Pharmacophore Guided Generation of Novel Molecules using 3D Diffusion Models

Chris Woodley
13/04/23

Stable Diffusion - Background

AI generated human faces from cutting-edge models over the past decade

- Over the past decade advancements in the availability of computing power and understanding of neural networks has increased exponentially
- An example of this is in the use of neural networks to generate photorealistic images
- Similar technologies have been applied to medical imaging and scientific simulations
- For better or for worse, a lot of the underlying scripts for these models is open source making them available for messing about with

"Dramatic watercolour painting of big bird from sesame street piloting a boat in a storm in the style of Rembrandt"

Stable Diffusion - Background

"90s cooking show hosted by Kermit the Frog"

"Jabba the Trump"

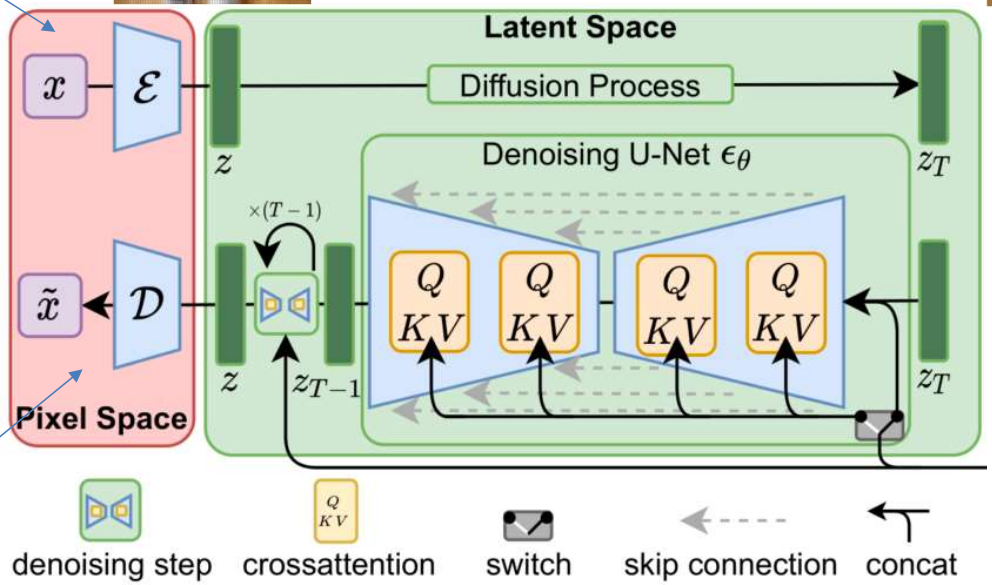
"Will Smith eating spaghetti"

- One of the most promising methods of image generation is stable diffusion
- Based on a network which unblurs blurred images
  - Give the model some noise and "trick" the model into producing an image that matches your guidance
- Widely adopted as the tool of choice for AI artists

Stable Diffusion - Background

- Image is encoded by a defined process

- Encoded representation is “blurred” – forward diffusion



- Conditioning can be used to nudge deblurring process each step

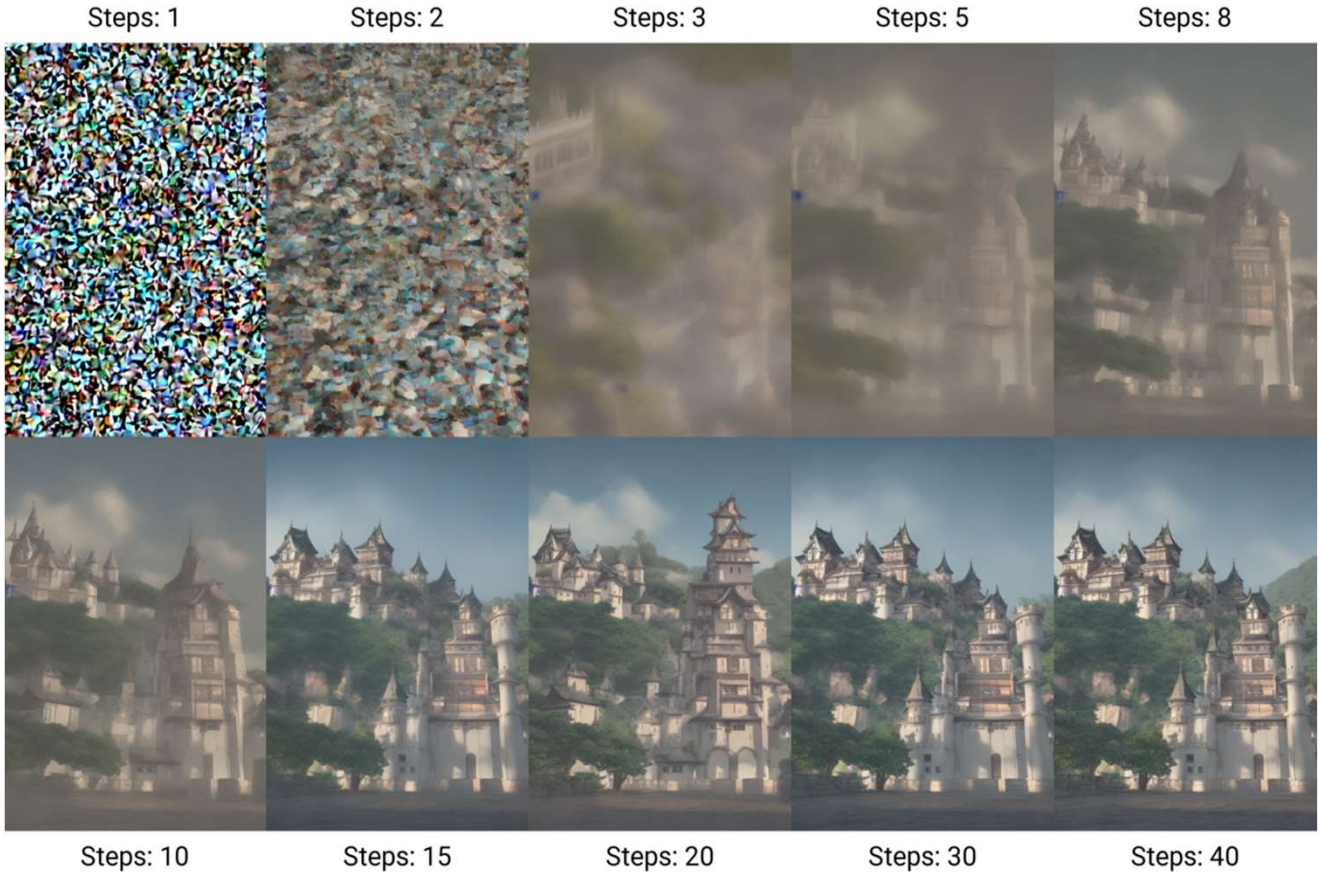
- Encoded image is reconstructed by a defined process

- U-net deblurs the representation step-by-step – reverse diffusion

arXiv:2112.10752

Stable Diffusion Sampling- Background

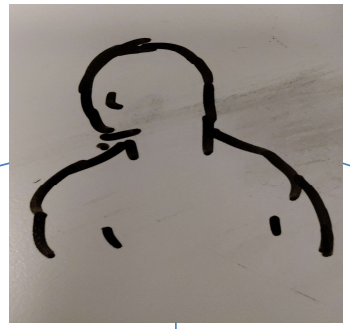
- In training the model denoises a noised image
- When we want to generate images (sampling) we input noise to the model
- Output of one pass through denoising sample mixed with previous step to move towards the generated image
- To guide the generation a second model can be used to gauge how closely an image matches the guidance
  - This model is used to modify the output to make it more closely match the guidance



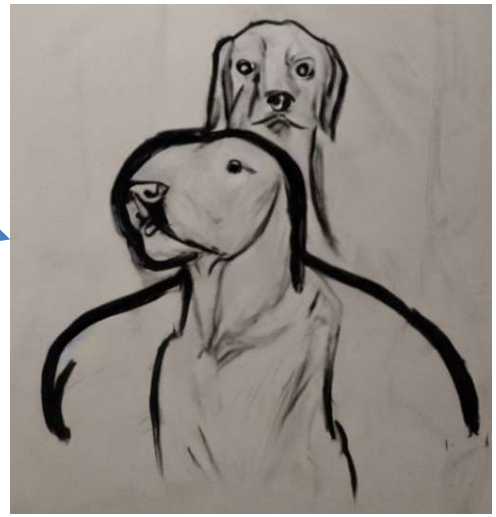
"European style castle in Japan digital artwork"

Stable Diffusion Img2Img - Background

"A charcoal drawing of Bruce Forsythe"



"A Charcoal drawing of Adrian Chiles"



"A Charcoal drawing of a dog"

- Can use an image as an input and treat it as a step in the denoising process
- Creates images which are similar to the input but with the characteristics of the guidance

Chemistry Applications?

- In another project we used Ligdream to generate novel molecules for virtual screening
  - Variational autoencoder (VAE) based
  - Encodes a voxel representation of a molecule
  - Adjusts the encoded representation
  - Decodes to produce a new representation
  - A second network is used to convert this to a SMILES string
- Useful because molecules of similar shapes with similar distributions of pharmacophores are likely to have similar bioactivities
- For image generation unet based diffusion networks have been shown to outperform VAE networks
- Can we use stable diffusion to directly generate the voxel representation from noise?

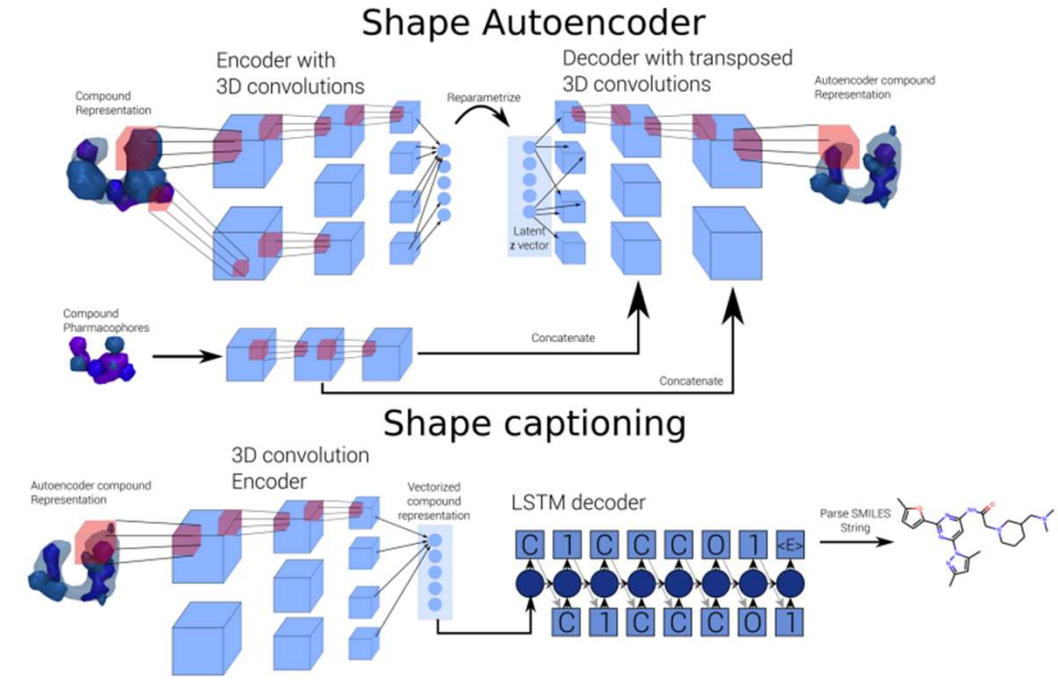


Figure 1. Proposed compound generating pipeline consisting of (top) a shape autoencoder and (bottom) a shape captioning network.

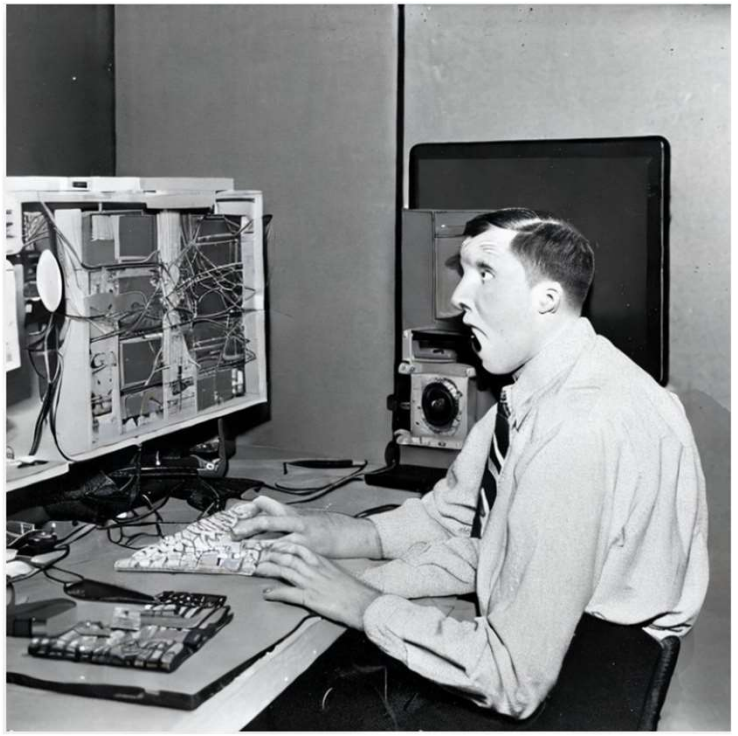
J. Chem. Inf. Model. 2019, 59, 3, 1205−1214

ChemMedChem **2019**, *14*, 1610

Current Medicinal Chemistry, 15, 10, 2008, pp. 1018-1024(7)d

Why bother?

- Ligdream relies on the original shape to generate similar/druglike shapes
  - Novelty comes from modifying latent representation
- Using diffusion we can use guidance to nudge the denoising process
  - Novelty comes from uncertainty in the denoising process
- Using diffusion we can also generate from a true voxel representation with noise added
  - Similar to image to image generation
- Guided generation of 3D volumes using diffusion networks has not yet been reported
- (I'd already wasted an entire afternoon messing generating funny pictures and wanted a way to make this relevant to my job to make myself feel better)



"Computer programmer annoyed at his useless AI, photograph 1940s"

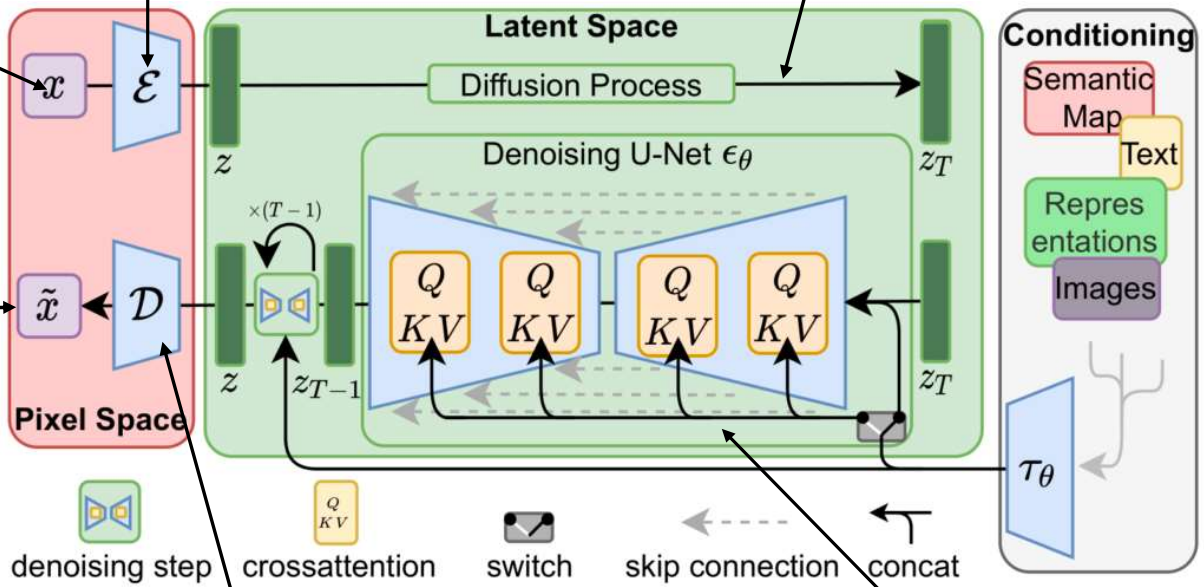
This work – analogy to stable diffusion

Generate representation and pharmacophore voxels using RDKIT and moleculekit

Add noise with ddpm scheduler

SMILES input

Use pharmacophore representation to guide denoising



Predicted smiles

Analysis

- Novelty
- Synthesisability
- Docking scores
- Similarity metrics

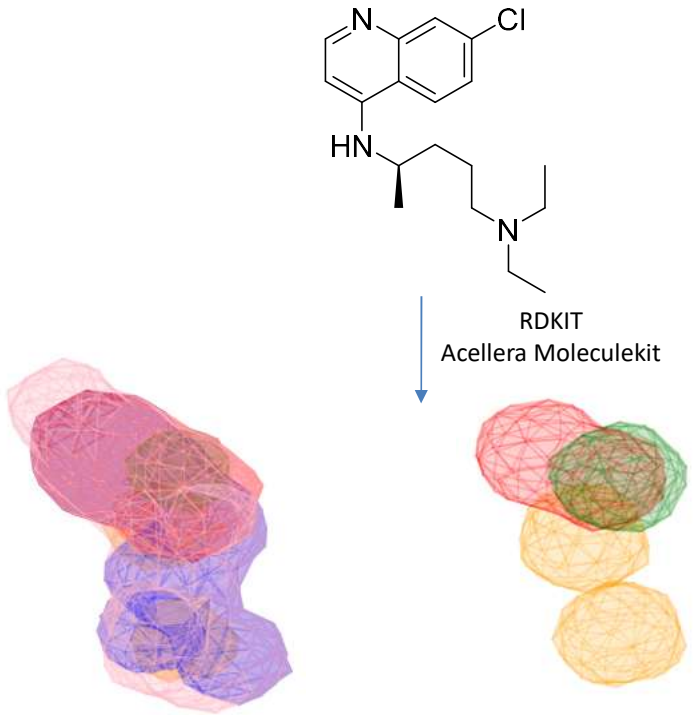
Can use captioning networks from ligdream (and weights in the short term)

Denoise representation

arXiv:2112.10752

Voxel Representations of molecules

- 3D volumes are just images with an extra dimension
- 5 Channel molecule representation
  - Hydrophobic (blue)
  - Aromatic (red)
  - H-acceptors (green)
  - H-donors (orange)
  - VdW occupancy (pink)
- 3 Channel pharmacophore representation (2 Å diameter spheres)
  - Aromatic ring centres
  - H-acceptors
  - H-donors
- Molecule coordinates randomly rotated and translated

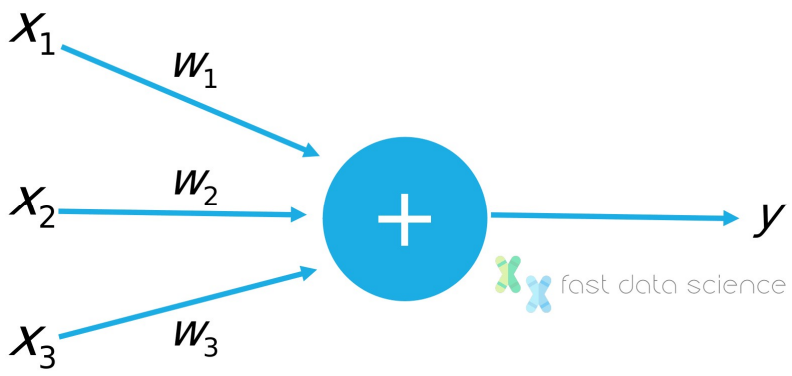


Isosurface representation of molecule and pharmacophore voxels

J. Chem. Theory Comput. 2016, 12, 4, 1845−1852

Training – Overview

- U-nets are a type of neural network
  - Neural networks have weights which determine whether or not the neuron will “fire”
  - Need to train neural network to set these weights for a specific purpose
  - Weights are set with the objective of minimizing a loss function (a measure of how good a network is at doing its task)
- In this work we repurpose a 3D U-net architecture used for medical imaging analysis
- To use our U-net to generate molecules we need to train it to denoise corrupted molecule representations
  - We use a loss function to compare the reconstructed representation to the original representation



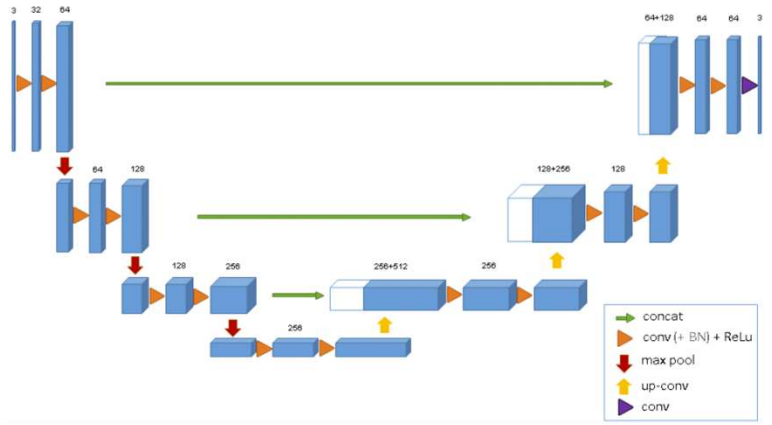


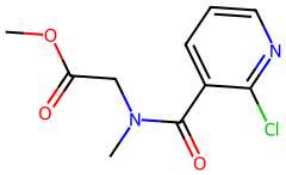
Fig. 2: The 3D u-net architecture. Blue boxes represent feature maps. The number of channels is denoted above each feature map.

https://arxiv.org/abs/1606.06650

Training – Results

Blue - hydrophobic, red – aromatic, green – H-acceptors, orange – H-donors, pink – VdW occupancy

- Initial hurdle was to see whether our network can denoise 3D molecule representations
- Models trained on druglike molecules from the Zinc15 database
- Right – model performance at reconstructing corrupt representations after training on different numbers of molecules
- We see that after 2000 batches (256,000 unique drug molecules) the model does a reasonable job of reconstruction
- Not perfect, but this doesn't matter
- Supports use of these models in generative tasks



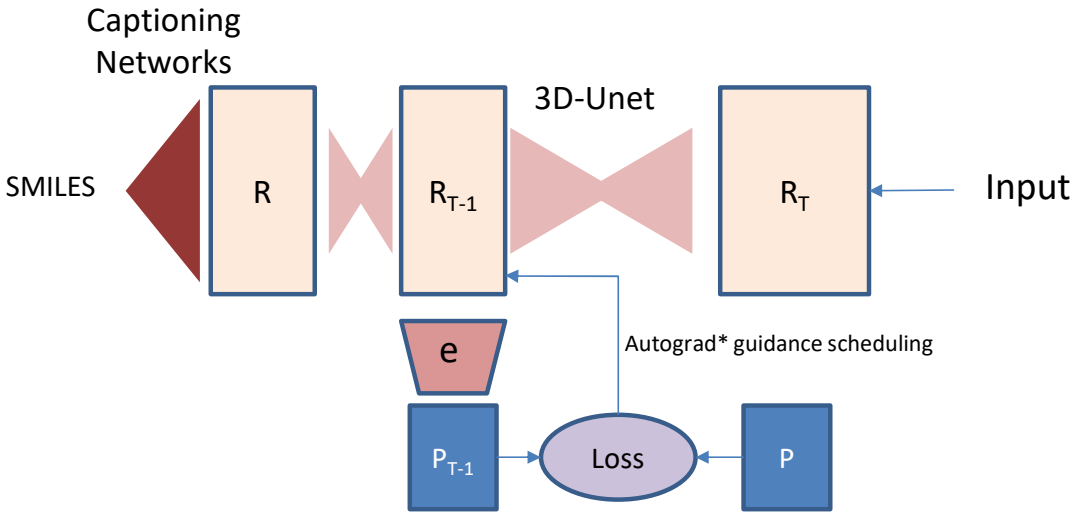
Generation of Molecular Voxels

- Two guidance methods
  - Using pharmacophores to nudge the denoising process (denoise noise, guide in denoising process)
  - Embedding the pharmacophore into the model input (denoise noise given the pharmacophores)

“a renaissance painting by Caravaggio entitled guidance”

Pharmacophore Nudging

Sampling guided by predicted pharmacophore loss

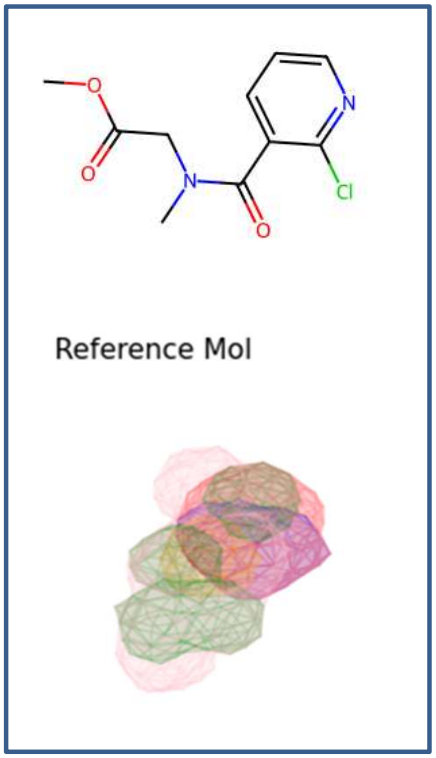


1) Input noise/ corrupted rep
2) Denoise representation with timestep embedding
3) Predict pharmacophore from intermediate representation
4) Calculate BCE loss between predicted and true pharmacophore
5) Adjust original input with autograd and mix in with denoised rep
6) Generated representaiton

x num_time steps

- Moving parts:
  - 3D U-net - denoising
  - Pharmacophore encoder – predicts pharmacophore rep from U-net output
- Parameters
  - How many denoising steps we want to take (timesteps)
  - How strong our guidance is (guidance scale)

Pharmacophore Nudging - Results

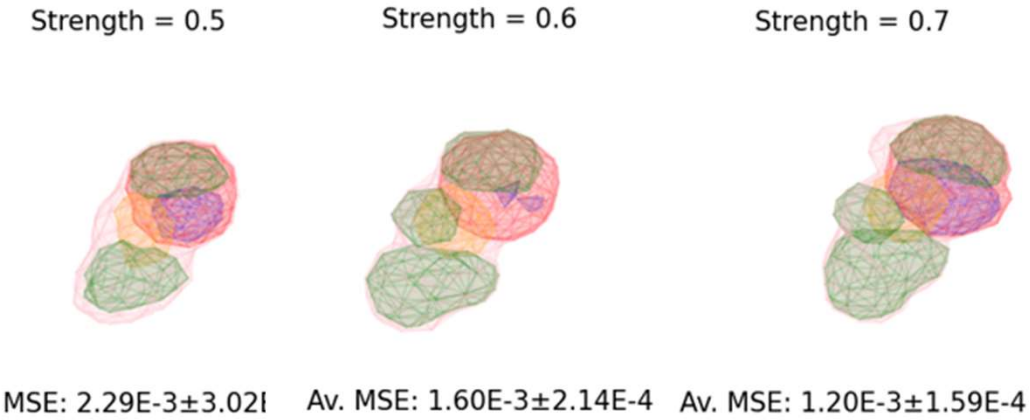


Blue - hydrophobic, red – aromatic, green – H-acceptors, orange – H-donors, pink – VdW occupancy

Generated From Noise

Three examples of generated representations at different settings

Reconstructed Corrupted Representation

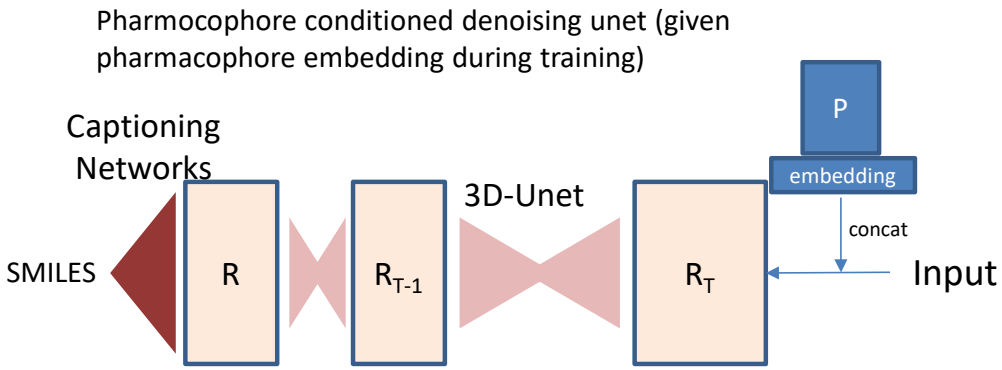


MSE = mean squared error compared to reference molecule (smaller is better)

- Generating from noise produces molecule like representations
- Struggles with hydrophobic channel
- Sensible placement and shape of pharmacophores suggest guidance methods works

- Reconstructing a corrupt voxel works better producing very similar molecules
- Not sure how necessary any guidance is for this
- Tunable similarity to parent molecule

Embedded Pharmacophore sampling

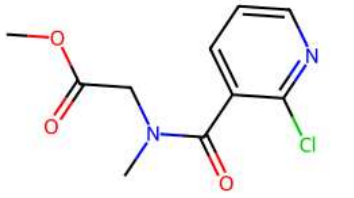


1) Input noise/ corrupted rep
2) Denoise representation with timestep and pharmacophore embedding
3) Mix in with original sample
4) Generated representaiton

X num_time steps

- Unet trained to denoise with embedding of pharmacophore representation – i.e. predicts denoised representation given pharmacophores
  - Only parameter is number of timesteps
  - Better denoising performance than network with just timestep embedding
- Sampling does not require a separate pharmacophore encoding network
  - Speeds up sampling as this removes the need for expensive loss calculations on 3D tensors
- When sampling from noise, number of timesteps is the only parameter
- When sampling from a corrupt voxel representation strength is an additional parameter

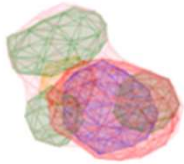
Embedded Pharmacophore- Results



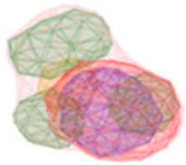
Reference Mol

Blue - hydrophobic, red – aromatic, green – H-acceptors, orange – H-donors, pink – VdW occupancy

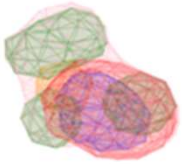
Generated From Noise



Av. MSE: 2.15E-3±1.92E-4



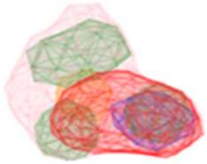
Av. MSE: 2.01E-3±1.24E-4



Av. MSE: 1.99E-3±1.09E-4

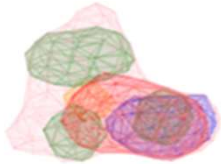
Reconstructed Corrupted Representation

Strength = 0.5



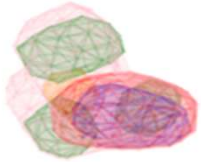
Av. MSE: 1.75E-3±2.18E-4

Strength = 0.6



Av. MSE: 1.47E-3±2.28E-4

Strength = 0.7



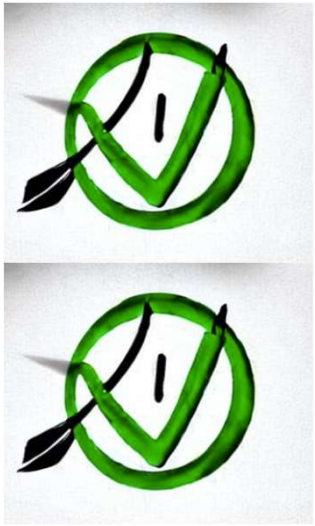
Av. MSE: 1.11E-3±1.66E-4

- With embedded pharmacophore, the models perform much better
- Hydrophobic channel better reconstructed

- Like with pharmacophore nudging, we see tunable similarity to reference molecule

Recap

- Does our repurposed U-net effectively deblur 3D volumes?
- Can we generated molecule like representations from noise?
- Can we generate useable SMILES strings from these representations?

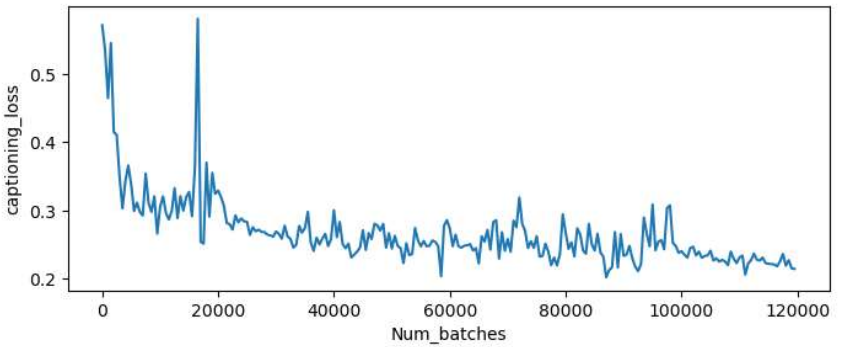


"check mark, green, drawn with a fountain pen"

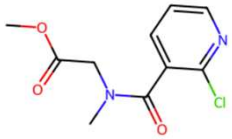
"cross symbol, red, drawn with a fountain pen using human blood as ink"

Captioning Networks

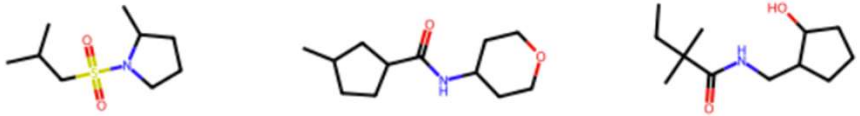
- Networks designed to turn a 3D shape into a useable SMILES string
- Same architecture as ligdream
  - Trained alongside other networks with generated representations
  - Objective is the minimize the difference between the SMILES string and predicted SMILES string
- Captioning networks trained on U-net generated molecules (~1.4 million molecules)
  - Trained on denoised reps with small amount of noise added
  - Generates valid SMILES strings and molecules
  - Not consistent with input molecule

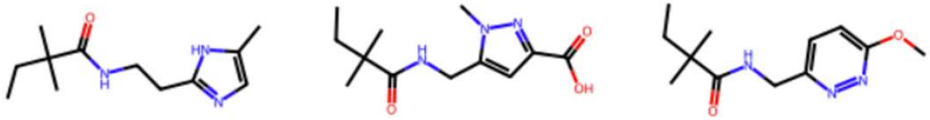


Ref mol



Generated molecules

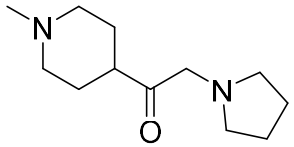




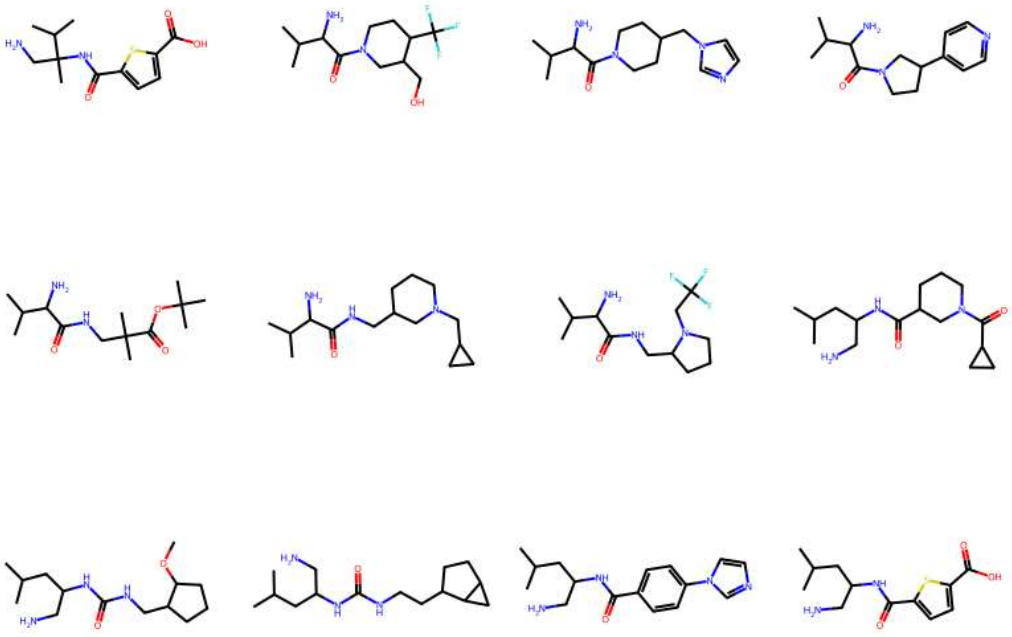
Generated molecules using networks with pharmacophore embedding

Captioning Networks - Examples

Input



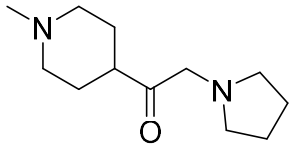
Generated Molecules



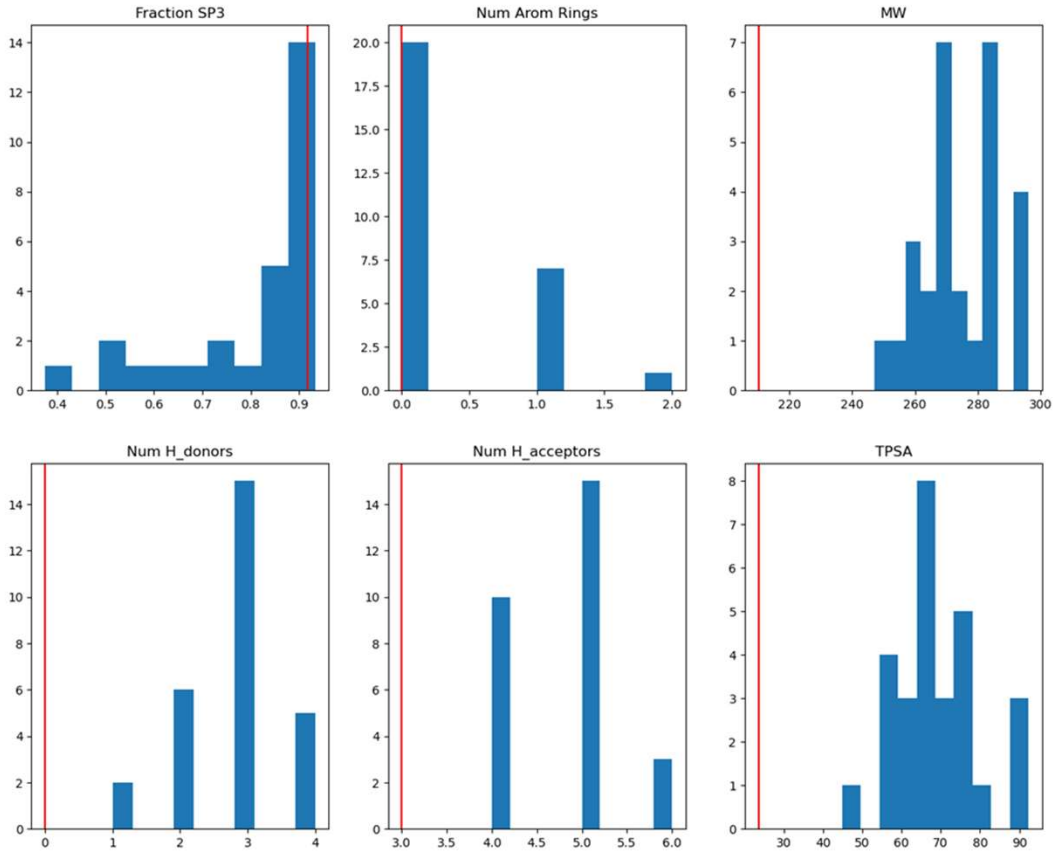
- Can see some similarities between input and generated molecules
- Not too convincing – ligdream produces much more similar molecules

Captioning Networks - Examples

Input



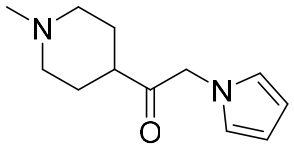
Distribution of Properties



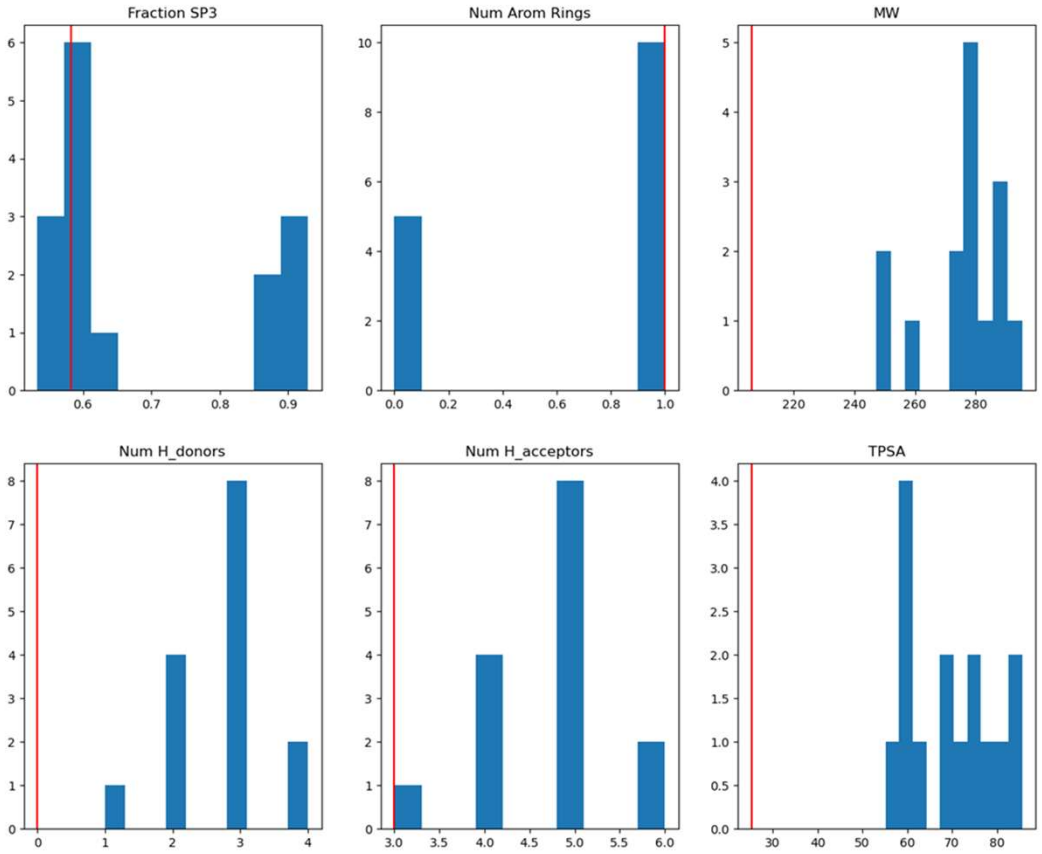
- Distribution of properties not too convincing
- Exception of fraction SP3 and number of aromatic rings

Captioning Networks - Examples

Input



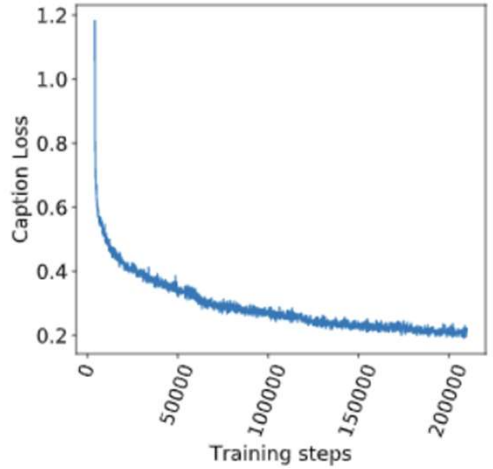
Distribution of Properties



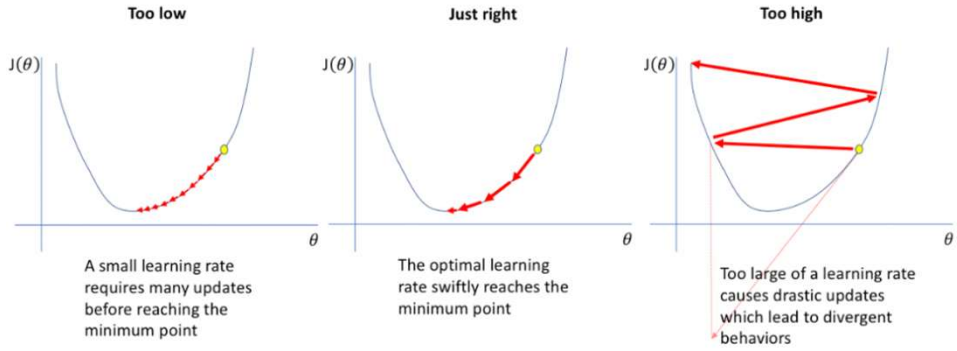
- When an aromatic ring is added the distribution of fraction SP3 and number of aromatic rings shifts

Why are these networks so bad?

- Lack of training - same networks from Ligdream's model were trained on ~29,000,000 compounds and appeared to still be "learning" after a large number of molecules
- This type of network (LSTM networks) are known to be sensitive to learning rate
  - When training the learning rate is adjusted after certain number of molecules have been seen by the model
  - Reducing the learning rate over time allows "fine tuning" of the model
  - At this point in training the learning rate may be too high to pick up on nuanced differences in representation
- Human error?



Plot of Ligdream captioning loss by number of training steps



J. Chem. Inf. Model. 2019, 59, 3, 1205−1214

Summary

- Does our repurposed U-net effectively deblur 3D volumes?
  - We have shown that our repurposed medical imaging U-net is capable of deblurring volumes
- Can we generate molecule-like representations from noise?
  - We have been able to generate reasonable, molecule-like representations from noise and corrupted representations
  - We have shown that from corrupted representations we can tune the similarity between the parent structure and the generated representation
- Can we generate useable SMILES strings from these representations?
  - At present the captioning networks are unable to produce reasonable analogues of parent drug molecules

Future Work

- Double check training scripts for the captioning networks and train more
- Transformer networks (e.g. Chat GPT) have broadly replaced LSTM networks in sentence predictions tasks
  - Currently working on replacing the shape captioning networks with transformers
- Alternative methods of generating molecules
  - Can we use a representation to generate molecules from noise directly without a captioning networks?
- Can we repurpose this code for other volume generation tasks?
  - No published work on the generation of shapes using denoising U-nets
  - Labelled datasets exist with labelled 3D structures; papers have been published using other methods to train models which can effectively generate chairs, cars and planes from noise

Thanks For Listening!

“A man made entirely from ears, high resolution photograph”