy. In contrast, TG-Min and ABCluster slightly improved diversity while achieving a significantly lower MAT (within 0.3 Å), demonstrating their efficiency in generating accurate structures. However, both TG-Min and ABCluster require additional DFT calculations, increasing their computational cost. Uni-Gen outperforms all other methods in both metrics, achieving the highest COV value that reflects exceptional diversity, while maintaining a slightly higher MAT but still competitive in accuracy. This comparison suggests that Uni-Gen is the most effective method overall, offering both high diversity and accuracy without the need for extensive post-processing calculations.

3.2 New uranyl clusters generated by Uni-Gen

After a thorough evaluation, the Uni-Gen model demonstrated a strong capability in generating uranyl peroxide clusters. Using this model, we built upon existing fullerene cluster structures to generate three types of uranyl peroxide clusters (U_n, $n \in (28, 38, 44)$) and subsequently optimized their geometries using the CP2K software package. This process provided stable geometries and corresponding energies for various uranyl peroxide clusters, including both previously studied clusters and their isomers, as well as novel structures yet to be synthesized. These results were compared with available experimental data. Additionally, by leveraging the symmetry of the fullerene input, the final generated structures, where uranium atoms form the vertices of the skeletal framework (excluding oxygen and cation coordinates), were classified according to their point groups. The optimized structures and their corresponding data are presented below.

3.2.1 *Results of U_{28}*

The theoretical structures and relative energies of the isomers of U_{28} ($K_{28}[(UO_2)_{28}(O_2)_{42}]$) are shown in Fig. 3(a). The calculations

Table 1 COV and MAT values to evaluate performance of generation on two distinct uranyl systems between traditional MCG methods and Uni-Gen

Method	$Li_6[(UO_2)_2(O_2)_5]$				$K_{28}[(UO_2)_{28}(O_2)_{42}]$			
	COV (↑, %)		MAT (↓, Å)		COV (↑, %)		MAT (↓, Å)	
	Mean	Medium	Mean	Medium	Mean	Medium	Mean	Medium
RD-Kit	75.33	75.66	1.694	1.687	71.67	72.00	2.345	2.455
TG-Min	83.33	83.67	0.113	0.122	72.33	73.33	0.189	0.190
ABCluster	86.33	89.67	0.181	0.199	73.00	74.33	0.274	0.288
Uni-Gen	99.33	98.67	0.231	0.211	92.67	93.00	0.342	0.322

indicate that the most stable structure among the investigated isomers is U_{28}-[T_d]. This structure is composed of 12 (U5) pentagons and 4 (U6) hexagons. Four uranyl pentagons are positioned at the four vertices of a regular tetrahedron, resulting in a highly symmetrical configuration of T_d. Notably, U_{28}-[T_d] has been experimentally synthesized and demonstrated to exhibit remarkable stability under various conditions. Similarly, the corresponding C_{28}-[T_d] is the most stable among fullerene isomers. These findings validate the reliability of the Uni-Gen model, as the experimentally realized clusters align with the predicted low-energy configurations. Moreover, the energy difference between the two predicted U_{28} isomers is relatively small, with U_{28}-[D_2] being only +2.85 eV higher, suggesting that the U_{28}-[D_2] cluster could potentially be synthesized under appropriate experimental conditions. This study underscores the capability of the Uni-Gen model in predicting novel cluster structures.

3.2.2 *Results of U_{44}*

Figure 3(b) displays the theoretical structures and energies of the three most stable U_{44} ($K_{44}[(UO_2)_{44}(O_2)_{66}]$) isomers predicted by the previously described workflow. All the three anionic framework of these isomers consists of 12 (U5) pentagons and 11 (U6) hexagons. Notably, U_{44}-[D_{3d}] adopts a bispherical topology corresponding to the uranyl anion spontaneously self-assembling in an aqueous solution, as experimentally observed and as the experimental synthesis of this structure has been reported in the literature. Taking this experimentally identified structure as the energy zero point, the energy differences of the other two calculated U_{44} isomers range from −0.83 to +1.83 eV, indicating their plausibility as alternative isomers. All identified isomers are detailed in the ESM.

Furthermore, U_{44}-[D_2] exhibits a lower energy than U_{44}-[D_{3d}], suggesting its theoretical stability and potential for further study. However, the relatively small energy difference of 0.83 eV indicates that the experimentally observed structure remains significantly stable. Additionally, CP2K frequency calculations for U_{44}-[D_2] and U_{44}-[D_{3d}] yield results of 26/916 and 24/916 (number of imaginary frequencies/total vibrational modes), with all imaginary frequencies below 50 cm^{-1}. These minor deviations confirm the reliability of the computational results.

Despite the strong topological resemblance between fullerene clusters and the corresponding stable uranyl peroxide clusters, notable differences exist in specific bond lengths, bond angles, and skeletal geometries. For instance, C_{44}-[D_2] exhibits a more spherical shape, whereas U_{44}-[D_2] adopts an ellipsoidal form. Similarly, while C_{44}-[D_{3d}] maintains a standard ellipsoidal structure, U_{44}-[D_{3d}] features a bispherical shape. These structural variations highlight the importance of diversity in uranyl peroxide cluster generation based on fullerene frameworks for discovering novel cluster architectures. The findings further validate the reliability of the Uni-Gen model and underscore its significance in predicting and analyzing new uranyl peroxide cluster structures.

3.2.3 *Results of U_{38}*

The U_{38} cluster has not yet been experimentally synthesized, with the exception of the nitrate-bridged double spherical uranyl peroxide clusters composed of pyrophosphate and nitrate $U_{38}Py_{10}Nt_4$ [71]. Whereas numerous isomers of C_{38} have been identified through both experimental and theoretical investigations. Using the established workflow, we generated over 400 corresponding U_{38} ($K_{38}[(UO_2)_{38}(O_2)_{57}]$) isomers, each consisting of 12 (U5) pentagons and 9 (U6) hexagons. Following structural optimization, 17 stable isomers were ultimately obtained. The energy spectra and structures of these isomers are shown in Fig. 4. In the case of U_{38}, the fullerene framework exhibits a strong resemblance to the resulting stable uranyl peroxide clusters, demonstrating a higher degree of structural similarity compared to other clusters.

The energy differences among these 17 U_{38} isomers range from 0.00 to 5.45 eV. Among them, U_{38}-[C_2-1] is the most stable configuration, while U_{38}-[C_2-2] and U_{38}-[C_1] exhibit minor energy variations, differing by only 0.41 and 0.45 eV, respectively. Although several of the most stable configurations of U_{38} primarily feature a low-symmetry spherical cage structure, the cluster as a whole displays considerable structural diversity across its various isomers. All structural representations and coordinate files will be provided in the ESM. Given the complex electronic characteristics of uranyl peroxide clusters and potential model limitations, these findings indicate that most of these isomers could be promising candidates for further theoretical and experimental exploration.

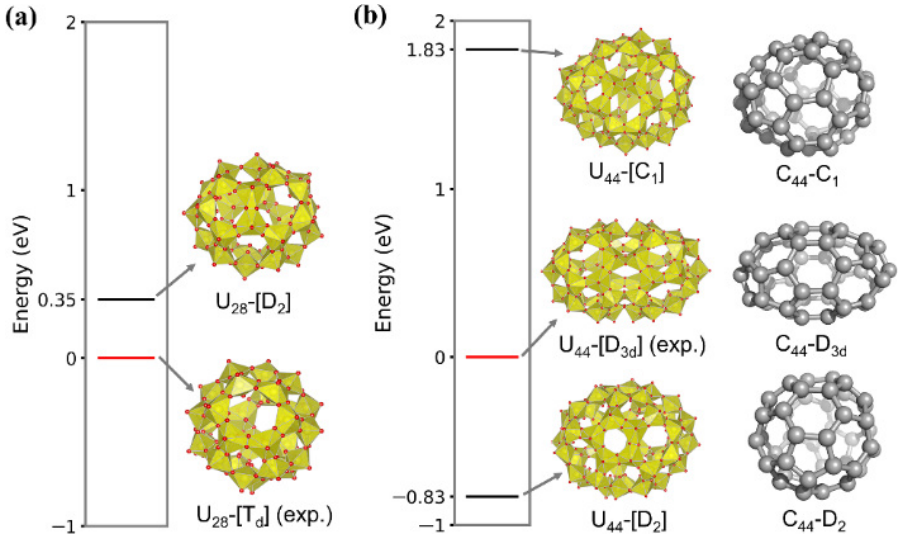


Figure 3 Application of the Uni-Gen model in U_{28} and U_{44}. (a) Energies and structures of the two most stable generated uranyl clusters with 28 vertices, using the energy of the experimentally observed conformation U_{28}-[T_d] as the energy zero point. (b) Energies and structures of the 3 most stable generated uranyl clusters with 44 vertices along with the structures of their corresponding fullerene backbones, using the energy of the experimentally observed conformation U_{44}-[D_{3d}] as the energy zero point.

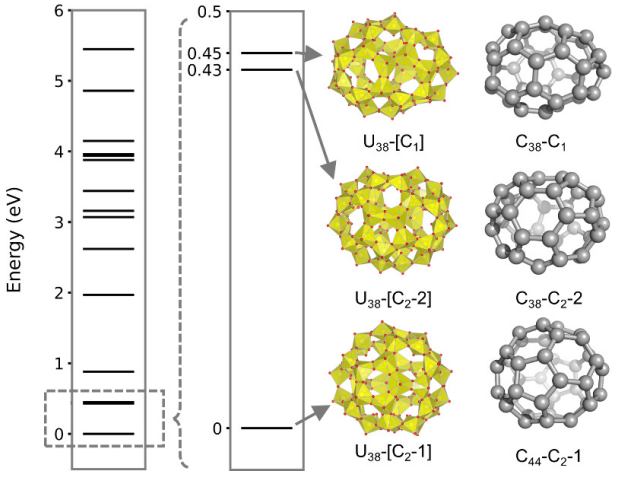


Figure 4 Application of the Uni-Gen model in U_{38}. Energy spectra of generated uranyl clusters with 38 vertices, where the energies and structures of the 3 most stable generated uranyl clusters with 38 vertices along with the structures of their corresponding fullerene backbones are shown in detail, using the energy of the most stable generated conformation U_{38}-[C_2-1] as the energy zero point.

Additionally, all these isomers generated by Uni-Gen successfully converged during CP2K optimization, further demonstrating the viability of using fullerene cluster frameworks as structural templates for uranyl peroxide cluster design. These results reinforce the reliability of the Uni-Gen model in predicting and analyzing novel uranyl peroxide cluster structures.

4 Conclusions

In conclusion, this work presents a novel approach for the theoretical design of uranyl peroxide clusters. It provides an efficient, precise, and scalable method for generating and optimizing complex molecular structures, and avoiding over-fitting. Facing the inherent complexity of modeling large uranyl clusters and the challenges posed by the system's high dimensionality, we utilized advanced deep learning techniques, including self-supervised denoising, two-stage pre-training and fine-tuning processes, as well as the SE(3)-equivariant transformer framework, leveraging the topological similarities between uranyl and carbon clusters.

Its ability to predict both known and previously unreported uranyl cluster isomers, as evidenced by the successful generation of clusters including U_{28}, U_{44}, and U_{38}, demonstrates a high degree of accuracy, closely aligning with experimentally observed clusters and accurately predicting new ones. Such predictive capabilities not only validate the effectiveness of the Uni-Gen model but also lay a solid foundation for its potential applications. These remarkable prediction results indicate that with further refinement, the Uni-Gen model could be crucial for designing uranyl-based materials, make significant contributions to the corresponding diversity, and enhance understanding of complex uranyl systems for nuclear chemistry and other applications.

Electronic Supplementary Material: Supplementary material (detailed composition of training dataset, detailed evaluation of model, and all optimized structures) is available in the online version of this article at https://doi.org/10.26599/NR.2025.94908047.

Data availability

All the data to train the model and the structures generated in this article are available at Zenodo (DOI: 10.5281/zenodo.15241895). All data needed to support the conclusions in the paper are presented in the manuscript and/or the Electronic Supplementary Material. Additional data related to this paper may be requested from the corresponding author upon request.

Acknowledgements

This work was supported by the National Key Research and Development Project (No. 2022YFA1503900) and the National Natural Science Foundation of China (Nos. 22222605 and 22076095).

Declaration of competing interest

All the contributing authors report no conflicts of interest in this work.

Author contribution statement

H.-S. H.: Conceptualization (lead); supervision (lead); writing – review & editing (lead). B. L.: Methodology (lead); investigation (lead); data curation (lead); literature review (lead); writing – original draft (lead). J. O. Y.: Data curation (supporting); visualization (supporting). Y. H. Z.: Literature review (lead); investigation (supporting); data curation (supporting). X. C. X.: Resources (supporting). All authors contributed to the general discussion of the results and approved the final manuscript.

Use of AI statement

None.

References

[1] Kubatko, K. A. H.; Helean, K. B.; Navrotsky, A.; Burns, P. C. Stability of peroxide-containing uranyl minerals. *Science* **2003**, *302*, 1191–1193.

[2] Burns, P. C.; Nyman, M. Captivation with encapsulation: A dozen years of exploring uranyl peroxide capsules. *Dalton Trans.* **2018**, *47*, 5916–5927.

[3] Zhao, X. K.; Cao, C. S.; Liu, J. C.; Lu, J. B.; Li, J.; Hu, H. S. Theoretical prediction of a graphene-like 2D uranyl material with p-orbital antiferromagnetism. *Chem. Sci.* **2022**, *13*, 8518–8525.

[4] Ostanin, S.; Zeller, P. *Ab initio* study of uranyl peroxides: Electronic factors behind the phase stability. *Phys. Rev. B* **2007**, *75*, 073101.

[5] Sigmon, G. E.; Weaver, B.; Kubatko, K. A.; Burns, P. C. Crown and bowl-shaped clusters of uranyl polyhedra. *Inorg. Chem.* **2009**, *48*, 10907–10909.

[6] Kubatko, K. A.; Burns, P. C. The crystal chemistry of uranyl peroxide nanospheres. *Geochim. Cosmochim. Acta* **2005**, *69*, A477.

[7] McGrail, B. T.; Sigmon, G. E.; Jouffret, L. J.; Andrews, C. R.; Burns, P. C. Raman spectroscopic and ESI-MS characterization of uranyl peroxide cage clusters. *Inorg. Chem.* **2014**, *53*, 1562–1569.

[8] Qiu, J.; Ling, J.; Jouffret, L.; Thomas, R.; Szymanowski, J. E. S.; Burns, P. C. Water-soluble multi-cage super tetrahedral uranyl peroxide phosphate clusters. *Chem. Sci.* **2014**, *5*, 303–310.

[9] Yang, G. P.; Li, K.; Hu, C. W. Recent advances in uranium-containing polyoxometalates. *Inorg. Chem. Front.* **2022**, *9*, 5408–5433.

[10] Miro, P.; Bo, C. Uranyl-peroxide nanocapsules: Electronic structure and cation complexation in $[(UO_2)_{20}(\mu\text{-}O_2)_{30}]^{20-}$. *Inorg. Chem.* **2012**,

51, 3840–3845.

[11] Sigmon, G. E.; Ling, J.; Unruh, D. K.; Moore-Shay, L.; Ward, M.; Weaver, B.; Burns, P. C. Uranyl-peroxide interactions favor nanocluster self-assembly. *J. Am. Chem. Soc.* **2009**, *131*, 16648–16649.

[12] Hickam, S.; Aksenov, S. M.; Dembowski, M.; Perry, S. N.; Traustason, H.; Russell, M.; Burns, P. C. Complexity of uranyl peroxide cluster speciation from alkali-directed oxidative dissolution of uranium dioxide. *Inorg. Chem.* **2018**, *57*, 9296–9305.

[13] Unruh, D. K.; Burtner, A.; Pressprich, L.; Sigmon, G. E.; Burns, P. C. Uranyl peroxide closed clusters containing topological squares. *Dalton Trans.* **2010**, *39*, 5807–5813.

[14] Sigmon, G. E.; Unruh, D. K.; Ling, J.; Weaver, B.; Ward, M.; Pressprich, L.; Simonetti, A.; Burns, P. C. Symmetry versus minimal pentagonal adjacencies in uranium-based polyoxometalate fullerene topologies. *Angew. Chem., Int. Ed.* **2009**, *48*, 2737–2740.

[15] Forbes, T. Z.; McAlpin, J. G.; Murphy, R.; Burns, P. C. Metal-oxygen isopolyhedra assembled into fullerene topologies. *Angew. Chem., Int. Ed.* **2008**, *47*, 2824–2827.

[16] Soltis, J. A.; Wallace, C. M.; Penn, R. L.; Burns, P. C. Cation-dependent hierarchical assembly of U60 nanoclusters into macro-ion assemblies imaged via cryogenic transmission electron microscopy. *J. Am. Chem. Soc.* **2016**, *138*, 191–198.

[17] Masci, B.; Thuéry, P. Uranyl complexes with the pyridine-2,6-dicarboxylato ligand: New dinuclear species with μ-η^2,η^2-peroxide, μ_2-hydroxide or μ_2-methoxide bridges. *Polyhedron* **2005**, *24*, 229–237.

[18] Kubatko, K. A.; Forbes, T. Z.; Klingensmith, A. L.; Burns, P. C. Expanding the crystal chemistry of uranyl peroxides: Synthesis and structures of Di- and triperoxodioxouranium(VI) complexes. *Inorg. Chem.* **2007**, *46*, 3657–3662.

[19] Qiu, J.; Burns, P. C. Clusters of actinides with oxide, peroxide, or hydroxide bridges. *Chem. Rev.* **2013**, *113*, 1097–1120.

[20] Kubatko, K. A.; Burns, P. C. Expanding the crystal chemistry of actinyl peroxides: Open sheets of uranyl polyhedra in $Na_5[(UO_2)_3(O_2)_4(OH)_3](H_2O)_{13}$. *Inorg. Chem.* **2006**, *45*, 6096–6098.

[21] Vlaisavljevich, B.; Gagliardi, L.; Burns, P. C. Understanding the structure and formation of uranyl peroxide nanoclusters by quantum chemical calculations. *J. Am. Chem. Soc.* **2010**, *132*, 14503–14508.

[22] Tiferet, E.; Gil, A.; Bo, C.; Shvareva, T. Y.; Nyman, M.; Navrotsky, A. The energy landscape of uranyl- peroxide species. *Chem. Eur. J.* **2014**, *20*, 3646–3651.

[23] Armstrong, C. R.; Nyman, M.; Shvareva, T.; Sigmon, G. E.; Burns, P. C.; Navrotsky, A. Uranyl peroxide enhanced nuclear fuel corrosion in seawater. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 1874–1877.

[24] Nyman, M.; Rodriguez, M. A.; Campana, C. F. Self-assembly of alkali-uranyl-peroxide clusters. *Inorg. Chem.* **2010**, *49*, 7748–7755.

[25] Kim, K. W.; Hyun, J. T.; Lee, K. Y.; Lee, E. H.; Lee, K. W.; Song, K. C.; Moon, J. K. Effects of the different conditions of uranyl and hydrogen peroxide solutions on the behavior of the uranium peroxide precipitation. *J. Hazard. Mater.* **2011**, *193*, 52–58.

[26] Mallon, C.; Walshe, A.; Forster, R. J.; Keyes, T. E.; Baker, R. J. Physical characterization and reactivity of the uranyl peroxide $[UO_2(\eta^2\text{-}O_2)(H_2O)_2]\bullet 2H_2O$: Implications for storage of spent nuclear fuels. *Inorg. Chem.* **2012**, *51*, 8509–8515.

[27] Wylie, E. M.; Peruski, K. M.; Weidman, J. L.; Phillip, W. A.; Burns, P. C. Ultrafiltration of uranyl peroxide nanoclusters for the separation of uranium from aqueous solution. *ACS Appl. Mater. Interfaces* **2014**, *6*, 473–479.

[28] Blanchard, F.; Ellart, M.; Rivenet, M.; Vigier, N.; Hablot, I.; Morel, B.; Grandjean, S.; Abraham, F. Role of ammonium ions in the formation of ammonium uranyl peroxides and uranyl peroxo-oxalates. *Cryst. Growth Des.* **2016**, *16*, 200–209.

[29] Gao, Y. Y.; Haso, F.; Szymanowski, J. E. S.; Zhou, J.; Hu, L.; Burns, P. C.; Liu, T. B. Selective permeability of uranyl peroxide nanocages to different alkali ions: Influences from surface pores and hydration shells. *Chem. Eur. J.* **2015**, *21*, 18785–18790.

[30] Nyman, M.; Alam, T. M. Dynamics of uranyl peroxide nanocapsules. *J. Am. Chem. Soc.* **2012**, *134*, 20131–20138.

[31] Miró, P.; Pierrefixe, S.; Gicquel, M.; Gil, A.; Bo, C. On the origin of the cation templated self-assembly of uranyl-peroxide nanoclusters. *J. Am. Chem. Soc.* **2010**, *132*, 17787–17794.

[32] Zhang, H. X.; Balasubramanian, K. Spectral moments of fullerene cages. *Mol. Phys.* **1993**, *79*, 727–745.

[33] Albertazzi, E.; Domene, C.; Fowler, P. W.; Heine, T.; Seifert, G.; Van Alsenoy, C.; Zerbetto, F. Pentagon adjacency as a determinant of fullerene stability. *Phys. Chem. Chem. Phys.* **1999**, *1*, 2913–2918.

[34] Kerim, A. A systematic investigation on the aromaticity and kinetic stability of fullerene C_{40} isomers. *J. Phys. Org. Chem.* **2012**, *25*, 379–384.

[35] Witek, H. A.; Irle, S. Diversity in electronic structure and vibrational properties of fullerene isomers correlates with cage curvature. *Carbon* **2016**, *100*, 484–491.

[36] Karasulu, B.; Leyssale, J. M.; Rowe, P.; Weber, C.; de Tomas, C. Accelerating the prediction of large carbon clusters via structure search: Evaluation of machine-learning and classical potentials. *Carbon* **2022**, *191*, 255–266.

[37] Xing, C. Y.; Chen, G. Y.; Zhu, X.; An, J. K.; Bao, J. C.; Wang, X.; Zhou, X. Q.; Du, X. L.; Xu, X. X. Synthesis of carbon dots with predictable photoluminescence by the aid of machine learning. *Nano Res.* **2024**, *17*, 1984–1989.

[38] Oglic, D.; Oatley, S. A.; Macdonald, S. J. F.; McInally, T.; Garnett, R.; Hirst, J. D.; Gärtner, T. Active search for computer-aided drug design. *Mol. Inform.* **2018**, *37*, 1700130.

[39] Ragoza, M.; Masuda, T.; Koes, D. R. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chem. Sci.* **2022**, *13*, 2701–2713.

[40] Schneuing, A.; Harris, C.; Du, Y. Q.; Didi, K.; Jamasb, A.; Igashov, I.; Du, W. T.; Gomes, C.; Blundell, T. L.; Lio, P. et al. Structure-based drug design with equivariant diffusion models. *Nat. Comput. Sci.* **2024**, *4*, 899–909.

[41] Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX - Evolutionary crystal structure prediction. *Comput. Phys. Commun.* **2006**, *175*, 713–720.

[42] Wang, Y. C.; Lv, J.; Zhu, L.; Ma, Y. M. CALYPSO: A method for crystal structure prediction. *Comput. Phys. Commun.* **2012**, *183*, 2063–2070.

[43] Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.

[44] Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80–85.

[45] Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Wang, Z. L.; Shysheya, A.; Crabbé, J.; Ueda, S. et al. A generative model for inorganic materials design. *Nature* **2025**, *639*, 624–632.

[46] Sriram, A.; Miller, B. K.; Chen, R. T. Q.; Wood, B. M. FlowLLM: Flow matching for material generation with large language models as base distributions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 2024, pp. 1464.

[47] Zhong, M.; Tran, K.; Min, Y. M.; Wang, C. H.; Wang, Z. Y.; Dinh, C. T.; De Luna, P.; Yu, Z. Q.; Rasouli, A. S.; Brodersen, P. et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. *Nature* **2020**, *581*, 178–183.

[48] Mok, D. H.; Back, S. Generative pretrained transformer for heterogeneous catalysts. *J. Am. Chem. Soc.* **2024**, *146*, 33712–33722.

[49] Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

[50] Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.;

Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089–1100.

[51] Wu, K. E.; Yang, K. K.; van den Berg, R.; Alamdari, S.; Zou, J. Y.; Lu, A. X.; Amini, A. P. Protein structure generation via folding diffusion. *Nat. Commun.* **2024**, *15*, 1059.

[52] Kouchehbaghi, N. H.; Yousefzadeh, M.; Gharehaghaji, A.; Khosravi, S.; Khorsandi, D.; Haghniaz, R.; Cao, K.; Dokmeci, M. R.; Rostami, M.; Khademhosseini, A. et al. A machine learning-guided design and manufacturing of wearable nanofibrous acoustic energy harvesters. *Nano Res.* **2024**, *17*, 9181–9192.

[53] Cho, Y. J.; Chaney, H.; Lu, K. Machine learning approach for predicting optical and photothermal properties of gold nanoparticle/polymer hybrid films: Effect of synthetic data. *Nano Res.* **2025**, *18*, 94907216.

[54] Hawkins, P. C. D. Conformation generation: The state of the art. *J. Chem. Inf. Model.* **2017**, *57*, 1747–1756.

[55] Wang, M. L.; Hu, X. Q.; Beratan, D. N.; Yang, W. T. Designing molecules by optimizing potentials. *J. Am. Chem. Soc.* **2006**, *128*, 3228–3232.

[56] Balamurugan, D.; Yang, W. T.; Beratan, D. N. Exploring chemical space with discrete, gradient, and hybrid optimization methods. *J. Chem. Phys.* **2008**, *129*, 174105.

[57] Zhao, Y. F.; Chen, X.; Li, J. TGMin: A global-minimum structure search program based on a constrained basin-hopping algorithm. *Nano Res.* **2017**, *10*, 3407–3420.

[58] Zhang, J.; Dolg, M. ABCluster: The artificial bee colony algorithm for cluster global optimization. *Phys. Chem. Chem. Phys.* **2015**, *17*, 24173–24181.

[59] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

[60] Mansimov, E.; Mahmood, O.; Kang, S.; Cho, K. Molecular geometry prediction using a deep generative graph neural network. *Sci. Rep.* **2019**, *9*, 20381.

[61] Ganea, O. E.; Pattanaik, L.; Coley, C. W.; Barzilay, R.; Jensen, K. F.; Green, W. H.; Jaakkola, T. S. GEOMOL: Torsional geometric generation of molecular 3D conformer ensembles. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021, pp. 1054.

[62] Zhu, J. H.; Xia, Y. C.; Liu, C.; Wu, L. J.; Xie, S. F.; Wang, Y. S.; Wang, T.; Qin, T.; Zhou, W. G.; Li, H. Q. et al. Direct molecular conformation generation. *Trans. Mach. Learn. Res.*, 2022.

[63] Zhou, G. M.; Gao, Z. F.; Ding, Q. K.; Zheng, H.; Xu, H. T.; Wei, Z. W.; Zhang, L. F.; Ke, G. L. Uni-Mol: A universal 3D molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[64] Küehne, T. D.; Iannuzzi, M.; Del Ben, M.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F. et al. CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.* **2020**, *152*, 194103.

[65] Kim, J.; Nguyen, T. D.; Min, S.; Cho, S.; Lee, M.; Lee, H.; Hong, S. Pure transformers are powerful graph learners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2022, pp. 1060.

[66] Wang, J. Q.; Liu, J. P.; Wang, H. S.; Zhou, M. S.; Ke, G. L.; Zhang, L. F.; Wu, J. Z.; Gao, Z. F.; Lu, D. N. A comprehensive transformer-based approach for high-accuracy gas adsorption predictions in metal-organic frameworks. *Nat. Commun.* **2024**, *15*, 1904.

[67] Xu, F. J.; Guo, W. T.; Wang, F.; Yao, L.; Wang, H. S.; Tang, F. J.; Gao, Z. F.; Zhang, L. F.; E, W. N.; Tian, Z. Q. et al. Toward a unified benchmark and framework for deep learning-based prediction of nuclear magnetic resonance chemical shifts. *Nat. Comput. Sci.* **2025**, *5*, 292–300.

[68] Hara, K.; Saito, D.; Shouno, H. Analysis of function of rectified linear unit used in deep learning. In *2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, 2015, pp. 1–8.

[69] Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[70] van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

[71] Ling, J.; Ozga, M.; Stoffer, M.; Burns, P. C. Uranyl peroxide pyrophosphate cage clusters with oxalate and nitrate bridges. *Dalton Trans.* **2012**, *41*, 7278–7284.