NEBULA: Neural Empirical Bayes Under LAtent Representations for Efficient and Controllable Design of Molecular Libraries

Ewa M. Nowara [1] **Pedro O. Pinheiro** [2] **Sai Pooja Mahajan** [1] **Omar Mahmood** [2] **Andrew Martin Watkins** [1] **Saeed Saremi** [1] **Michael Maser** [1]

Abstract

We present NEBULA, the first latent 3D generative model for scalable generation of large molecular libraries around a seed compound of interest. Such libraries are crucial for scientific discovery, but it remains challenging to generate large numbers of high quality samples efficiently. 3D-voxel-based methods have recently shown great promise for generating high quality samples *de novo* from random noise (Pinheiro et al., 2023). However, sampling in 3D-voxel space is computationally expensive and use in library generation is prohibitively slow. Here, we instead perform *neural empirical* Bayes sampling (Saremi & Hyvärinen, 2019) in the learned latent space of a vector-quantized variational autoencoder. NEBULA generates large molecular libraries nearly **an order of magnitude faster** than existing methods without sacrificing sample quality. Moreover, NEBULA generalizes better to unseen drug-like molecules, as demonstrated on two public datasets and multiple recently released drugs. We expect the approach herein to be highly enabling for machine learning-based drug discovery. The code is available at https://github.com/prescient-design/nebula.

1. Introduction

Computational generation of new molecules with desired properties is a key element of chemical research, especially in drug discovery (DD). While search-based methods have achieved some success (Ghorbani et al., 2023; Kowalski et al., 2023; Janda, 1994), they can only explore small portions of chemical space, which is often insufficient for modern DD. Machine learning (ML) generative models can potentially accelerate search by rapidly generating large molecular libraries closely resembling a seed molecule of interest (Bilodeau et al., 2022). However, doing so while maintaining high sample quality remains an open challenge.

Chemical space exploration for the *optimization* of molecular properties remains at the forefront of DD. Traditional experimental approaches can be slow and unsatisfactory in results, in part due to the economic infeasibility of examining large enough chemical spaces in the wet lab (Wang et al., 2024). Methods for the prediction of molecules' biological properties have enabled more experimentation to occur *in silico* and have increased the rate of discovery and triage (Konze et al., 2019; Khalak et al., 2022; Thompson et al., 2022; Maser et al., 2023a). However, these are still largely dependent on *ad hoc*, combinatorial, and/or commercially available definitions of search space, which are unlikely to yield precise compounds with globally optimal properties. Therefore, combined with property predictors, expanding virtual libraries (VLs) beyond such definitions with generative models could be highly enabling for MLDD (Wang et al., 2024).

Molecules exist in 3D space, and 3D representations, such as volumes (voxels), capture rich information about their structure, shape, and properties. 3D generative models (3DGMs) trained with such inputs (Hoogeboom et al., 2022; Pinheiro et al., 2023) have the potential to learn more complete representations of molecules compared to models trained with 1D sequences (Segler et al., 2018; Blaschke et al., 2018; Guimaraes et al., 2018) or 2D graphs (Jin et al., 2018; Li et al., 2018; You et al., 2018; Mahmood et al., 2021). However, generative sampling in discrete 3D-input space is very computationally intensive (Pinheiro et al., 2023), making it impractical for creating VLs around lead chemical matter in which to perform optimization.

We herein present an approach to overcome such challenges by training denoising autoencoders (DAEs) on the continuous, low-dimensional latent representations of 3DGMs, which we call NEBULA. We demonstrate that *neural empirical* Bayes (NEB) sampling (Saremi & Hyvärinen, 2019) around noisy latent vectors substantially reduces the cost

[1] Prescient Design, Genentech, South San Francisco, USA [2] Prescient Design, Genentech, New York City, USA. Correspondence to: Ewa M. Nowara <nowara.ewa@gene.com>.

Proceedings of the 41^{st} International Conference on Machine Learning AI for Science workshop, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

and increases control of sample generation in the neighborhood surrounding input seed molecules, which we find highly desirable. Further, we show that training under our approach improves generalizability to chemical spaces well outside of training distributions, including for building-block and fragment generation. Finally, we provide real-world examples of candidate VLs generated around recently disclosed small-molecule drugs.

The main contributions of this work are the following:

1. Our latent approach, NEBULA, generates new molecules up to an order of magnitude faster than the state-of-the-art (SOTA) 3DGMs (Pinheiro et al., 2023; Xu et al., 2023) while maintaining high sample quality;
2. NEBULA scales to generation of very large molecular libraries with stable, unique, and valid (SUV) molecules similar to a seed compound;
3. Finally, NEBULA generalizes better than SOTA to generation around molecules from new chemical spaces, including multiple examples of real drugs that were just released in March 2024.

2. Related Work

2.1. Molecular Library Generation

The majority of modern methods for large molecular-library generation are based on non-ML approaches involving cheminformatics or combinatorial enumeration (Janda, 1994; Konze et al., 2019; Ghorbani et al., 2023; Kowalski et al., 2023). These methods are very often restricted to molecules within a commercially available chemical space. This is a significant limitation as most advanced DD programs require exploring new molecules outside of the commercially available space, which ML generative methods excel at as they able to produce completely novel samples.

2.2. Generation of Molecules in 3D Space

Generation of molecules represented in 3D space has been successful in creating high quality molecules. Wang et al. (2022) used a generative adversarial network (Goodfellow et al., 2014) on voxelized electron densities for pocket-based generation of molecules resembling known ligands. Skalic et al. (2019) and Ragoza et al. (2020) generated 3D voxelized molecules with CNNs and variational autoencoders (VAEs) (Kingma & Welling, 2014). VoxMol (Pinheiro et al., 2023) similarly generates molecules as 3D voxels with a *walk-jump sampling* (WJS) generative approach (Saremi & Hyvärinen, 2019), which currently holds SOTA performance in terms of sample quality and efficiency of generation. GSchNet (Gebauer et al., 2019) used autoregressive models to iteratively sample atoms and bonds in 3D space. Hoogeboom et al. (2022) proposed E(3) Equivariant Diffusion Model (EDM) that learns to generate molecules by iteratively applying a denoising network to a noise initialization. Finally, Vignac et al. (2023) improved EDM by jointly generating the 3D conformation and the 2D connectivity graph of molecules.

2.3. Latent-based Generation

Latent diffusion models (LDMs) (Rombach et al., 2022) have enabled efficient generation of 2D images by compressing input data to a learned latent space where denoising and sampling is less computationally intensive. Xu et al. (2023) applied EDM in the latent space instead of the input atomic coordinates to generate molecular *geometries*, dubbed GeoLDM. Finally, Mahajan et al. (2023) demonstrated the effectiveness of NEB sampling in the latent space of pre-trained language models for generation of novel 1D protein sequences.

It is worth noting that EDM and other geometric methods require node matrices of a predetermined size for generation; i.e., the number of atoms N in a generated sample must be given. This is a significant limitation for our setting of seeded generation, where N may take on a large array of values around an input molecule with N_0, especially when, e.g., adding a new fragment or functional group. We avoid this limitation by leveraging 3D voxel representations, while still retaining the advantages of latent-based generation and NEB. Our work is the first latent-based 3DGM of voxelized molecular denisities, and the first 3DGM suitable for library generation relative to lead chemical matter.

3. Proposed Approach

We present, NEBULA, Neural Empirical Bayes Under LAtent Representations. NEBULA is an efficient latent score-based generative model which compresses voxelized molecules to a lower dimensional latent space and generates new molecules by sampling around noised latent embeddings using NEB (Saremi & Hyvärinen, 2019). See Figure 1 for an overview of our proposed framework.

3.1. Compressing Voxels to a Latent Representation

We obtain 3D voxel representations of molecules by treating each atom as a continuous Gaussian density in 3D space centered around its atomic coordinates on a voxel grid by following Pinheiro et al. (2023). Each molecule is represented as a 4D tensor of $[c \times l \times l \times l]$, where c is the number of atom types and l is the length of the voxel grid edge. The values of each voxel range between 0 and 1.

Following common practice (Rombach et al., 2022; Dai & Wipf, 2019; Esser et al., 2021; Ramesh et al., 2021; Razavi et al., 2019; Van Den Oord et al., 2017), we separate the

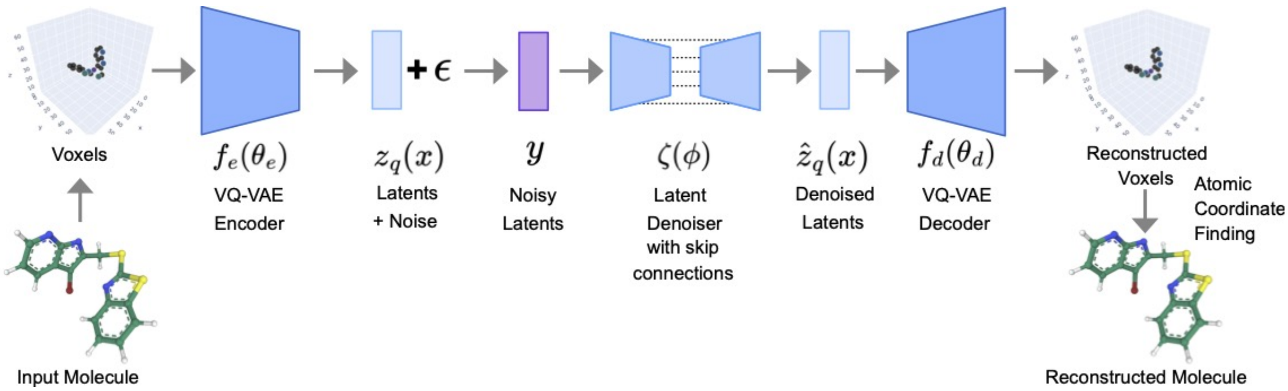

Figure 1. Overview of the proposed latent generative model, NEBULA. A 3D molecular graph is represented as a voxel grid and is passed through a VQ-VAE encoder to obtain latent embeddings. Noise is added and denoised by a latent U-Net, which is used to generative sampling. Denoised latents are passed through a VQ-VAE decoder to reconstruct the voxel grid, and the molecular graph is obtained via peak finding and sanitization.

training of a compression model to learn a latent representation and the training of a latent denoising model used for generation in two independent stages. As in Stable (latent) Diffusion (LDMs), we train a vector-quantized VAE (VQ-VAE) (Van Den Oord et al., 2017) to compress voxels to a latent space and then decode back to the original input. VQ-VAEs have been shown to aid in producing stable and valid samples for images (Rombach et al., 2022) and molecules (Wang et al., 2022). The input voxelized molecules x are encoded with a network $f_e(x)$ parameterized by θ_e to continuous latent embeddings z_e (Eq. 1). Each latent dimension is quantized to discrete vector z_q by matching it with one of k vectors in a learned "codebook" of embeddings e via nearest neighbor look-up (Eq. 2). The quantized latent embeddings z_q are then passed through the decoder f_d (θ_d) to reconstruct the input voxels $\hat{x}$ (Eq. 3).

$$f_e(x) : x \rightarrow z_e \tag{1}$$

$$z_q \leftarrow e_k, \text{where } k = \text{argmin}_j ||z_e - e_j||_2 \tag{2}$$

$$f_d(z_q) : z_q \rightarrow \hat{x} \tag{3}$$

The VQ-VAE is trained with a loss (Eq. 4) comprised of 1) a mean-squared error (MSE) reconstruction term between the input x and reconstructed voxels $\hat{x}$, 2) loss to learn the codebook of the embeddings e by moving the learned quantized embedding vectors e_i towards the continuous latent embeddings z_e, 3) and a "commitment cost" term which ensures the outputs of the encoder do not grow arbitrarily, where β is a hyperparameter and "sg" denotes a `stopgradient` operation.

$$\mathcal{L}_V = ||\hat{x} - x||_2^2 + ||\text{sg}[z_e] - e||_2^2 + \beta ||z_e - \text{sg}[e]||_2^2 \tag{4}$$

3.2. Denoising Latent Embeddings

For generation, we train a DAE on the learned VQ-VAE latents by adding isotropic Gaussian noise to the quantized embeddings $z_q \rightarrow z'_q$ with identity covariance matrix scaled by a fixed large noise level σ (Eq. 5). We normalize the latent embeddings before adding noise by subtracting the mean μ_c and dividing by the standard deviation σ_c computed across the $[c * l * l * l]$ channels over the training set (the empirical latent posterior). We train the latent model ζ_ϕ to denoise z'_q with a MSE loss computed between the input and predicted embeddings z_q and $\hat{z}_q$, respectively (Eq. 6). The denoised latents are then unnormalized with μ_c, σ_c before passing them to the decoder f_d for voxel reconstruction.

$$z'_q = z_q + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I_d) \tag{5}$$

$$\mathcal{L}_\mathcal{D} = ||z_q - \hat{z}_q||_2^2 \, ; \, \hat{z}_q = \zeta_\phi(z'_q) \tag{6}$$

3.3. Generation with Walk Jump Sampling

We generate new molecules with a *neural empirical* Bayes sampling scheme (Saremi & Hyvärinen, 2019) performed in the latent space in a two step process, referred to as *walk-jump sampling* (WJS). The WJS generative approach is a score-based method similar to diffusion models but it is more efficient as it only requires a single denoising step and noise scale compared to iterative denoising and complex noise schedules required in DMs. WJS is well suited for structured and textureless data, such as molecules and amino acid sequences (Frey et al., 2023), but it has also been successfully used for generation of images (Saremi & Srivastava, 2022; Saremi et al., 2023). First, we sample new noisy latent embeddings from the smooth (noisy) latent distribution using Langevin Markov chain Monte Carlo (MCMC) sampling (Cheng et al., 2018) with multiple *k walk* *steps* along the initialized manifold, where B_t is standard Brownian motion, γ and u are hyperparameters (referred to as friction and inverse mass, respectively), and g_ζ is the learned latent denoising score function.

$$\begin{aligned} d\upsilon_t = -\gamma\upsilon_t dt - u g_\zeta(z'_{qt})dt + (\sqrt{2\gamma u})dB_t \\ dz'_{qt} = \upsilon_t dt \end{aligned} \tag{7}$$

Second, we denoise the latent embeddings in a *jump step* with a forward pass of the denoising model for an arbitrary step k.

$$z_{qk} = \zeta_\phi(z'_{qk}) \tag{8}$$

We obtain the generated molecules by reconstructing the 3D voxels from the newly generated latent embeddings by passing them through the decoder of the frozen compression VQ-VAE model. Finally, we obtain the molecular graphs from the voxels by using the approach in Pinheiro et al. (2023), which is an optimization-based peak finding used to locate the atomic coordinates as the center of each voxel.

4. Data and Methods

Datasets. We train NEBULA on GEOM-drugs (GEOM) (Axelrod & Gomez-Bombarelli, 2022), which is a standard dataset used for molecule generation. We follow Vignac et al. (2023) to split it into the train, validation, and test sets with 1.1M/146K/146K molecules each. We use soft random subsampling (Cui et al., 2023) and train on 10% of the training set in each epoch. We test the cross-dataset generalizability of NEBULA and VoxMol trained on GEOM to a different large public dataset called PubChem Quantum (PCQM) (Nakata & Maeda, 2023; Nakata et al., 2020; Nakata & Shimazaki, 2017) and only use its test set containing 10,000 molecules. We additionally test the generalizability of NEBULA to several real drug molecules disclosed earlier this year (ACS). We use voxel grids of dimension 64, 8 atom channels ([C, H, O, N, F, S, Cl, Br]), and atomic radii of 0.25 Å resolution for all experiments.

Training Details. We use a VQ-VAE 3D autoencoder with no skip connections as the compression model architecture and a 3D U-Net with skip connections for the latent denoiser. Both models use a 3D convolutional architecture similar to (Pinheiro et al., 2023), with 4 levels of resolution and self-attention on the lowest two resolutions. We found that a latent dimension of 1024 worked best with higher noise levels. We train the compression and latent denoising models for about 150 epochs. We augment all models during training by randomly rotating and translating every training sample. We use a noise level of $\sigma = 1.8$ for all generations. See Appendix for details of the architecture, architecture ablations, training and sampling.

Baselines. The majority of existing ML generative models for molecules were only evaluated on *de novo* generation from random noise (Gebauer et al., 2019; Hoogeboom et al., 2022; Pinheiro et al., 2023). Therefore, to compare our method to existing approaches, we implemented and evaluated VoxMol (Pinheiro et al., 2023) for generation around a seed molecule, as it is the SOTA molecule generation method. Similar to NEBULA, VoxMol is a voxel-based 3D CNN approach, however, it generates molecules in the large input voxel space.

Evaluation Metrics. We evaluate the quality of the generated molecules with multiple metrics used by Vignac et al. (2023) and Pinheiro et al. (2023), such as molecular stability and validity (details in the Appendix). We also quantify how similar the generated molecules are to the seed molecule with the Tanimoto similarity computed between the seed and the generated molecular graphs (Bender & Glen, 2004). Following (Vignac et al., 2023) we report the results by only keeping the largest generated molecule in cases where multiple molecule fragments are generated for each generation.

5. Results and Discussion

5.1. Molecular Library Generation

We show within-dataset results on *seeded* generation where we aim to generate new molecules similar to a provided seed for hypothesis generation. We pass the seed molecule through the VQ-VAE encoder to construct latent embeddings and add noise in the latent space. We generate new molecules by taking different numbers of WJS steps in the noisy latent space and denoising the newly sampled embeddings after the last step. We generate molecules around 1,000 randomly selected seeds from the test set of GEOM (see Table 1 and Figure 2 for quantitative results and Figure 3 for examples of generations).

Both NEBULA and VoxMol generate stable and valid molecules that are similar to the seed, especially with few WJS steps. We experimentally determined the number of WJS steps for NEBULA and VoxMol that yielded comparable Tanimoto similarity and molecular stability on GEOM. As can be seen, NEBULA maintains significantly higher similarity to seed for much of the duration of WJS trajectories. Though similarity and stability ultimately decay at later steps, it is straightforward to filter unwanted compounds and the efficiency of generation yields overall gains for our task (see Sec 5.2).

Molecule Sanitization. Generating molecules via latent WJS requires the removal of skip connections in NEBULA's compression model. We find this makes it more difficult to train the VQ-VAE, in particular to reconstruct the 3D *geometries* of input molecules, as activations are not transferred from encoder to decoder. As a result, while NEBULA generates valid molecules, some heavy atoms are found to be missing attached hydrogens, which leads to an overall low molecular stability metric despite high validity and high atomic stability. Adding missing hydrogens can be solved trivially with fast cheminformatics toolkits such as RDKit (Landrum et al., 2016) via a process known as "sanitization". In practice, generating molecules with hydrogens is not crucial for discovering new molecules and many models are in fact trained with implicit (removed) hydrogens for efficiency (Vignac et al., 2023).

5.2. Scalability

Sampling in NEBULA's latent space is on average 6 times faster than the baselines (see the last column of Table 1 and in Figure 2) allowing us to quickly generate large molecular libraries with hundreds of thousands of molecules. Drug-like molecules require a voxel grid size of at least $[8 \times 64 \times 64 \times 64]$, where 8 is the number of atom types, requiring ~ 2 million points to represent each molecule and to sample new molecules in the input voxel space as is done in VoxMol. In contrast, NEBULA compresses the voxel space to a much smaller latent representation of $[1024 \times 8 \times 8 \times 8]$, where 1024 is the latent dimension, requiring only $\sim$ 500 k points per molecule which is 4 times smaller than the input voxel space representation. As can be seen in Figure 2, NEBULA trajectories require very minimal compute time, even at higher WJS steps.

5.3. Cross-Dataset Generalizability

We compare the cross-dataset generalizability of NEBULA and VoxMol trained on GEOM by generating libraries around 1,000 randomly selected seeds from the PCQM test set (see Table 2 and Figure 2 for quantitative results and Figure 4 for example generations). NEBULA generates molecules that are much closer to the seed at all WJS steps as demonstrated by ***tan. sim. %***. VoxMol tends to generate molecules which are quite different from PCQM, even at few WJS steps. Figure 4 shows that NEBULA makes incremental changes to seeds at early steps and maintains their overall scaffolds through entire trajectories, while VoxMol tends to generate very different molecule not resembling the seed. NEBULA also generates molecules with ***atom TV*** and ***bond TV*** that are very close to the raw PCQM metrics (step 0), while VoxMol seems to be reverting back to molecules that are more similar to its training distribution, GEOM.

We hypothesize that NEBULA achieves robustness to new chemical spaces due to the removal of skip connections in the VQ-VAE (see Sec 4). While this makes MSE loss slower to converge, it may also prevent "shortcuts" such as fitting geometric features that are specific to a certain structure-computation method but that do not generalize (Maser et al., 2023b). A similar behavior was observed for U-Net architectures used in image segmentation (Wilm et al., 2024; Kamath et al., 2023). This is visible in ***bond ang W_1***, which shows the distribution of bond angles diverges from the training set at long WJS steps for NEBULA but not for VoxMol. Future work will seek to identify training and data strategies that maximize chemical-space robustness (Maser et al., 2023a) while also producing faithful geometries.

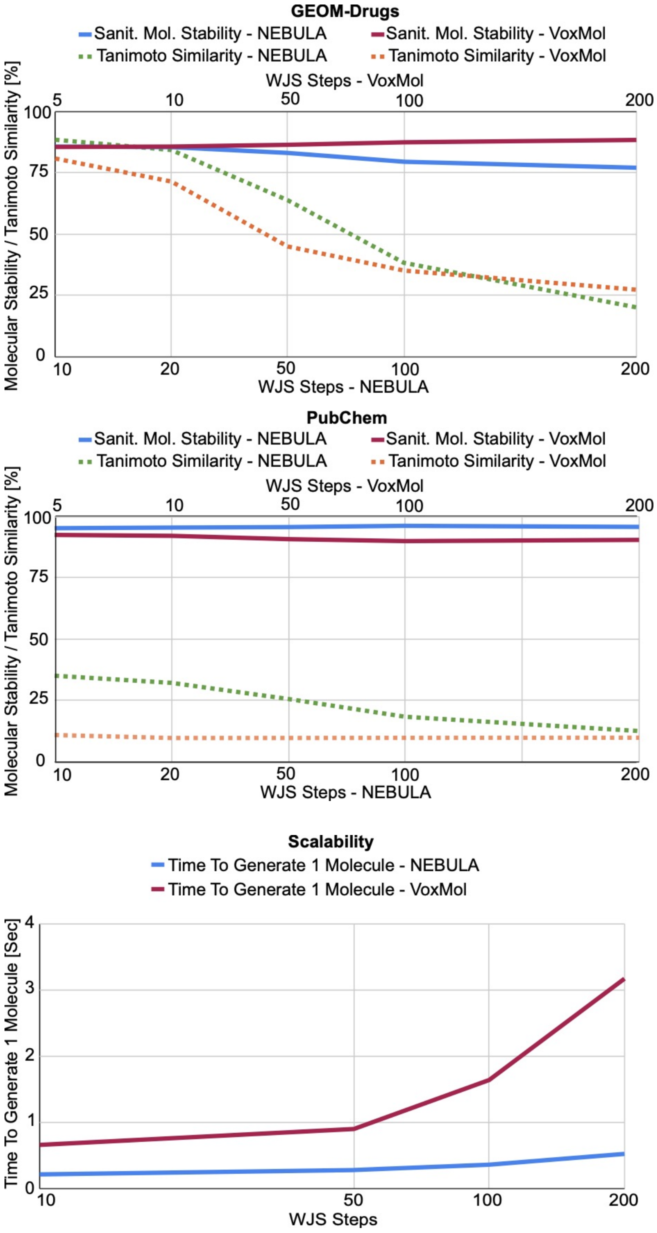

Figure 2. Molecular stability and Tanimoto similarity over WJS steps for molecules generated with NEBULA and VoxMol on **GEOM** (top) and **PCQM** (middle). (bottom) Scalability of each method plotted as the amount of time needed to generate one molecule at different WJS steps.

WJS Steps	tan.	stable	stable	valid	valency	atom	bond	bond	bond	avg. t
	sim. %$_{\uparrow}$	sanit. %$_{\uparrow}$	atom %$_{\uparrow}$	%$_{\uparrow}$	$W1_{\downarrow}$	$TV_{\downarrow}$	$TV_{\downarrow}$	len $W1_{\downarrow}$	ang $W1_{\downarrow}$	[s/mol.]$_{\downarrow}$
0 (*data*)	100.	99.90	99.90	99.80	0.001	0.001	0.025	0.00	0.05	-
5 VoxMol	80.84 (±0.96)	85.54 (±0.15)	99.43 (±0.04)	90.12 (±0.36)	0.26 (±0)	**0.02** (±0)	**0.03** (±0)	0.00 (±0)	0.67 (±0.02)	0.38
10	71.44 (±1.82)	85.71 (±0.36)	99.36 (±0.04)	89.86 (±0.28)	0.25 (±0)	0.02 (±0)	0.03 (±0)	0.00 (±0)	0.67 (±0.03)	0.66
50	44.99 (±1.03)	86.42 (±0.94)	99.35 (±0.02)	90.15 (±0.76)	0.25 (±0)	0.03 (±0)	0.03 (±0)	0.00 (±0)	0.54 (±0.02)	0.90
100	35.18 (±0.34)	87.44 (±1.06)	99.37 (±0.04)	90.52 (±0.81)	0.25 (±0)	0.03 (±0)	0.03 (±0)	0.00 (±0)	**0.50** (±0.00)	1.64
200	27.37 (±0.38)	**88.40** (±0.80)	99.35 (±0.03)	**90.86** (±0.51)	0.25 (±0)	0.04 (±0)	0.03 (±0)	0.00 (±0)	0.54 (±0.02)	3.17
10 **NEBULA**	**88.47** (±0.18)	85.73 (±0.12)	**99.46** (±0.01)	90.29 (±0.11)	0.26 (±0)	0.03 (±0)	**0.03** (±0)	0.00 (±0)	0.77 (±0)	**0.21**
20	84.3 (±0.46)	85.46 (±0.31)	99.42 (±0.01)	89.82 (±0.31)	0.26 (±0)	0.03 (±0)	0.03 (±0)	0.00 (±0)	0.85 (±0.01)	0.23
50	63.85 (±0.46)	83.11 (±0.31)	99.16 (±0.01)	86.61 (±0.31)	0.25 (±0)	0.03 (±0)	0.03 (±0)	0.00 (±0)	1.15 (±0.02)	0.28
100	38.33 (±0.8)	79.52 (±0.85)	98.45 (±0.04)	81.42 (±0.75)	0.23 (±0)	0.04 (±0)	0.03 (±0)	0.00 (±0)	1.80 (±0.01)	0.36
200	20.18 (±0.29)	77.08 (±0.64)	96.74 (±0.02)	77.96 (±0.52)	**0.20** (±0)	0.07 (±0)	0.04 (±0)	0.00 (±0)	3.56 (±0.06)	0.52

Table 1. Seeded generation results on **GEOM**. All results are averaged over 5 repeats of each experiment.

WJS Steps	tan.	stable	stable	valid	valency	atom	bond	bond len	bond ang
	sim. %$_{\uparrow}$	sanit. %$_{\uparrow}$	atom %$_{\uparrow}$	%$_{\uparrow}$	$W1_{\downarrow\Delta}$	$TV_{\downarrow\Delta}$	$TV_{\downarrow\Delta}$	$W1_{\downarrow\Delta}$	$W1_{\downarrow\Delta}$
0 (*data*)	100.00	95.51	99.62	99.40	2.98	0.90	0.24	0.01	3.40
5 VoxMol	10.95 (±1.85)	92.45 (±0.29)	95.60 (±0.07)	97.11 (±0.17)	0.21 (±0)	0.43 (±0.01)	0.14 (±0)	0.00 (±0)	4.04 (±0.07)
10	9.74 (±0.62)	92.08 (±0.83)	96.89 (±0.11)	96.49 (±0.63)	0.22 (±0)	0.35 (±0.00)	0.14 (±0)	0.00 (±0)	**3.49** (±0.05)
50	9.75 (±0.38)	90.73 (±0.54)	98.44 (±0.07)	95.17 (±0.30)	0.24 (±0)	0.15 (±0.01)	0.13 (±0)	0.00 (±0)	2.36 (±0.04)
100	9.81 (±0.21)	89.93 (±0.47)	98.62 (±0.04)	94.81 (±0.43)	0.25 (±0)	0.10 (±0.01)	0.12 (±0)	0.00 (±0)	1.98 (±0.07)
200	9.86 (±0.22)	90.44 (±0.85)	**98.77** (±0.07)	94.68 (±0.56)	**0.25** (±0)	0.06 (±0)	0.12 (±0)	0.00 (±0)	1.62 (±0.05)
10 **NEBULA**	**35.1** (±0.57)	95.18 (±0.26)	75.64 (±0.13)	**98.86** (±0.23)	0.23 (±0)	**0.84** (±0)	**0.24** (±0)	**0.01** (±0)	4.59 (±0.03)
20	32.19 (±0.38)	95.44 (±0.27)	76.43 (±0.24)	98.80 (±0.13)	0.22 (±0)	0.83 (±0)	0.23 (±0)	0.01 (±0)	4.72 (±0.05)
50	25.62 (±0.26)	95.64 (±0.26)	78.35 (±0.18)	98.36 (±0.12)	0.20 (±0)	0.79 (±0)	0.23 (±0)	0.01 (±0)	5.19 (±0.06)
100	18.38 (±0.27)	**96.16** (±0.19)	80.04 (±0.20)	97.88 (±0.35)	0.18 (±0)	0.74 (±0)	0.23 (±0)	0.01 (±0)	6.39 (±0.04)
200	12.58 (±0.09)	95.72 (±0.49)	79.77 (±0.51)	96.72 (±0.26)	0.19 (±0.01)	0.68 (±0)	0.23 (±0)	0.00 (±0)	9.30 (±0.09)

Table 2. Cross-dataset generalizability: *seeded* generation results on **PCQM**. NEBULA is able to generate molecules with atom and bond distributions closer to the new dataset (step 0). Bolded distribution metrics are those closest to the seed (↓ Δ). All results are averaged over 5 repeats of each experiment.

Real-world Application. We additionally demonstrate that NEBULA can generalize to unseen real drug molecules that were recently publicly released (March 2024) (ACS). These molecules represent significant leaps in complexity from GEOM and PCQM and have real biological relevance; all are currently in clinical trials as advanced cancer therapeutics. While similarity and stability measures are lower than for other datasets (Table 3), NEBULA is able to generate molecules that remain very faithful to the seed scaffold and introduce functional-group changes that typify a medicinal chemistry process (Figure 5).

See the Appendix for more examples of qualitative generation results on GEOM, PCQM, and drug molecules (ACS).

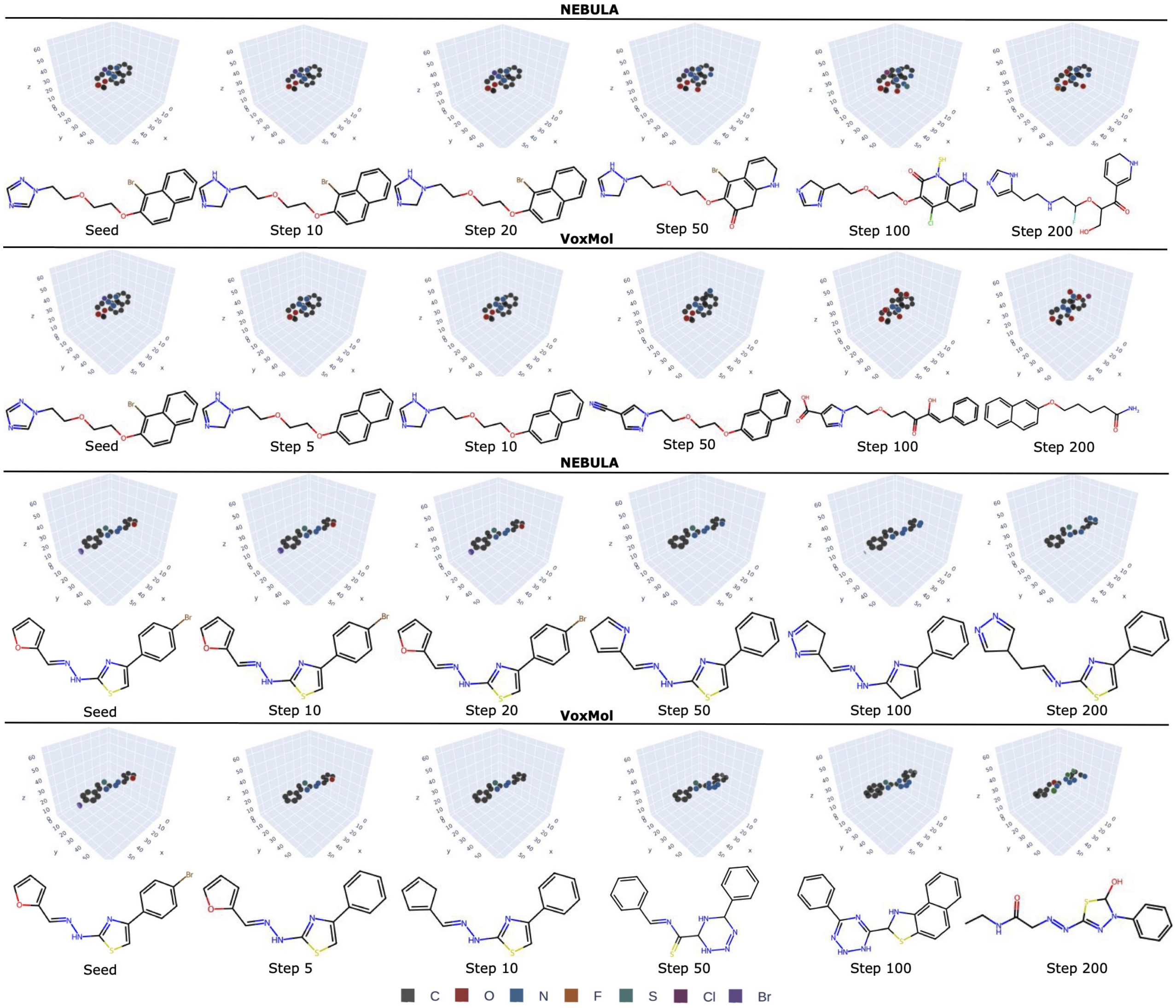

Figure 3. Seeded generation on **GEOM** with NEBULA and VoxMol at different WJS steps with the corresponding voxels. Both methods can generate molecules close to the seed in within-dataset generation.

6. Conclusions

We present the first latent generative model based on 3D-voxel representations of molecules, called NEBULA. NEBULA produces high quality, stable, and valid molecules around a seed molecule and is scalable to generation of very large molecular libraries needed for drug discovery via efficient *walk-jump* sampling in latent space. Moreover, NEBULA generalizes well to new chemical spaces, including fragments and real drug molecules. We expect to see great utility of our approach for accelerating ML-based drug discovery.

WJS Steps	tan. sim. $\%_\uparrow$	stable sanit. $\%_\uparrow$	stable atom $\%_\uparrow$	valid $\%_\uparrow$
10	30.62	42.22	42.20	42.22
20	31.13	40.00	42.44	40.00
50	23.37	43.33	42.93	43.33
100	22.95	42.22	43.48	42.22
200	19.17	43.33	43.26	43.33

Table 3. Seeded generation results averaged over 10 experiments for 5 real recently released drugs (ACS) (for a total of 50 generations).

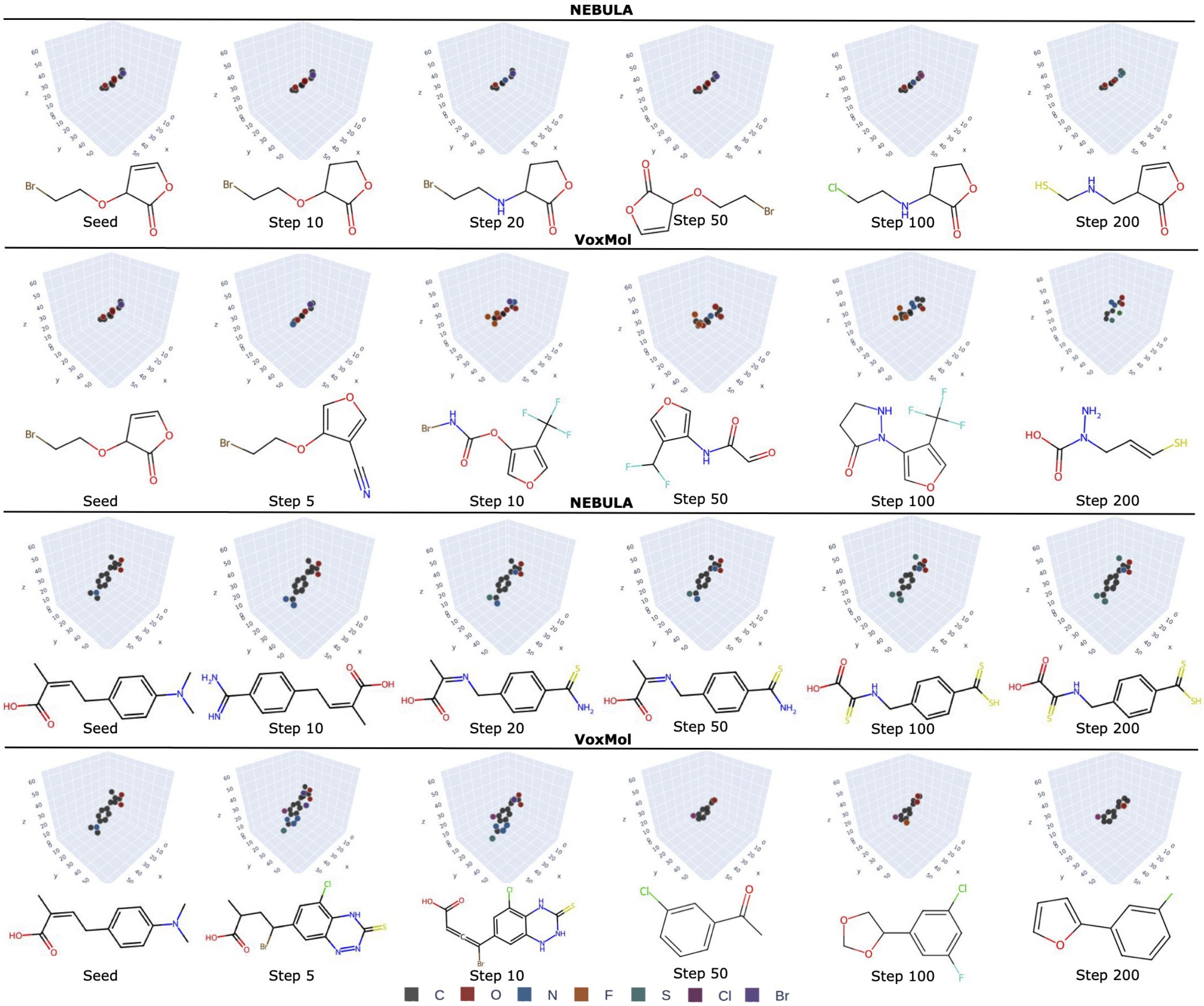

Figure 4. Seeded generation on **PCQM** with NEBULA and VoxMol at different WJS steps with the corresponding voxels. NEBULA is able to generate molecules that maintain the seed scaffold in all cases in cross-dataset generation, while VoxMol tends to diverge from the seed compound.

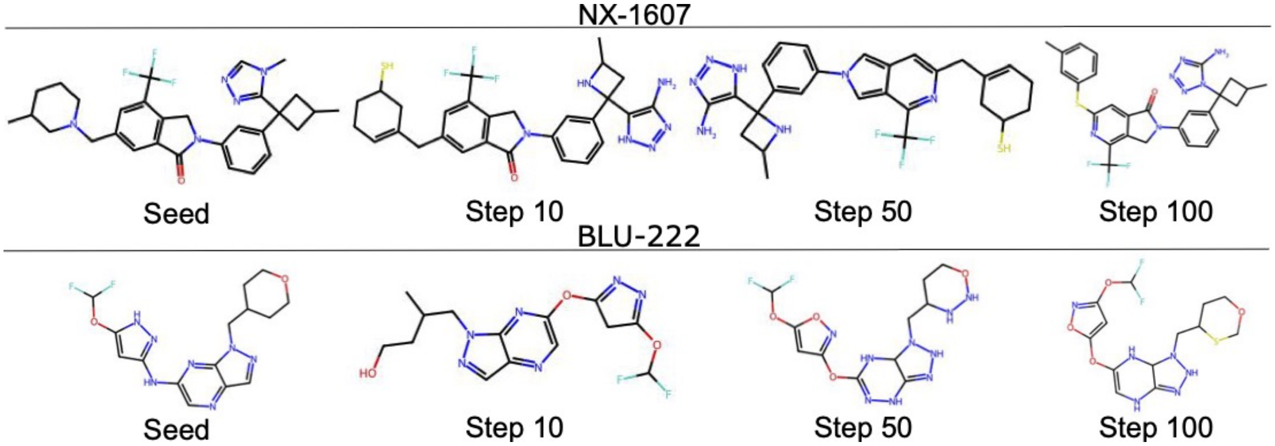

Figure 5. Seeded Generation on real drugs just released in March 2024 (ACS) with NEBULA at different WJS steps.

References

Acs spring 2024: First time disclosures. https://drughunter.com/articles/acs-spring-2024-first-time-disclosures/. Accessed: 2024-05-20.

Axelrod, S. and Gomez-Bombarelli, R. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Bender, A. and Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2 22:3204–18, 2004. URL https://api.semanticscholar.org/CorpusID:16399588.

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., and Jensen, K. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12, 03 2022. doi: 10.1002/wcms.1608.

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., and Chen, H. Application of generative autoencoder in de novo molecular design. *Molecular Informatics*, 37, 11 2018. doi: 10.1002/minf.201700123.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on learning theory*, pp. 300–323. PMLR, 2018.

Cui, X., Mittal, A., Lu, S., Zhang, W., Saon, G., and Kingsbury, B. Soft random sampling: A theoretical and empirical analysis. *arXiv preprint arXiv:2311.12727*, 2023.

Dai, B. and Wipf, D. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

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

Frey, N. C., Berenberg, D., Kleinhenz, J., Ra, S., Hotzel, I., Lafrance-Vanasse, J., Kelly, R. L., Wu, Y., Rajpal, A., Bonneau, R., et al. Learning protein family manifolds with smoothed energy-based models. In *ICLR Workshop on Physics for Machine Learning*, 2023.

Gebauer, N., Gastegger, M., and Schütt, K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.

Ghorbani, M., Gendelev, L., Beroza, P., and Keiser, M. J. Autoregressive fragment-based diffusion for pocket-aware ligand design. *arXiv preprint arXiv:2401.05370*, 2023.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C., and Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv 1705.10843*, 2018.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887, 2022.

Janda, K. D. Tagged versus untagged libraries: methods for the generation and screening of combinatorial chemical libraries. *Proceedings of the National Academy of Sciences*, 91(23):10779–10785, 1994.

Jin, W., Barzilay, R., and Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2323–2332, 10–15 Jul 2018.

Kamath, A., Willmann, J., Andratschke, N., and Reyes, M. Do we really need that skip-connection? understanding its interplay with task complexity. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 302–311. Springer, 2023.

Khalak, Y., Tresadern, G., Hahn, D. F., de Groot, B. L., and Gapsys, V. Chemical space exploration with active learning and alchemical free energies. *Journal of Chemical Theory and Computation*, 18(10):6259–6270, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.

Konze, K. D., Bos, P. H., Dahlgren, M. K., Leswing, K., Tubert-Brohman, I., Bortolato, A., Robbason, B., Abel, R., and Bhat, S. Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. *Journal of chemical information and modeling*, 59(9):3782–3793, 2019.

Kowalski, D. J., MacGregor, C. M., Long, D.-L., Bell, N. L., and Cronin, L. Automated library generation and serendipity quantification enables diverse discovery in coordination chemistry. *Journal of the American Chemical Society*, 145(4):2332–2341, 2023.

Landrum, G. et al. Rdkit: Open-source cheminformatics software. 2016.

Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324*, 2018.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mahajan, S. P., Frey, N. C., Berenberg, D., Kleinhenz, J., Bonneau, R., Gligorijevic, V., Watkins, A., and Saremi, S. Exploiting language models for protein discovery with latent walk-jump sampling. *Machine Learning for Structural Biology Workshop, NeurIPS*, 2023.

Mahmood, O., Mansimov, E., Bonneau, R., and Cho, K. Masked graph modeling for molecule generation. *Nature communications*, 2021.

Maser, M., Park, J. W., Lin, J. Y.-Y., Lee, J. H., Frey, N. C., and Watkins, A. Supsiam: Non-contrastive auxiliary loss for learning from molecular conformers. *arXiv preprint arXiv:2302.07754*, 2023a.

Maser, M., Tagasovska, N., Lee, J. H., and Watkins, A. Moleclues: Molecular conformers maximally in-distribution for predictive models. *arXiv preprint arXiv:2306.11681*, 2023b.

Nakata, M. and Maeda, T. Pubchemqc b3lyp/6-31g*//pm6 data set: The electronic structures of 86 million molecules using b3lyp/6-31g* calculations. *Journal of Chemical Information and Modeling*, 63(18):5734–5754, 2023. doi: 10.1021/acs.jcim.3c00899. URL https://doi.org/10.1021/acs.jcim.3c00899. PMID: 37677147.

Nakata, M. and Shimazaki, T. Pubchemqc project: A large-scale first-principles electronic structure database for data-driven chemistry. *Journal of Chemical Information and Modeling*, 57(6):1300–1308, 2017. doi: 10.1021/acs.jcim.7b00083. URL https://doi.org/10.1021/acs.jcim.7b00083. PMID: 28481528.

Nakata, M., Shimazaki, T., Hashimoto, M., and Maeda, T. Pubchemqc pm6: Data sets of 221 million molecules with optimized molecular geometries and electronic properties. *Journal of Chemical Information and Modeling*, 60(12):5891–5899, 2020. doi: 10.1021/acs.jcim.0c00740. URL https://doi.org/10.1021/acs.jcim.0c00740. PMID: 33104339.

Pinheiro, P. O., Rackers, J., Kleinhenz, J., Maser, M., Mahmood, O., Watkins, A. M., Ra, S., Sresht, V., and Saremi, S. 3d molecule generation by denoising voxel grids. *Advances in Neural Information Processing Systems*, 2023.

Ragoza, M., Masuda, T., and Koes, D. R. Learning a continuous representation of 3d molecular structures with deep generative models. *arXiv preprint arXiv:2010.08687*, 2020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

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

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Saremi, S. and Hyvärinen, A. Neural empirical bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019.

Saremi, S. and Srivastava, R. K. Multimeasurement generative models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=QRX0nCX_gk.

Saremi, S., Srivastava, R. K., and Bach, F. Universal smoothed score functions for generative modeling. *arXiv preprint arXiv:2303.11669*, 2023.

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4(1):120–131, 2018. doi: 10.1021/acscentsci.7b00512. URL https://doi.org/10.1021/acscentsci.7b00512. PMID: 29392184.

Skalic, M., Jiménez, J., Sabbadin, D., and De Fabritiis, G. Shape-based generative modeling for de novo drug design. *Journal of chemical information and modeling*, 59(3):1205–1214, 2019.

Thompson, J., Walters, W. P., Feng, J. A., Pabon, N. A., Xu, H., Maser, M., Goldman, B. B., Moustakas, D., Schmidt, M., and York, F. Optimizing active learning for free energy calculations. *Artificial Intelligence in the Life Sciences*, 2:100050, 2022.

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

Vignac, C., Osman, N., Toni, L., and Frossard, P. Midi: Mixed graph and 3d denoising diffusion for molecule generation. *arXiv preprint arXiv:2302.09048*, 2023.

Wang, L., Bai, R., Shi, X., Zhang, W., Cui, Y., Wang, X., Wang, C., Chang, H., Zhang, Y., Zhou, J., et al. A pocket-based 3d molecule generative model fueled by experimental electron density. *Scientific reports*, 12(1):15100, 2022.

Wang, L., Zhou, Z., Yang, X., Shi, S., Zeng, X., and Cao, D. The present state and challenges of active learning in drug discovery. *Drug Discovery Today*, pp. 103985, 2024.

Wilm, F., Ammeling, J., Öttl, M., Fick, R. H., Aubreville, M., and Breininger, K. Rethinking u-net skip connections for biomedical image segmentation. *arXiv preprint arXiv:2402.08276*, 2024.

Xu, M., Powers, A. S., Dror, R. O., Ermon, S., and Leskovec, J. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

You, J., Ying, R., Ren, X., Hamilton, W. L., and Leskovec, J. Graphrnn: Generating realistic graphs with deep autoregressive models, 2018.

A. Additional Implementation Details

A.1. Training and Architecture Details

We use a VQ-VAE 3D autoencoder with no skip connections as the compression model architecture and a 3D U-Net with skip connections for the latent denoiser. Both models use a 3D convolutional architecture similar to (Pinheiro et al., 2023), with 4 levels of resolution and self-attention on the lowest two resolutions. We found that a latent dimension of 1024 worked best with higher noise levels. We reduce the 1024 latent embedding to 32 in the first layer of the latent denoiser and gradually increase it with subsequent layers. We use a commitment cost of 0.25 and 256 latent codebook embeddings of 256 for the compression VQ-VAE. We train the compression and latent denoising models for about 150 epochs on GEOM-Drugs (until the mean intersection over union between the input and reconstructed voxels is above 0.90). We train the models with a batch size 32, AdamW optimizer (Loshchilov & Hutter, 2017) with β_1=0.9, β_2=0.999, learning rate 10^{-5}, weight decay 10^{-2}, dropout 0.1, SiLU activation function, and we update the weights with exponential moving average (EMA) with decay of 0.999. We augment all models during training by randomly rotating and translating every training sample. We use a noise level of $\sigma = 1.8$ for all generations, friction of 1.0, lipschitz of 1.0 and step size of 0.25 (VoxMol uses step size of 0.5 in the voxel space) for the MCMC sampling to generate new molecules with the trained latent denoising model.

A.2. Evaluation Metrics

We evaluate the models using the metrics following (Vignac et al., 2023). ***stable mol*** and ***stable atom*** are molecular and atomic stability; an atom is stable when the number of bonds with other atoms matches their valence and a molecule is stable only if 100 % of its atoms are stable (Hoogeboom et al., 2022) (see Sec 5.1). ***validity*** is measured as the percent of molecules passing the RDKit's (Landrum et al., 2016) sanitization. ***valency*** $\boldsymbol{W}_1$ is the Wasserstein distances between valences in the distributions of the generated molecules and the molecules in the dataset. ***atoms TV*** and ***bonds TV*** are the total variation between the atom and bond types distributions. ***bond length*** $\boldsymbol{W}_1$ and ***bond angle*** $\boldsymbol{W}_1$ are the Wasserstein distances between the distributions of bond lengths and angles of the generated molecules and the molecules in the dataset.

B. Additional Generation Results (Within-Dataset)

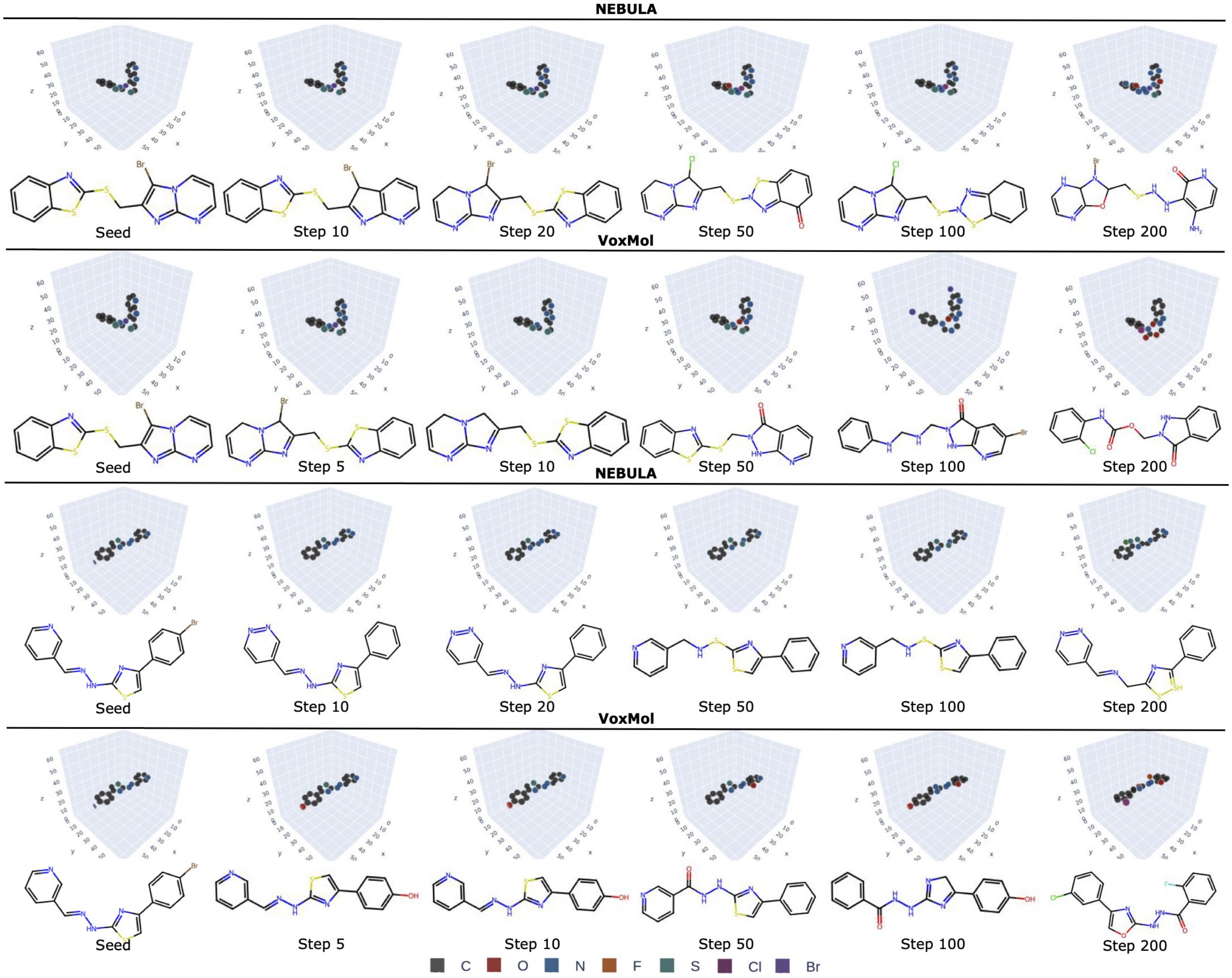

Figure 6. Additional examples of seeded generation on GEOM-Drugs (Axelrod & Gomez-Bombarelli, 2022).

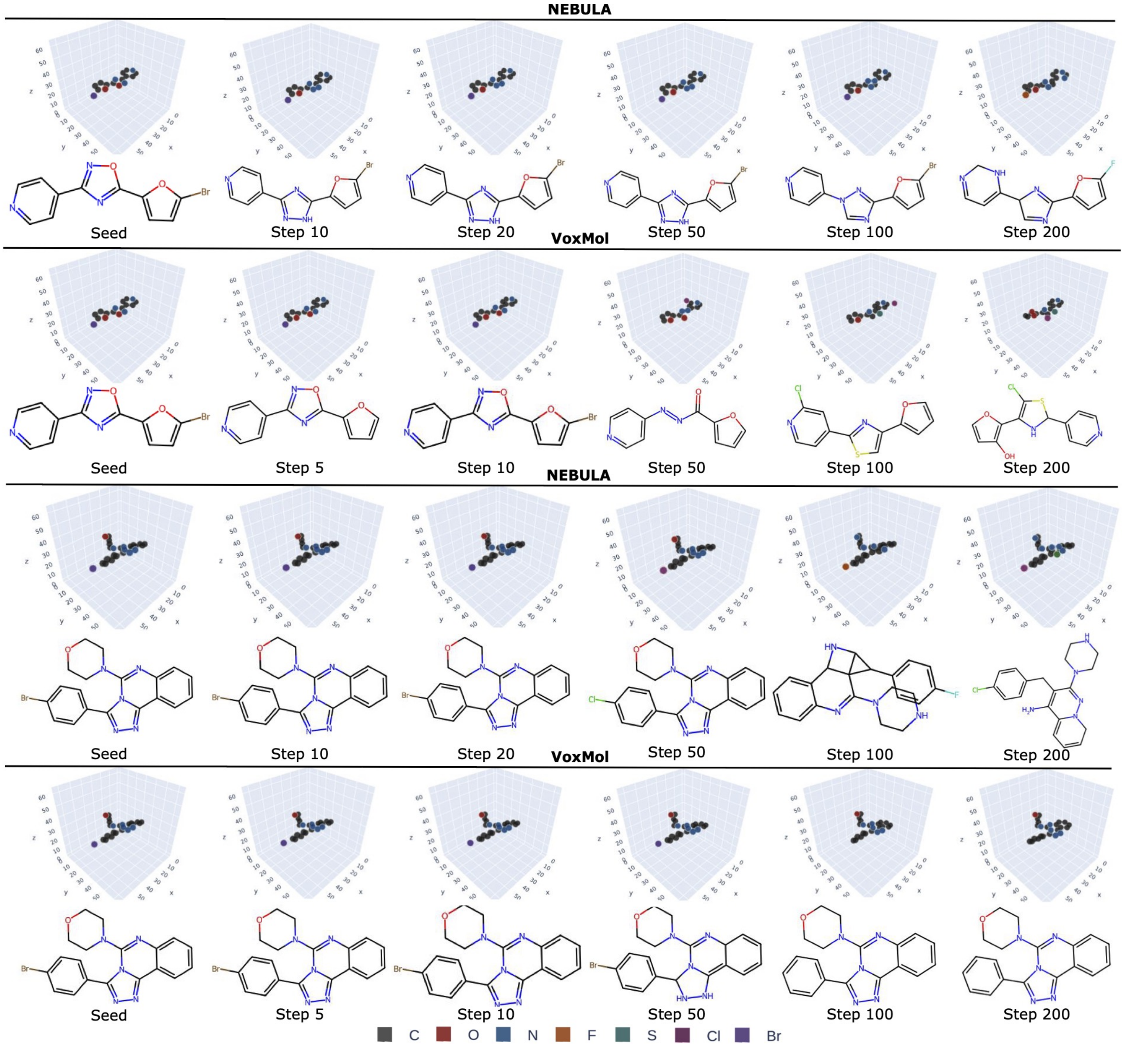

Figure 7. Additional examples of seeded generation on GEOM-Drugs (Axelrod & Gomez-Bombarelli, 2022).

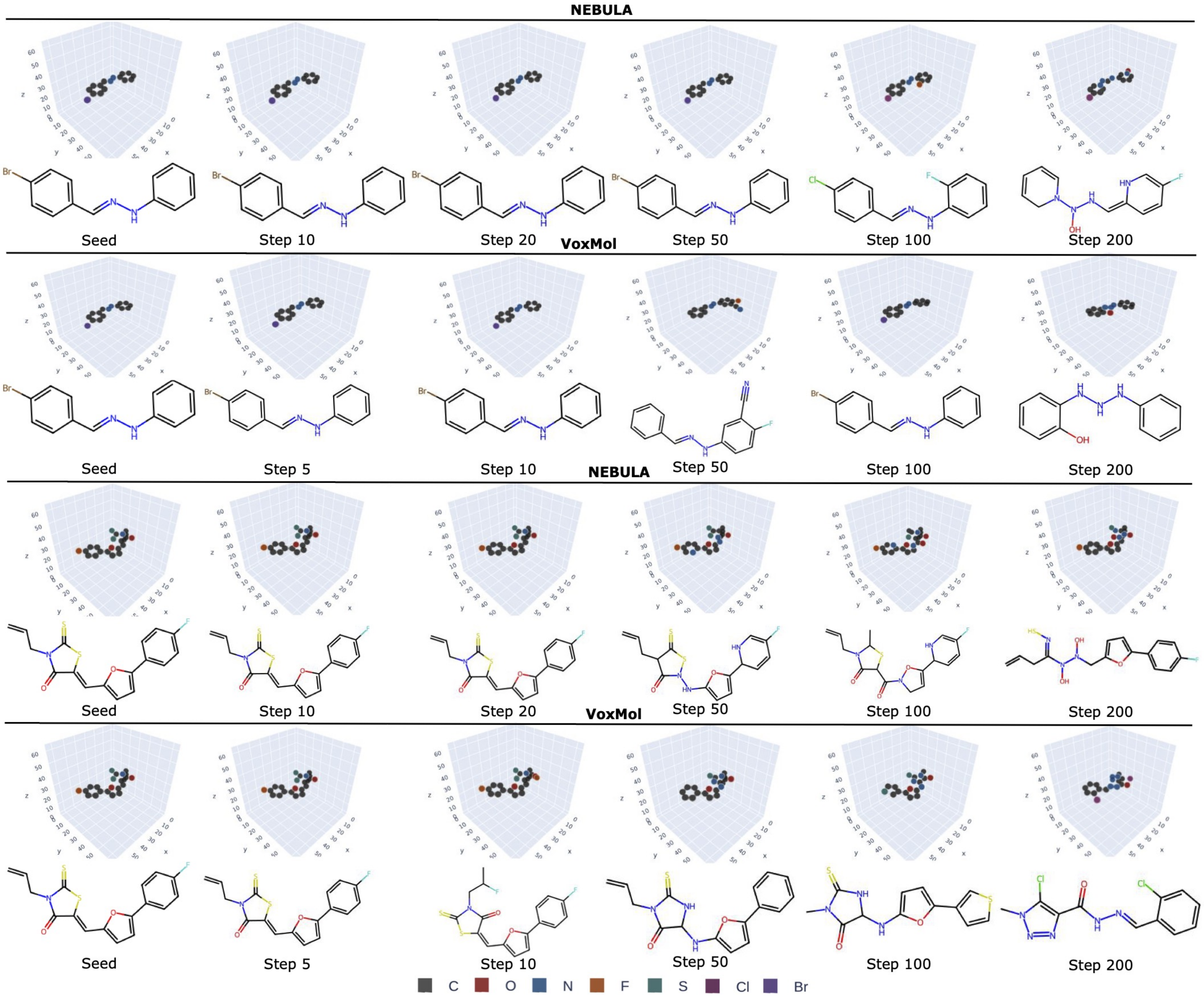

Figure 8. Additional examples of seeded generation on GEOM-Drugs (Axelrod & Gomez-Bombarelli, 2022).

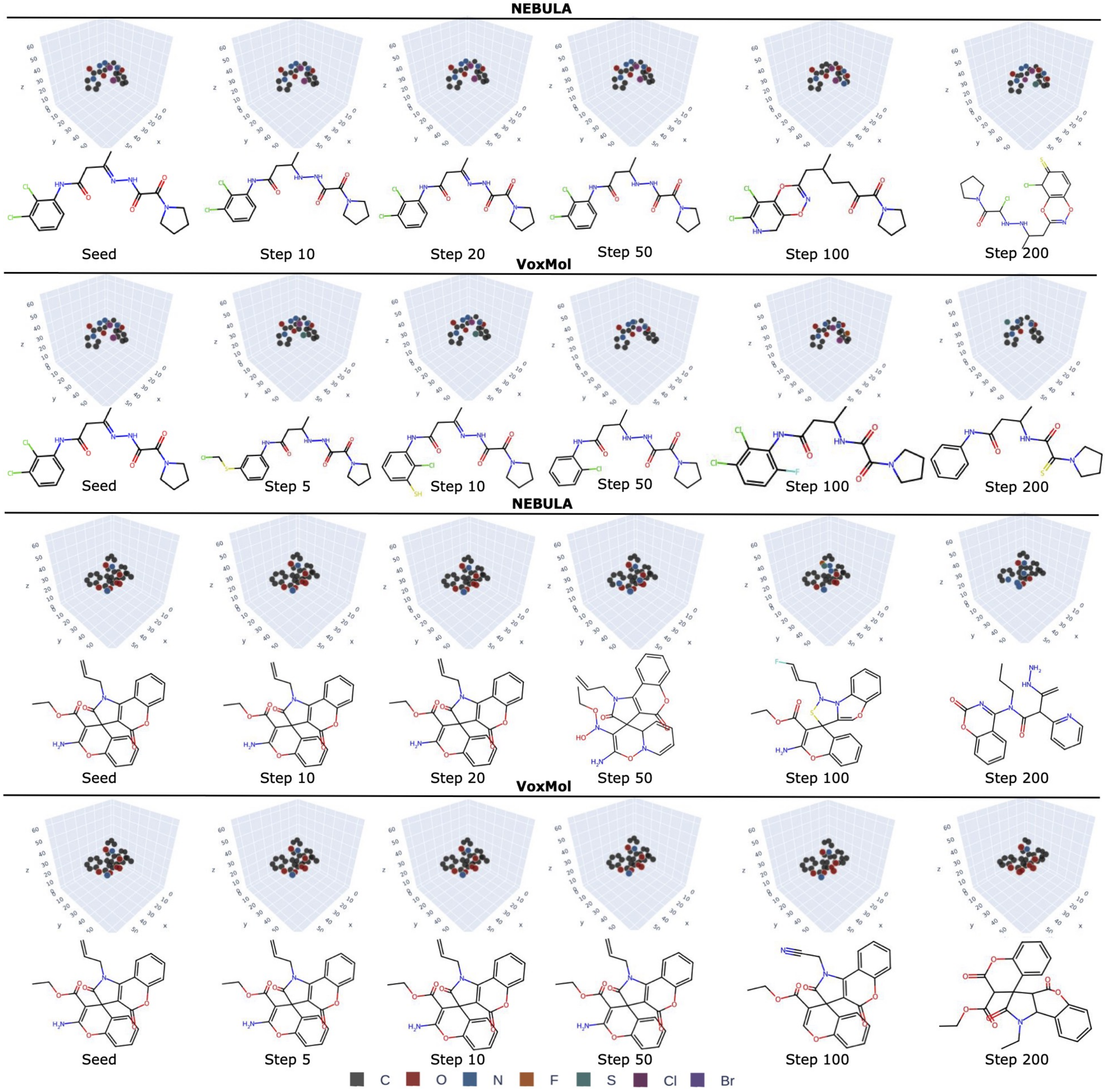

Figure 9. Additional examples of seeded generation on GEOM-Drugs (Axelrod & Gomez-Bombarelli, 2022).

C. Additional Generation Results (Cross-Dataset)

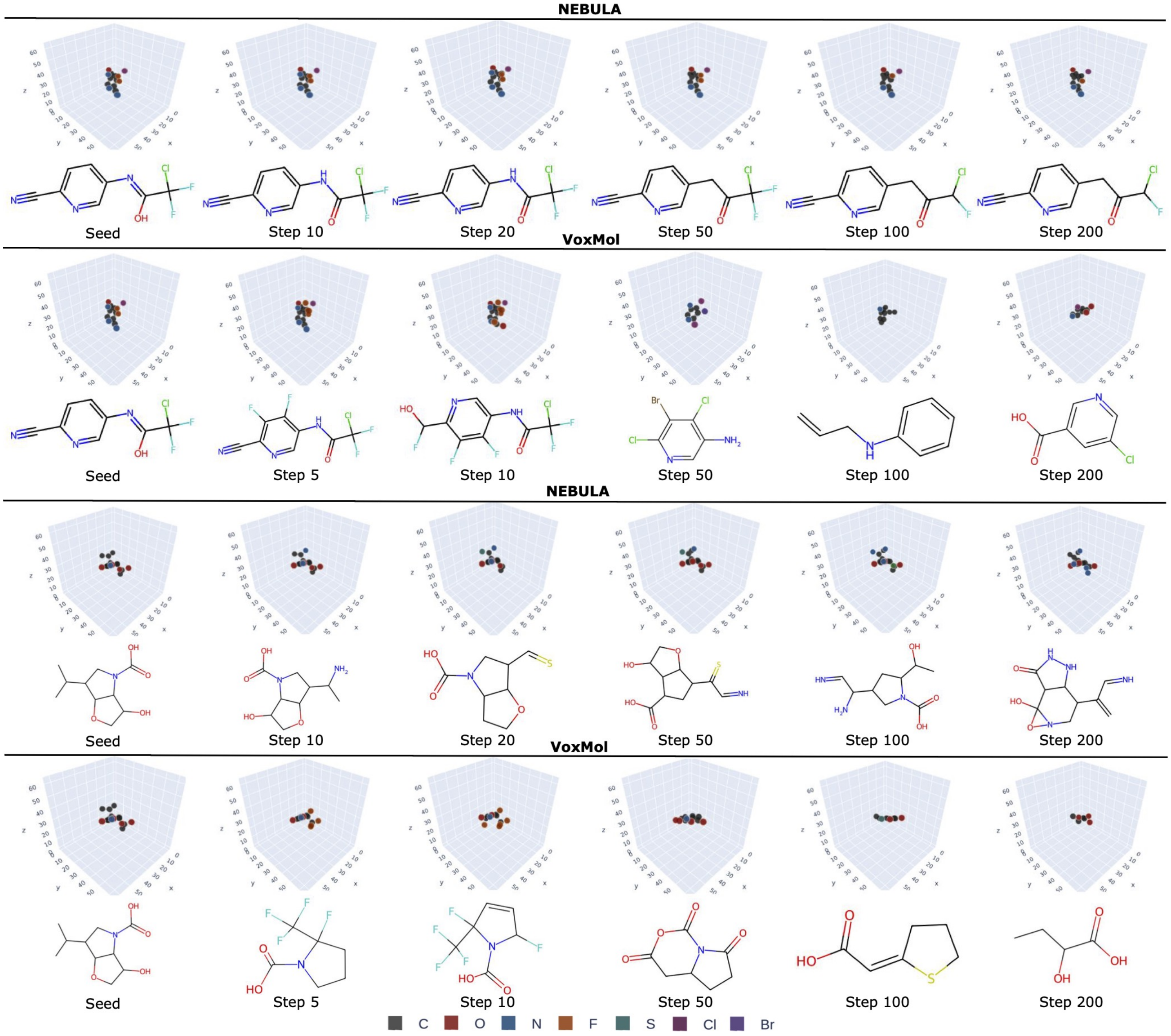

Figure 10. Additional examples of seeded generation on PubChem (Nakata & Maeda, 2023).

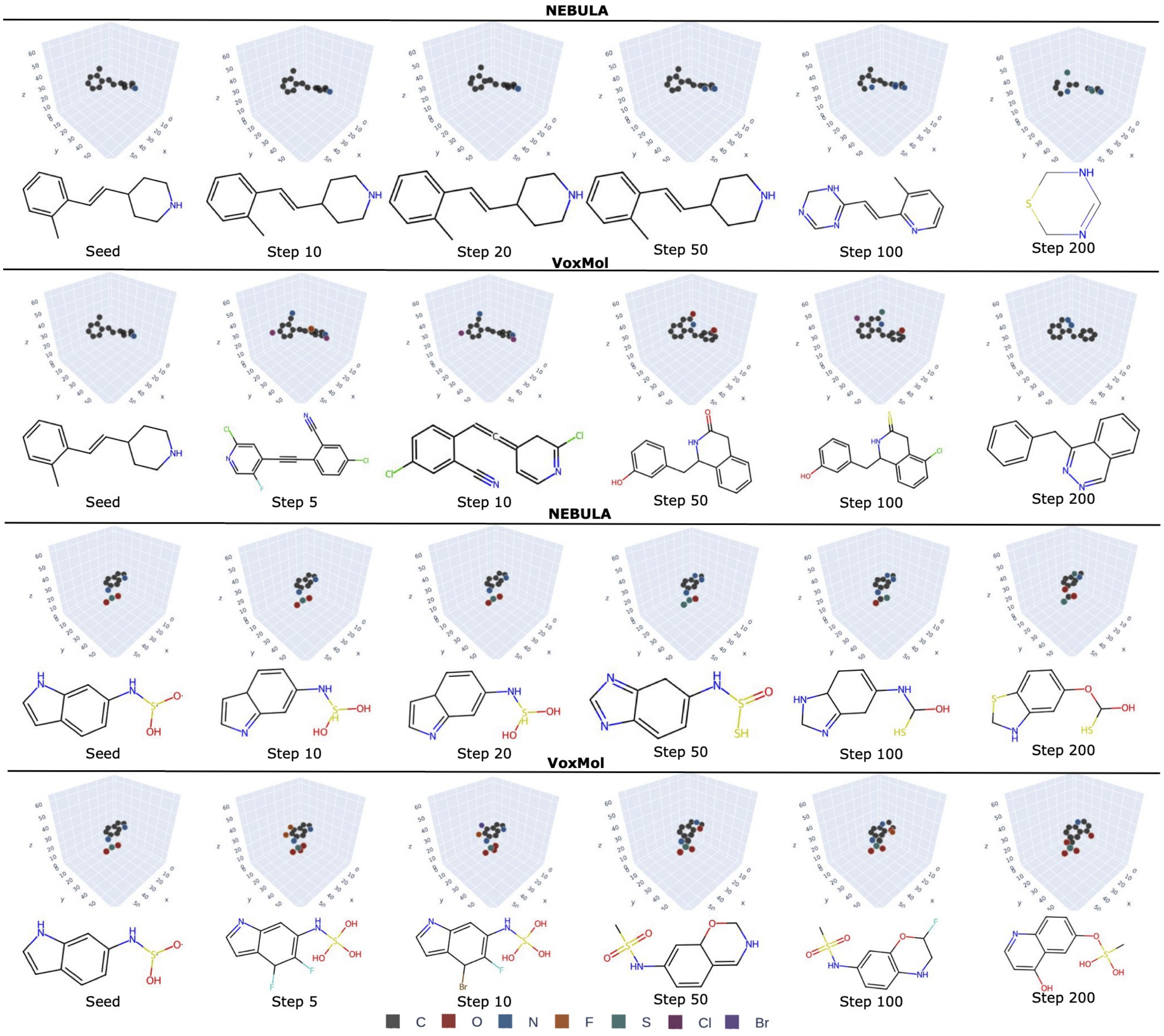

Figure 11. Additional examples of seeded generation on PubChem (Nakata & Maeda, 2023).

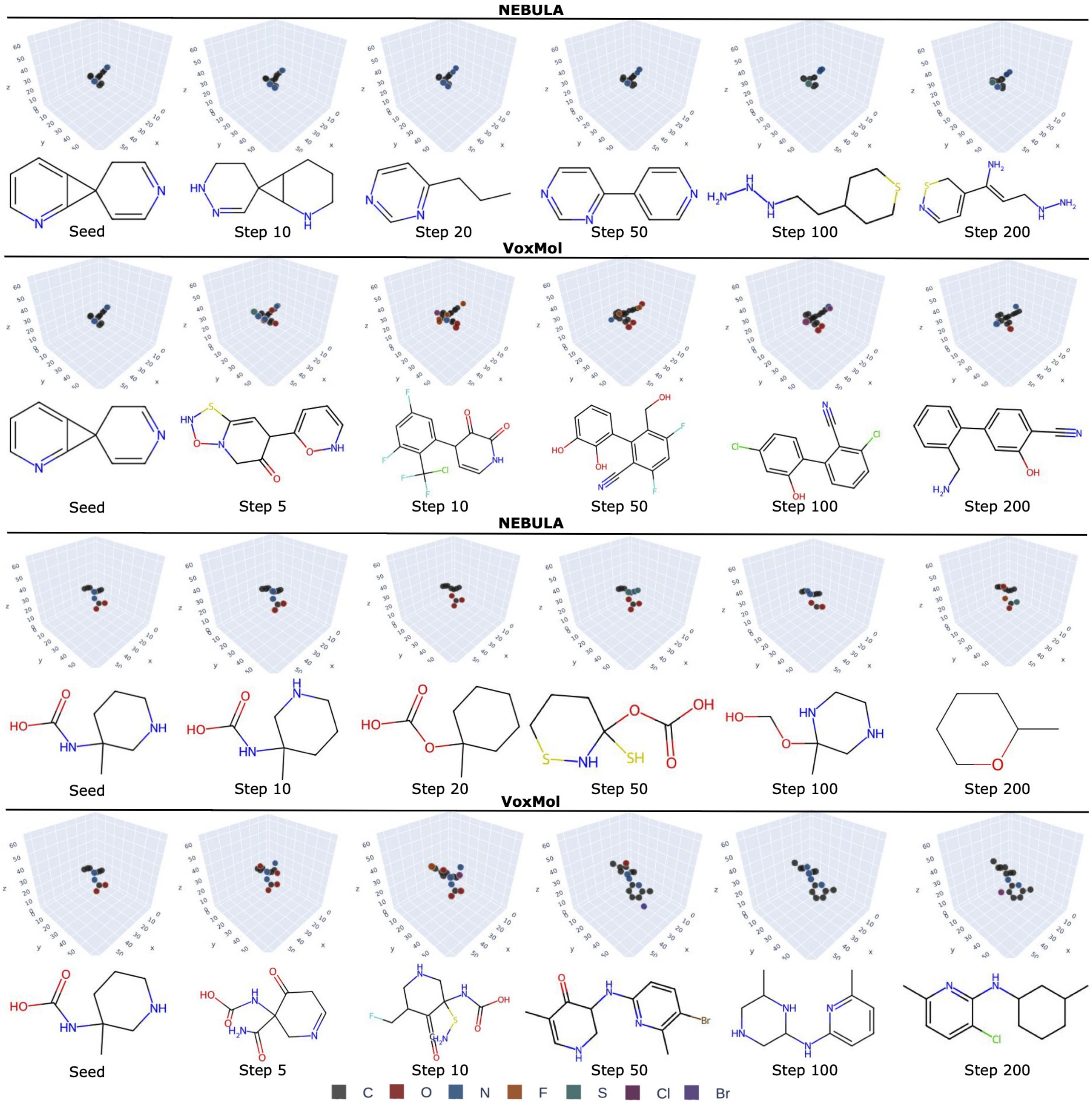

Figure 12. Additional examples of seeded generation on PubChem (Nakata & Maeda, 2023).

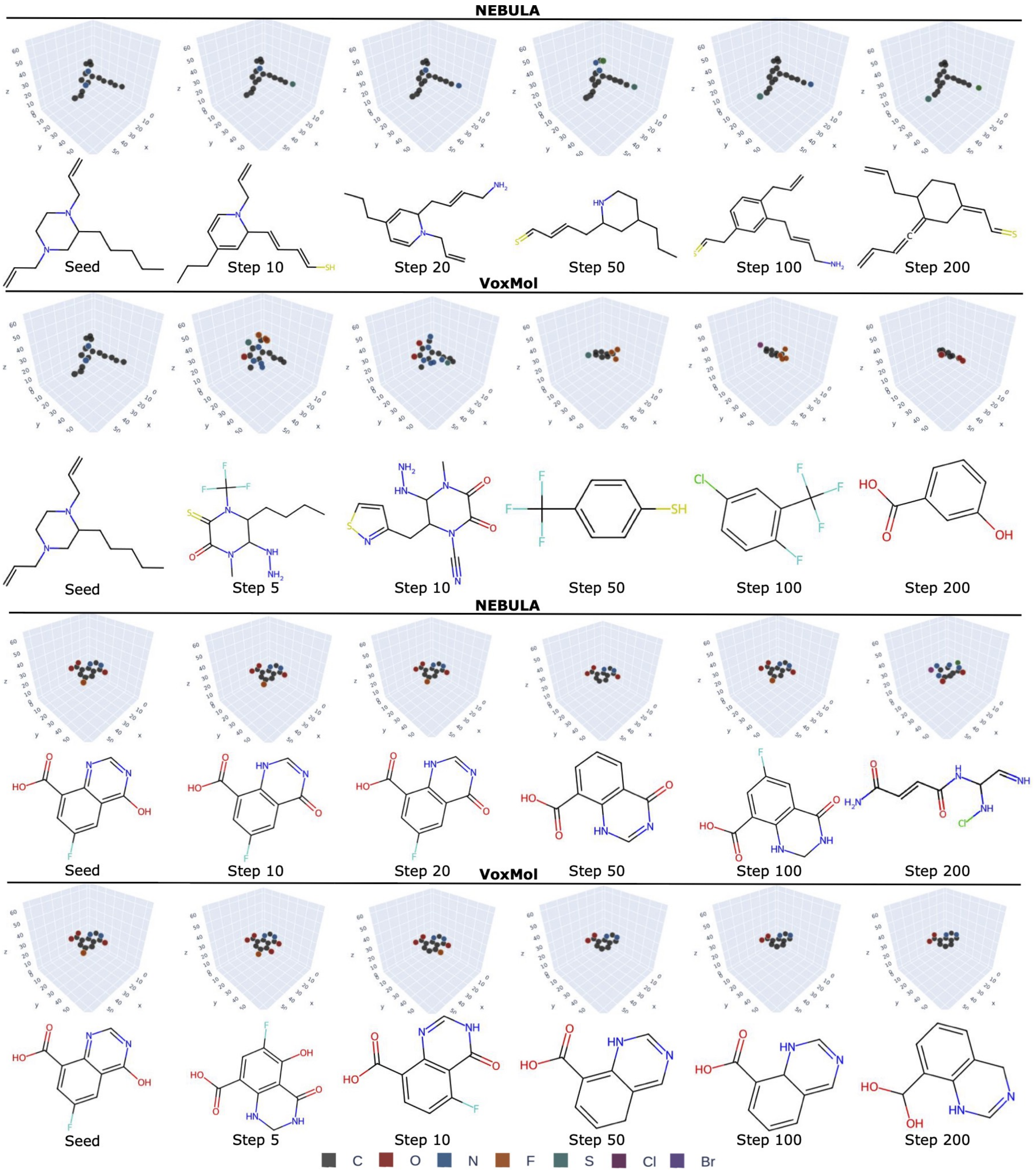

Figure 13. Additional examples of seeded generation on PubChem (Nakata & Maeda, 2023).

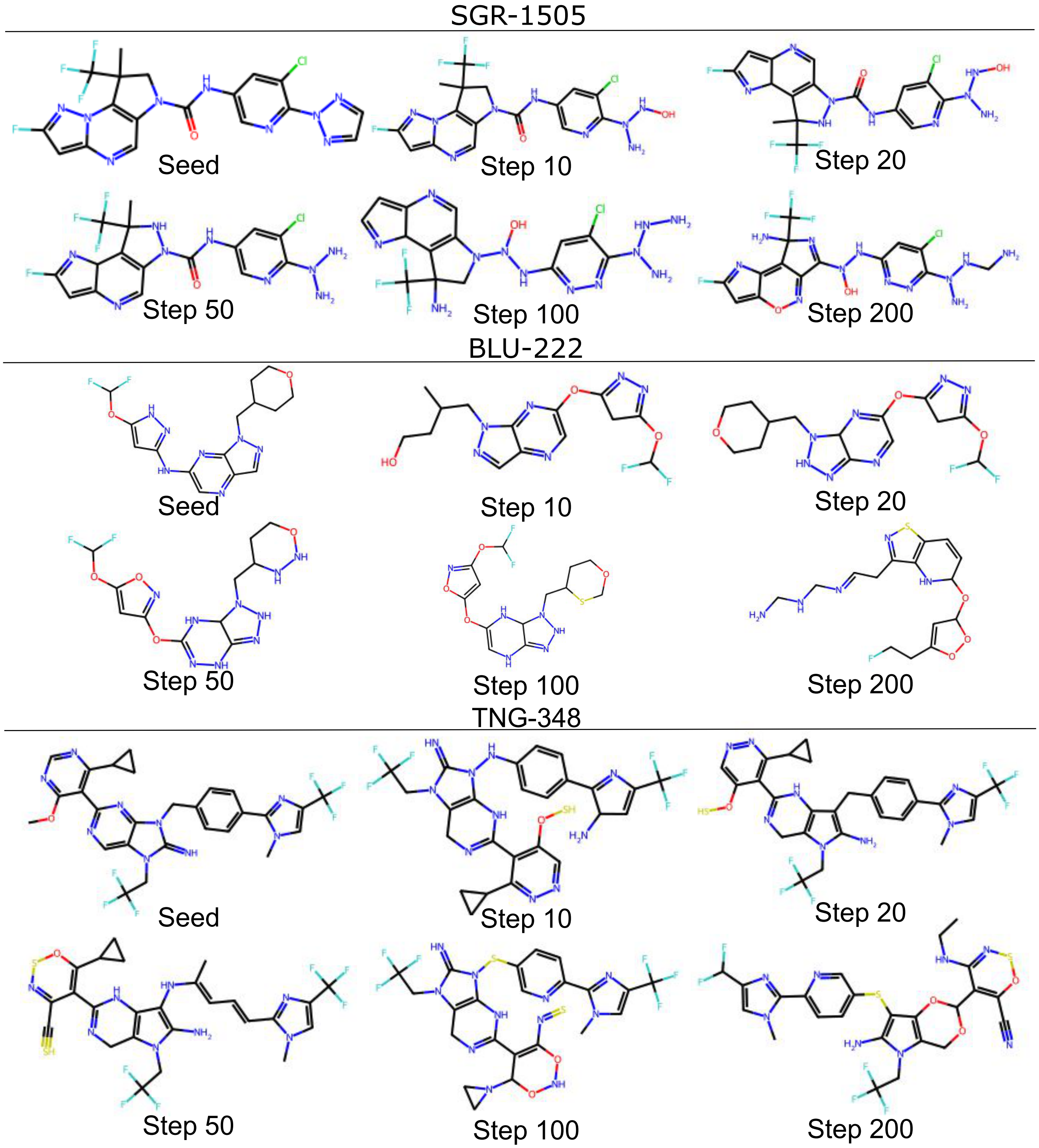

Figure 14. Additional examples of seeded generation on recently released drugs (ACS) with additional WJS steps.

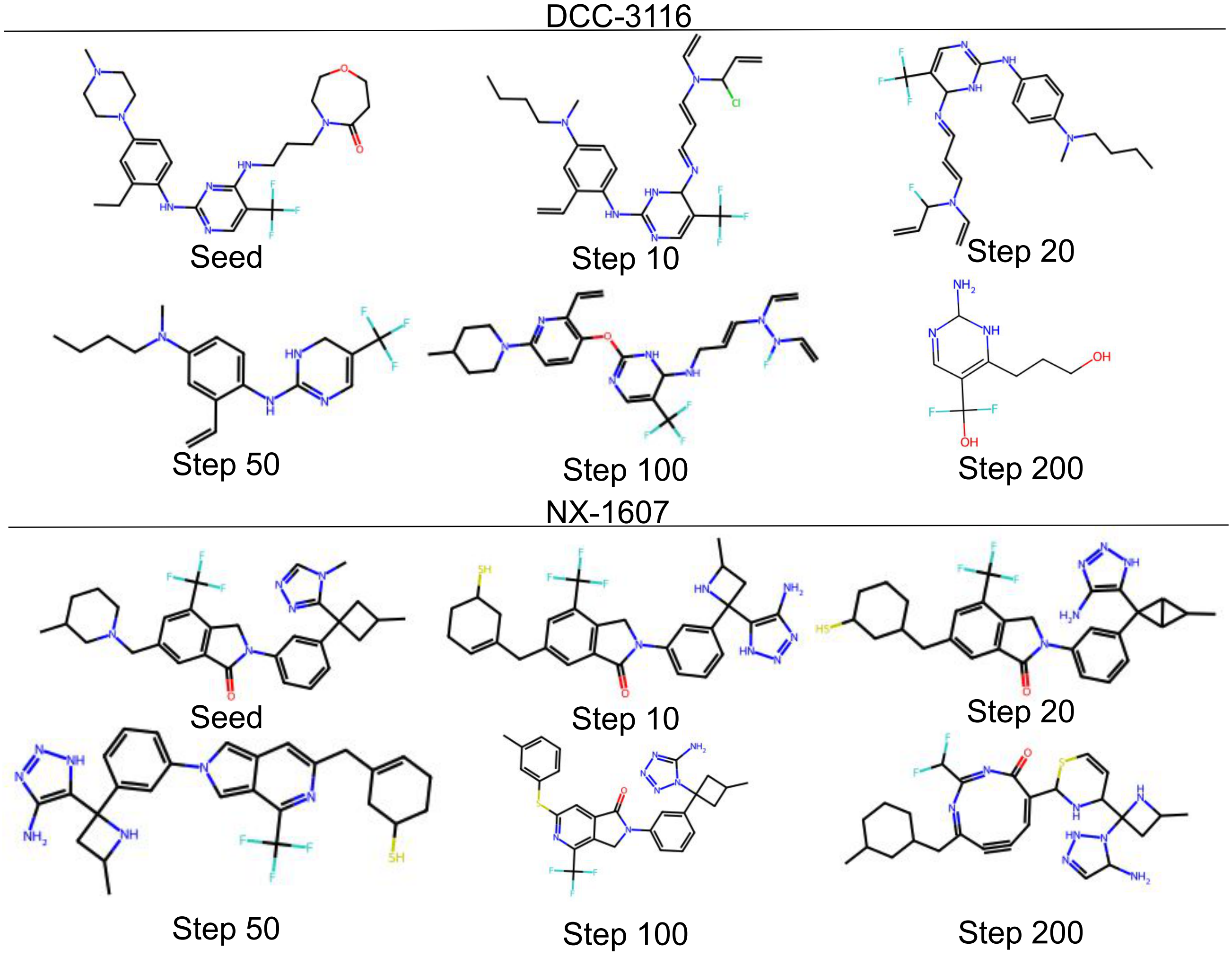

Figure 15. Additional examples of seeded generation on recently released drugs (ACS) with additional WJS steps.