A FOUNDATION MODEL FOR SIMULATION-GRADE MOLECULAR ELECTRON DENSITIES

Eduardo Soares, Victor Shirasuna, Emilio Vital Brazil, Breno Carvalho, Caio Gama & Daniel Briquez
IBM Research - Brazil
{eduardo.soares}@ibm.com
{evital}@br.ibm.com

Dmitry Zubarev, Brandi Ransom, Holt Bui, & Krystelle Lionti
IBM Research - Almaden Lab {dmitry.zubarev}@ibm.com

ABSTRACT

This paper introduces 3DGrid-VQGAN, a generative framework for the representation and reconstruction of molecular electronic structures as electron charge densities on 3D grid produced by quantum chemical simulations. The model efficiently encodes high-dimensional data into compact latent representations, enabling downstream tasks such as molecular property prediction with enhanced accuracy. Evaluation on the QM9 dataset, which contains quantum chemical properties computed at the density functional theory (DFT) level, demonstrates the model's ability to capture essential features of the electronic structure. The reconstructed charge densities achieve high fidelity, preserving critical details such as electron density cusps at nuclear positions. This is quantified using metrics such as Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Fréchet Inception Distance (FID). These metrics, alongside visual and numerical analyses, demonstrate the model's robustness across diverse molecular structures, including complex geometries and chemical environments. The results suggest that generative approaches like 3DGrid-VQGAN could significantly reduce reliance on computationally intensive quantum chemical simulations, offering simulation-grade data derived directly from learned representations. Future work will focus on extending the model to larger and more complex molecular systems, improving interpretability of latent representations, and integrating the framework into workflows for molecular property prediction and generative design.

1 INTRODUCTION

Generative modeling, particularly in multimodal domains, has recently demonstrated substantial advancements in tackling complex problems across various disciplines (Anstine & Isayev, 2023; Ingraham et al., 2023; Sankaran & Holmes, 2023). The ability to integrate text and visual modalities has progressed to a level where physical phenomena can be accurately represented (Liang et al., 2024). For instance, GameNGen, a recent model built on stable diffusion, enables real-time, interactive simulations of first-person shooter games (Valevski et al., 2024). Similar to video game simulations, molecular simulations aim to construct accurate and realistic representations of complex, dynamic environments. This raises a critical question: can generative AI produce simulation-grade molecular data without relying on computationally expensive simulations?

Such a capability would be fundamentally different from the popular track of learning neural potentials that still implies running simulations to obtain results informed by physics. It would establish new theoretical frameworks for the computational treatment of molecular systems, significantly reducing the dependence on specialized software and hardware, such as quantum chemical simulation codes on high-performance computing infrastructures (Liu et al., 2024; Shang et al., 2023). Moreover, the envisioned generative AI capability would be fully compatible with existing computational chemistry approaches, enabling direct validation of results or seamless progression to simulations using higher

levels of theory. This paradigm shift would also lower costs and reduce barriers to entry, alleviating the need for specialized expertise in computational chemistry (Gangwal et al., 2024; Ananikov, 2024).

We present preliminary results of a generative approach capable of producing simulation-grade molecular electronic structures at the level of Hartree-Fock (HF) approximation (Lykos & Pratt, 1963) and density functional theory (DFT) (Hohenberg & Kohn, 1964). These simulations, which involve solving the electronic Schrödinger equation (Hermann et al., 2020), yield the electron density — a probabilistic representation of electron distribution in three-dimensional space (March, 1982). The electron density is the central and most information-rich artifact of quantum chemical simulations, making it an ideal candidate for generative modeling. The electron density encodes detailed molecular information enabling reconstruction of the atomic structure by identifying charge density cusps at nuclear positions, partitioning the electron density into atomic volumes and reconstructing atomic charges and types, using methods such as the atoms-in-molecules formalism (Bader et al., 1987; Alibakhshi & Schäfer, 2024). Moreover, per holographic electron density theorem (Mezey, 1999) a nonzero segment of the ground-state electron density uniquely determines the full molecular electron density thus opening a possibility of generative design of molecular electronic structure constrained by a fractional segment of the electron charge density. Our work demonstrates the potential of generative AI to bypass traditional simulation-based computational pipelines, paving the way for efficient and accessible simulation-grade molecular modeling.

2 Overview of the Proposed Approach

This section provides an overview of the 3DGrid-VQGAN foundation model developed for representing 3D electron density grids. The methodology begins with a description of the processes involved in collecting, curating, and pre-processing the data used for pre-training. This is followed by an explanation of the encoder-decoder architecture employed for learning compact and expressive representations of the 3D data. Key aspects of the data pipeline and model architecture are outlined to contextualize the approach within the broader framework of 3D representation learning.

2.1 Pre-training Data

The dataset for VQGAN included approx. 855K molecules from PubChem database, such that the molecules in the constructed dataset contained only main-group elements up to Barium, the number of heavy atoms did not exceed 30, the molecules were charge-neutral and did not have charge separation. For each molecule represented as a SMILES string we generated and optimized 50 conformations using distance geometry algorithm and molecular force field as implemented in rdkit Landrum et al. (2024). We reoptimized 5 lowest-energy conformations using MINDO3 semi-empirical level of theory as implemented in pyscf electronic structure library Sun et al. (2020). The conformation with the lowest energy at MINDO3 level of theory went into restricted Hartree-Fock evaluation with the minimal basis set STO-3G to produce ab initio electron density and output it in volumetric grid format, aka voxeled charge density.

2.2 Model Architecture

To transform 3D electron density grids into a meaningful latent representation, vector quantized autoencoders have proven to be a viable solution, particularly for addressing the issue of blurry outputs commonly encountered in variational autoencoders (Van Den Oord et al., 2017; Razavi et al., 2019). These models achieve this by mapping the latent feature vectors at the bottleneck of the autoencoder to a quantized representation drawn from a learned codebook (Esser et al., 2021). The VQ-GAN architecture, a class of vector quantized autoencoders, further enhances image reconstruction quality by incorporating a discriminator loss at its output (Yu et al., 2021).

In this framework, 3D electron density grids are passed through the encoder to generate a latent code $z_e \in \mathbb{R}^{(\frac{H}{s})\times(\frac{W}{s})\times k}$, where H represents the height, W represents the width, C is the number of channels, k is the number of latent feature maps, and s is a compression factor. During the vector quantization step, these latent feature vectors are quantized by substituting each vector with its closest match from the learned codebook Z (see Fig. 1). The image is then reconstructed by feeding the quantized feature vectors into the decoder G. The learning objective is to minimize a combination

of three losses: the reconstruction loss L_{rec}, the codebook loss L_{codebook}, and the commitment loss L_{commit}. To extend this architecture to support 3D inputs, we follow the (Ge et al., 2022; Khader et al., 2022) and replace the 2D convolutions by 3D convolutions.

2.3 PRE-TRAINING STRATEGIES

The training of the 3DGrid-VQGAN model was conducted using a batch size of 1, necessitated by the large size of the input 3D grids. The training process was distributed across 600 NVIDIA A100 GPUs, running for a total of 43 epochs and consuming approximately 5000 GPU node hours. The model pre-training was performed on the Polaris cluster at the Argonne National Laboratory using discretionary allocation of the INCITE program.

The input to the model consists of 3D electron density grids with a resolution of $128 \times 128 \times 128$. During the vector quantization step, the latent features generated by the encoder are mapped to a codebook with 16,384 entries. Each codebook entry (or prototype vector) represents a distinct cluster or region within the latent feature space. This quantization step enforces the alignment of latent features with predefined discrete prototypes, replacing the continuous representation of the latent space with a structured and interpretable discrete representation. To manage the complexity of the data and enhance the efficiency of representation learning, an internal downsampling scheme with a factor of $[4, 4, 4]$ is employed. This approach reduces the spatial dimensions of the encoded representation, preserving essential features while lowering computational costs. The model is optimized using a learning rate of 3×10^{-4}, facilitating convergence over the training period.

2.4 FINE-TUNING STRATEGIES

To further enhance the 3D energy grid VQ-GAN model on downstream tasks, we conducted a fine-tuning strategy. The model architecture includes a feature extractor and feature predictor. The feature extractor reuses the pre-trained encoder weights from the 3DGrid-VQGAN model (see Fig 2). This module receives the 3D electron density grids to obtain the grid embeddings of size $32 \times 32 \times 32 \times 256$, which contains the learned representations of the input. Then, we used a mean pooling to obtain an embedding representing a unique molecule, resulting in an embedding size of 2048. The feature predictor is a fully connected network with 2 hidden layers and Dropout (dropout=0.1) to prevent overfitting. This module receives the molecular embeddings from the encoder and estimates the downstream tasks. The fine-tuning experiments were conducted using a NVIDIA A100 GPU with a batch size of 8, since only the weights of feature extractor and predictor are updated. To achieve a better converge through the fine-tuning, the model was optimized using a learning rate of 3×10^{-5} for the first 30 epochs and a learning rate of 1×10^{-5} for the remaining epochs.

3 EXPERIMENTS

To evaluate the effectiveness of the latent space representations learned by our 3DGrid-VQGAN model, we conducted a series of experiments using the QM9 dataset, obtained through MoleculeNet (Wu et al., 2018), as outlined in Table 4. QM9 is a benchmark dataset in computational chemistry, providing quantum chemical properties computed at the density functional theory (DFT) level using the B3LYP/6-31G(2df,p) functional. The dataset consists of approximately 134,000 small organic molecules, each containing up to nine heavy atoms selected from a predefined set of elements (H, C, O, N, F). To ensure an unbiased assessment, we adopted identical train/validation/test splits for all tasks as in Wu et al. (2018).

To evaluate the reconstruction capability of the 3DGrid-VQGAN model, a subset of 4,141 samples was randomly selected from the 854,919 samples used during pre-training. This sample size corresponds to a confidence level of 99% and a margin of error of 2%, ensuring statistical validity for the analysis. The selection process aimed to provide a representative distribution of the dataset without introducing sampling bias.

4 RESULTS AND DISCUSSION

The performance of the 3DGrid-VQGAN model for predicting quantum chemical properties is analyzed using the QM9 benchmark dataset. Table 1 presents a comparative evaluation of the mean absolute error (MAE) achieved by the 3DGrid-VQGAN model and several state-of-the-art methods, as reported in prior studies. The evaluation adheres to the protocol established in the original MoleculeNet work by Wu et al. (2018), where the average MAE is used as metric for this task.

Table 1: Performance for quantum chemical properties prediction.

Method	**QM9 Dataset**
D-MPNN Yang et al. (2019)	3.241
GC Altae-Tran et al. (2017)	4.356
A-FP Xiong et al. (2019)	4.353
MPNN Gilmer et al. (2017)	3.189
N-Gram Liu et al. (2019)	2.510
MolCLR$_{\text{GIN}}$ Wang et al. (2022)	2.357
Hu et al. Hu et al. (2020)	4.349
GEM Fang et al. (2022)	2.970
Uni-Mol Zhou et al.	1.830
MoLFormer-XL Ross et al. (2022)	1.589
SELFIES-TED$_{289M}$ Priyadarsini et al. (2024)	4.263
SMI-TED$_{289M}$ Soares et al. (2024)	1.324
3DGrid-VQGAN	**1.217**

The QM9 dataset provides a diverse set of molecular quantum chemical properties derived from density functional theory calculations, offering a testing ground for evaluating machine learning methods. The comparison includes models employing graph-based neural architectures, such as D-MPNN (Yang et al., 2019) and MolCLRGIN (Wang et al., 2022), as well as transformer-based architectures, including MoLFormer-XL (Ross et al., 2022) and SMI-TED$289M$ (Soares et al., 2024). These methods represent different approaches to modeling molecular systems, ranging from graph convolutional frameworks to self-supervised pre-training strategies.

The results in Table 1 indicate that the 3DGrid-VQGAN model achieves the lowest MAE of 1.2197, outperforming all competing methods. For comparison, SMI-TED$_{289M}$, a fine-tuned large-scale transformer model, achieves an MAE of 1.3246, while MoLFormer-XL achieves an MAE of 1.5894. The performance of 3DGrid-VQGAN can be attributed to its ability to encode the 3D electron density grids directly, preserving critical spatial and electronic information that underpins the molecular properties.

The performance of 3DGrid-VQGAN suggests that generative models operating on 3D data representations offer distinct advantages over methods relying on 2D graph-based or SMILES representations. By incorporating spatial and electronic context into the latent representations, the model captures essential features of the molecular structure and dynamics that are often lost in lower-dimensional representations.

3D ELECTRON DENSITY RECONSTRUCTION EVALUATION

To evaluate the reconstruction capability of the 3DGrid-VQGAN model, a subset of 4,141 samples was randomly selected from the 854,919 samples used during pre-training. This subset size was determined to achieve a confidence level of 99% with a margin of error of 2%, ensuring statistical robustness and a representative evaluation of the model's reconstruction performance. The random sampling process was designed to preserve the distributional characteristics of the dataset, minimizing sampling bias and providing reliable insights into the model's behavior across the broader data distribution.

The results summarized in Table 2 demonstrate the model's progressive improvement in reconstruction quality as training epochs advance. At early stages (e.g., Epoch 1), the reconstructions show relatively low SSIM (0.049297) and high MAE (0.020376), reflecting the model's initial learning phase where latent space representations are not yet optimized. By Epoch 10, significant improvements are observed across all metrics, with SSIM increasing to 0.102990 and MAE dropping to 0.005006, indicating better alignment between the reconstructed and original grids.

Table 2: 3D electron density reconstruction evaluation.

Epoch	SSIM ↑	MS-SSIM ↑	PSNR (dB) ↑	FID ↓	KID ↓	LPIPS ↓	Reconstruction (MAE) ↓
VQGAN - Epoch 01	0.049297	0.718320	28.266564	2.111260	0.008810	0.224641	0.020376
VQGAN - Epoch 10	0.102990	0.781453	33.195561	0.308362	0.000661	0.174354	0.005006
VQGAN - Epoch 20	**0.480144**	**0.913850**	34.866451	0.141509	0.000155	0.169709	0.003550
VQGAN - Epoch 30	0.095842	0.778469	34.751817	0.195592	0.000335	0.138630	0.003838
VQGAN - Epoch 40	0.350813	0.902409	34.747153	0.185152	0.000363	**0.127490**	0.003994
VQGAN - Epoch 43	0.055365	0.724635	**35.515505**	**0.099327**	**0.000027**	0.134057	**0.003300**

The steady improvement in performance metrics indicates that the 3DGrid-VQGAN model effectively learns a compact and meaningful latent representation of the 3D electron density grids. The ability to accurately reconstruct these grids is critical for preserving the spatial and electronic features essential for downstream applications, such as quantum chemical property prediction or generative tasks. The diverse set of evaluation metrics provides a comprehensive assessment, ensuring that the model's capabilities are validated from both numerical and perceptual perspectives.

The reconstructed grids demonstrate a high degree of structural fidelity, with spatial distributions and key features, such as the electron density cusps at nuclear positions, closely matching those in the original grids (see Fig. 3). These results indicate that the model effectively encodes and reconstructs high-dimensional electron density data, maintaining essential properties for tasks requiring accurate quantum chemical representations. The preservation of these features is particularly significant as they carry critical information about atomic positions and molecular geometry.

The selected molecules include a range of chemical configurations, such as functional groups, unsaturated bonds, and cyclic structures, providing evidence of the model's generalizability across diverse molecular environments. The consistency of reconstruction quality across these examples suggests that the 3DGrid-VQGAN latent space effectively captures the inherent variability in the dataset. However, subtle deviations in grid intensities and spatial details highlight potential challenges, particularly for molecules with complex electronic delocalization or pronounced spatial anisotropy. These challenges warrant further investigation to ensure robust performance across all regions of chemical space.

5 Conclusion

This paper presents the 3DGrid-VQGAN model, a generative approach designed to encode and reconstruct high-dimensional 3D electron density grids with high fidelity. Through a series of evaluations, the model demonstrates its ability to capture and preserve critical spatial and electronic features, as evidenced by the accurate reconstruction of diverse molecular configurations from the QM9 dataset. Quantitative metrics and visual analyses highlight the effectiveness of the latent space learned by the model in representing complex molecular structures.

The results indicate that 3DGrid-VQGAN outperforms several state-of-the-art methods in tasks such as quantum chemical property prediction and 3D electron density reconstruction. By leveraging the intrinsic spatial and electronic properties encoded in the electron density grids, the model provides a robust framework for applications in molecular modeling, property prediction, and generative tasks in quantum chemistry. Additionally, the ability to reconstruct simulation-grade molecular data directly from the learned representations underscores the potential for reducing computational reliance on traditional quantum chemical methods.

Despite these findings, challenges remain, particularly in reconstructing highly complex molecular systems or handling datasets with larger molecular geometries. Future work should focus on expanding the evaluation to include broader datasets, refining latent space representations to improve interpretability, and exploring the integration of the model into practical workflows for predictive and generative applications.

References

Amin Alibakhshi and Lars V Schäfer. Electron iso-density surfaces provide a thermodynamically consistent representation of atomic and molecular surfaces. *Nature Communications*, 15(1):6086,

2024.

Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.

Valentine P Ananikov. Top 20 influential ai-based technologies in chemistry. *Artificial Intelligence Chemistry*, pp. 100075, 2024.

Dylan M Anstine and Olexandr Isayev. Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society*, 145(16):8736–8750, 2023.

Richard FW Bader, Marshall T Carroll, James R Cheeseman, and Cheng Chang. Properties of atoms in molecules: atomic volumes. *Journal of the American Chemical Society*, 109(26):7968–7979, 1987.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Amit Gangwal, Azim Ansari, Iqrar Ahmad, Abul Kalam Azad, Vinoth Kumarasamy, Vetriselvan Subramaniyan, and Ling Shing Wong. Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities. *Frontiers in Pharmacology*, 15: 1331062, 2024.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pp. 102–118. Springer, 2022.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Jan Hermann, Zeno Schätzle, and Frank Noé. Deep-neural-network solution of the electronic schrödinger equation. *Nature Chemistry*, 12(10):891–897, 2020.

Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136(3B):B864, 1964.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1857–1867, 2020.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baessler, Sebastian Foersch, et al. Medical diffusion: denoising diffusion probabilistic models for 3d medical image generation. *arXiv preprint arXiv:2211.03364*, 2022.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, Peter Gedeck, Gareth Jones, NadineSchneider, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, tadhurst cdd, Vincent F. Scalfani, Rachel Walker, Kazuya Ujihara, Daniel Probst, Juuso Lehtivarjo, guillaume godin, Axel Pahl, François Bérenger, and Hussein Faara. rdkit/rdkit: 2024_09_4 (q3 2024) release, December 2024. URL `https://doi.org/10.5281/zenodo.14535873`.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10): 1–42, 2024.

Jie Liu, Huan Ma, Honghui Shang, Zhenyu Li, and Jinlong Yang. Quantum-centric high performance computing for quantum chemistry. *Physical Chemistry Chemical Physics*, 26(22):15831–15843, 2024.

Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32, 2019.

P Lykos and GW Pratt. Discussion on the hartree-fock approximation. *Reviews of Modern Physics*, 35(3):496, 1963.

Norman Henry March. Electron density theory of atoms and molecules. *The Journal of Physical Chemistry*, 86(12):2262–2267, 1982.

P B Mezey. Holographic electron density shape theorem and its role in drug design and toxicological risk assessment. *J Chem Inf Comput Sci.*, 39(2):224–30, 1999.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4602–4609, 2019.

Indra Priyadarsini, Seiji Takeda, Lisa Hamada, Emilio Vital Brazil, Eduardo Soares, and Hajime Shinohara. Self-bart: A transformer-based molecular representation model using selfies. *arXiv preprint arXiv:2410.12348*, 2024.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

Kris Sankaran and Susan P Holmes. Generative models: An interdisciplinary perspective. *Annual Review of Statistics and Its Application*, 10(1):325–352, 2023.

Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1): 13890, 2017.

Honghui Shang, Yi Fan, Li Shen, Chu Guo, Jie Liu, Xiaohui Duan, Fang Li, and Zhenyu Li. Towards practical and massively parallel quantum computing emulation for quantum chemistry. *npj Quantum Information*, 9(1):33, 2023.

Eduardo Soares, Victor Shirasuna, Emilio Vital Brazil, Renato Cerqueira, Dmitry Zubarev, and Kristin Schmidt. A large encoder-decoder family of foundation models for chemical language. *arXiv preprint arXiv:2407.20267*, 2024.

Qiming Sun, Xing Zhang, Samragni Banerjee, Peng Bao, Marc Barbry, Nick S. Blunt, Nikolay A. Bogdanov, George H. Booth, Jia Chen, Zhi-Hao Cui, Janus J. Eriksen, Yang Gao, Sheng Guo, Jan Hermann, Matthew R. Hermes, Kevin Koh, Peter Koval, Susi Lehtola, Zhendong Li, Junzi Liu, Narbe Mardirossian, James D. McClain, Mario Motta, Bastien Mussard, Hung Q. Pham, Artem Pulkin, Wirawan Purwanto, Paul J. Robinson, Enrico Ronca, Elvira R. Sayfutyarova, Maximilian Scheurer, Henry F. Schurkus, James E. T. Smith, Chong Sun, Shi-Ning Sun, Shiv Upadhyay, Lucas K. Wagner, Xiao Wang, Alec White, James Daniel Whitfield, Mark J. Williamson, Sebastian Wouters, Jun Yang, Jason M. Yu, Tianyu Zhu, Timothy C. Berkelbach, Sandeep Sharma,

Alexander Yu. Sokolov, and Garnet Kin-Lic Chan. Recent developments in the pyscf program package. *The Journal of Chemical Physics*, 153(2):024109, 07 2020. ISSN 0021-9606. doi: 10.1063/5.0006074. URL https://doi.org/10.1063/5.0006074.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8): 3370–3388, 2019.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*.

6 Supplementary Materials

6.1 3DGrid-VG-Gan general architecture

Fig. 1 illustrates the general architecture of the 3D energy grid VQ-GAN model.

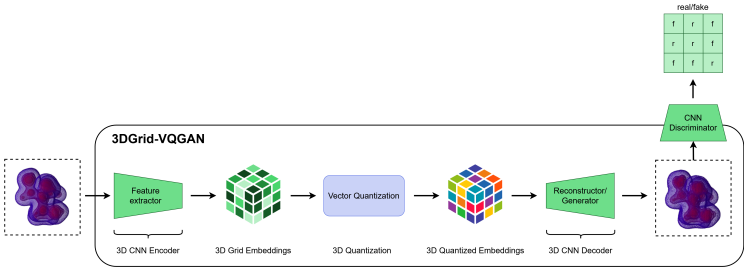

Figure 1: This figure illustrates the general architecture of the 3D electron density grids VQ-GAN model.

6.2 3DGrid-VG-Gan fine-tuning architecture

Fig. 2 illustrates the general architecture of fine-tuning the 3D energy grid VQ-GAN model.

Table 3 provides a detailed overview of the hyper-parameters considered for the fine-tuning of 3D energy grid VQ-GAN model.

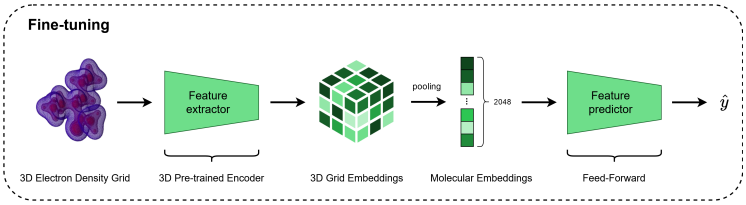

Figure 2: This figure illustrates the general architecture of fine-tuning the 3D electron density grids VQ-GAN model.

6.3 QM9 DETAILS

The QM9 dataset encompasses a range of quantum chemical properties, including dipole moments, polarizability, HOMO and LUMO energies, HOMO-LUMO gaps, and thermodynamic quantities such as internal energy, enthalpy, and heat capacity. These properties make QM9 a robust and comprehensive benchmark for evaluating machine learning models' ability to encode, reconstruct, and predict quantum mechanical features of molecular systems. By leveraging these diverse and information-rich properties, we aim to rigorously assess the fidelity and utility of the learned latent space in capturing the underlying quantum mechanical characteristics of molecular systems.

6.4 METRICS FOR RECONSTRUCTION EVALUATION

The reconstruction quality was assessed using a range of metrics, each providing a distinct perspective on performance (see Table 2):

- **Structural Similarity Index Measure (SSIM)** and **Multi-Scale SSIM (MS-SSIM)** quantify the structural fidelity of the reconstructed grids relative to the original, emphasizing perceptual and spatial coherence.
- **Peak Signal-to-Noise Ratio (PSNR)** measures the pixel-wise accuracy of the reconstructions, with higher values indicating closer agreement to the original grids.
- **Fréchet Inception Distance (FID)** and **Kernel Inception Distance (KID)** evaluate the similarity between the distributions of the reconstructed and original grids, capturing higher-order statistical consistency.
- **Learned Perceptual Image Patch Similarity (LPIPS)** assesses perceptual similarity, providing a complementary perspective on the quality of reconstructions.
- **Mean Absolute Error (MAE)** reflects the average absolute deviation between reconstructed and original grid values, offering a straightforward measure of numerical accuracy.

The analysis demonstrates the model's ability to reconstruct key features of the 3D electron density grids while preserving critical structural and numerical properties.

A deeper analysis over the QM9 benchmark: In this subsection, we present a detailed analysis of the results for the QM9 dataset. Table 5 summarizes the performance of state-of-the-art (SOTA) approaches across the molecular properties included in QM9. Our comparative analysis benchmarks the proposed encoder-decoder foundation model against SOTA models from four different materials representations: (i) Graph-based, (ii) Geometry-based, and (iii) SMILES-based approaches for

Table 3: 3DGrid-VQGAN fine-tuning architecture specificity.

Grid hidden size	256.	# batch	8
Molecular hidden size	2048	# epochs	100
Codebook size	16,384	# seeds	5
Downsample	[4, 4, 4]	# GPUs	1 NVIDIA A100 (80G)
Dropout	0.1	Total params	27M
Learning rate	3e-5 and 1e-5		

Table 4: Properties Included in the QM9 Dataset.

Property	Description
Dipole moment (μ)	Measure of the molecular polarity, representing the separation of positive and negative charges within the molecule.
Polarizability (α)	Ability of a molecule to be polarized in the presence of an external electric field.
HOMO energy (ϵ_{HOMO})	Energy of the highest occupied molecular orbital, indicative of a molecule's ability to donate electrons.
LUMO energy (ϵ_{LUMO})	Energy of the lowest unoccupied molecular orbital, indicative of a molecule's ability to accept electrons.
HOMO-LUMO gap ($\Delta\epsilon$)	Energy difference between the HOMO and LUMO orbitals, related to a molecule's electronic excitation energy.
Electronic spatial extent ($\langle R^2 \rangle$)	Spatial spread of the electron density, providing insights into molecular size and shape.
Zero-point vibrational energy (ZPVE)	Quantum mechanical energy of a molecule at 0 K due to vibrational motion.
Internal energy at 0 K (U_0)	Total electronic energy of the molecule at 0 K.
Internal energy at 298.15 K (U)	Total energy of the molecule at standard temperature (298.15 K).
Enthalpy (H)	Total energy of the molecule, including internal energy and the energy required to displace the environment.
Free energy (G)	Energy available to perform work at constant temperature and pressure.
Heat capacity (C_v)	Amount of heat energy required to raise the temperature of the molecule by one degree under constant volume.

molecular property prediction. The baseline models considered in this comparison include 123-gnn (Morris et al., 2019), a multitask neural network utilizing the Coulomb Matrix (CM) (Rupp et al., 2012), its graph neural network (GNN) extension, the deep tensor neural network (DTNN) (Schütt et al., 2017), MoLFormer-XL (Ross et al., 2022), SELFIES-TED (Priyadarsini et al., 2024), and SMI-TED$_{289M}$ (Soares et al., 2024).

Table 5: Comparing state-of-the-art models performance over the QM9 dataset.

	Graph-based			**Geometry-based**			**SMILES**		**SELFIES**	**3D grids**
Measure	A-FP	123-gnn	GC	CM	DTNN	MPNN	MoLFormer-XL	SMI-TED$_{289M}$	SELFIES-TED	3DGrid-VQGAN
α	0.49	0.27	1.37	0.85	0.95	0.89	0.33	0.27	0.66	0.93
C_v	0.25	0.09	0.65	0.39	0.27	0.42	0.14	0.12	0.43	0.34
G	0.89	0.05	3.41	2.27	2.43	2.02	0.34	0.11	2.29	0.89
$\Delta\epsilon$	0.0052	0.0048	0.01126	0.0086	0.0112	0.0066	0.0038	0.0036	0.0084	0.0088
H	0.89	0.04	3.41	2.27	2.43	2.02	0.25	0.09	2.70	1.31
ϵ_{HOMO}	0.0036	0.0034	0.0072	0.0051	0.0038	0.0054	0.0029	0.0027	0.0054	0.0058
ϵ_{LUMO}	0.0041	0.0035	0.0092	0.0064	0.0051	0.0062	0.0027	0.0026	0.0071	0.0057
μ	0.451	0.476	0.583	0.519	0.244	0.358	0.361	0.384	0.6223	0.206
$\langle R^2 \rangle$	26.84	22.90	35.97	46.00	17.00	28.5	17.06	14.72	38.83	8.35
U_0	0.898	0.0427	3.41	2.27	2.43	2.05	0.3211	0.0850	2.9195	1.14
U	0.89	0.111	3.41	2.27	2.43	2.00	0.25	0.0905	2.6551	1.39
ZPVE	0.00207	0.0002	0.00299	0.00207	0.0017	0.00216	0.0003	0.0002	0.0032	0.0003
Avg. MAE	2.6355	1.9995	4.3536	4.7384	2.3504	3.1898	1.5894	1.3246	4.263	**1.2148**

The performance comparison in Table 5 highlights the strengths of the 3DGrid-VQGAN model in predicting quantum chemical properties using the QM9 dataset. The model, which utilizes 3D electron density grids as input, achieves the lowest average mean absolute error (MAE) of 1.2148, outperforming state-of-the-art methods across graph-based, geometry-based, and SMILES-based approaches.

One of the key strengths of 3DGrid-VQGAN lies in its ability to accurately predict spatially dependent properties, such as the electronic spatial extent ($\langle R^2 \rangle$) and the dipole moment (μ). For $\langle R^2 \rangle$, the model achieves an MAE of 8.35, which is significantly lower than the second-best model, SMI-TED$_{289M}$ (14.72), and geometry-based approaches like MPNN (28.5). This performance underscores the model's ability to leverage the volumetric data of 3D electron density grids, effectively capturing spatial distributions that are critical for this property. Similarly, the model achieves the lowest MAE for μ (0.206), outperforming all other methods. These results highlight the advantage of utilizing 3D grids, which provide rich spatial and electronic context that traditional methods relying solely on atomic coordinates or SMILES strings lack.

For thermodynamic properties such as enthalpy (H), internal energy (U), and zero-point vibrational energy (ZPVE), the model performs competitively, though it does not outperform SMILES-based methods like SMI-TED$_{289M}$. This suggests that while 3DGrid-VQGAN captures a substantial amount

of the underlying electronic and spatial information, certain features relevant to these properties might be better captured by complementary representations, such as those derived from SMILES. Similarly, for electronic properties like HOMO and LUMO energies, the model's performance is comparable but slightly behind SMILES-based approaches, which excel in encoding the localized electronic environments.

The consistency of 3DGrid-VQGAN's performance, as evidenced by its lower MAE metric, highlights its reliability across diverse molecular properties. This robustness can be attributed to its ability to represent high-dimensional electron density data in a compact latent space, ensuring effective generalization across various chemical environments and molecular geometries. Furthermore, the model surpasses geometry-based methods, such as DTNN and MPNN, which rely on atomic coordinates but lack the detailed electronic information inherent in 3D density grids.

While the 3DGrid-VQGAN model excels in spatially sensitive properties and demonstrates overall superior performance, there is room for improvement in properties where SMILES-based methods, such as SMI-TED$_{289M}$, perform better. Hybrid approaches that combine the strengths of SMILES-derived features with the rich spatial data from 3D grids could further enhance its predictive power, particularly for electronic and thermodynamic properties.

6.5 RECONSTRUCTION ILUSTRATION

Figure 3 investigates the reconstruction capabilities of the 3DGrid-VQGAN model by comparing original and reconstructed 3D electron density grids for selected molecular examples. The analysis focuses on the model's ability to preserve essential spatial and electronic features that are critical for downstream quantum chemical applications.

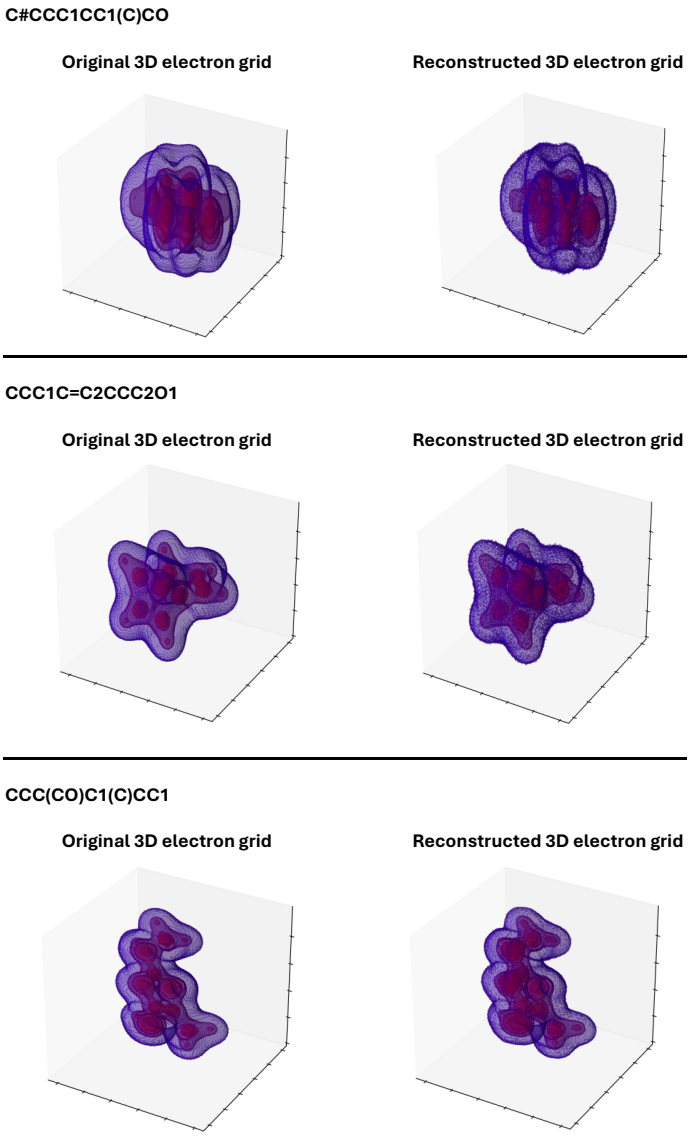

Figure 3: This figure illustrates the original and reconstructed 3D electron density grids for selected molecular examples.