Generative AI Design and Exploration of Nucleoside Analogs

Damien A. Dablain[+], Geoffrey H. Siwo[+]*, Nitesh V. Chawla*

[1]Department of Computer Science and Engineering and the Lucy Family Institute for Data and Society, University of Notre Dame; Notre Dame, IN, 46556, USA.

[2]Department of Biological Sciences, Center for Research Computing, Eck Institute for Global Health, University of Notre Dame; Notre Dame, IN 46556, USA.

[+]Authors with equal contribution

*Corresponding authors. Email: gsiwo@nd.edu; nchawla@nd.edu

Abstract: Nucleosides are fundamental building blocks of DNA and RNA in all life forms and viruses. In addition, natural nucleosides and their analogs are critical in prebiotic chemistry, innate immunity, signaling, antiviral drug discovery and artificial synthesis of DNA / RNA sequences. Combined with the fact that quantitative structure activity relationships (QSAR) have been widely performed to understand their antiviral activity, nucleoside analogs could be used to benchmark generative molecular design. Here, we undertake the first generative design of nucleoside analogs using an approach that we refer to as the Conditional Randomized Transformer (CRT). We also benchmark our model against five previously published molecular generative models. We demonstrate that AI-generated molecules include nucleoside analogs that are of significance in a wide range of areas including prebiotic chemistry, antiviral drug discovery and synthesis of oligonucleotides. Our results show that CRT explores distinct molecular spaces and chemical transformations, some of which are similar to those undertaken by nature and medicinal chemists. Finally, we demonstrate the potential application of the CRT model in the generative design of molecules conditioned on Remdesivir and Molnupiravir as well as other nucleoside analogs with in vitro activity against SARS-CoV-2.

One-Sentence Summary: Generative design of nucleoside analogs relevant to antiviral drug discovery, prebiotic chemistry and synthetic biology.

Introduction: Genetic information of all life forms and viruses is stored in nucleic acid sequences which are linear polymers synthesized in cells by the polymerization of four distinct nucleotide triphosphates - adenosine (ATP), guanosine (GTP), cytidine (CTP), thymidine (TTP) and uridine (UTP, which replaces TTP in RNA). Each nucleotide consists of a purine or pyrimidine nitrogenous base attached to a ribose sugar (deoxyribose for DNA) together forming a nucleoside, linked to a triphosphate group. In addition to their role as the building blocks of DNA and RNA,

nucleotides and nucleosides play critical roles in several biological processes including energy storage (ATP) (*1*), signaling pathways (e.g. cyclic adenosine monophosphate, cAMP) (*2*, *3*), innate immunity (e.g. cyclic guanosine monophosphate-adenosine monophosphate, cGAMP) (*4*, *5*), as cofactors for specific enzymes (e.g. s-adenosylmethionine, SAM) (*6*, *7*) and as neurotransmitters (e.g. adenosine) (*8*). Chemical analogs of nucleotides and nucleosides therefore have a wide range of applications, for example as substrates for artificial synthesis of DNA/RNA sequences with increased resistance to nucleases (*9*, *10*) or reduced effect on innate immune responses (*11*, *12*), inhibition of viral DNA and RNA polymerases (*13*, *14*), inhibition of tumor cell replication (*15*, *16*) and treatment of autoimmune diseases (*17*). Nucleoside analogs are also critical in understanding the early evolution of life on earth under prebiotic conditions (*18–20*) and could provide a means of engineering synthetic organisms with expanded genetic alphabets (*21*). Thus, exploration of molecular space of nucleoside analogs could have a huge impact on several areas of fundamental and translational biology.

To that end, we undertake, to the best of our knowledge, the first computational design of nucleoside analogs in which the resulting molecules are designed de novo using artificial intelligence (AI) models, an approach that could potentially accelerate the design of new nucleoside analogs. Recently, deep generative models that use neural networks have been employed to search chemical space(s) and identify promising drug candidates (*22*). The family of deep generative models consist of generative adversarial networks (GAN) (*23*, *24*), adversarial autoencoders (AAE) (*25*), variational autoencoders (VAE) (*26*), Long Short-Term Memory (LSTM) (*27*) and Gated Recurrent Unit (GRU) models (*28*). We present a benchmarking study of five previously reported, general purpose deep generative models to assess their performance in generating nucleoside analogs when conditioned on a small set of natural nucleosides.

Deep generative models are typically trained on a large corpus of molecules. These models can either use a SMILES or graphical representation of molecules to learn the rules of chemical construction. For example, given the first four atoms and bonds of a molecule in SMILES format, such as "Cc1c", the model learns to generate a chemically valid molecule or "Cc1ccccc1" (Toluene). A deep generative model trained on a large corpus of molecules may learn to generate valid molecules that are diverse; however, the molecules may not contain specific drug-like properties. For example, in drug discovery, chemists are often interested in exploring a distinct area of chemical space, where molecules have desired properties and structure. We refer to the task of generating molecules with specific structural attributes as conditional generation.

When faced with the task of conditional molecule generation, two potential issues arise: generating molecules that are structurally close to a target class (here, the nucleosides), but that are diverse. These twin goals – generating diverse molecules with desired properties – are diametrically opposed. Models that yield highly unique compounds may not produce molecules that are similar to the reference set.

The task of conditional molecule generation can be approached with transfer learning (also known as fine-tuning) or, alternatively, direct steering of desired chemical properties *(22)*. The transfer learning approach trains a model on a large corpus of data, and then fine-tunes the model with a smaller, more focused dataset. The idea behind transfer learning is that a model can learn the general rules of molecular construction from a large, widely available dataset and then transfer this general knowledge to a specific task. For example, a deep generative model can first be trained on a large corpus of molecules and then "fine-tuned" on a smaller dataset of compounds (the nucleosides), where the goal is to generate chemically valid molecules (learned from the larger dataset) that are similar in structure to the molecules in the smaller dataset (the nucleosides). This approach is potentially constrained by the problem of catastrophic forgetting (*29*), or the tendency of the weights of a neural network to "forget" the patterns learned from the training dataset when they are subsequently fine-tuned on another dataset, which may contain a different distribution of examples.

Another approach to conditional molecular generation is direct steering. Direct steering simultaneously trains a model to learn the rules of valid chemical generation and to focus on a specific property, using a large dataset. During training, this approach provides two inputs to a generative model – a representation of the molecule and one of its properties. The model learns to associate certain properties (e.g. drug-likeness) with a particular molecular representation and to generate valid molecules at the same time.

To deal with the dual issues of generating focused, yet diverse molecules, we develop a transformer-based model, which we call Conditional Randomized Transformer (CRT). CRT is a decoder-only model that is based on the original architecture proposed by Radford et al. (60) and modified by Bagal et al. (35). (See Fig. 1 for a high-level depiction of CRT's architecture and the Materials and Methods section for a more complete description.)

We incorporate *both* fine-tuning and direct steering into our training pipeline to address the issue of generating molecules with desired properties and demonstrate that CRT successfully addresses the catastrophic forgetting issue. CRT incorporates direct steering by conditioning on molecular structure with Morgan fingerprints. This is important because Morgan fingerprints capture important chemical relationships between molecules that are correlated with bioactivity. CRT also uses a training pipeline that first learns the rules of chemical construction from a large dataset and then transfers this learning via fine-tuning to smaller datasets consisting of only nucleoside molecules. Unlike other models, which freeze neural network layers when fine-tuning, we show that CRT works best when the layers are not frozen. We demonstrate that the reason why this technique works is because of the way in which the model updates network parameters.

To address the issue of generating chemically diverse molecules, we infuse diversity into the generation process by randomizing the Morgan fingerprints of targeted molecules. Application of

CRT to design nucleoside analogs conditioned on natural or synthetic nucleosides shows that it generates molecules that are similar or identical to nucleoside analogs with a wide range of biological significance, including in basic biology and medicinal chemistry. To facilitate further exploration and critique by the community, we publicly provide the most promising nucleoside analogs generated by our model.

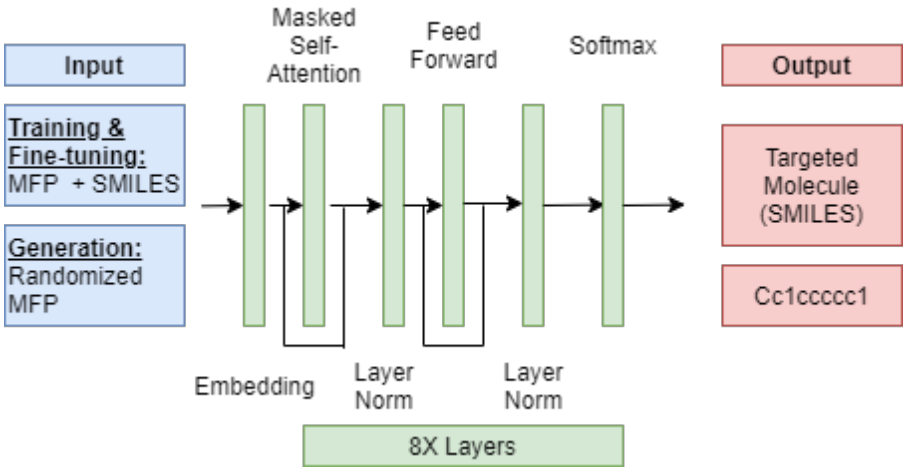


Fig. 1. CRT high-level architecture. The CRT implementation consists of 3 phases: (1) training, (2) fine-tuning, and (3) generation. All 3 phases share a similar architecture, except as follows. (1) During training, a Morgan fingerprint (MFP) and SMILES representation are input. A large, augmented dataset is used. The model is trained with cross-entropy loss. During training, CRT learns the rules of valid chemical construction and to associate a Morgan fingerprint (molecular structure) with a SMILES representation of a compound. (2) Once trained, the model is fine-tuned on a smaller dataset that is focused on a certain area of chemical space (e.g., the nucleosides). During fine-tuning, both Morgan fingerprints and SMILES representations of the targeted molecules are input. (3) During generation, only Morgan fingerprints are used. The Morgan fingerprints are randomized, subject to a hyper-parameter, to encourage diversity. The combination of specific structural cues (the Morgan fingerprints), the fine-tuning process and randomization allows CRT to generate focused molecules that are also diverse.

Results

Model benchmarking on parent natural nucleosides and synthetic nucleoside analogs demonstrates variable performance of models

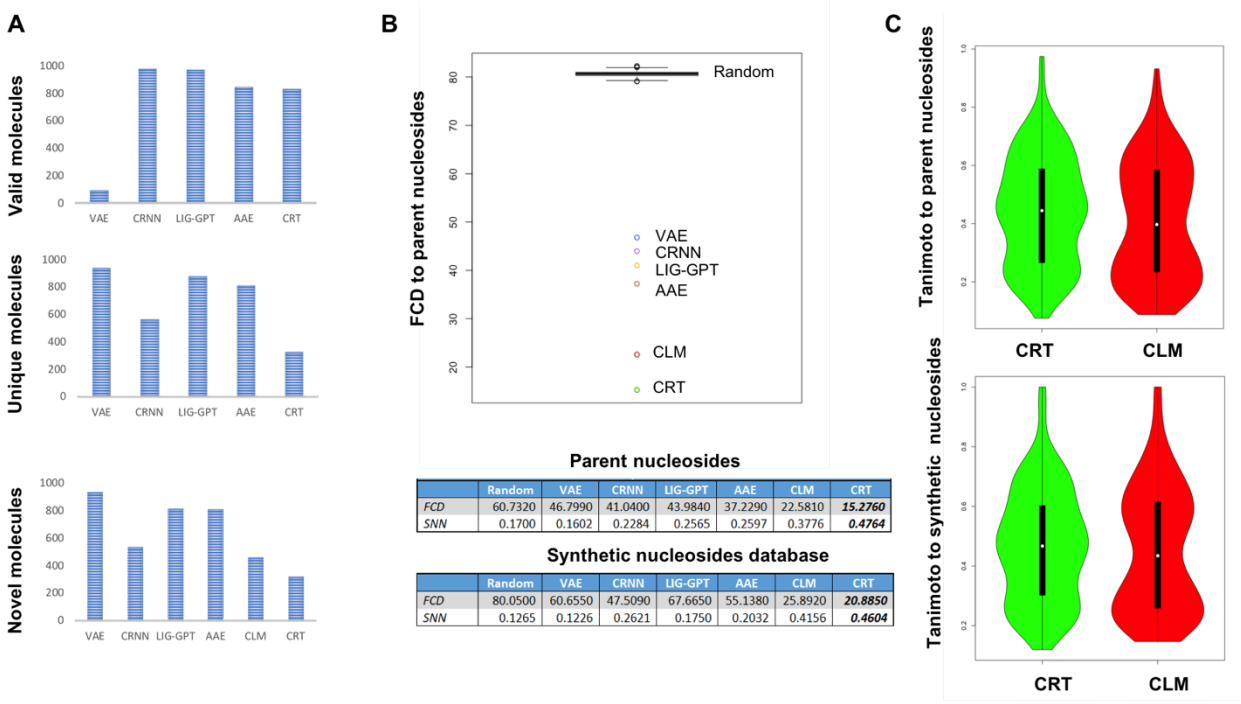
We first benchmarked five previously reported generative molecular design models namely: a Variational Autoencoder (VAE) (*30*, *31*), two Long Short Term Memory models (LSTM) – CLM (*32*) and CRNN (*33*), an Adversarial Autoencoder (AAE) (*34*) and a base Transformer model, LIG-GPT (*35*) (see detailed descriptions in Methods). The five benchmark models were originally tested or applied to non-nucleoside datasets (e.g., TRPM8 ligands, MEGX, *Plasmodium falciparum*, or optimized on specified properties, such as the quantitative estimate of drug-

likeness). In addition, we developed a transformer-based model - the Conditional Randomized Transformer (CRT), which during the fine-tuning step, conditions molecular generation using extended-connectivity fingerprints (commonly referred to as ECFPs or Morgan fingerprints) (*36*). Each of the models were first trained on a large database of small molecules that was filtered based on bioactivity (Low Dataset, Moret et al (*33*)). The models were then fine-tuned on five natural nucleosides - henceforth referred to as the Parent Nucleosides (adenosine, guanosine, cytidine, uridine, thymidine). After training and fine-tuning, each of the models was applied in generating 1,000 molecules (see Supplementary Tables S1 to S6), followed by an assessment of their novelty, validity, uniqueness and chemical similarity to the Parent Nucleosides relative to randomly sampled molecules from the training set (Fig. 2).

Out of the six models, CRNN, LIG-GPT and AAE generated the highest proportion of valid, unique and novel molecules (see Materials and Methods section for metrics; Fig. 2 A; Supplementary Table 8). However, their generated molecules are less chemically similar to the Parent Nucleosides, as measured by the Similarity of Nearest Neighbor (SNN) and Frechet ChemNet Distance (FCD) metrics (Fig. 2 B; see Materials and Methods section for details on SNN and FCD estimation). Therefore, in terms of generation conditioned on the Parent Nucleosides, CRT and CLM are the top performing models, with CRT producing the best similarity metrics (i.e., highest SNN and lowest FCD).

Because we are interested in generating molecules that are chemically similar to the Parent Nucleosides, we focused on a deeper assessment of the molecules produced by the CRT and CLM models. For the Parent Nucleosides, both models produced approximately the same number of molecules with a Tanimoto coefficient greater than ~ 0.7 (34 and 30 for CLM and CRT, respectively); although CRT produced a higher number of molecules similar to molecules in the PubChem database (16 vs. 10). (Fig. 2 C; See Supplementary Table 9 and 10 for a listing of unique molecules generated by CRT and CLM based on the Parent Nucleosides, the individual Tanimoto coefficient of each molecule compared to their nearest neighbor in the reference set, and their PubChem identifier - CID, if applicable). In general, even though CRT and CLM generated structurally distinct molecules, the generated molecules in both cases showed a similar profile to the fine-tuning sets (Fig. 2 C; Wilcoxon test P-value = 0.01 for comparison between CRT and CLM based on similarity of generated molecules to parent nucleosides.

To provide a more rigorous assessment of the nucleoside analogs generated by CRT and CLM, we obtained a set of 188 synthetic nucleoside analogs from Selleckchem (the Synthetic Nucleosides) that have a high structural diversity, medicinal activity, cell permeability and have rich structure and bioactivity information (*37*). These nucleoside analogs are frequently used in high throughput screening and high content screening. At a threshold of Tanimoto coefficient greater than ~ 0.7, there were 61 CLM generated molecules that were similar to the set of synthetic nucleoside analogs compared to 41 CRT generated molecules (P = 0.19 for synthetic nucleosides; Fig. 2 C).



	Random	VAE	CRNN	LIG-GPT	AAE	CLM	CRT
FCD	60.7320	46.7990	41.0400	43.9840	37.2290	22.5810	***15.2760***
SNN	0.1700	0.1602	0.2284	0.2565	0.2597	0.3776	***0.4764***

	Random	VAE	CRNN	LIG-GPT	AAE	CLM	CRT
FCD	80.0500	60.6550	47.5090	67.6650	55.1380	25.8920	***20.8850***
SNN	0.1265	0.1226	0.2621	0.1750	0.2032	0.4156	***0.4604***

Fig. 2. Summary of performance of models. **A**. Validity, uniqueness and novelty of molecules generated by each model. CLM validity and uniqueness were not directly determinable (see Methods for details). **B.** Assessment of chemical similarity of generated molecules to the parent nucleosides using FCD and SNN relative to random background distribution. **C.** Tanimoto coefficient distributions to Parent Nucleosides and Synthetic Nucleosides.

CRT generated molecules conditioned on Parent Nucleosides contain chemical alterations observed in natural and synthetic nucleosides analogs

To gain deeper insights into the relationships between the CRT generated molecules and the Parent Nucleosides, we focused on generated molecules with a high similarity to the Parent Nucleosides (30 molecules, Tanimoto coefficient > 0.7). At this threshold, 28 of the CRT generated molecules were identical or similar to molecules with an existing structure in PubChem, including to naturally occurring and synthetic analogs of adenosine, guanosine, uridine, thymidine, hypoxanthine and inosine, as well as potential prebiotic nucleosides (Supplementary Table 9). We manually inspected the generated structures to identify chemical differences between the generated molecules to the Parent Nucleosides (Fig. 3).

Among the generated molecules, we observed chemical transformations involving both changes in the ribose moiety of nucleosides as well as some involving altered purine or pyrimidine rings (Fig. 3). For example, in five of the molecules, the ribose moiety was replaced with an oxetane

ring. Two of these oxetane containing nucleoside analogs are shown in Fig. 3 in which the oxetane is connected to a purine (Fig. 3 A) or pyrimidine ring (Fig. 3 B). The generated molecule 2-(6-aminopurin-9-yl)-4-(hydroxymethyl)oxetan-3-ol (Fig 3 A) is similar to oxetanocin A, an antiviral drug first isolated in bacteria (*38*) while 1-[4-(hydroxymethyl)oxetan-2-yl]-5-methyl-pyrimidine-2,4-dione (Fig. 3 B) is similar to A-73209, an oxetanocin with antiviral activity against Herpes Simplex Virus and Varicella Zoster Virus (*39*). It is interesting to note that the oxetane ring occurs in relatively few natural products, though in many cases when it occurs plays an important role in biological activity (*40*). Furthermore, due its small, polar nature, medicinal chemists have on several occasions incorporated oxetane rings into specific drug candidates to enhance their drug-likeness and gain intellectual property (*40*). Thus, the ability of the CRT model to generate oxetane containing nucleoside analogs given a fine-tuning set of natural nucleosides demonstrates its potential in drug discovery.

In addition to alterations in the ribose moiety, we observed changes in purine and pyrimidine rings (Fig. 3 C and D). One of the generated molecules (2-(2,6-diaminopurin-9-yl)-5-(hydroxymethyl)tetrahydrofuran-3,4-diol; Fig. 3 C) is identical to 2,6-diaminopurine riboside, which occurs naturally in cyanophage S-2L genome in place of adenine and is unusual in that it forms complementary base pairs with thymine involving 3 hydrogen bonds instead of the usual 2 when thymine pairs with adenine (*41*, *42*). This additional hydrogen bond may contribute to stability of synthetic oligonucleotides containing 2,6-diaminopurine (*43*). Furthermore, this nucleoside is of particular interest in prebiotic chemistry because it can be formed abiotically alongside adenine and although is rare terrestrially, it has been isolated from meteorites (*44*), increases the rate of non-enzymatic RNA formation (*45*) and has unique electron donating properties that allow it to undergo self-repair upon UV damage (*46*). One of the generated molecules is identical to a modified guanosine residue- N2-methyl-guanosine (Fig. 3 D)- that occurs in some natural RNAs including tRNAs of archaea and eukaryotes where it stabilizes the sequences (*47*).

Collectively, these results indicate that CRT generated molecules are of significance in a wide range of areas including antiviral drug discovery, oligonucleotide synthesis and prebiotic chemistry studies.

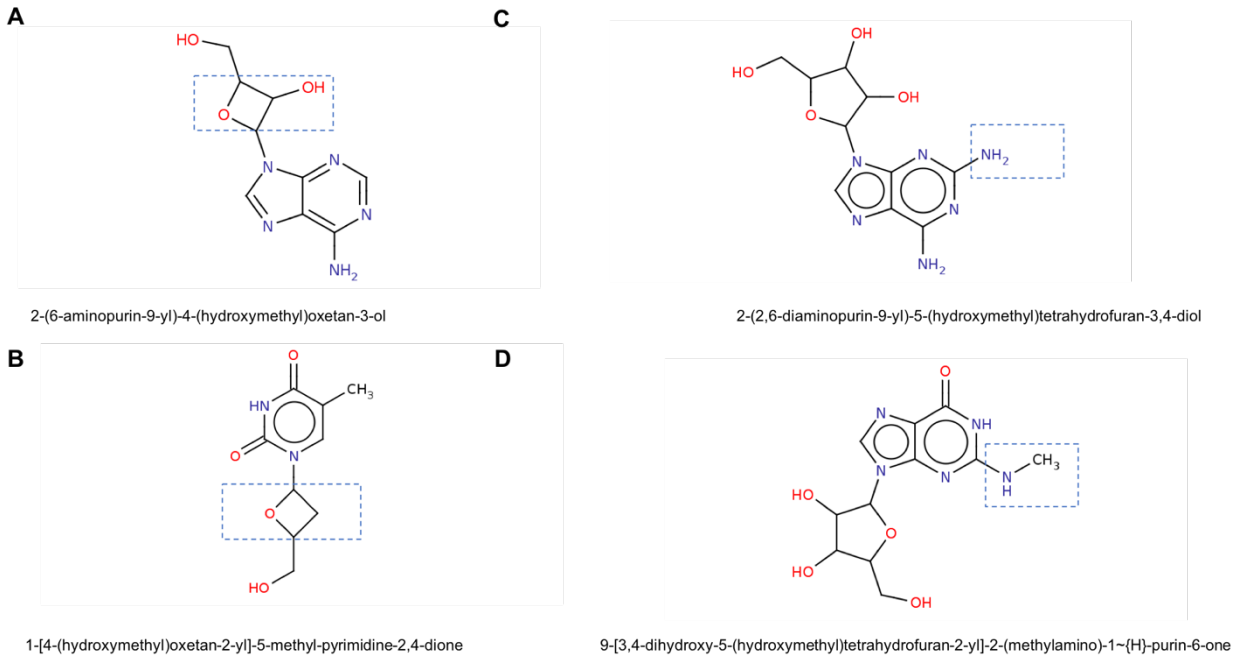


Fig. 3. Depiction of select CRT generated nucleoside analogs of distinct functional significance in fundamental and translational biology. **A.** Generated nucleoside analog similar to adenosine but with an oxetane ring instead of ribose which makes it more related to Oxetanocin A. **B.** Generated nucleoside analog similar to A-73209, an oxetanocin with antiviral activity. **C.** Generated molecule identical to 2,6-diaminopurine Riboside. **D.** Generated molecule identical to N2-methyl-guanosine which is present in some natural RNAs. Highlighted region of each molecule shows the location of the alteration relative to parental natural nucleosides. Structures generated using the ChemDB portal (*48*) and visualizations based on ChemAXon's MarvinView and MarvinSketch. Standard names generated using OpenEye's LexiChem.

Application of CRT to focused generation of nucleoside analogs fine-tuned on anti-SARS-CoV-2 Nucleosides produces diverse yet structurally similar molecules

Next, we applied CRT to condition the generation of molecules on a set of nucleoside analogs that have reported in vitro activity against SARS-CoV-2 (*49*) (the SARS-CoV-2 Nucleosides – see Methods and Materials section). We hypothesize that nucleoside analogs targeting the SARS-CoV-2 RNA-dependent RNA polymerase contain implicit chemical information that contributes to their activity against the enzyme. Thus, searching the molecular space that optimizes the properties of the anti-SARS-CoV-2 Nucleosides could lead to more inhibitors against the viral enzyme.

For the anti-SARS-CoV-2 Nucleosides, CRT produced 74 molecules with Tanimoto coefficient greater than 0.7 when compared to the reference set versus 17 for CLM (Fig. 3). Nine of the CRT

generated molecules were similar to molecules listed in PubChem compared to 4 for CLM (See Supplementary Table 11 and 12). In contrast, 29 CRT generated molecules were similar to the Synthetic Nucleosides at this threshold while 28 CLM generated molecules were similar to the Synthetic Nucleosides (Supplementary Table 13 and 14). Notably, CRT generated molecules following fine-tuning with the SARS-CoV-2 Nucleosides were highly similar to this fine-tuning set than to the broad Synthetic Nucleosides set (Wilcoxon test P-value = 1.75e-14; Fig. 3 A).

To gain better insight into the diversity of the properties of the generated molecules, we compared the distributions of certain compound properties to the training and fine-tuning sets. Specifically, we compared the distribution of quantitative estimates of drug-likeness (QED) and the logarithm of the partition function (logP). Figs. 4 B and 2 C show a clear difference between the distribution of QED and logP properties between the generated (for both CRT and CLM denoted in green) and training sets (denoted in blue). The distribution of the properties of the generated molecules bears closer resemblance to the reference sets (denoted in orange), although with some distinctive differences. Thus, although the molecules generated by CRT are structurally similar to the fine-tuning sets, they show diversity in terms of other properties, such as QED and logP.

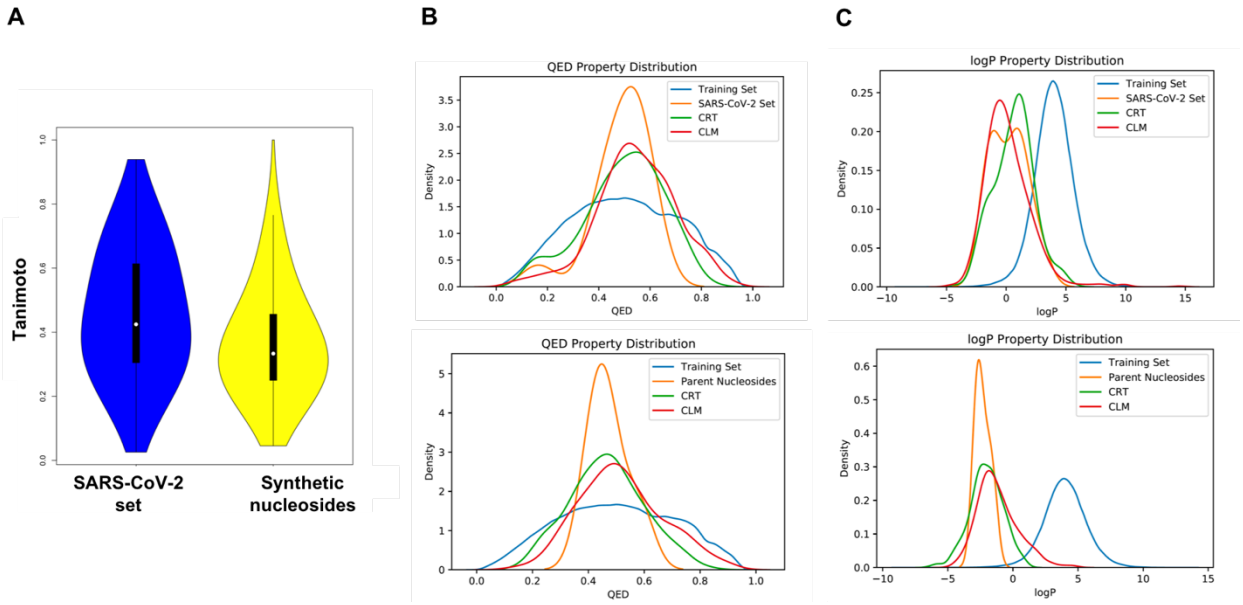
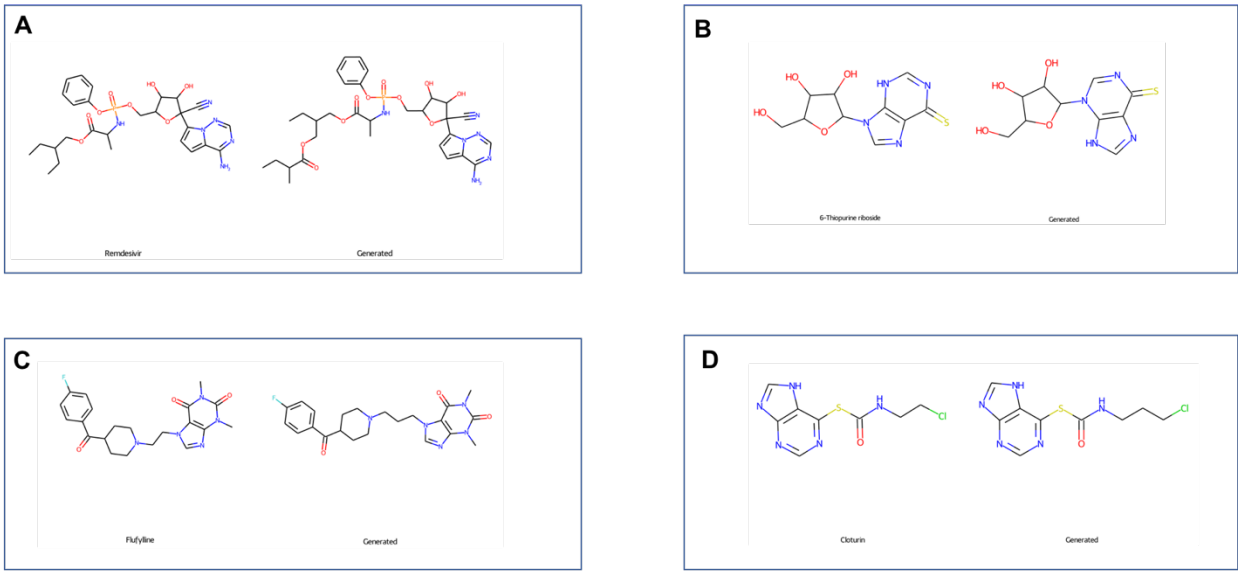


Fig. 4. Assessment of CRT and CLM generated molecules following fine-tuning on the SARS-CoV-2 active nucleoside set. **A.** Similarity between CRT generated molecules to SARS-CoV-2 active nucleosides and Synthetic Nucleosides when CRT was fine-tuned on the SARS-CoV-2 Nucleosides. **B.** QED property distributions of generated molecules using the SARS-CoV-2 Nucleosides as a fine-tuning set (upper panel) or Parent Nucleosides (lower panel). **C.** logP property distributions of generated molecules using the SARS-CoV-2 Nucleosides as a fine-tuning set (upper panel) or Parent Nucleosides (lower panel).

In addition to assessing the similarity of molecules generated by CRT to the reference sets, based on their Tanimoto coefficients, we also performed visual inspection of their chemical structure. For this purpose, we examined the RDKit visualizations of certain molecules. As seen in Figs. 5, the CRT generated molecules bear close structural resemblance to their respective Parent Nucleosides and SARS-CoV-2 Nucleosides; and yet they contain unique and distinctive variations. For the SARS-CoV-2 Nucleosides, we show a generated sample of two of the more prominent members of this class - Remdesivir (Fig. 5 A) and Molnupiravir (Fig. 5 C) - as well as 6-Thiopurine riboside, Flufylline, and Cloturin (see Fig. 4 B to D). We are able to trace the generated molecules to the source molecules because CRT conditions on Morgan fingerprints; unlike other models, which only use transfer learning to generate from chemical space. We believe that the close structural resemblance of the generated molecules to the respective source compounds, as well as their uniqueness, indicates that CRT is capable of generating structurally similar molecules to the Parent Nucleosides and the SARS-CoV-2 Nucleosides.



E

Input Molecule		Generated Molecule	
Name	SMILES	CID	SMILES
Flufylline	Cn1c(=O)c2c(ncn2CCN2CCC[C(=O)c3ccc(F)cc3)CC2)n(C)c1=O	65636	Cn1c(=O)c2c(ncn2CCCN2CCC(C(=O)c3ccc(F)cc3)CC2)n(C)c1=O
7-Deazadenosine	Nc1ncnc2c1ccn2C1OC(CO)C(O)C1O	54079154	NCC1OC(n2ccc3c(N)ncnc32)C(O)C1O
Remdesivir	CCC(CC)COC(=O)C(C)NP(=O)(OCC1OC(C#N)(c2ccc3c(N)ncnn23)C(O)C1O)Oc1ccccc1	77401118	CCC(CC)COC(=O)C(C)NP(=O)(OCC1OC(C#N)(c2ccc3c(N)ncnn23)CC1O)Oc1ccccc1
6-Thio-dG	Nc1nc(=S)c2ncn(C3CC(O)C(CO)O3)c2[nH]1	4317458	Nc1nc(=S)c2ncn(C3CC(O)C(COP(=O)(O)O)O3)c2[nH]1
Thiamiprine	Cn1cnc([N+](=O)[O-])c1Sc1nc(N)nc2nc[nH]c12	15225846	Cn1cnc(Sc2nc(N)nc3nc[nH]c23)c1[N+](=O)[O-]

Fig. 5. Visualization of some of the CRT generated molecules using the SARS-CoV-2 Nucleosides as the fine-tuning set. **A.** Remdesivir and an analog generated by CRT. **B.** 6-thiopurine riboside and an analog generated by CRT. **C.** Flufylline and an analog generated by CRT. **D.** Cloturin and an analog generated by CRT.

CRT model learns to focus on Morgan fingerprints to guide generation

To better visualize the manifold searched by the CRT model, we performed a principal component analysis (PCA) of the Morgan fingerprints of the training set, fine-tuning sets and the generated

molecules for the Natural Nucleosides and the SARS-CoV-2 Nucleosides. In Figs. 6 A and B, the PCA of the Morgan fingerprints of each training example is indicated in yellow, the fine-tuning instances are in green and the generated molecules are shown in red for CRT and blue for CLM. The PCA diagrams are of generated molecules with Tanimoto coefficient of 0.7 or greater. The figures show that the CRT generated molecules (red) search the manifold near and around the Parent Nucleosides fine-tuning set (Fig. 6 A) as well as near the SARS-CoV-2 fine tuning set (Fig. 6 B). In both cases, CRT generated molecules are also diverse (spread out in the molecular space shown in the PCA plots). Notably, a virtual walk in this molecular space starting from CRT generated molecules passes through paths that include molecules drawn from the SARS-CoV-2 fine tuning set (Fig. 6 B). This implies that CRT extrapolates the molecular regions lying between molecules in the fine-tuning set that are separated spatially in this space.

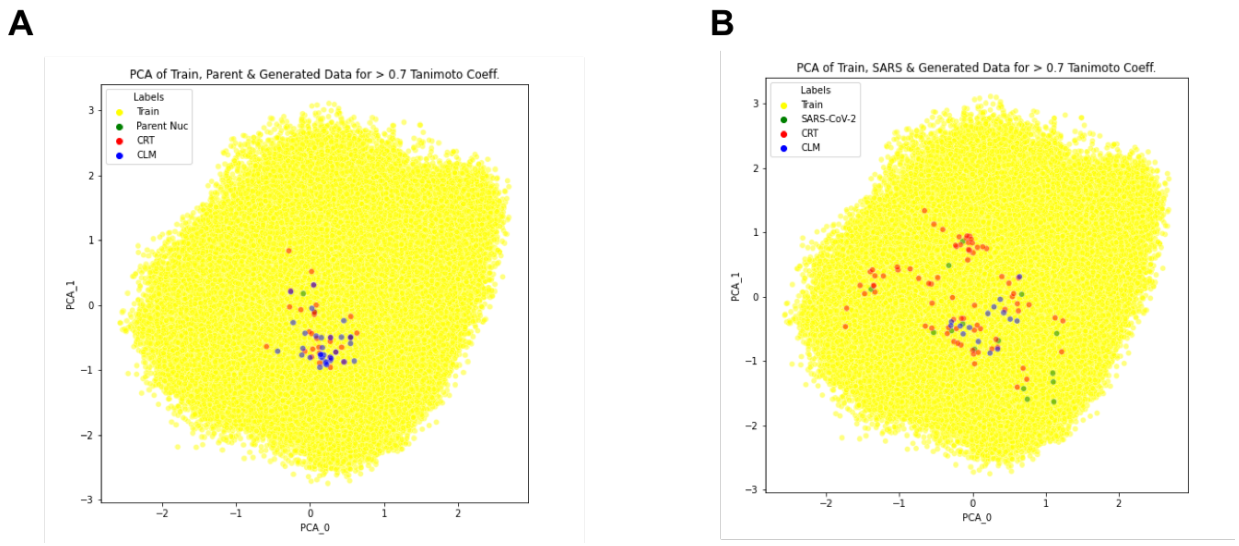


Fig. 6. Molecular space visualization of generated molecules relative to the fine-tuning sets. Visualization of CRT and CLM generated molecules based on ECFP fingerprints of the training sets alongside the Parent Nucleosides used as fine-tuning sets **(A)** or using the SARS-CoV-2 Nucleosides for fine tuning **(B)**.

To demonstrate that CRT learns to focus on the Morgan fingerprint of the input molecule when generating new molecules, we examine the model's attention maps. In general, Transformer attention maps show the correlation of tokens in a sequence. A Transformer model uses these correlations to predict the next token. CRT has eight layers with eight attention blocks per layer. We selected the final attention block in the final layer of two models: a Transformer model that did not use fingerprint conditioning and CRT (which uses conditioning). The Transformer model without conditioning generated a valid molecule - **C**1(=O)N**c**2c(cc(N**S**(=O)(C)=O)cc2)C1=Cc1[nH]ccc1 (Fig. 6 A). Based on the attention map in the final layer / final attention block, the three tokens with the highest correlations (the ones that

other tokens focused on the most when generating the next character) are: the first token, C (27.4%), the 16th token, S (7.8%), and the 8th token, c (4.2%). (The relevant characters are bolded in the above SMILES representation).

For the second model, we condition on the Morgan fingerprint of Molnupiravir, which generated the following valid molecule: C1(n2c(=O)nc(NO)c3ccc(Cl)nc32)OC(COC(C(C)C)=O)C(O)C1O. For purposes of illustration, we insert a token, MFP, at the beginning of the sequence to indicate that the generated molecule is conditioned on a Morgan fingerprint (MFP): **MFP**C1(n2c(=O)nc(NO)c3ccc(Cl)nc32)OC(COC(C(C)C)=O)**C**(O)C1O. In this case, the top 3 tokens that the model focus on (in the final layer / final attention block) are: the first token (MFP - 11.2%), the second token (C - 41.0%) and the 47th token (C - 3.5%). In both sequences, the models largely focus on the first atom token (C in both cases). However, in the case of CRT, the model clearly also focuses on the conditioning token (MFP).

To better visualize the process by which CRT attends to different tokens as it generates characters, we select four layers in the model to see how self-attention gradually changes. We create a visualization based on Clark et al. (*50*). We intentionally select the early layers in the models (Layer 1, Block 1; Layer 2, Block 1) and the final layers in the models (Layer 8, Block 1; and Layer 8, Block 8). We show how the model attends to each character in the molecule sequence for two models: (1) a Transformer model without Morgan fingerprint conditioning (see Fig. 7 A) and (2) CRT, which conditions generation on Morgan fingerprints (see Fig. 7 B). In both illustrations, the initial layers of the models focus or attend almost equally to each atom and bond in the sequence when predicting the next token in the sequence; however, by the final layer and block, both models learn to focus on specific characters when generating the next token. As can be seen in Fig. 7 A, the Transformer model without conditioning heavily focuses on the first token in the sequence; however, CRT focuses on both the first and second tokens, with the first token representing the Morgan fingerprint, when generating subsequent tokens.

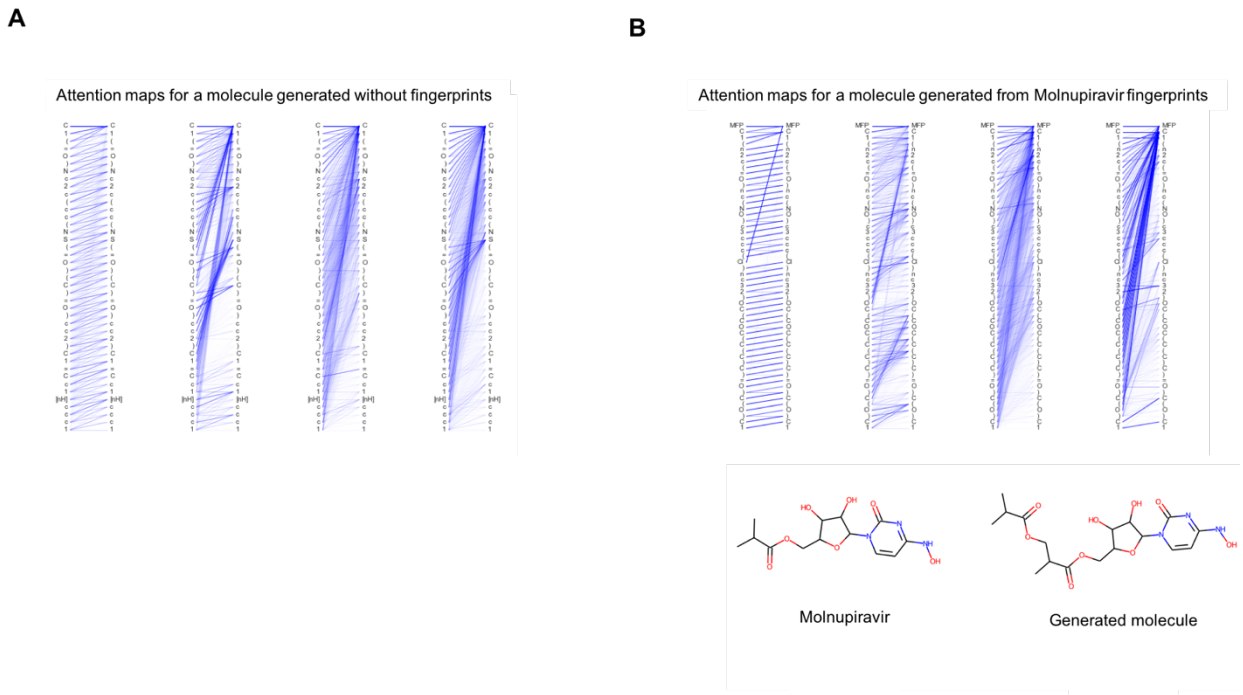


Fig. 7. Visualization of CRT attention maps without conditioning on Morgan fingerprints **(A)** or with conditioning **(B).**

CRT model avoids catastrophic forgetting of neural network models

A common problem encountered in neural network transfer learning is catastrophic forgetting (*29*). This issue occurs when a neural network is trained on a new task, causing the model weights to be updated based on the distribution of the new data set. Catastrophic forgetting can occur during the fine-tuning process, when the model is re-trained on new data. In our case, CRT is first trained to learn how to construct valid molecules. This process involves learning the "grammar" by which atom, bond and branch characters interact and correlate. It is critical that the learning of this grammar is not lost when the model learns the distribution of the fine-tuning set. Clearly, CRT does not suffer from catastrophic forgetting because it is able to decode valid molecules after fine-tuning. Thus, it does not forget the grammar rules that it learned.

In order to gain insight into how the model transfers learning from the training to the fine-tuning process, we examine the weights of 3 trained models: a Transformer without ECFP (Morgan Fingerprints) conditioning, a Transformer with ECFP conditioning, and finally a Transformer with both fingerprint conditioning and fine-tuning (CRT). To determine how the weights changed due to conditioning and fine-tuning, we calculated the Wasserstein distance between the weights by layer in the model without Morgan fingerprint conditioning (denoted as "No Condition" in Supplementary Table 17) to the model with conditioning ("MFP Condition" in Supplementary Table 18). We also compared the change in model weights in a model that included Morgan

fingerprint conditioning to a model that included *both* Morgan fingerprint conditioning *and* fine-tuning (denoted as "CRT" in Supplementary Table 18). Since the models contained between 134 and 136 layers, we aggregate and average the layers into 8 categories for ease of reference. The categories are: layer normalization weights and bias ("Layer Norm Weights" and "Layer Norm Bias"), attention layers for key, query and value ("Attention Key Weights" and "Attention Key Bias"), attention projection layers ("Projection Weights" and "Projection Bias"), and multilayer perceptron layers ("MLP Weights" and "MLP Bias"). As can be seen in Supplementary Table 18, the greatest change in model weights occurs between the "no conditioning" and the conditioned models, with very little change due to fine-tuning. This small change implies that the model did not lose much of its learned "knowledge" due to fine-tuning, even though we did not freeze any layer parameters, as is sometimes done in transfer learning.

When comparing the conditioned to the no conditioned models, the layers that exhibit the greatest change tend to be the layer normalization and bias layers. The change in these layers is distributed across all layers and does not occur only, for example, in the final layers. The fact that the changes in the layer norm and bias layers occurs throughout all eight layers has implications for transfer learning. Because all such layers change, it would be detrimental to freeze the weights of the early layers. This may explain why the Transformer model generated better results when no layers were frozen and the learning rate is reduced, instead of freezing layers (which apparently works with LSTM models as shown by Moret et al. (*33*)) We also hypothesize that the CRT model retains its grammar knowledge in the regular weight layers (which change very little) and it changes the bias layers (which vary more substantially) to account for conditioning.

Discussion

This work presents one of the first studies in which generative molecular models are focused on nucleoside analogs design. As one of the most defining molecules of all living organisms and viruses given their role in DNA and RNA synthesis, exploration of the molecular space of nucleosides and nucleotides could impact a broad range of areas including fundamental and applied biology. In addition to the five core nucleosides that are basic components of DNA and RNA (A, T, U, G, C), there are several other naturally occurring nucleosides and nucleotides that play diverse roles in biology. Over several decades, nucleoside analogs have also become one of the most important molecular classes for antiviral drug design with several approved medicines from this class already in the market and many more in development for a wide range of diseases including emerging viruses (*13*, *14*). Indeed, Remdesivir - the first FDA approved antiviral drug against the recently emerged SARS-CoV-2 virus is a nucleoside analog (*51–53*) and one of the most promising orally administered antivirals against the virus (Molnupiravir), also belongs to this molecular class (*54–57*).

Deep generative molecular designs in principle could lead to exploration of regions in molecular space that currently remain underexplored or rare in nature or inaccessible to medicinal chemists. Perhaps more importantly, deep generative models could accelerate the pace at which large amounts of molecules could be designed, which could provide a starting point for identifying molecules with new properties including new drugs. Hence, to explore the potential for generative design, in this study we focused on their application in the design of nucleoside analogs. Due to the critical role of nucleosides in biology, a vast amount of knowledge on their quantitative structure activity relationships (QSAR), diverse biological targets and natural or synthetic analogs has been accumulated (*13*, *14*). Therefore, generative design of these molecules could also play an important role in benchmarking computational tools for de novo molecular design.

In this study, we first benchmarked five deep generative molecular models alongside our approach (CRT). In summary, our main contributions are:

- Focused Molecule Generation. We generate focused molecules that are structurally similar to the nucleosides and make them publicly available in an effort to further research into potential therapeutics.
- Condition on Molecular Structure. Instead of conditioning molecule generation on a single property (e.g., QED, TPSA, molecular weight, etc.), we condition generation on chemical structure - Morgan fingerprints.
- Combine Direct Steering and Fine-Tuning. Unlike other models that generate focused molecules by conditioning on properties *or* by using transfer learning, we combine both approaches into a single pipeline.
- Encourage Diversity. Transformer models sometimes generate repetitive sequences. This issue has been dealt with by introducing variance into the token selection process via temperature sampling, top-k and top-p sampling. All of these methods work at the end of the generation process. We, instead, introduce diversity at the beginning of the generation process by partially randomizing Morgan fingerprint *inputs*.

The significance of our study in the fields of fundamental and translational biology is demonstrated by several observations. We show that our model (CRT) given only a fine-tuning set of the five parent nucleosides, generates several molecules that are similar or identical to natural or synthetic nucleoside analogs. These generated nucleoside analogs have chemical alterations that involve either the ribose or nucleobase moiety, in one case mimicking molecular transformations observed in oxetanocin A- a naturally occurring antiviral nucleoside analog first isolated from a bacterium (*38*). Interestingly, one of the generated molecules -N2-methylguanosine- based on fine tuning on parent nucleosides occurs naturally in eukaryotic and archeal tRNAs where it is a product of a base modification introduced by specific enzyme (*47*). We also observe a rare nucleoside- 2,6-diaminopurine, so far only reported naturally in a single genome (*41*, *42*) and that has implications

for understanding early evolution of life on earth (*44*, *46*). Finally, we show that focused molecular generation could also be directly leveraged to explore the molecular space around antiviral nucleosides, specifically those active against SARS-CoV-2. In conclusion, we show that generative models could aid in molecular design of nucleosides with a wide range of applications from prebiotic chemistry to drug discovery and synthetic biology.

It is important to highlight some limitations of our study. First, while focusing on nucleoside analogs in itself is of high significance biologically, the benchmarking of deep generative models on a single class of molecules cannot capture their performance in other molecular design problems. Second, metrics for assessing generative design models continue to evolve and it is not feasible to explore all metrics that have been reported in literature. It is critical that performance assessments of generative models be always taken in the context of the metrics that were applied and the goals of a given project. Third, gold-standards validation sets for generative molecular design problems do not exist. Thus, in assessing the similarity between the generated molecules to the reference sets, we recognize that some otherwise biologically significant molecules may be missed.

Materials and Methods

Methods overview

Because our goal is to search a limited space of potential molecules that are similar to the nucleosides, we selected three datasets to fine-tune and benchmark our model (CRT, details below) and five other leading deep generative models. The three datasets are: the Parent Nucleosides, the SARS-CoV-2 Nucleosides, and the Synthetic Nucleosides. The Parent Nucleosides consist of adenosine, guanosine, cytidine, thymidine, and uridine. The SARS-CoV-2 Nucleosides were identified by Schultz et al. (*49*) as a group of fifteen nucleoside analogs that show in vitro activity against SARS-CoV- 2. The SARS-CoV-2 Nucleosides include: Molnupiravir, Remdesivir, 6-Mercaptopurine, 6-Thio-dG, 6-Thiopurine riboside, 8-Azaguanine, Azathiopurine, BCNA, Cloturin, Flufylline, Gemcitabine, GS-441524, Thiamiprine, Thioguanine, and Tubercidin. In addition, we selected a group of approximately 188 synthetic nucleoside analogs (*37*) (the Synthetic Nucleosides).

As discussed in more detail below, we first train the CRT model and 5 leading deep generative molecular models on a large dataset of diverse molecules, which contain mostly non-nucleoside compounds. The goal of this process is to train the models to learn how to generate chemically valid molecules. We then "fine-tune" the models by training them on smaller datasets - the Parent Nucleosides and the SARS-CoV-2 Nucleosides. The fine-tuning process involves freezing the initial layers of each of the deep generative models after they are first trained on the larger dataset that is made up of mostly non-nucleoside molecules. Freezing the initial layers of a neural network prevents parameters in those layers from being updated by data contained in the fine-tuning sets.

This process is commonly used in transfer learning. Then, the models are fine-tuned on smaller datasets consisting of the parent nucleosides and the SARS-CoV-2 Nucleosides. Once the models are fine-tuned on the smaller datasets, we generate molecules from each of the models and compare them to the Parent Nucleosides, the SARS-CoV-2 Nucleosides and the Synthetic Nucleosides for similarity.

The fine-tuning datasets are small in size, with only 5 and 15 molecules. Our goal is to determine if the models can learn how to generate chemically valid molecules from the larger training dataset (i.e., learn the rules of chemical construction) and then transfer this learning to smaller datasets (i.e., the nucleosides). Other key goals are to assess the diversity of the generated molecules, their similarity to the nucleoside class (i.e., the ability of the models to conditionally generate molecules), and whether the generated molecules bear resemblance to chemicals undertaken by medicinal chemists.

Existing generative chemistry models leveraged in this work

We benchmark molecules generated by the CRT model against molecules randomly drawn from the training set and against molecules produced by five leading generative models. The five generative models include: a VAE (*30*, *31*), two LSTMs (*32*, *33*), an AAE (*34*) and a base Transformer model (*35*). The VAE is based on architectures developed by Gomez-Bombarelli et al. (*30*) and Blaschke et al. (*31*). The VAE consists of bidirectional Gated Recurrent Units (GRU) with an encoder / decoder structure. The decoder is a 3-layer GRU with 512 hidden dimensions with intermediate dropout layers (probability of 0.2). The LSTM models were developed by Segler et al. (*32*) and Moret et al. (*33*). The Segler et al. model, referred to as CRNN, uses three LSTM layers structured as a decoder-only with a hidden dimension of 600 and dropout probability of 0.2 (*32*). The Moret et al. model (CLM) consists of two LSTM layers, with hidden sizes of 1,024 and 256. The model uses batch normalization and 0.4 dropout (*33*). The AAE model was developed by Kadurin et al. and consists of an encoder, decoder and a discriminator (*34*). The base Transformer is the LIG-GPT (*35*) model described below, under CRT. The VAE, CRNN and AAE models are based on implementations contained in MOSES (*58*). All of the models are trained on the training set described under the Training datasets section below and then fine-tuned on the parent nucleosides and the SARS-CoV-2 Nucleoside datasets. Each of the models represent molecules as a sequence of characters using the Simplified Molecular Input Line Entry Specification (SMILES) (*59*).

Conditional Randomized Transformer (CRT)

Our method is based on the Transformer decoder model developed by Bagal et al., LIG-GPT (*35*), which, in turn, is constructed from the GPT model originally designed by Radford et al. (*60*). Unlike the original GPT language model or its variants (GPT-2 and GPT-3), which can contain

from 100 million to over 1 billion parameters (*60*), LIG-GPT is substantially more compact, containing approximately 6 million parameters (*35*). As such, the model is more interpretable and easily understood. The LIG-GPT model consists of 8 layers, each with 8 self-attention blocks, with embedding size of 256, and is trained with maximum likelihood (cross-entropy), 0.1 dropout and a .0006 learning rate with annealing (*35*).

We made several modifications to the LIG-GPT structure. We call our version of the model, CRT. Our modifications consist of the following. First, the base model is trained to condition on a single property or a series of individual properties (e.g., logP, QED, SA). We, instead, condition molecule generation based on extended-connectivity fingerprints (commonly referred to as ECFPs or Morgan fingerprints) (*36*). ECFPs encode structural and functional features of molecules. Second, the base model is designed to search chemical space based on specific conditions. We train CRT both with conditions *and* fine-tune it on smaller nucleoside datasets, which consist of only 5 to 15 examples. Third, during the inference phase, we introduce variance into the generative process. Without additional randomization measures, we found that the Transformer model, when conditioned on Morgan fingerprints, tended to produce repetitive molecules. To encourage diversity, and to search the chemical space around a given molecular fingerprint, we introduce randomization into the generative process.

The generation of repetitive patterns (here, each molecule represents a chemical pattern) is a common problem with Transformer models due to their deterministic nature (*61*). A variety of techniques have been designed to overcome this problem, including the use of top-k sampling (*62*) , top-p sampling (*63*), the use of so-called temperature sampling, and the introduction of complex variational encodings into a Transformer model's latent space (*64*). Top-k, top-p and temperature sampling introduce variance at the end of the token generation process, by sampling various probable next token candidates. We instead introduce variance into the front-end of the Transformer decoding process by adding random noise to the Morgan fingerprint, subject to a hyper-parameter. Morgan fingerprints are binary encodings (zeros and ones to indicate the absence or presence of a structure) that are both sparse and discrete. Simply changing 0s and 1s does not impart significant variance. Therefore, we insert random noise after the Morgan fingerprint is processed by a dense neural network layer. We found that multiplying the random noise by a hyper-parameter (0.3) before adding it to the model's encoding of the Morgan fingerprint achieved a balance between diversity and chemical validity. This step generally causes the model to generate 3 to 5 times as many unique molecules. We also use a modest temperature sampling parameter (0.9), following Bagal et al (*35*). We train our models for 5 to 10 epochs on a single RTX-2080 GPU.

Training datasets

We experimented with several datasets to train the deep generative models, including ones developed by Brown et al. (GUACAMOL dataset) (*65*), Polykovskiy et al. (MOSES dataset) (*58*), and Moret et al. (Low Data dataset) (*33*). The MOSES dataset was compiled from ZINC (*66*, *67*), while the GUACAMOL and Low Data datasets were compiled from ChEMBL (*68*). We found that the Low Data dataset produced molecules that were more similar to the Parent Nucleosides and the SARS-Cov-2 Nucleosides, perhaps because it was filtered for bioactive compounds. The Low Data dataset consists of approx. 365,000 training molecules. Because there are numerous ways in which a compound can be expressed in the SMILES format (*69*, *70*),this dataset is augmented by a factor of 10x by randomizing each molecule, whereby each molecular string is instantiated with a different non-hydrogen atom. This methodology follows the implementation of Moret et al (*33*).

Metrics

To assess the similarity of generated molecules to the reference sets (Parent Nucleosides, the SARS-CoV-2 Nucleosides and the Synthetic Nucleosides), we draw on metrics offered in three recent papers (*58*, *65*, *71*). Although each paper generally contained different metrics to some extent, they all generally agreed on the need for five key performance indicators. First, they agreed that the number of valid molecules generated by a model should be measured. A popular way to implement this metric is to determine the number of canonical SMILES, as calculated by RDKit (*72*). This metric confirms that a model is generating atom, bond and ring combinations that correspond to valid chemical sequences. This ensures that the model learns the proper chemical grammar. Second, the number of unique molecules should be tracked. This metric ensures that a model is not generating the same compound repeatedly. In other words, the model must not only be able to learn the proper chemical grammar, but it must also generate diverse chemical combinations. Third, the molecules should be novel. This measures whether the generated molecules are different from the molecules in the training and fine-tuning sets. Fourth, the fidelity of the model in generating molecules that are similar to the training and fine-tuning sets should be determined. In our case, we are interested in generating molecules that are similar to the fine-tuning or reference sets (the parent Nucleosides and the SARS-CoV-2 active nucleosides). Preuer et al. developed the Frechet ChemNet Distance (FCD) (*73*), which is commonly used to measure model fidelity. The FCD is determined based on the hidden representation of molecules in a neural network named ChemNet, which is similar to the Frechet Inception Distance (*74*), used to assess GANs in the image domain. Similar molecule distributions have low FCDs. Fifth, the similarity of generated molecules and molecules in the reference sets can (Parent Nucleosides, the SARS-CoV-2 Nucleosides and the Synthetic Nucleosides) should be measured, which is commonly done with the Tanimoto coefficient. An adaptation of the Tanimoto coefficient is the Similarity of Nearest Neighbor (SNN) measure, which calculates the average Tanimoto similarity of standard Morgan fingerprints of a molecule in the generated set as compared to its nearest neighbor in the reference (*58*). This metric is bounded by 0 and 1, with a 1 indicating perfect identity of the

generated and reference sets. We applied SNN to measure the similarity of generated molecules to those in the parent nucleoside, SARS-CoV-2 active nucleoside and synthetic nucleoside datasets.

Code availability

Code, pre-trained models and supplementary information are available on Github at https://github.com/SiwoResearch/generative_AI_nucleosides

Conflicts of interest

A provisional patent application is in preparation in relation to some of the molecules generated in this work with GHS, DD and NVC listed as inventors.

Acknowledgements

This project is part of the Pandemic Response Emergency Preparedness Therapeutics Initiative (PREEMPTIVE) and is funded by a grant from the Bill and Melinda Gates Foundation to GHS.

References

1. J. R. Knowles, Enzyme-catalyzed phosphoryl transfer reactions. *Annu. Rev. Biochem.* (1980), , doi:10.1146/annurev.bi.49.070180.004305.
2. J. A. Beavo, L. L. Brunton, Cyclic nucleotide research - Still expanding after half a century. *Nat. Rev. Mol. Cell Biol.* (2002), , doi:10.1038/nrm911.
3. E. W. Sutherland, T. W. Rall, Fractionation and characterization of a cyclic adenine ribonucleotide formed by tissue particles. *J. Biol. Chem.* (1958), doi:10.1016/s0021-9258(19)77423-7.
4. J. Wu, L. Sun, X. Chen, F. Du, H. Shi, C. Chen, Z. J. Chen, Cyclic GMP-AMP is an endogenous second messenger in innate immune signaling by cytosolic DNA. *Science (80-.).* (2013), doi:10.1126/science.1229963.
5. A. Ablasser, M. Goldeck, T. Cavlar, T. Deimling, G. Witte, I. Röhl, K. P. Hopfner, J. Ludwig, V. Hornung, CGAS produces a 2′-5′-linked cyclic dinucleotide second messenger that activates STING. *Nature* (2013), doi:10.1038/nature12306.
6. A. W. Struck, M. L. Thompson, L. S. Wong, J. Micklefield, S-Adenosyl-Methionine-Dependent Methyltransferases: Highly Versatile Enzymes in Biocatalysis, Biosynthesis and Other Biotechnological Applications. *ChemBioChem* (2012), , doi:10.1002/cbic.201200556.
7. G. L. Cantoni, Biological methylation: selected aspects. *Annu. Rev. Biochem.* (1975), , doi:10.1146/annurev.bi.44.070175.002251.
8. Z. Huang, N. Xie, P. Illes, F. Di Virgilio, H. Ulrich, A. Semyanov, A. Verkhratsky, B. Sperlagh, S. G. Yu, C. Huang, Y. Tang, From purines to purinergic signalling: molecular functions and human diseases. *Signal Transduct. Target. Ther.* (2021), , doi:10.1038/s41392-021-00553-z.

9. D. A. Barawkar, T. C. Bruice, Synthesis, biophysical properties, and nuclease resistance properties of mixed backbone oligodeoxynucleotides containing cationic internucleoside guanidinium linkages: Deoxynucleic guanidine/DNA chimeras. *Proc. Natl. Acad. Sci. U. S. A.* (1998), doi:10.1073/pnas.95.19.11047.
10. B. P. Monia, J. F. Johnston, H. Sasmor, L. L. Cummins, Nuclease resistance and antisense activity of modified oligonucleotides targeted to Ha-ras. *J. Biol. Chem.* (1996), doi:10.1074/jbc.271.24.14533.
11. A. Khvorova, J. K. Watts, The chemical evolution of oligonucleotide therapies of clinical utility. *Nat. Biotechnol.* (2017), , doi:10.1038/nbt.3765.
12. K. Karikó, M. Buckstein, H. Ni, D. Weissman, Suppression of RNA recognition by Toll-like receptors: The impact of nucleoside modification and the evolutionary origin of RNA. *Immunity* (2005), doi:10.1016/j.immuni.2005.06.008.
13. M. K. Yates, K. L. Seley-Radtke, The evolution of antiviral nucleoside analogues: A review for chemists and non-chemists. Part II: Complex modifications to the nucleoside scaffold. *Antiviral Res.* (2019), , doi:10.1016/j.antiviral.2018.11.016.
14. K. L. Seley-Radtke, M. K. Yates, The evolution of nucleoside analogue antivirals: A review for chemists and non-chemists. Part 1: Early structural modifications to the nucleoside scaffold. *Antiviral Res.* (2018), , doi:10.1016/j.antiviral.2018.04.004.
15. C. M. Galmarini, J. R. Mackey, C. Dumontet, Nucleoside analogues and nucleobases in cancer treatment. *Lancet Oncol.* (2002), , doi:10.1016/S1470-2045(02)00788-X.
16. L. P. Jordheim, D. Durantel, F. Zoulim, C. Dumontet, Advances in the development of nucleoside and nucleotide analogues for cancer and viral diseases. *Nat. Rev. Drug Discov.* (2013), , doi:10.1038/nrd4010.
17. T. Robak, E. Lech-Maranda, A. Korycka, E. Robak, Purine Nucleoside Analogs as Immunosuppressive and Antineoplastic Agents: Mechanism of Action and Clinical Activity. *Curr. Med. Chem.* (2009), doi:10.2174/092986706778742918.
18. W. D. Fuller, R. A. Sanchez, L. E. Orgel, Studies in prebiotic synthesis. VII - Solid-State Synthesis of Purine Nucleosides. *J. Mol. Evol.* (1972), doi:10.1007/BF01660244.
19. L. E. Orgel, Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.* (2004), , doi:10.1080/10409230490460765.
20. M. W. Powner, B. Gerland, J. D. Sutherland, Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* (2009), doi:10.1038/nature08013.
21. D. A. Malyshev, K. Dhami, T. Lavergne, T. Chen, N. Dai, J. M. Foster, I. R. Corrêa, F. E. Romesberg, A semi-synthetic organism with an expanded genetic alphabet. *Nature* (2014), doi:10.1038/nature13314.
22. P.-C. Kotsias, J. Arús-Pous, H. Chen, O. Engkvist, C. Tyrchan, E. J. Bjerrum, Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* (2020), doi:10.1038/s42256-020-0174-5.
23. I. Goodfellow, J. Poughet-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, General adversarial networks. *Adv. Neural Inf. Process. Syst.* (2014).
24. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks. *Commun. ACM* (2020), doi:10.1145/3422622.
25. A. Makhzani, Unsupervised Representation Learning With Autoencoders. *ProQuest Diss.*

Theses (2018).
26. D. P. Kingma, M. Welling, in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* (2014).
27. S. Hochreiter, J. Schmidhuber, Long Short-Term Memory. *Neural Comput.* (1997), doi:10.1162/neco.1997.9.8.1735.
28. J. Chung, C. Gulcehre, K. Cho, Y. Bengio, in *NIPS'2014 Deep Learning workshop* (2014).
29. J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. U. S. A.* (2017), doi:10.1073/pnas.1611835114.
30. R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* (2018), doi:10.1021/acscentsci.7b00572.
31. T. Blaschke, M. Olivecrona, O. Engkvist, J. Bajorath, H. Chen, Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inform.* (2018), doi:10.1002/minf.201700123.
32. M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* (2018), doi:10.1021/acscentsci.7b00512.
33. M. Moret, L. Friedrich, F. Grisoni, D. Merk, G. Schneider, Generative molecular design in low data regimes. *Nat. Mach. Intell.* (2020), doi:10.1038/s42256-020-0160-y.
34. A. Kadurin, A. Aliper, A. Kazennov, P. Mamoshina, Q. Vanhaelen, K. Khrabrov, A. Zhavoronkov, The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* (2017), doi:10.18632/oncotarget.14073.
35. V. Bagal, R. Aggarwal, P. K. Vinod, U. D. Priyakumar, LigGPT: Molecular Generation using a Transformer-Decoder Model. *ChemRxiv*, 1–30 (2021).
36. D. Rogers, M. Hahn, Extended-connectivity fingerprints. *J. Chem. Inf. Model.* (2010), doi:10.1021/ci100050t.
37. Selleckchem Nucleoside Analog Library, (available at https://www.selleckchem.com/screening/Nucleoside-Analogue-Library.html).
38. N. Shimada, S. Hasegawa, T. Harada, T. Tomisawa, A. Fujii, T. Takita, Oxetanocin, a novel nucleoside from bacteria. *J. Antibiot. (Tokyo).* (1986), doi:10.7164/antibiotics.39.1623.
39. J. Alder, M. Mitten, D. Norbeck, K. Marsh, E. R. Kern, J. Clement, Efficacy of A-73209, a potent orally active agent against VZV and HSV infections. *Antiviral Res.* (1994), doi:10.1016/0166-3542(94)90037-X.
40. J. A. Bull, R. A. Croft, O. A. Davis, R. Doran, K. F. Morgan, Oxetanes: Recent Advances in Synthesis, Reactivity, and Medicinal Chemistry. *Chem. Rev.* (2016), , doi:10.1021/acs.chemrev.6b00274.
41. M. D. Kirnos, I. Y. Khudyakov, N. I. Alexandrushkina, B. F. Vanyushin, 2-Aminoadenine is an adenine substituting for a base in S-2L cyanophage DNA. *Nature* (1977), doi:10.1038/270369a0.
42. I. Y. Khudyakov, M. D. Kirnos, N. I. Alexandrushkina, B. F. Vanyushin, Cyanophage S-

2L contains DNA with 2,6-diaminopurine substituted for adenine. *Virology* (1978), doi:10.1016/0042-6822(78)90104-6.
43. C. Bailly, M. J. Waring, The use of diaminopurine to investigate structural properties of nucleic acids and molecular recognition between ligands and DNA. *Nucleic Acids Res.* (1998), doi:10.1093/nar/26.19.4309.
44. M. P. Callahan, K. E. Smith, H. J. Cleaves, J. Ruzicka, J. C. Stern, D. P. Glavin, C. H. House, J. P. Dworkin, Carbonaceous meteorites contain a wide range of extraterrestrial nucleobases. *Proc. Natl. Acad. Sci. U. S. A.* (2011), doi:10.1073/pnas.1106493108.
45. C. Hartel, M. W. Göbel, Substitution of adenine by purine-2,6-diamine improves the nonenzymatic oligomerization of ribonucleotides on templates containing thymidine. *Helv. Chim. Acta* (2000), doi:10.1002/1522-2675(20000906)83:9<2541::AID-HLCA2541>3.0.CO;2-8.
46. R. Szabla, M. Zdrowowicz, P. Spisz, N. J. Green, P. Stadlbauer, H. Kruse, J. Šponer, J. Rak, 2,6-diaminopurine promotes repair of DNA lesions under prebiotic conditions. *Nat. Commun.* (2021), doi:10.1038/s41467-021-23300-y.
47. S. L. Ginell, R. Parthasarathy, Conformation of N2-methylguanosine, A modified nucleoside of tRNA. *Biochem. Biophys. Res. Commun.* (1978), doi:10.1016/0006-291X(78)91666-2.
48. J. Chen, S. J. Swamidass, Y. Dou, J. Bruand, P. Baldi, ChemDB: A public database of small molecules and related chemoinformatics resources. *Bioinformatics* (2005), doi:10.1093/bioinformatics/bti683.
49. D. C. Schultz, R. M. Johnson, K. Ayyanathan, J. Miller, K. Whig, B. Kamalia, M. Dittmar, S. Weston, H. L. Hammond, C. Dillen, L. Castellana, J. S. Lee, M. Li, E. Lee, S. Constant, M. Ferrer, C. A. Thaiss, M. B. Frieman, S. Cherry, Pyrimidine biosynthesis inhibitors synergize with nucleoside analogs to block SARS-CoV-2 infection. *bioRxiv Prepr. Serv. Biol.* (2021), , doi:10.1101/2021.06.24.449811.
50. K. Clark, U. Khandelwal, O. Levy, C. D. Manning, (2019).
51. E. de Wit, F. Feldmann, J. Cronin, R. Jordan, A. Okumura, T. Thomas, D. Scott, T. Cihlar, H. Feldmann, Prophylactic and therapeutic remdesivir (GS-5734) treatment in the rhesus macaque model of MERS-CoV infection. *Proc. Natl. Acad. Sci. U. S. A.* (2020), doi:10.1073/pnas.1922083117.
52. J. H. Beigel, K. M. Tomashek, L. E. Dodd, A. K. Mehta, B. S. Zingman, A. C. Kalil, E. Hohmann, H. Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R. W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T. F. Patterson, R. Paredes, D. A. Sweeney, W. R. Short, G. Touloumi, D. C. Lye, N. Ohmagari, M. Oh, G. M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M. G. Kortepeter, R. L. Atmar, C. B. Creech, J. Lundgren, A. G. Babiker, S. Pett, J. D. Neaton, T. H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, H. C. Lane, Remdesivir for the Treatment of Covid-19 — Final Report. *N. Engl. J. Med.* (2020), doi:10.1056/nejmoa2007764.
53. J. Grein, N. Ohmagari, D. Shin, G. Diaz, E. Asperges, A. Castagna, T. Feldt, G. Green, M. L. Green, F.-X. Lescure, E. Nicastri, R. Oda, K. Yo, E. Quiros-Roldan, A. Studemeister, J. Redinski, S. Ahmed, J. Bernett, D. Chelliah, D. Chen, S. Chihara, S. H. Cohen, J. Cunningham, A. D'Arminio Monforte, S. Ismail, H. Kato, G. Lapadula, E. L'Her, T. Maeno, S. Majumder, M. Massari, M. Mora-Rillo, Y. Mutoh, D. Nguyen, E. Verweij, A. Zoufaly, A. O. Osinusi, A. DeZure, Y. Zhao, L. Zhong, A. Chokkalingam, E. Elboudwarej, L. Telep, L. Timbs, I. Henne, S. Sellers, H. Cao, S. K. Tan, L.

Winterbourne, P. Desai, R. Mera, A. Gaggar, R. P. Myers, D. M. Brainard, R. Childs, T. Flanigan, Compassionate Use of Remdesivir for Patients with Severe Covid-19. *N. Engl. J. Med.* (2020), doi:10.1056/nejmoa2007016.

54. A. Wahl, L. E. Gralinski, C. E. Johnson, W. Yao, M. Kovarova, K. H. Dinnon, H. Liu, V. J. Madden, H. M. Krzystek, C. De, K. K. White, K. Gully, A. Schäfer, T. Zaman, S. R. Leist, P. O. Grant, G. R. Bluemling, A. A. Kolykhalov, M. G. Natchus, F. B. Askin, G. Painter, E. P. Browne, C. D. Jones, R. J. Pickles, R. S. Baric, J. V. Garcia, SARS-CoV-2 infection is effectively treated and prevented by EIDD-2801. *Nature* (2021), doi:10.1038/s41586-021-03312-w.
55. W. P. Painter, W. Holman, J. A. Bush, F. Almazedi, H. Malik, N. C. J. E. Eraut, M. J. Morin, L. J. Szewczyk, G. R. Painter, Human safety, tolerability, and pharmacokinetics of molnupiravir, a novel broad-spectrum oral antiviral agent with activity against SARS-CoV-2. *Antimicrob. Agents Chemother.* (2021), doi:10.1128/AAC.02428-20.
56. F. Kabinger, C. Stiller, J. Schmitzová, C. Dienemann, G. Kokic, H. S. Hillen, C. Höbartner, P. Cramer, Mechanism of molnupiravir-induced SARS-CoV-2 mutagenesis. *Nat. Struct. Mol. Biol.* (2021), doi:10.1038/s41594-021-00651-0.
57. NCT04575597, Efficacy and Safety of Molnupiravir (MK-4482) in Non-Hospitalized Adult Participants With COVID-19 (MK-4482-002). *https://clinicaltrials.gov/show/NCT04575597* (2020).
58. D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik, A. Zhavoronkov, Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* (2020), doi:10.3389/fphar.2020.565644.
59. D. Weininger, SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* (1988), doi:10.1021/ci00057a005.
60. A. Radford, T. Salimans, Improving Language Understanding by Generative Pre-Training, 1–12.
61. R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, in *Advances in Neural Information Processing Systems* (2019).
62. A. Fan, M. Lewis, Y. Dauphin, in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* (2018).
63. A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, in *CEUR Workshop Proceedings* (2019).
64. Z. Lin, G. I. Winata, P. Xu, Z. Liu, P. Fung, Variational Transformers for Diverse Response Generation (2018).
65. N. Brown, M. Fiscato, M. H. S. Segler, A. C. Vaucher, GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* (2019), doi:10.1021/acs.jcim.8b00839.
66. J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, R. G. Coleman, ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* (2012), , doi:10.1021/ci3001277.
67. J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield, R. A. Sayle, ZINC20 - A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* (2020), doi:10.1021/acs.jcim.0c00675.
68. A. Gaulton, A. Hersey, M. L. Nowotka, A. Patricia Bento, J. Chambers, D. Mendez, P.

Mutowo, F. Atkinson, L. J. Bellis, E. Cibrian-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magarinos, J. P. Overington, G. Papadatos, I. Smit, A. R. Leach, The ChEMBL database in 2017. *Nucleic Acids Res.* (2017), doi:10.1093/nar/gkw1074.
69. J. Arús-Pous, S. V. Johansson, O. Prykhodko, E. J. Bjerrum, C. Tyrchan, J. L. Reymond, H. Chen, O. Engkvist, Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminform.* (2019), doi:10.1186/s13321-019-0393-0.
70. E. J. Bjerrum, SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules.
71. M. A. Skinnider, R. G. Stacey, D. S. Wishart, L. J. Foster, Chemical language models enable navigation in sparsely populated chemical space. *Nat. Mach. Intell.* (2021), doi:10.1038/s42256-021-00368-1.
72. G. Landrum, Getting Started with the RDKit in Python — The RDKit 2020.03.1 documentation. *RDKit* (2016).
73. K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter, G. Klambauer, Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J. Chem. Inf. Model.* (2018), doi:10.1021/acs.jcim.8b00234.
74. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, in *Advances in Neural Information Processing Systems* (2017).

Supplementary Tables

Supplementary Table 18: Change in CRT Model Weights Based on Conditioning and Fine-Tuning