A deep generative model enables automated structure elucidation of novel psychoactive substances

Michael A. Skinnider[1*], **Fei Wang**[2,3], **Daniel Pasin**[4], **Russell Greiner**[3,5], **Leonard J. Foster**[1,6], **Petur W. Dalsgaard**[4], and **David S. Wishart**[2,3,8,9*]

[1] Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada

[2] Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

[3] Department of Computing Science, University of Alberta, Edmonton, AB, Canada

[4] Section of Forensic Chemistry, Department of Forensic Medicine, University of Copenhagen, Copenhagen, Denmark

[5] Alberta Machine Intelligence Institute, Edmonton, AB, Canada

[6] Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, BC, Canada

[7] Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, AB, Canada

[8] Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB, Canada

[9] Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA

[*] email: michael.skinnider@msl.ubc.ca, david.wishart@ualberta.ca

Over the past decade, the illicit drug market has been reshaped by the proliferation of clandestinely produced designer drugs. These agents, referred to as new psychoactive substances (NPSs), are designed to mimic the physiological actions of better-known drugs of abuse while skirting drug control laws. The public health burden of NPS abuse obliges toxicological, police, and customs laboratories to screen for them in law enforcement seizures and biological samples. However, the identification of emerging NPSs is challenging due to the chemical diversity of these substances and the fleeting nature of their appearance on the illicit market. Here, we present DarkNPS, a deep learning-enabled approach to automatically elucidate the structures of unidentified designer drugs using only mass spectrometric data. Our method employs a deep generative model to learn a statistical probability distribution over unobserved structures, which we term the structural prior. We show that the structural prior allows DarkNPS to elucidate the exact chemical structure of an unidentified NPS with an accuracy of 51%, and a top-10 accuracy of 78%. Our generative approach has the potential to enable *de novo* structure elucidation for other types of small molecules that are routinely analyzed by mass spectrometry.

The past decade has witnessed an explosive increase in the availability of new psychoactive substances (NPSs), also known as 'designer drugs' or 'legal highs'[1,2]. NPSs are typically created by slight modifications to the chemical structure of existing drugs of abuse, generating derivatives that circumvent drug control legislation while retaining their psychoactive properties[3]. Examples of well-known NPSs include synthetic cannabinoids ('spice'), synthetic cathinones ('bath salts'), psychedelic tryptamines and phenethylamines, and, more recently, synthetic opioids[4].

NPSs are synthesized by clandestine chemists, who mine the scientific and patent literature to identify compounds targeting the same receptors as existing psychoactive drugs[5]. The ingenuity of these chemists, and the lack of controls on the distribution of these compounds, means that new NPSs are continuously entering the 'grey market,' at a rate of roughly one compound per week[1]. At the same time, established drugs may rapidly disappear from the market in response to legislation[6]. The pharmacology and toxicology of NPSs have not been well characterized, and many have been associated with life-threatening toxidromes and fatalities[7]. Patients intoxicated with a NPS thus present a significant burden to healthcare systems[1,8,9]. This public health burden obliges forensic laboratories around the globe to screen for NPSs in law enforcement seizures or biological samples. However, the chemical diversity of these substances, and the

fleeting nature of their appearance on the illicit market, poses a profound challenge to the detection and identification of novel compounds, pitting forensic scientists against clandestine chemists in a cat-and-mouse game[10].

Identifying a new designer drug within a seizure or biological sample is challenging for several reasons. First is the high degree of structural similarity between candidate NPSs, which are often analogues from the same medicinal chemistry series[11,12]. A second challenge is the rapid rate at which novel compounds emerge onto the grey market, which necessitates the development of new assays for previously unknown substances[13,14]. Assay development requires substantial time and effort, and the inherent novelty of NPSs means that analytical reference materials are rarely available for NPSs that have recently entered the market[15].

A number of analytical methods have been developed to overcome these challenges. Historically, screening was accomplished predominantly by immunochemical approaches, but these are limited by their low sensitivity, inability to provide component-resolved drug profiles, and the time and effort required to establish new assays[16,17]. More recently, mass spectrometry (MS) has emerged as the method of choice for NPS detection and identification[18]. High-resolution mass spectrometry (HR-MS) can provide highly accurate mass measurements for a given analyte, narrowing the list of potential candidates and allowing for compari-

son against a reference database. Tandem mass spectrometry (MS/MS) provides additional information in the form of diagnostic product ions, allowing for higher-confidence molecule identification. However, a key shortcoming of mass spectrometric approaches is that, in order to identify a NPS by its exact mass or tandem mass spectrum, investigators minimally require its chemical structure to be present in a reference database. This presents an obstacle to the identification of new designer drugs that have just emerged on the market, and whose structures are, by definition, unknown to law enforcement or forensic laboratories. Elucidating the complete chemical structures of these novel compounds is generally thought to require an orthogonal technique—most commonly, nuclear magnetic resonance spectroscopy (NMR)[19], which necessitates large amounts of NPS material as input, is labour-intensive, and requires additional expertise. Moreover, due to its low sensitivity, NMR cannot be applied to screen human tissues in cases of suspected NPS intoxication.

Here, we present DarkNPS, a deep learning-enabled system to automatically elucidate the chemical structures of unidentified NPSs using only mass spectrometric data. Our approach is based on the use of a deep generative model of chemical structures. Models of this family have attracted intense interest within the fields of chemistry and deep learning for their potential to generate molecules with arbitrary physicochemical or biological properties on demand[20–24], thereby solving what has been termed the 'inverse design' problem[25]. Much of this work has focused on the possibility of generating ligands active against a particular receptor[26]. Here, we seek instead to generate NPS-like molecules that match one or more analytically measured properties. We achieve this by using strategies adapted to the low-data regime[27,28] to learn a robust generative model of designer drugs from only ~1,700 examples[15]. Sampling from this model allows us to stochastically generate new molecules that populate the same chemical space as existing designer drugs. We validate DarkNPS using a held-out set of 194 NPSs that were received by forensic laboratories after our training set was finalized, demonstrating that our model successfully anticipated >90% of NPSs that subsequently appeared on the illicit market. We then show that the frequency at which novel molecules are sampled from the model can be used to suggest the chemical structure most likely to explain an observed exact mass. Integration of the generated structures with tandem mass spectrometry data further improves the accuracy of structure elucidation. We demonstrate the application of DarkNPS to elucidate the structure of a novel designer drug that first appeared in Europe in February 2021, and which at the time of writing had not been described in the peer-reviewed literature.

Results

A deep generative model of novel psychoactive substances.
A number of computational tools have been developed to enable the automated identification of drugs and their metabolites within mass spectrometric data[29]. However, all of these tools require a database of known chemical structures as input, against which to compare the observed mass spectrometric data. As a result, these tools cannot be used to identify newly synthesized designer drugs that are not found in existing databases. We reasoned that by generating a database of novel, NPS-like chemical structures, we could automate the identification of entirely unknown NPSs. We therefore set out to learn a deep generative model of NPS chemical structures, from which we could then stochastically sample novel NPS structures (**Fig. 1a-b**).

We obtained a training dataset of NPS chemical structures from HighResNPS, a database developed to facilitate NPS screening using mass spectrometry[15]. Contributors from dozens of forensic laboratories around the world submit data to HighResNPS when new substances are detected in biological samples or law enforcement seizures, making this database arguably the most up-to-date and comprehensive resource of NPS structures. Despite this crowdsourced effort, however, the database contained only 1,753 unique NPS structures at the beginning of June 2020.

The limited size of this dataset reflects the number of NPSs that have appeared on the illicit market and subsequently been detected by forensic laboratories. However, it is orders of magnitude smaller than the datasets that have typically been used to train generative models of chemical structures, which are generally thought to require training datasets comprising hundreds of thousands—if not millions—of examples[26].

We hypothesized that this small training dataset could nonetheless provide a basis to learn a robust generative model of NPS chemical structures. We recently carried out a systematic analysis of deep generative models of molecules in the low-data regime[27], and showed that it is possible to learn robust models from far smaller datasets than has been widely assumed. We also identified strategies that facilitate learning from a small number of examples. One of the most effective such strategies takes advantage of the fact that a single molecule can be represented by multiple SMILES strings, depending on the order in which the atoms in the graph are traversed. This redundancy opens up an opportunity for data augmentation, by enumerating multiple non-canonical SMILES for each molecule in the training dataset (**Fig. 1c**)[28]. However, we also identified a risk of 'over-augmentation,' in which excessive non-canonical SMILES enumeration actually degrades the performance of the trained model.

To empirically determine the optimal degree of data augmentation, we trained deep generative models on the HighResNPS dataset after subjecting it to varying degrees of non-canonical SMILES enumeration. We also experimented with two different recurrent neural network-based architectures, including gated recurrent units (GRUs) and long short-term memory networks (LSTMs). We evaluated model performance using five metrics that we had previously found to be robust indicators of model quality[27]. These metrics generally suggested that a high degree of SMILES enumeration markedly improved model performance, and that LSTM models slightly outperformed GRUs (**Fig. 1d** and **Supplementary Fig. 1a-c**). Integrating all five metrics into a sin-
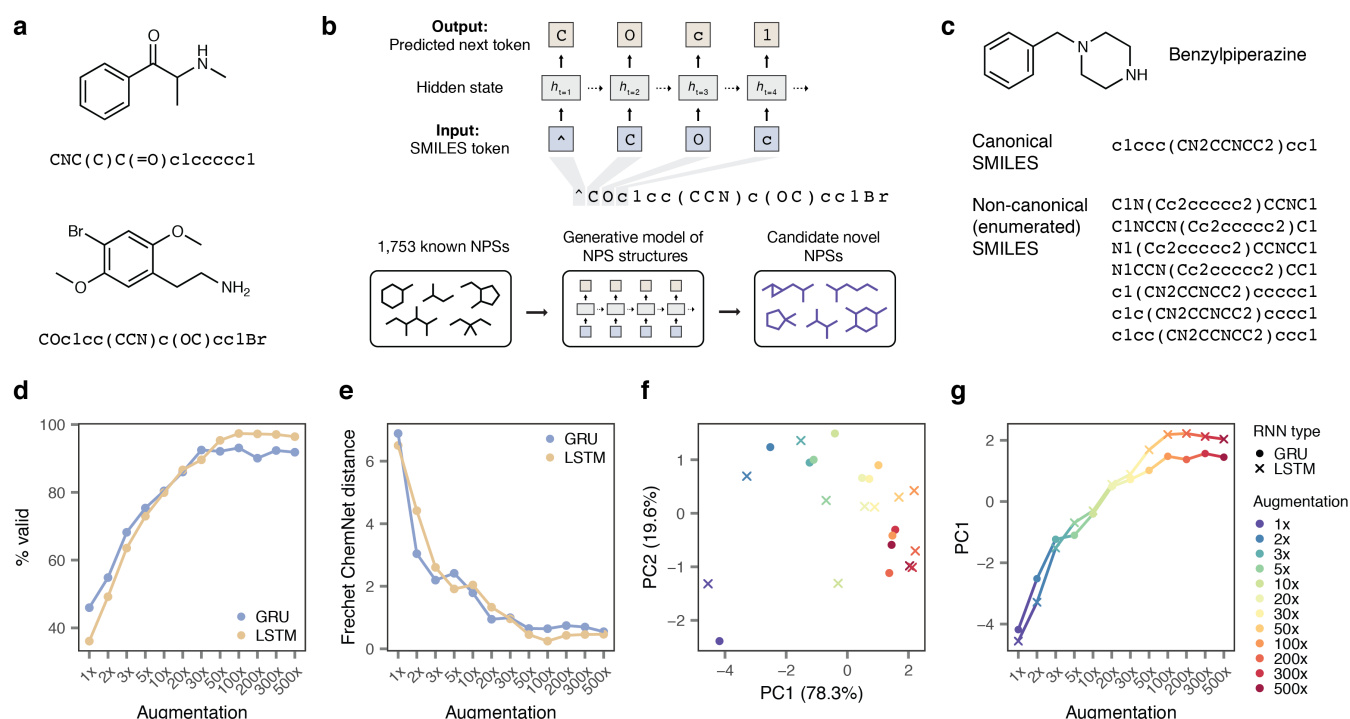
Fig. 1 | A deep generative model of novel psychoactive substances.
a, Chemical structures and canonical SMILES representations of two exemplary designer drugs, methcathinone (top) and 2C-B (bottom).
b, Top, schematic overview of the recurrent neural network-based generative model. A SMILES string is split into tokens, and the start-of-string token (ˆ) is prepended to the tokenized SMILES. The model is trained to predict the next token, given the sequence of tokens that has already appeared. Bottom, the generative model is trained on the SMILES representations of known designer drugs. SMILES strings are then stochastically sampled from the trained model by providing only the start-of-string token as input, enabling generation of novel candidate NPSs.
c, Canonical SMILES and seven enumerated non-canonical SMILES for an example designer drug, benzylpiperazine.
d, Proportion of valid SMILES strings generated by recurrent neural network-based models trained on the HighResNPS database after varying degrees of non-canonical SMILES enumeration.
e, Fréchet ChemNet distances to the training set for recurrent neural network-based models trained on the HighResNPS database after varying degrees of non-canonical SMILES enumeration.
f, Principal component analysis of top-performing metrics for molecules generated by recurrent neural network-based models trained on HighResNPS database after varying degrees of non-canonical SMILES enumeration.
g, PC1 scores for GRU and LSTM models trained on the HighResNPS database after varying degrees of non-canonical SMILES enumeration.

gle consensus measure of model performance using principal component analysis[27] confirmed the trends that were apparent from inspection of individual metrics (**Fig. 1e-f** and **Supplementary Fig. 1d**). Based on these results, we selected a LSTM model, trained on a dataset in which 100 non-canonical SMILES were enumerated for each unique molecule, for further analysis.

Generated molecules closely resemble known designer drugs. We next sought to characterize the molecules generated by our model in more detail. As a first step, we asked whether the structural and physicochemical properties of the generated molecules were similar to those of known NPSs. To address this question, we sampled 500,000 SMILES strings from our trained model. Of these, 62,354 were syntactically valid and corresponded to molecules that were not found within the training set. We compared these generated molecules to the 1,753 known NPSs that comprised the training set.

We computed a series of chemical properties for each known NPS and generated molecule, including its atomic composition, the number of ring systems it con-

tained, its molecular weight, its topological complexity[30], its octanol-water partition coefficient[31], and measures of drug-likeness[32], natural product-likeness[33], and synthetic accessibility[34]. Strikingly, despite the limited amount of training data, we found that the generated molecules had property distributions that were almost indistinguishable from those of known NPSs (**Fig. 2a-e** and **Supplementary Fig. 2a-d**).

To gain a more holistic perspective on the molecules generated by the trained model, we sought to visualize the chemical spaces occupied by known and generated NPSs. We embedded known NPSs and a random sample of generated molecules of equal size into two dimensions using the non-linear dimensionality reduction algorithm UMAP[35]. We then plotted the resulting two-dimensional embeddings, with either the known or generated NPSs overlaid on top of one another. These plots demonstrated that the generated molecules almost perfectly reproduced the chemical space of known NPSs, with very few regions of chemical space occupied exclusively by either known or generated drugs (**Fig. 2f**).

We also asked how the generated NPSs fit into the categories of designer drugs assigned by HighResNPS, which
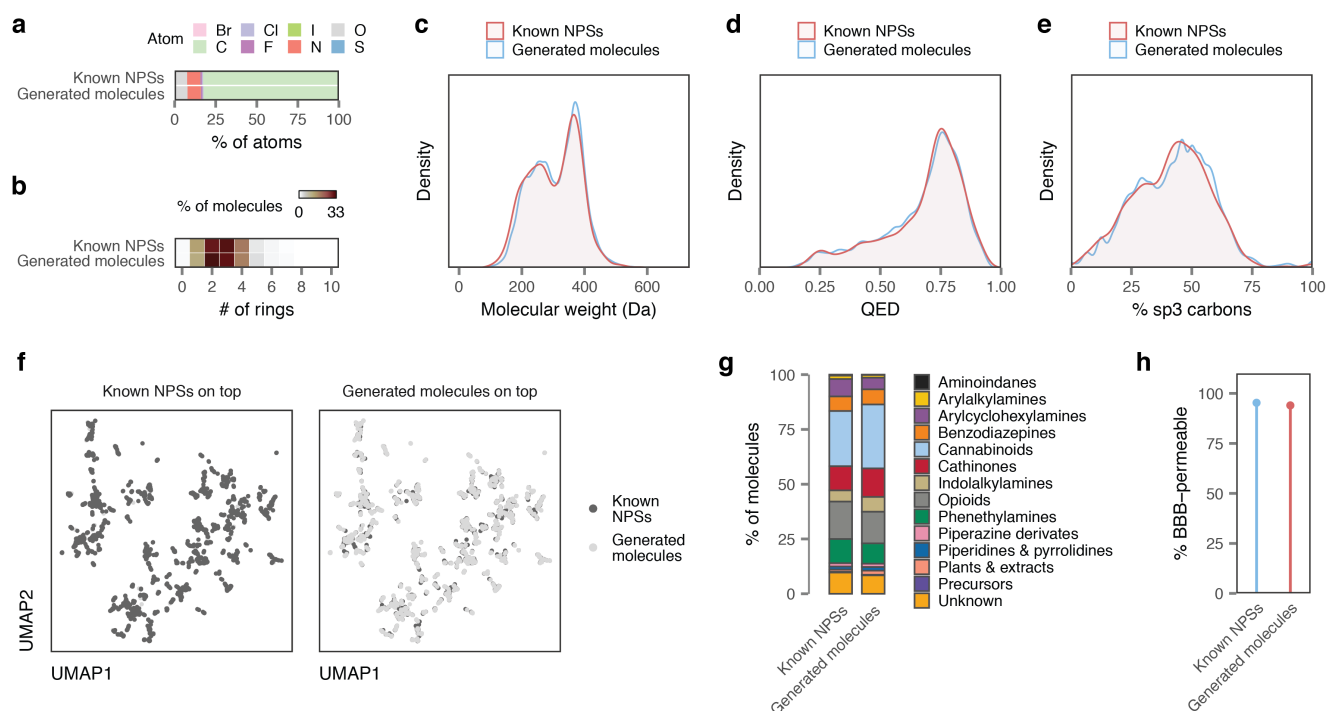
Fig. 2 | Generated molecules closely resemble known designer drugs.
a, Atomic composition of known NPSs and generated molecules.
b, Number of ring systems in known NPSs and generated molecules.
c, Molecular weights of known NPSs and generated molecules.
d, QED scores of known NPSs and generated molecules.
e, Proportion of carbons that are sp_3-hybridized within known NPSs and generated molecules.
f, UMAP visualization of known NPSs and an equal number of generated molecules sampled at random from the trained generative model. Left, known NPSs superimposed over generated molecules. Bottom, generated molecules superimposed over known NPSs.
g, EMCDDA categorizations of known NPSs and generated molecules.
h, Proportions of known NPSs and generated molecules predicted to cross the blood-brain barrier.

are based on those established by the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA). Overall, we observed a close correspondence between the EMCDDA categorizations of known and generated NPSs (**Fig. 2g** and **Supplementary Fig. 3a**). Only two categories were generated at frequencies significantly different from the training set, with cannabinoids being modestly enriched in the generative model output, and arylcyclohexylamines being moderately depleted (odds ratio, p = 0.040 and 1.9×10^{-4}, respectively; **Supplementary Fig. 3b**).

NPSs exert their psychoactive effects by acting at receptors in the brain, which they must cross the blood-brain barrier (BBB) to access. To validate the potential psychoactive properties of the generated NPSs, we used LightBBB[36] to predict the likelihood that they would cross the BBB. As a baseline, we also used LightBBB to predict the BBB permeability of known NPSs. We found that 95.3% of known NPSs were predicted to cross the BBB, consistent with the estimated false-negative rate of ~7% for this tool[36]. Among generated molecules, a very similar proportion (93.2%) were predicted to cross the BBB (**Fig. 2h**). This suggests that the generated molecules have the potential to access the same receptors in the brain at which known NPSs act.

Together, these results suggest that, with appropriate adjustments for the low-data regime, it is possible to learn a ro-

bust generative model of NPS chemical structures from only ~1,750 training examples. This model generated molecules whose physicochemical properties were nearly identical to those of known NPSs, and which populated overlapping regions of chemical space. These results support the notion that a library of generated molecules could be used to search for previously unknown NPSs within mass spectrometric data.

Sampling frequency defines a structural prior for the annotation of unknown NPSs. While inspecting the molecules generated by our model, we noticed that some molecules appeared repeatedly in the model output. To investigate this phenomenon further, we sampled a total of 1 billion SMILES strings from the generative model, and tabulated the frequency at which each unique chemical structure was found in this sample (**Fig. 3a-b**). After removing syntactically invalid SMILES strings and known NPSs, we identified a total of 8.9 million unique molecules within this sample. The vast majority of these molecules appeared just once, or at most a handful of times, in the model output. However, a long tail of molecules were repeatedly sampled tens or hundreds of thousands of times (**Fig. 3c** and **Supplementary Fig. 4a**).

We were surprised to observe that the model generated molecules at dramatically different frequencies, and sought to explain this unexpected finding. We hypothesized that
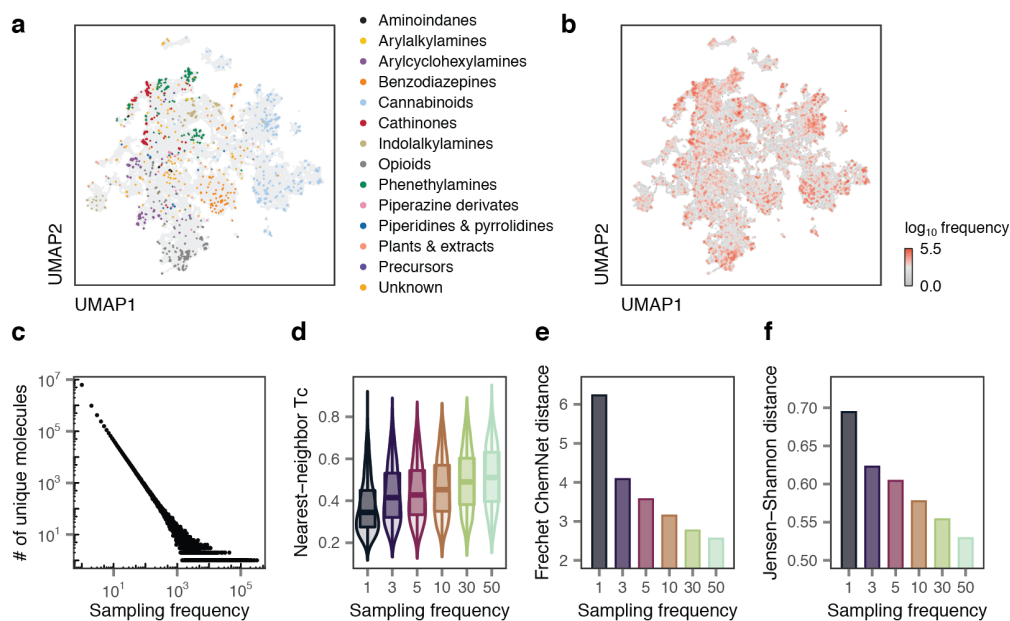
Skinnider *et al.* | Automated structure elucidation of designer drugs

Fig. 3 | Sampling frequency defines a structural prior over unseen molecules.

a, UMAP visualization of known NPSs and a random sample of up to 5,000 generated molecules at each sampling frequency in a sample of 1 billion SMILES strings. Known NPSs are colored by their EMCDDA categorizations, with generated molecules in grey.

b, As in **a,** but showing only generated molecules colored by their sampling frequency.

c, Distribution of sampling frequencies within a sample of 1 billion SMILES strings from the trained generative model.

d, Tanimoto coefficients between generated molecules and their nearest neighbor in the set of known NPSs, for molecules generated with progressively increasing frequencies.

e, Fréchet ChemNet distances between generated molecules and the set of known NPSs, for molecules generated with progressively increasing frequencies.

f, Jensen-Shannon distance between the Murcko scaffold compositions of generated molecules and the set of known NPSs, for molecules generated with progressively increasing frequencies.

the generative model had learned to implicitly evaluate the likelihood of novel NPSs, based on the structural properties of known designer drugs. In other words, we posited that molecules sampled more frequently by the model would exhibit a higher degree of structural similarity to known NPSs, and would be more likely to subsequently appear on the 'grey market.'

To test this hypothesis, we assessed the structural similarity of generated molecules and known NPSs, using the Tanimoto coefficient (Tc) as a quantitative measure of similarity[37,38]. We then compared the Tc between each generated molecule and its nearest neighbor among the set of known NPSs, for molecules generated at progressively increasing frequencies by the trained model. Molecules sampled more frequently exhibited significantly greater similarity to an existing NPS ($p < 10^{-15}$, Jonckheere-Terpstra test), supporting the hypothesis that the sampling frequency reflects the implicit likelihood of observing a novel NPS structure (**Fig. 3d**).

To further corroborate this notion, we computed a range of physicochemical properties for molecules sampled at increasing frequencies from the generative model. We then compared these properties to those of known NPSs. We found that molecules sampled more frequently from the generative model had a lower Fréchet ChemNet distance to the training set[39], and better matched the distribution of Murcko scaffolds found in known NPSs[40] (**Fig. 3e-f**). Moreover, fre-

quently sampled molecules also better matched the molecular weights, partition coefficients, drug-likenesses, and stereochemical complexities of known NPSs (**Supplementary Fig. 4b-g**).

Taken together, these findings demonstrate that novel molecules generated frequently by our model are more similar to known NPSs than those generated infrequently. In turn, this raises the possibility that the sampling frequency could be used to prioritize the most likely structures of novel NPSs.

Anticipating the structures of unidentified designer drugs. Our experiments established that frequently sampled molecules are more similar to known NPSs. This finding led us to ask whether these frequently sampled molecules are also more likely to subsequently appear on the grey market. In other words, we asked whether we could leverage the implicit likelihood learned by the generative model to anticipate the chemical structures of as-of-yet unsynthesized drugs.

To test this possibility, we assembled a held-out set of 194 NPSs, which were identified by forensic laboratories and added to the HighResNPS database only after our training set was finalized. We then asked what proportion of these held-out NPSs were successfully anticipated by our model. A total of 176, or 90.7%, appeared at least once within our sample of 1 billion SMILES strings (**Fig. 4a**). The 18 held-out molecules that were never sampled by the generative model exhibited significantly less structural similarity to any known NPS in the training set, as quantified by the Tc ($p = 2.9 \times$
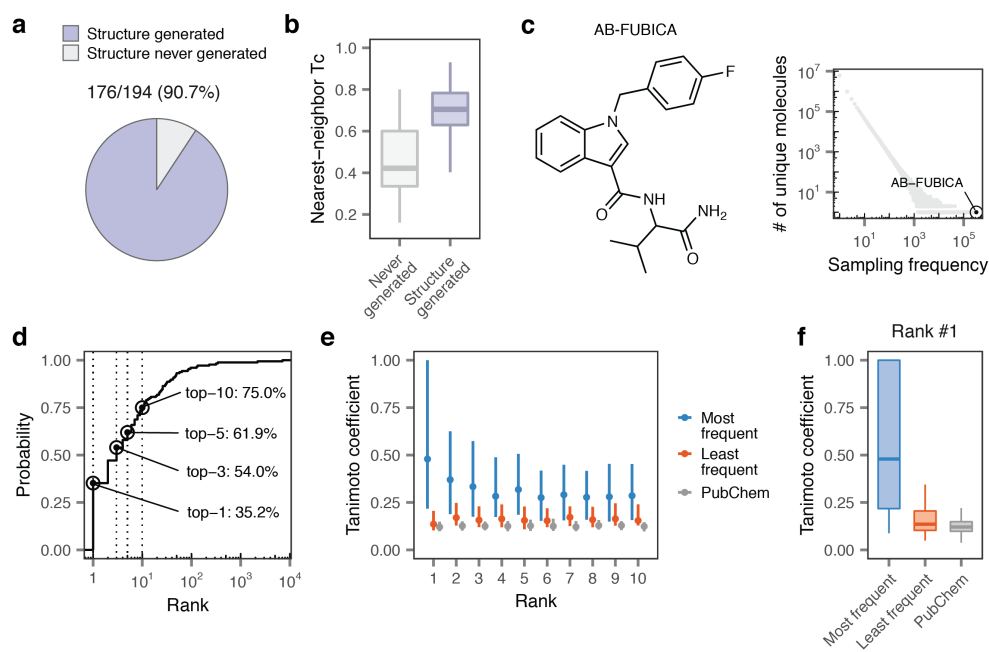
Fig. 4 | Automated structure elucidation of unidentified NPSs.

a, Proportion of molecules within the set of 194 NPSs added to the HighResNPS database between October 2020 and April 2021 that appeared at least once within a sample of 1 billion SMILES strings from the generative model.

b, Tanimoto coefficients between held-out NPSs and their nearest neighbor in the training set, for molecules in the held-out that were generated at least once vs. molecules in the held-out set that were never generated.

c, Example of a molecule in the held-out set, AB-FUBICA, that was correctly anticipated by the generative model. Left, structure of AB-FUBICA. Right, sampling frequency of AB-FUBICA.

d, Ranks of held-out NPSs among generated molecules matching their exact masses within a window of 10 ppm, arranged in descending order by sampling frequency.

e, Median Tanimoto coefficient between held-out NPSs and generated molecules matching their exact masses ($\pm$ 10 ppm), arranged in descending order by sampling frequency ("most frequent"), ascending order by sampling frequency ("least frequent"), or a random sample of molecules with matching exact masses from PubChem. Error bars show the interquartile range.

f, Distribution of Tanimoto coefficients between held-out NPSs and generated molecules matching their exact masses ($\pm$ 10 ppm), taking either the single most frequently sampled generated molecule, the single least frequently sampled generated molecule, or a random molecule with a matching exact mass from PubChem.

10^{-13}; **Fig. 4b**). This reflects an inherent limitation of our model: namely, it can only generate novel molecules that are structurally similar to known designer drugs. However, closer inspection revealed that some of these 18 molecules were not actually designer drugs at all. For example, some of the molecules in the held-out set included the dietary supplement citicoline, the antipsychotic clozapine, or the alcohol dependence medication nalmefene (**Supplementary Fig. 5**). After curating the held-out set to remove these questionable entries, the proportion of structures anticipated by our model climbed to 93.1% (176/189).

Interestingly, although a handful of the held-out NPSs were sampled only once or twice from the model, the vast majority were among the relatively small subset of generated molecules that appeared 50 or more times in our sample of 1 billion SMILES (**Supplementary Fig. 6**). This observation further supported the possibility that the sampling frequency could be used to prioritize candidate NPSs most likely to emerge on the grey market in the future. As one striking example, the single most frequently generated novel molecule, appearing 323,299 times within our sample, was the synthetic cannabinoid AB-FUBICA, which was added to the HighResNPS database in October, 2020 (**Fig. 4c**).

Structure elucidation of unidentified NPSs from accurate mass measurements. Encouraged by these results, we asked whether we could leverage the sampling frequency to anticipate the most likely chemical structure for an unidentified NPS subjected to mass spectrometric analysis. When analyzing a law enforcement seizure by mass spectrometry, the first clue to the identity of the seized compound that investigators receive is its mass. We therefore devised an experiment to test the feasibility of elucidating the structure of a novel NPS from an accurate mass measurement alone. For each of the NPSs in the held-out set, we searched in our sample of 1 billion SMILES strings to identify all generated molecules matching the exact mass of the held-out NPS, allowing for a window of $\pm$ 10 ppm to account for the accuracy of modern HR-MS instrumentation. We then sorted these matches by their sampling frequency in descending order and calculated the frequency with which the correct molecule was ranked first, or within the top 3, top 5, or top 10 candidates. We dubbed this workflow the 'structural prior,' on the basis that it provides a prior probability distribution over all possible chemical structures matching the exact mass of the unidentified molecule.
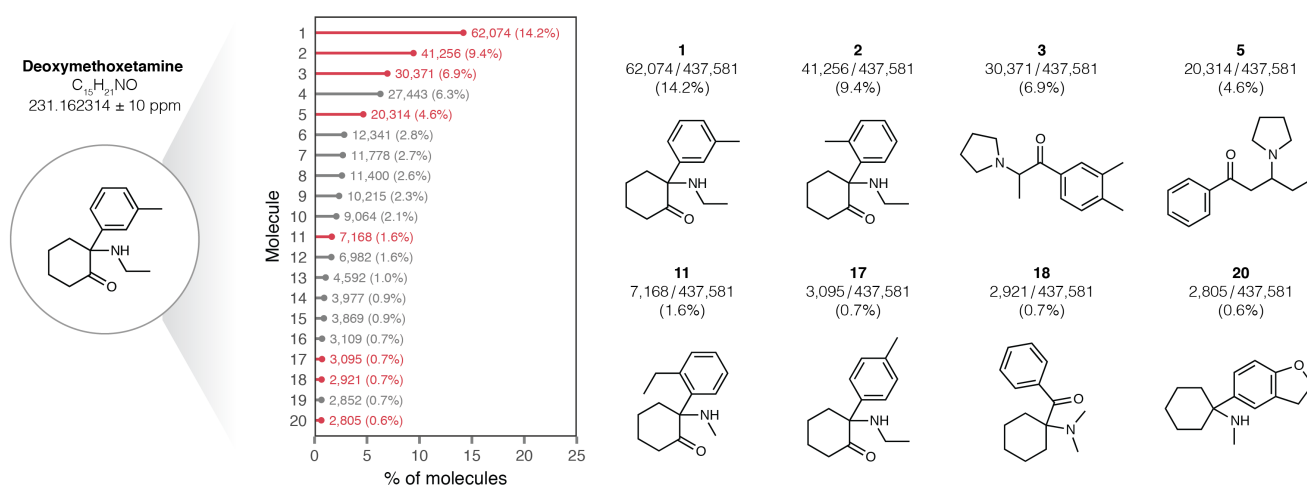
Remarkably, using only an accurate mass as input, we

Fig. 5 | Application of the structural prior to the designer dissociative deoxymethoxetamine.
Left, the chemical structure, molecular formula, and exact mass of deoxymethoxetamine. Middle, sampling frequencies of the 20 most frequently sampled molecules matching the exact mass of deoxymethoxetamine ($\pm$ a 10 ppm window). An illustrative subset of the generated molecules, highlighted in red, are shown at the right.

found our structural prior could predict the chemical structure of the unidentified molecule with perfect accuracy 35.2% of the time (**Fig. 4d**). Moreover, the correct structure was ranked among the top 3 candidates 54.0% of the time, and among the top 10 candidates 75.0% of the time. This performance is remarkable, given that only a single piece of information is provided to the model as a basis for the prediction of complete chemical structures of entirely novel molecules.

We reasoned that in the cases where the structural prior failed to rank the correct molecule first, or even among the top 10 candidates, it would still be very helpful to forensic scientists if the top prediction was a closely related analog. To evaluate whether the top-ranked molecule was at least structurally similar to the correct structure, we computed the Tc between the candidates nominated by the structural prior and the unidentified NPS. As a baseline, we also computed the Tc for the molecules sampled less frequently by the generative model. As a second baseline, we searched for the exact mass of the unidentified NPS against the PubChem database, which is commonly used as a reference for unidentified mass spectrometric signals[41].

Interestingly, we noted that molecules sampled infrequently from the generative model were more similar to the correct structure than isobaric molecules from PubChem (**Fig. 4e**). This observation likely reflects the fact that even infrequently sampled molecules populate the chemical space of known designer drugs, unlike molecules sampled at random from PubChem. However, the molecules nominated by the structural prior were dramatically more similar to the unidentified NPS than either baseline (**Fig. 4e**). This similarity was particularly apparent when inspecting the Tc for only the single top-ranked molecules in more detail (p $\leq$ 1.9 $\times$ 10^{-55}; **Fig. 4f**). These analyses indicate that even when the structural prior does not perfectly annotate the structure of an unidentified NPS, it tends to at least prioritize molecules that are highly similar.

A limitation of the Tc in evaluating chemical similarity

is that its range of possible values scales with the sizes of the molecules being compared[42]. As a second, orthogonal measure of chemical similarity, we computed the Euclidean distance between continuous molecule embeddings derived from a neural machine translation task[43]. We reproduced our finding that the structural prior markedly outperformed both baselines when using continuous embeddings to quantify chemical similarity (**Supplementary Fig. 7a-b**).

To illustrate the power of the structural prior, we focused on an illustrative example of a new designer drug, deoxymethoxetamine (DXME). DXME is a dissociative hallucinogen of the arylcyclohexylamine class, which includes well-known drugs of abuse such as ketamine and phencyclidine (PCP). It appears to have first emerged on the illicit market in late 2020, and was added to the HighResNPS database in February 2021 after being identified in a law enforcement seizure in Denmark. At the time of writing, it had not been described in a peer-reviewed article, rendering this a representative prospective application of the structural prior. Querying the structural prior with the exact mass of DXME returned a list of 11,479 candidate structures. This enormous number of candidates illustrates the difficulty of predicting complete chemical structures from only an accurate mass. Yet, despite having never seen this molecule during training, the structure of DXME was correctly ranked as the single most frequent match to the exact mass, appearing 62,074 times in our sample of 1 billion SMILES (**Fig. 5**). Moreover, the second-most frequently sampled compound, appearing 41,256 times, was a closely related isomer, differing only in the position of a methyl group on the aromatic ring. Interestingly, several other candidates ranked within the top 20 were arylcyclohexylamines structurally related to DXME, suggesting the model deemed it likely that the exact mass in question belonged to the arylcyclohexylamine category, despite the fact that these were generally underrepresented in the model output (**Supplementary Fig. 3b**). As a second example, we found that the structural prior correctly

elucidated the chemical structure of ADB-HEXINACA, the most recent synthetic cannabinoid to have emerged on the US market at the time of writing, selecting the most likely structure from among a series of closely related analogues (**Supplementary Fig. 8**).

Tandem mass spectrometry enables high-confidence annotation of unidentified NPSs. Our experiments to this point have shown that the structural prior can generate remarkably accurate annotations of the structure of an unidentified NPS from an accurate mass measurement alone. However, accurate mass measurements are fundamentally limited in their ability to distinguish structural isomers with the same molecular formula. The limitations of accurate mass measurements are especially apparent in cases where several analogues from the same medicinal chemistry series with identical chemical formulas could plausibly represent novel designer drugs. Such isomers can, however, be differentiated using tandem mass spectrometry (MS/MS)[18]. We therefore asked whether integrating MS/MS data into the predictions made by the structural prior could further improve the accuracy of structure elucidation.

To test this notion, we used CFM-ID[44,45] to predict tandem mass spectra for all 8.9 million generated molecules. We then compared the accuracy of structure annotations assigned by three different approaches: (i) the structural prior alone, (ii) CFM-ID alone, or (iii) the combination of the two. To combine CFM-ID predictions with the structural prior, we re-weighted the scores assigned by CFM-ID according to the prior probabilities assigned by the generative model. We then evaluated the accuracy of each approach in our held-out dataset, restricting our analysis to the 79 NPSs in our held-out set for which MS/MS data had been deposited to HighResNPS.

Integrating tandem mass spectrometry data yielded substantially more accurate predictions than those made by the structural prior alone. The combined approach successfully elucidated the complete chemical structures of 40 unidentified NPSs (51%), as compared to 30 correctly elucidated by the generative model (38%) and only one (1%) by CFM-ID alone (**Fig. 6a**). Similar improvements in the top-k accuracy were apparent for many values of k. For instance, the combined approach ranked the correct chemical structure within the top-3 68% of the time, compared to 57% for the generative model alone and 8% for CFM-ID alone (**Fig. 6b**).

An example of an NPS for which the automated elucidation of the complete chemical structure relied on the integration of MS/MS data is the 5-hydroxyindole analogue of JWH-122, as shown in **Fig. 6d**. The structural prior selected a closely related analogue from among 2,599 generated molecules matching the exact mass, but with a methoxy group misplaced, yielding a Tc of 0.41. Incorporating the predicted mass spectra for all 2,599 possible matches into the structural prior rescued the correct structure.

Even when the correct molecule was not the top-ranked hit, integrating MS/MS data yielded structural annotations that were more chemically similar to the unidentified NPS than those generated by the structural prior alone, as quan-
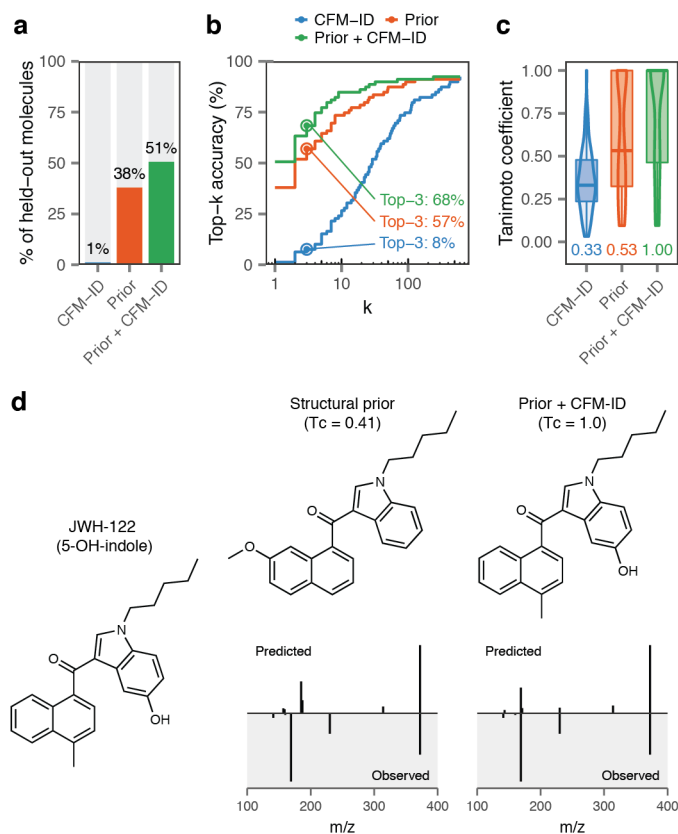
Fig. 6 | High-confidence structure elucidation using tandem mass spectrometry.
a, Top-1 accuracy with which the complete chemical structures of unidentified NPSs in the held-out set were correctly elucidated by CFM-ID alone, the structural prior alone, or the combination of the two.
b, Top-k accuracy curve of structure elucidation of unidentified NPSs in the held-out set by CFM-ID alone, the structural prior alone, or the combination of the two.
c, Tanimoto coefficients between the held-out set of unidentified NPSs and the top-ranked structures suggested by CFM-ID alone, the structural prior alone, or the combination of the two.
d, Automated structure elucidation of an unidentified NPS using tandem mass spectrometry. Left, the chemical structure of the 5-hydroxyindole analogue of JWH-122. Middle, the top-ranked molecule suggested by the structural prior (top) and mirror plot comparing the observed tandem mass spectrum for the 5-hydroxyindole analogue of JWH-122 with the tandem mass spectrum predicted by CFM-ID. Right, the top-ranked molecule after integrating the structural prior with MS/MS evidence (top) and mirror plot comparing observed and predicted tandem mass spectra.

tified either by the Tc (**Fig. 6c**) or the Euclidean distance between CDDD embeddings (**Supplementary Fig. 9a**). For instance, given the accurate mass of α-hydroxyetizolam as input, the structural prior selected a structure with relatively little resemblance to the ground truth (**Supplementary Fig. 9b**). However, after incorporating MS/MS data into the structural annotation, the top-ranked molecule was almost perfectly correct, with the lone exception of a misplaced hydroxyl group.

Collectively, these experiments demonstrate that integrating MS/MS data into DarkNPS enables high-confidence structural annotation, yielding a system that is capable of automatically elucidating complete chemical structures from mass spectrometry data alone.

Discussion

The proliferation of unregulated substances marketed as legal alternatives to established drugs of abuse presents a major challenge to public health. However, the identification of designer drugs that have recently emerged on the illicit market is a low-throughput and labour-intensive endeavour. Here, we describe a system capable of anticipating the chemical structures of the NPSs most likely to emerge on the illicit market in the future, and annotating the most likely structure of an unidentified NPS using mass spectrometric data. We prospectively validated our model in a held-out set of 194 NPSs that were identified by forensic laboratories around the globe after our training set was finalized. We demonstrate that our method generates highly accurate annotations of the structure of an unidentified NPS from its exact mass alone, using the concept of the structural prior, and that these annotations are further improved by the integration of MS/MS data. Our final model was able to perform automated structure elucidation of complete chemical structures with an accuracy over 50%. Moreover, in cases where the model did not correctly identify the exact structure of the unknown NPS, it typically suggested a closely related analogue. This performance is remarkable given that *de novo* structure elucidation is typically thought to require experimental techniques that are entirely orthogonal to mass spectrometry, most notably NMR. Our method thus has the potential to dramatically accelerate the pace at which emerging designer drugs can be identified by forensic, toxicological, police, and customs laboratories.

Many of the challenges faced by investigators seeking to determine the structure of an unknown NPS are ubiquitous throughout the field of analytical chemistry. Current computational approaches to mass spectrometric data make use of experimentally measured information such as exact masses, fragmentation patterns, and isotopic distributions[29]. In this work, we posited that the chemical space of interest itself provides a highly informative prior that can be used to nominate the most likely structures matching an experimentally observed property. In other words, given the structures of known observed designer drugs as input, we demonstrate that we can learn a statistical probability distribution over unobserved designer drug structures, and define those that are more or less likely to be observed in the future. This represents a conceptually new approach to the interpretation of mass spectrometric data. Moreover, we show that this paradigm is complementary to existing approaches for searching tandem mass spectrometry data against a database of chemical structures. Indeed, we find that our approach can dramatically improve the accuracy of chemical database search. Our approach may therefore find broad application to other problems in the field of analytical chemistry: for instance, the study of xenobiotic metabolism or the identification of environmental pollutants.

Critical to the success of this effort was our ability to learn a robust generative model of chemical structures from a small number of examples. Remarkably, we were able to train an excellent generative model from only ~1,700 known NPSs.

This dataset is orders of magnitude smaller than those that have conventionally been used to train generative models[26]. What factors underlie the surprisingly good performance of our model from such a small amount of training data? Data augmentation by non-canonical SMILES enumeration had a dramatic impact on model performance, consistent with previous results[27,46,47]. Another factor that likely contributed to our success is that the chemical space of NPSs is relatively homogenous. These substances are derived from a small number of core structures, using a finite vocabulary of medicinal chemistry transformations. This notion is consistent with our finding that generative models are dramatically more likely to succeed in low-data settings when the training set is less diverse[27], and suggests it might be possible to learn generative models for many restricted chemical spaces of biomedical interest.

A limitation of our approach is that it requires us to draw a very large sample from the generative model, in order to tabulate the frequency with which each unique molecule appears in the model output. This is due both to the redundancy of the SMILES format (that is, many different SMILES strings can correspond to the same molecule), and the fact that the model does not know in advance what the exact mass of a given SMILES string will be while generation is still in progress. Future efforts could conceivably improve the computational efficiency by conditioning molecule generation on one or more experimentally observed properties.

Methods

Training dataset. We obtained a training dataset of 1,753 chemical structures corresponding to known NPSs, their metabolites, and common drugs of abuse from HighResNPS (https://highresnps.forensic.ku.dk)[15]. HighResNPS is a free, on-line, crowdsourced database of NPS structures and accompanying high-resolution mass spectrometry data, initiated and managed by researchers at the Section of Forensic Chemistry at the University of Copenhagen. Forensic toxicology and chemistry laboratories from around the world submit data to HighResNPS when novel designer drugs are detected and analyzed by a mass spectrometer. New compounds may also be added when they have been reported by drug monitoring agencies such as, but not limited to, the United Nations Office of Drugs and Crime (UNODC), European Monitoring Centre of Drugs and Drug Addiction (EMCDDA) and the Drug Enforcement Administration (DEA). Entries minimally include the unambiguous chemical structures of the detected molecules, and may also include tandem mass spectrometry data, or diagnostic product ions derived from theoretical bond dissociations.

The training set was obtained from the HighResNPS database in June, 2020. At that time, the total number of entries in the database corresponded to 2,065 unique molecules. These entries had been contributed by 57 laboratories located in 21 different countries. Each molecule in the database had also been assigned to a class of designer drugs based on the EMCDDA categorizations. All 2,065 molecules were parsed by the RDKit, after which charged moieties were neutralized, using code provided in the RDKit documentation, and molecules were converted into their canonical SMILES forms with stereochemistry removed. After this preprocessing step, redundant SMILES representations (e.g., stereoisomers or alternatively charged forms of the same molecule) were discarded, leaving a total of 1,761 unique canonical SMILES. We then removed a further eight molecules containing characters (the phosphorus symbol, P, and the token for a fifth ring atom, 5) that were each found in less than 0.05% of the 1,761 molecules, reasoning that it was unlikely the model would be able to learn how to use these tokens from such a small number of examples. Collectively, these preprocessing steps yielded a dataset of 1,753 canonical SMILES that formed the basis for all further analysis.

Generative models. Recurrent neural network-based models of SMILES strings were trained on canonical SMILES or non-canonical SMILES after varying degrees of data augmentation, using either LSTM or GRU architectures. The Python source code used to train the model was derived from our recent benchmarking analysis of generative models of molecules in the low-data regime[27] (https://github.com/skinnider/low-data-generative-models), which was itself adapted from the REINVENT package[24,48] (http://github.com/MarcusOlivecrona/REINVENT). Briefly, each SMILES was converted into a sequence of tokens by splitting the SMILES string into its constituent characters, except for atomic symbols composed of two characters (Br, Cl) and environments within square brackets, such as [nH]. The vocabulary of the RNN consisted of all unique tokens detected in the training data, as well as start-of-string and end-of-string characters and a padding token. Enumeration of non-canonical SMILES was performed using the SmilesEnumerator class available from http://github.com/EBjerrum/SMILES-enumeration. We experimented with varying degrees of non-canonical SMILES enumeration, assembling training sets in which between one and 500 non-canonical SMILES strings were enumerated for each unique molecule in the training dataset. The final model was a LSTM trained on a dataset with an augmentation factor of 100x.

The architecture of the recurrent neural networks consisted of three-layer GRU or LSTM models, with a hidden layer of 512 dimensions, an embedding layer of 128 dimensions, and no dropout layers. Models were trained using the Adam optimizer with β_1 = 0.9 and β_1 = 0.999, with a batch size of 128 and a learning rate of 0.001, using teacher forcing. 10% of the molecules in the training set were reserved as a validation set and used to perform early stopping with a patience of 50,000 minibatches. A total of 500,000 SMILES strings were sampled from each trained model after completion of model training.

Model evaluation. To select an optimal recurrent neural network architecture and degree of SMILES enumeration, we evaluated the trained models using a set of five metrics that we had previously found to be robust indicators of the quality of generative models of molecules[27]. Each of these five metrics seeks to quantify the degree to which the generated molecules resemble the training set (in this case, known NPSs). The five metrics in question included:

- The proportion of valid molecules generated by the model, where "valid" molecules are those that can be parsed by the RDKit ("% valid").

- The Fréchet ChemNet distance[39] between the training and generated molecules ("FCD"). The PyTorch implementation available from http://github.com/insilicomedicine/fcd_torch was used to calculate the FCD.

- The Jensen-Shannon distance between the distributions of Murcko scaffolds[40] of known NPSs and generated molecules.

- The Jensen-Shannon distance between the natural product-likeness score[33] distributions of known NPSs and generated molecules.

- The Jensen-Shannon distance between the distribution of the proportion of atoms in each molecule that were stereocenters in known NPSs and generated molecules.

The Murcko scaffolds, natural product-likeness, and proportion of stereocenters were calculated using the RDKit, and the Jensen-Shannon distance was calculated using scipy.

In addition to considering each of these metrics individually, we also integrated them into a single measure of model performance using principal component analysis to account for the covariance between metrics, as previously described[27]. PCA was on the centered and scaled matrix of model performance metrics, using the R function 'princomp', and the loadings of each model on the first principal component (PC1) were used for model evaluation.

Physicochemical properties. After selecting a LSTM-based generative model with an augmentation factor of 100x for further exploration, we sought to characterize the molecules sampled from the trained model in greater detail. To this end, we computed a series of physicochemical or structural properties for each generated molecule. A sample of 500,000 SMILES strings was drawn from the trained model, and these SMILES were parsed using the RDKit to remove syntactically invalid strings or molecules that were found in the training set. We then used the RDKit to compute the NP-likeness and proportion of stereocenters for each generated molecule, both as described above, as well as six additional properties, including (i) the molecular weight; (ii) the calculated octanol-water partition coefficient[31]; (iii) the topological complexity[30]; (iv) the synthetic accessibility score[34]; (v) the quantitative estimate of drug-likeness (QED) score[32]; and (vi) the proportion of carbons in the molecule that were sp^3-hybridized. These calculations were then repeated for

the known NPSs in the training set in order to provide a basis for comparison.

Chemical space analysis. To obtain a more holistic perspective on the chemical spaces occupied by known NPSs and the generated molecules, we used a previously described pipeline to visualize both sets of molecules within a two-dimensional space[27]. Briefly, we computed a continuous, 512-dimensional representation of each molecule using the Continuous and Data-Driven Descriptors (CDDD) package[43] (available from http://github.com/jrwnter/cddd). We then sampled CDDD descriptors for a subset of 1,753 generated molecules, to match the number of NPSs in the training set, and embedded both sets of descriptors into two dimensions with UMAP[35], using the implementation provided in the R package 'uwot' and the following parameters: n_neighbors = 20, alpha = 2, and beta = 1.

EMCDDA drug categorizations. To place the generated molecules into the context of the NPS categorizations established by the EMCDDA, we assigned each generated molecule to the category of its nearest neighbor among known NPSs. Briefly, we computed extended connectivity fingerprints[49] with a diameter of 3 (ECFP6) and a length of 1,024 bits for each known and generated molecule. The ECFP6 fingerprint was selected on the basis of its excellent performance in benchmarks of chemical similarity search and ligand-based virtual screening[38,50,51]. Each generated molecule was then compared to each known NPS in the training set, using the Tanimoto coefficient to quantify the similarity between their chemical fingerprints, and the generated molecule was assigned the EMCDDA category of the known NPS with the single highest Tc. EMCDDA categories that were significantly enriched or depleted among generated molecules were identified with a z-test of the log-odds ratio.

Blood-brain barrier permeability. LightBBB[36] was used to predict the blood-brain barrier permeability of the generated molecules, using the prediction server available at http://ssbio.cau.ac.kr/software/BBB. As a baseline, we also applied LightBBB to the set of known NPSs.

Structural prior. To investigate the relationship between sampling frequency and the chemical properties of generated molecules, we drew a sample of 1 billion SMILES strings from the trained model. After removing invalid SMILES and known NPSs, we obtained a set of 8,928,701 unique molecules that represented candidate novel NPSs, each of which was sampled between one and 323,299 times. To visualize the complete set of generated NPSs, we drew a random sample of at most 5,000 molecules per sampling frequency. We then embedded these into two dimensions alongside the training dataset of known NPSs using UMAP, as described above and with identical parameters. The nearest-neighbor Tanimoto coefficient between generated molecules and known NPSs was likewise calculated as described above for the assignment of EMCDDA drug categories. Finally, we computed the same eight physicochemical parameters described above for molecules sampled between one and 50 times, then computed the similarity of the property distributions for known NPSs and generated molecules using the Jensen-Shannon distance.

Model validation in a held-out set. To test the performance of our generative model on a held-out set of NPS structures, we assembled a database of 194 unique chemical structures that were added to the HighResNPS database between October, 2020 and April, 2021. These molecules comprised both previously described NPSs that had never been submitted to HighResNPS, as well as novel NPSs that had only emerged on the illicit market over the time frame in

question. These structures were preprocessed in the same manner as the training set using RDKit. We then used this held-out set to evaluate several aspects of model performance. Initially, we asked what proportion of held-out structures appeared at least once within the sample of 1 billion SMILES strings drawn from the generative model. We compared the chemical similarity to a known NPS (that is, we calculated nearest-neighbor Tanimoto coefficients, as described above) for held-out structures that were generated at least once by the model to those that were never generated.

We also investigated whether the sampling frequency of the generated molecules could be used to automatically annotate the most likely structure of an unidentified NPS whose exact mass had been determined using mass spectrometry. To this end, for each held-out structure in turn, we identified all generated molecules matching the exact mass of the held-out structure within a mass window of $\pm$ 10 ppm, and ranked them in descending order by their sampling frequency. We then quantified the frequency with which the held-out structure was correctly identified as the single top-ranked structure, or else appeared among the top-3, top-5, or top-10 structures ranked by sampling frequency.

Finally, in cases where the single most frequently sampled molecule was not a perfect match to the structure of the held-out NPS, we reasoned that generating a close structural analogue would nevertheless provide highly useful information to investigators. We therefore computed the chemical similarity between the top-ranked generated molecules and the held-out NPS using the Tanimoto coefficient, as described above. As a baseline, we also ranked the list of generated molecules by their sampling frequency ascending order (that is, we selected the least frequently sampled molecules from the generative model output), or obtained a set of matching molecules at random from PubChem.

Integration with MS/MS. In practice, investigators would generally have access not just to an accurate mass measurement for an unidentified NPS, but also to its tandem mass spectrum. A large body of work has shown that tandem mass spectra can be queried against databases of known chemical structures, even if the structures in these databases are not themselves associated with MS/MS data[52]. Accordingly, we posited that incorporating tandem mass spectrometry data into DarkNPS would further improve the accuracy of structure elucidation. To test this possibility, we applied CFM-ID (version 4.0.8) to predict tandem mass spectra for all 8.9 million unique molecules that appeared within our sample of 1 billion generated SMILES strings, using an ionization energy of 20 eV. Of these 8.9 million unique molecules, CFM-ID was unable to predict a tandem mass spectrum for approximately 400,000, which were assigned a score of zero. Each NPS in our held-out set was then compared to all generated molecules matching its exact mass ($\pm$ 10 ppm), using the dot product to quantify the similarity of predicted and observed spectra. This framework allowed us to perform MS/MS-based chemical structure search for novel molecules not present in any chemical structure database, at a scale of millions of candidates.

To integrate the spectral similarity scores assigned by CFM-ID with the generative model, we exploited the probabilistic interpretation of the structural prior. Specifically, we conjectured that the relative frequency at which a given molecule was sampled by the generative model could be interpreted as the prior probability that the molecule in question accounted for the observed mass spectrometric signal. Accordingly, we weighted the CFM-ID score according to the relative frequency with which each potential matching molecule was sampled by the generative model, considering only the subset of molecules matching the exact mass of the unidentified NPS. We

then compared this weighted spectral similarity score to the rankings assigned by CFM-ID alone, or by the structural prior alone. The three methods were evaluated within the subset of held-out NPSs for which tandem mass spectra had been deposited to HighResNPS, comprising 79 of the 189 molecules in the held-out set. The Tanimoto coefficient and Euclidean distance between CDDD embeddings were calculated for the top-ranked molecule nominated by each method as described above.

References

1. Peacock, A. *et al.* New psychoactive substances: challenges for drug surveillance, control, and public health responses. *Lancet* **394,** 1668–1684 (2019).

2. Baumann, M. H. *et al.* Baths salts, spice, and related designer drugs: the science behind the headlines. *J. Neurosci.* **34,** 15150–15158 (2014).

3. Underwood, E. A new drug war. *Science* **347,** 469–473 (2015).

4. Brandt, S. D., King, L. A. & Evans-Brown, M. The new drug phenomenon. *Drug Test. Anal.* **6,** 587–597 (2014).

5. Nichols, D. Legal highs: the dark side of medicinal chemistry. *Nature* **469,** 7 (2011).

6. Bijlsma, L. *et al.* Mass spectrometric identification and structural analysis of the third-generation synthetic cannabinoids on the UK market since the 2013 legislative ban. *Forensic Toxiocol.* **35,** 376–388 (2017).

7. Baumann, M. H. & Volkow, N. D. Abuse of new psychoactive substances: threats and solutions. *Neuropsychopharmacology* **41,** 663–665 (2016).

8. Johnson, L. A., Johnson, R. L. & Portier, R.-B. Current "legal highs". *J. Emerg. Med.* **44,** 1108–1115 (2013).

9. Luciano, R. L. & Perazella, M. A. Nephrotoxic effects of designer drugs: synthetic is not better! *Nat. Rev. Nephrol.* **10,** 314–324 (2014).

10. Gebel Berg, E. Designer drug detective work. *ACS Cent. Sci.* **2,** 363–366 (2016).

11. Carroll, F. I., Lewin, A. H., Mascarella, S. W., Seltzman, H. H. & Reddy, P. A. Designer drugs: a medicinal chemistry perspective. *Ann. N. Y. Acad. Sci.* **1248,** 18–38 (2012).

12. Lewin, A. H., Seltzman, H. H., Carroll, F. I., Mascarella, S. W. & Reddy, P. A. Emergence and properties of spice and bath salts: a medicinal chemistry perspective. *Life Sci.* **97,** 9–19 (2014).

13. Von Cüpper, M., Dalsgaard, P. W. & Linnet, K. Identification of new psychoactive substances in seized material using UHPLC-QTOF-MS and an online mass spectral database. *J. Anal. Toxicol.* **44,** 1047–1051 (2021).

14. Firman, J. W. *et al.* Chemoinformatic consideration of novel psychoactive substances: compilation and preliminary analysis of a categorised dataset. *Mol. Inform.* **38,** e1800142 (2019).

15. Mardal, M. *et al.* HighResNPS.com: an online crowd-sourced HR-MS database for suspect and non-targeted screening of new psychoactive substances. *J. Anal. Toxicol.* **43,** 520–527 (2019).

16. Wohlfarth, A. & Weinmann, W. Bioanalysis of new designer drugs. *Bioanalysis* **2,** 965–979 (2010).

17. Bell, C., George, C., Kicman, A. T. & Traynor, A. Development of a rapid LC-MS/MS method for direct urinalysis of designer drugs. *Drug Test. Anal.* **3,** 496–504 (2011).

18. Pasin, D., Cawley, A., Bidny, S. & Fu, S. Current applications of high-resolution mass spectrometry for the analysis of new psychoactive substances: a critical review. *Anal. Bioanal. Chem.* **409,** 5821–5836 (2017).

19. Reitzel, L. A., Dalsgaard, P. W., Müller, I. B. & Cornett, C. Identification of ten new designer drugs by GC-MS, UPLC-QTOF-MS, and NMR as part of a police investigation of a Danish internet company. *Drug Test. Anal.* **4,** 342–354 (2012).

20. Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37,** 1038–1040 (2019).

21. Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15,** 1120–1127 (2016).

22. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4,** 120–131 (2018).

23. Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4,** 268–276 (2018).

24. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9,** 48 (2017).

25. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361,** 360–365 (2018).

26. Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **4,** 828–849 (2019).

27. Skinnider, M. A., Stacey, R. G., Wishart, D. S. & Foster, L. J. Deep generative models enable navigation in sparsely populated chemical space. Preprint at https://doi.org/10.26434/chemrxiv.13638347.v1 (2021).

28. Bjerrum, E. J. SMILES enumeration as data augmentation for neural network modeling of molecules. Preprint at http://arxiv.org/abs/1703.07076 (2017).

29. Scheubert, K., Hufsky, F. & Böcker, S. Computational mass spectrometry for small molecules. *J. Cheminform.* **5,** 12 (2013).

30. Bertz, S. H. The first general index of molecular complexity. *J. Am. Chem. Soc.* **103,** 3599–3601 (1981).

31. Wildman, S. A. & Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **39,** 868–873 (1999).

32. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4,** 90–98 (2012).

33. Ertl, P., Roggo, S. & Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.* **48,** 68–74 (2008).

34. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1,** 8 (2009).

35. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. Preprint at http://arxiv.org/abs/1802.03426 (2018).

36. Shaker, B. *et al.* LightBBB: Computational prediction model of blood-brain-barrier penetration based on LightGBM. *Bioinformatics.* doi:10.1093/bioinformatics/btaa918 (2020).

37. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7,** 20 (2015).

38. Skinnider, M. A., Dejong, C. A., Franczak, B. C., McNicholas, P. D. & Magarvey, N. A. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. *J. Cheminform.* **9,** 46 (2017).

39. Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S. & Klambauer, G. Fréchet ChemNet distance: A metric for generative

models for molecules in drug discovery. *J. Chem. Inf. Model.* **58,** 1736–1741 (2018).

40. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39,** 2887–2893 (1996).

41. Blaženović, I. *et al.* Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy. *J. Cheminform.* **9,** 32 (2017).

42. Skinnider, M. A. *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11,** 6058 (2020).

43. Winter, R., Montanari, F., Noé, F. & Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **10,** 1692–1701 (2019).

44. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11,** 98–110 (2015).

45. Djoumbou-Feunang, Y. *et al.* CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. *Metabolites* **9,** 72 (2019).

46. Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2,** 171–180 (2020).

47. Arús-Pous, J. *et al.* Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminform.* **11,** 71 (2019).

48. Blaschke, T. *et al.* REINVENT 2.0: an AI tool for de novo drug design. *J. Chem. Inf. Model.* **60,** 5918–5922 (2020).

49. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50,** 742–754 (2010).

50. O'Boyle, N. M. & Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminform.* **8,** 36 (2016).

51. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform.* **5,** 26 (2013).

52. Böcker, S. Searching molecular structure databases using tandem MS data: are we there yet? *Curr. Opin. Chem. Biol.* **36,** 1–6 (2017).
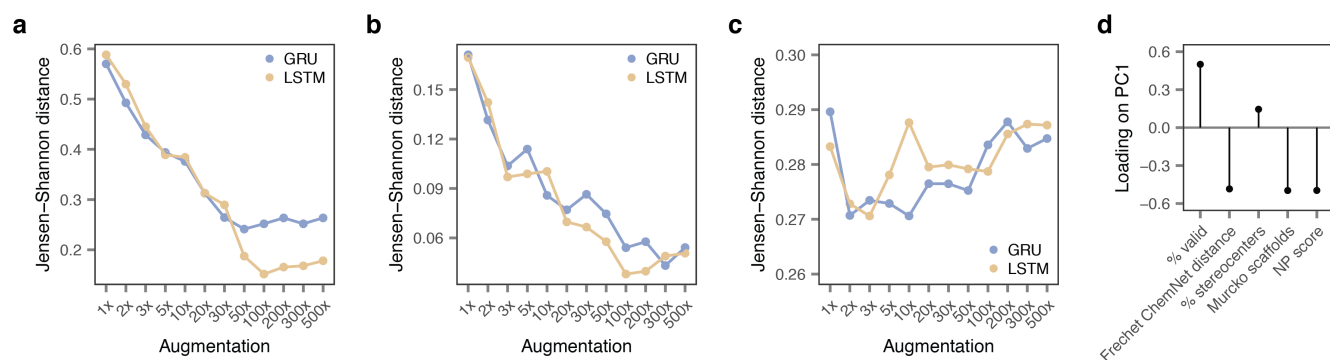
Supplementary Fig. 1 | Model selection and hyperparameter optimization.
a, Jensen-Shannon distance between the distribution of Murcko scaffolds in the training set and generated molecules, for recurrent neural network-based models trained on the HighResNPS database after varying degrees of non-canonical SMILES enumeration.
b, Jensen-Shannon distance between the natural product-likeness scores of the training set and generated molecules, for recurrent neural network-based models trained on the HighResNPS database after varying degrees of non-canonical SMILES enumeration.
c, Jensen-Shannon distance between the proportion of stereocenters in the training set and generated molecules, for recurrent neural network-based models trained on the HighResNPS database after varying degrees of non-canonical SMILES enumeration.
d, Factor loadings onto the first principal component in a principal component analysis of recurrent neural network-based models trained on the High-ResNPS database after varying degrees of non-canonical SMILES enumeration.
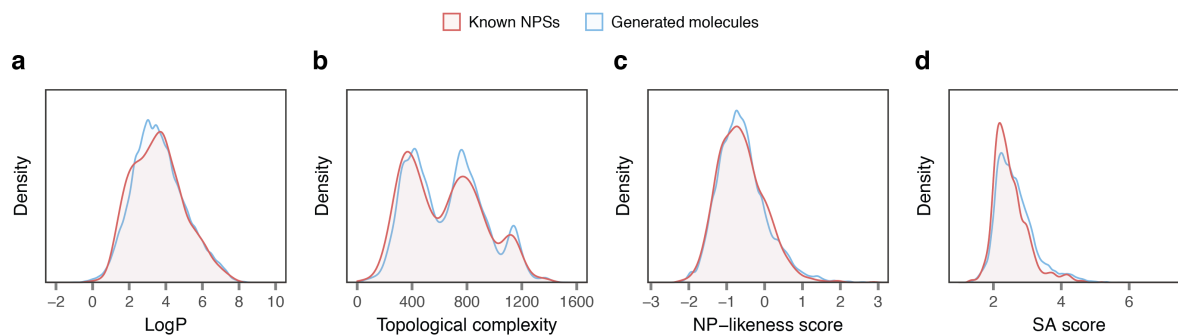
Supplementary Fig. 2 | Physicochemical properties of generated molecules.
a, Calculated octanol-water partition coefficients (LogP) of known NPSs and generated molecules.
b, Topological complexities of known NPSs and generated molecules.
c, Natural product-likeness scores of known NPSs and generated molecules.
d, Synthetic accessibility scores of known NPSs and generated molecules.
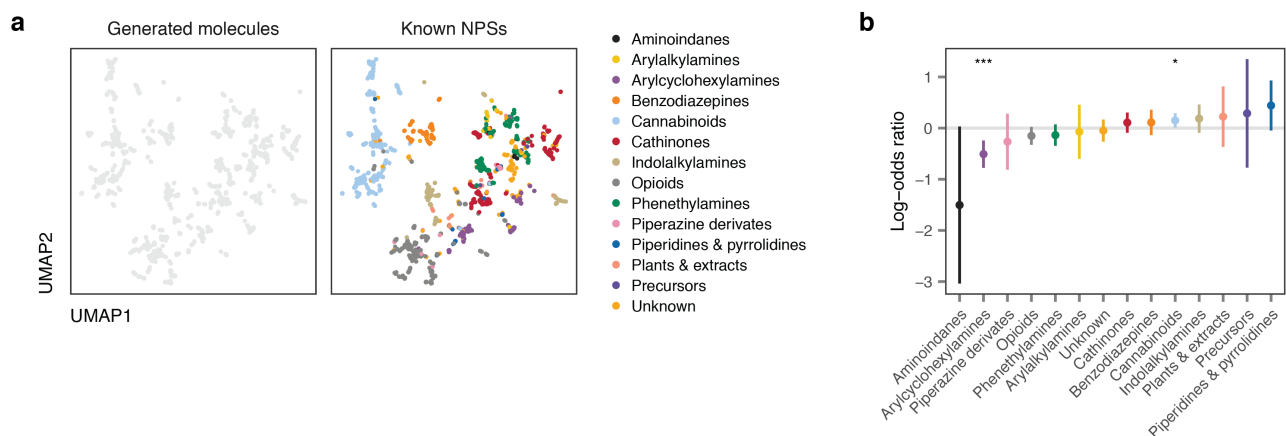
Supplementary Fig. 3 | EMCDDA drug categorizations of generated molecules.

a, UMAP visualization of known NPSs and an equal number of generated molecules sampled at random from the trained generative model, with the known NPSs colored by their EMCDDA drug categorizations.

b, Log-odds ratios of EMCDDA drug category frequencies among generated molecules, as compared to the training set. *, $p < 0.05$; ***, $p < 0.001$.
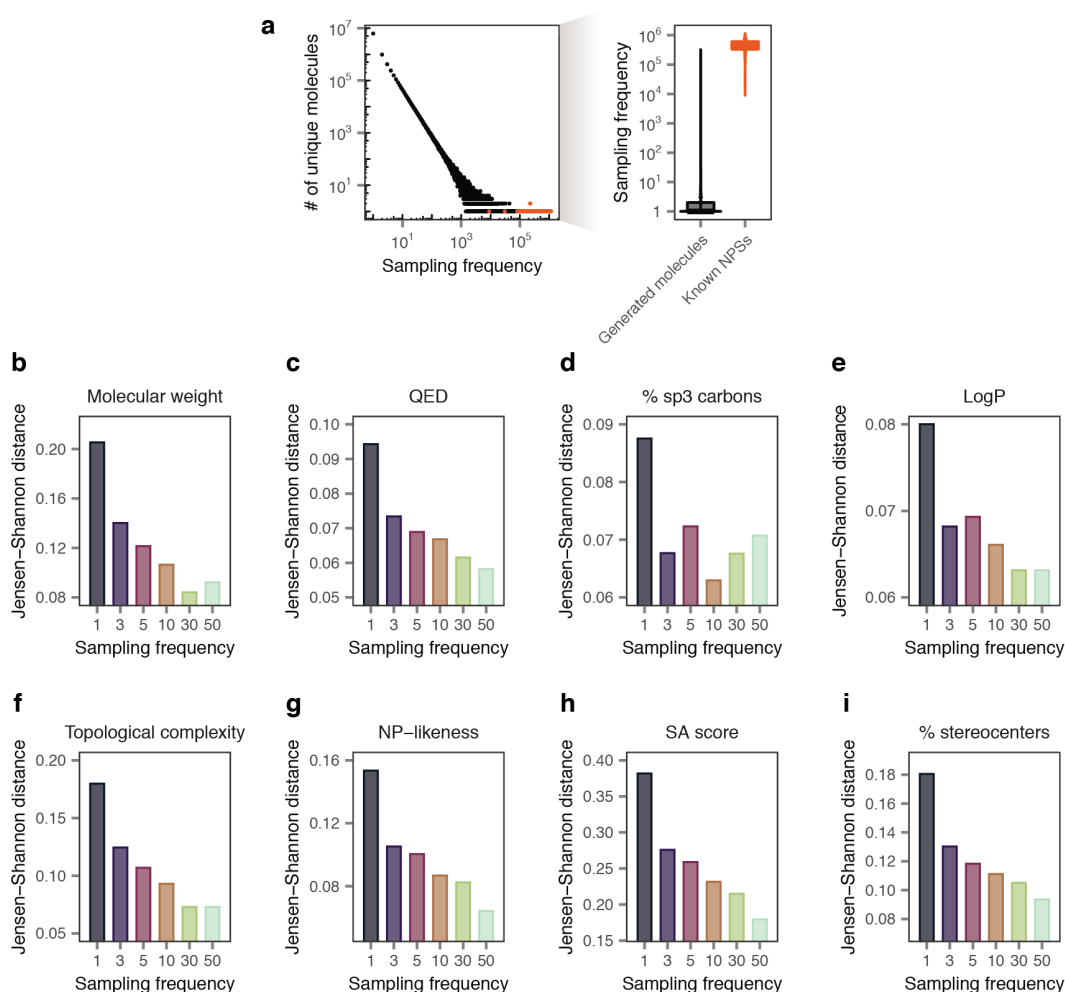
Supplementary Fig. 4 | Sampling frequency of known and generated molecules.

a, Distribution of sampling frequencies within a sample of 1 billion SMILES strings from the trained generative model, with known NPSs from the training set shown in red.

b, Jensen-Shannon distance between the molecular weights of generated molecules and the set of known NPSs, for molecules generated with progressively increasing frequencies.

c, Jensen-Shannon distance between the quantitative estimate of drug-likeness (QED) score of generated molecules and the set of known NPSs, for molecules generated with progressively increasing frequencies.

d, Jensen-Shannon distance between the proportion of carbons that are sp^3-hybridized in generated molecules and the set of known NPSs, for molecules generated with progressively increasing frequencies.
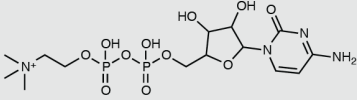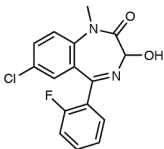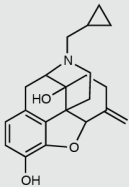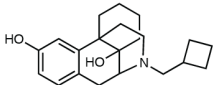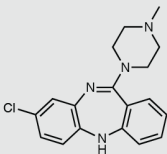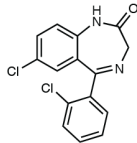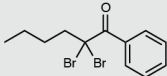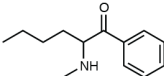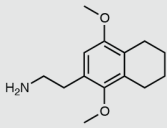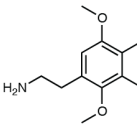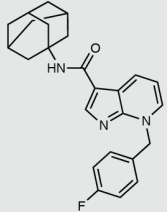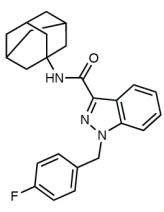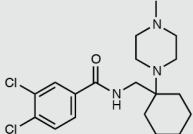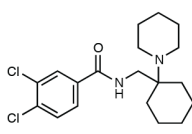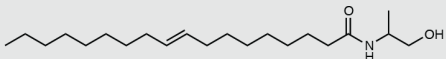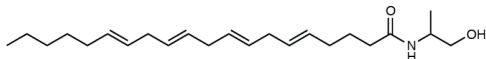
e, Jensen-Shannon distance between the partition coefficients of generated molecules and the set of known NPSs, for molecules generated with progressively increasing frequencies.

f, Jensen-Shannon distance between the topological complexities of generated molecules and the set of known NPSs, for molecules generated with progressively increasing frequencies.

g, Jensen-Shannon distance between the natural product-likeness scores of generated molecules and the set of known NPSs, for molecules generated with progressively increasing frequencies.

h, Jensen-Shannon distance between the synthetic accessibility scores of generated molecules and the set of known NPSs, for molecules generated with progressively increasing frequencies.

i, Jensen-Shannon distance between the proportion of stereocenters in generated molecules and the set of known NPSs, for molecules generated with progressively increasing frequencies.

Supplementary Fig. 5 | Examples of molecules from the held-out set that were not generated by DarkNPS.
Chemical structures of an illustrative subset of the 18 molecules in the held-out set that were never produced by the generative model in a sample of 1 billion SMILES strings, and their nearest neighbors among structures that were generated by the model. Many of these molecules either are not designer drugs at all (e.g., clozapine, citicoline, nalmefene, 2,2-dibromo-1-phenylhexan-2-one), or had a very closely related molecule appear in the model output.

Sampling frequency

☐ 0 ☐ 1 ☐ 2 ☐ 3 ▨ 4
▨ 5–9 ▨ 10–19 ▨ 20–49 ■ 50+
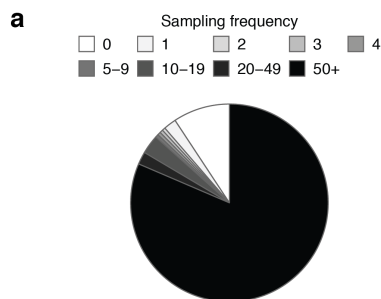
b

N-methyl-N-allyltryptamine (MALT)
$n = 160,665$
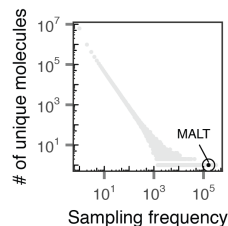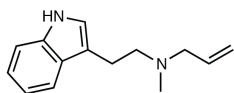
JWH-200 (5-OH-indole)
$n = 52,424$
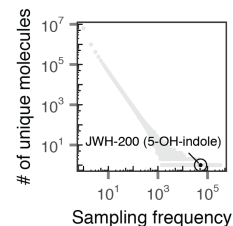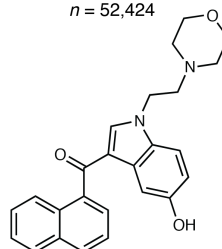
25CN-NBOMe
$n = 32,880$
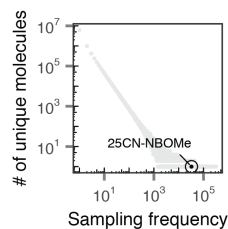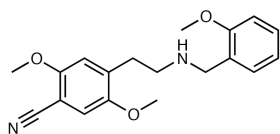
ALEPH-2
$n = 25,275$
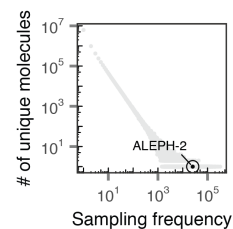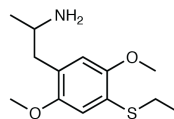
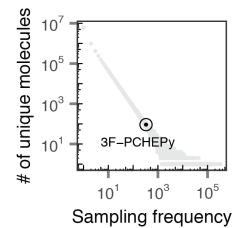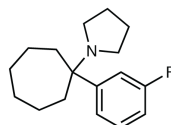### Cyclopropyl norfentanyl
$n = 8,735$
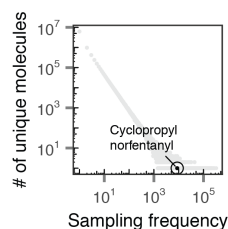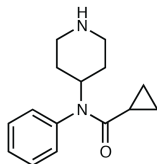
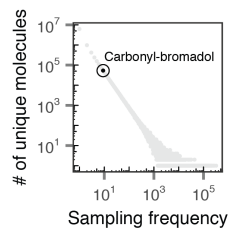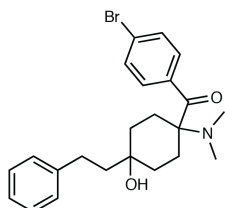### Lysergic acid N-methyl-N-propylamide (LAMPA)
$n = 2,407$
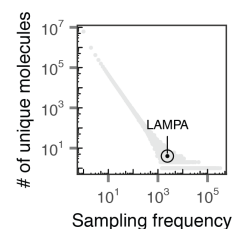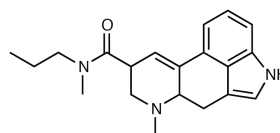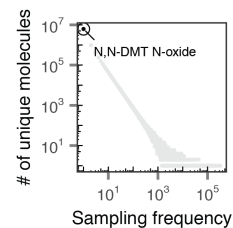

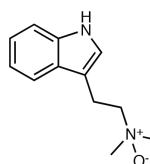### 4-bromo-4-methylaminorex (4Br-MAR)
$n = 1,304$
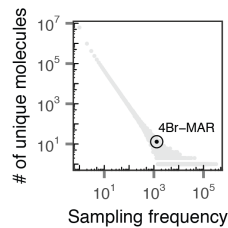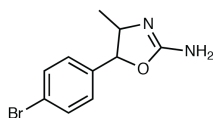
3F-PCHEPy
$n = 331$

Carbonyl-bromadol
$n = 9$

N,N-DMT N-oxide
$n = 1$

Skinnider *et al.* | Automated structure elucidation of designer drugs

Supplementary Fig. 6 | Examples of molecules from the held-out set that were correctly anticipated by DarkNPS.
a, Frequency with which each of the 194 molecules in the held-out set were sampled from the generative model.
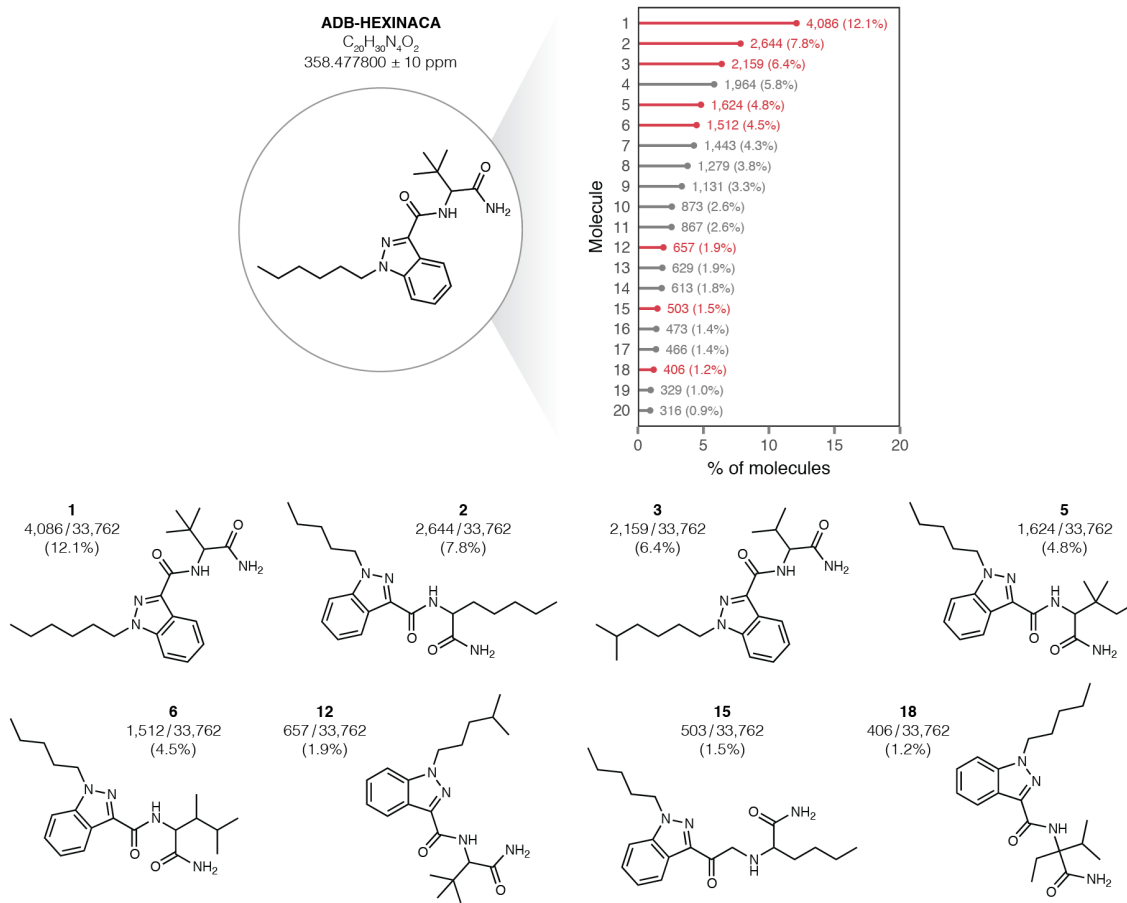**b,** Chemical structures, left, and sampling frequencies, right, for an illustrative subset of molecules in the held-out set that were correctly anticipated by the generated molecule. The molecules were selected from across the spectrum of sampling frequency in order to illustrate some of the major chemotypes captured by the generative model.
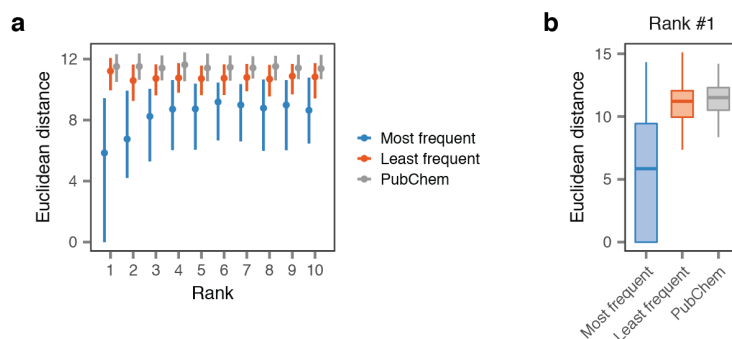
Supplementary Fig. 7 | Benchmarking the structural prior using continuous molecular embeddings.
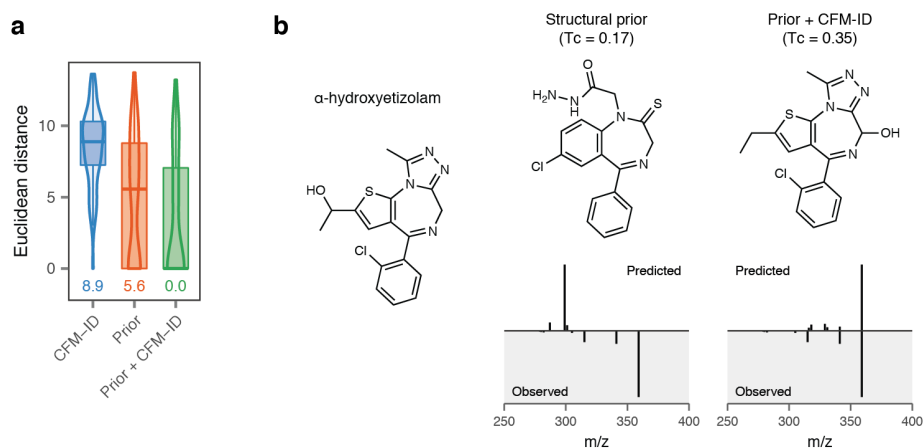
a, Median Euclidean distance between CDDD embeddings of held-out NPSs and generated molecules matching their exact masses ($\pm$ 10 ppm), arranged in descending order by sampling frequency ("most frequent"), ascending order by sampling frequency ("least frequent"), or a random sample of molecules with matching exact masses from PubChem. Error bars show the interquartile range.

b, Distribution of Euclidean distances between the CDDD embeddings of held-out NPSs and generated molecules matching their exact masses ($\pm$ 10 ppm), taking either the single most frequently sampled generated molecule, the single least frequently sampled generated molecule, or a random molecule with a matching exact mass from PubChem.

ADB-HEXINACA
$C_{20}H_{30}N_4O_2$
358.477800 ± 10 ppm

Supplementary Fig. 8 | Application of the structural prior to the synthetic cannabinoid ADB-HEXINACA.
Left, the chemical structure, molecular formula, and exact mass of ADB-HEXINACA. Middle, sampling frequencies of the 20 most frequently sampled molecules matching the exact mass of ADB-HEXINACA ($\pm$ a 10 ppm window). An illustrative subset of the generated molecules, highlighted in red, are shown at the right.

Supplementary Fig. 9 | Improved chemical similarity of automatically elucidated structures after MS/MS data integration.

a, Euclidean distances between CDDD embeddings for molecules in the held-out set of unidentified NPSs and the top-ranked structures suggested by CFM-ID alone, the structural prior alone, or the combination of the two.

b, Improvements in automated structure elucidation of an unidentified NPS using tandem mass spectrometry. Left, the chemical structure of α-hydroxyetizolam. Middle, the top-ranked molecule suggested by the structural prior (top) and mirror plot comparing the observed tandem mass spectrum of α-hydroxyetizolam with the tandem mass spectrum predicted by CFM-ID. Right, the top-ranked molecule after integrating the structural prior with MS/MS evidence (top) and mirror plot comparing the observed and predicted tandem mass spectra.