MegaSyn: Integrating Generative Molecule Design, Automated Analog Designer and Synthetic Viability Prediction

Fabio Urbina[1], Christopher T. Lowden[2], J. Christopher Culberson[2], and Sean Ekins[1]*

[1]Collaborations Pharmaceuticals, Inc., 840 Main Campus Drive, Lab 3510, Raleigh, NC 27606, USA.

[2]Workflow Informatics Corporation. 9316 Bramden Court, Wake Forest, NC 27587, USA.

*To whom correspondence should be addressed. E-mail:sean@collaborationspharma.com Phone: 215-687-1320

Short running tile: MegaSyn for De Novo Design

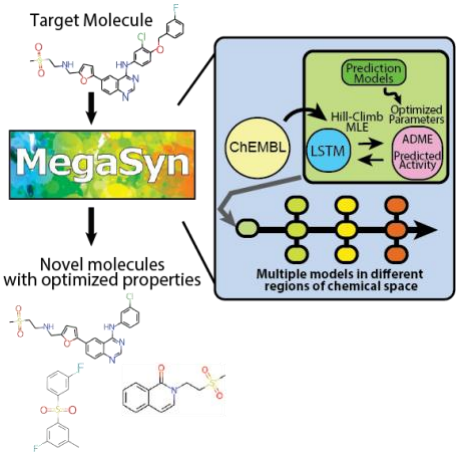
Keywords

Automated analog design, retrosynthesis, natural products, generative models, synthetic viability, recurrent neural networks

Abstract

Drug discovery is a multi-stage process, often beginning with the identification of active molecules from a high-throughput screen or machine learning model. Once structure activity relationship trends become well established, identifying new analogs with better properties is important. Synthesizing these new compounds is a logical next step, and is key to research groups that have a synthetic chemistry team or external collaborators. Generative machine learning models have become widely adopted to generate new molecules and explore molecular space, with the goal of discovering novel compounds with desires properties. These generative models have been composed from recurrent neural networks (RNNs), Variational Autoencoders (VAEs), and Generative Adversarial Networks (GANs) and are often combined with transfer learning or scoring of physicochemical properties to steer generative design. While these generative models have proven useful in generating new molecular libraries, often they are not capable of addressing a wide variety of potential problems, and often converge into similar molecular space when combined with a scoring function for desired properties. In addition, generated compounds are often not synthetically feasible, reducing their capabilities outside of virtual composition and limiting their usefulness in real-world scenarios. Here we introduce a suite of automated tools called MegaSyn representing 3 components: a new hill-climb algorithm which makes use of SMILES-based RNN generative models, analog generation software, and retrosynthetic analysis coupled with fragment analysis to score molecules for their synthetic feasibility. We now describe the development and testing of this suite of tools and propose how they might be used to optimize molecules or prioritize promising lead compounds using test case examples.

ToC Graphic


Target Molecule
MegaSyn
Novel molecules
with optimized properties
Prediction Models
Optimized Parameters
Hill-Climb MLE
ChEMBL
LSTM
ADME
Predicted Activity
Multiple models in different regions of chemical space

INTRODUCTION

We (and many other groups) have used machine learning models to propose molecules for testing and then validated them *in vitro* with vendor available molecules as a first step [1-3]. However, in order to optimize bioactivity, or maintain activity with improved absorption, distribution, metabolism, excretion and toxicity (ADME/tox) properties, vendor available compounds may not be sufficient. The most desirable chemical modifications are rarely available, and thus ways to generate and explore novel molecules are required.

In recent years, generative models have become commonly used to generate *de novo* molecules [4, 5]. These generative models have come from several different architectures, and have been shown to generate valid, novel molecules in the same chemical space as their training sets [6-8]. Molecular representation is varied in generative models, however the SMILES representation has seen widespread success and is favored due to the simplicity and ease of molecular representation [9]. However, many of these generative models have enjoyed limited success in real world drug discovery projects due to their narrow range of capabilities. The focus of drug discovery projects may be varied. For instance, in one project, a lead molecule scaffold may require iterative design to find the most suitable analog, and thus the generative model employed should only enumerate on the core structure. Conversely, in another case given a set of known active and inactive compounds against a target, the project may wish to discover entirely new scaffolds that do not exist in 'patent space' yet has similar desired molecular properties to the known active compounds. While most generative models can utilize the desired physicochemical properties in the training of the generative models, in practice the goals are often not achievable using generic, out-of-the-box generative models. To

control the “closeness” of the generated compounds to a molecule of interest, the Tanimoto similarity score [10] is often included in the training. Generative models retrained on the same parameters often end up in similar local minima of chemistry property space, reducing their usefulness past an initial run [11].

To address these limitations, we have now created MegaSyn, a suite of algorithms which takes a similar approach to weak learner ensemble methods such as Random Forests [12]. Instead of training one generative model, MegaSyn trains many “weaker” generative models, starting from a generic model trained on a drug-like library (ChEMBL) [13] and iterative generative models that are continuously “focused” down onto target molecule(s) and physicochemical properties of interest, as well as random branch models from each of these focused generative model nodes until multiple generative models have been created that explores many local minima. Depending on the desired outcome (i.e., completely new drug scaffolds or enumeration on a common structural core), MegaSyn allows flexibility and balance in the exploration of chemical property space versus. focused generative capabilities by traversing the “tree” of generative models based on the desired outcome.

In addition to machine-learning based generative models described above, there are numerous other algorithms for modifying existing candidate structures using a list of transforms (e.g. generating *bioisosteres*). Generating novel molecules is one thing but it is another to evaluate these proposed structures for *synthesizability* and suggest synthetic pathways for the synthesis of the compound. These are important concepts generally absent from most recently published generative models. The technologies involved in proposing, evaluating, planning and assessing the synthetic feasibility of

compound syntheses have been available within the cheminformatics industry for decades [14-18], but their implementation remains relatively state of the art (reviewed recently [19]). For example, the earliest efforts in synthesis planning, reaction prediction and synthetic feasibility assessment developed rule-based approaches such as LHASA, CAMEO and CAESA (as reviewed in [20]). These software required collaboration between the chemist and machine to get the best out of the relatively limited functionality [20].

In recent years we have seen considerable development in computer-aided synthesis planning with the collection of tens of thousands of manually curated reaction transformation rules to yield millions of chemical reactions as a network in Chematica [21] which can be used to select the most cost effective or chemically diverse synthetic pathways [22]. While the manual collection of such rules is not scalable there has also been a shift to use machine learning approaches. One has used deep neural networks trained on 3.5 million reactions from the Reaxys database with the extended connectivity fingerprints (ECFP4) [23]. When the data was split 70: 10: 20 (Training: development: testing) the top 10 accuracy was 95% in retrosynthesis and 97% for reaction prediction [23]. Another approach has used 15,000 reactions from the USPTO augmented by a set of over 5M reactions with non-recorded products to train a neural network [15]. Again, the data was split 70:10:20 and the top ten accuracy was 94.6% in the best case [15]. Others have developed proof of concept tools that they suggest are not ready for practical use such as CompRet which enumerates a chemical reaction network based on depth first proof number search, enumerating all synthetic routes and then recommending synthetic routes using simple scoring functions [24]. A template free self-corrected retrosynthesis predictor was a built using a transformer neural network architecture which improved on prior

accuracy rates using the USPTO-50K set [25]. Scientists at Pfizer have also demonstrated that a transformer-based retrosynthesis model generated with public USPTO training data could predict over 147,000 reactions from Pfizer electronic notebooks with top 1% accuracy of 69% and this number increases with their own data is used in training [26]. A more recent use of the transformer architecture using transfer learning for retrosynthesis prediction with literature data demonstrated a top 1% accuracy up to 60.7% [27]. Various methods have been used to predict synthetic accessibility such as using the probability of existence of substructures for the compound in question along with the number of symmetry atoms, graph complexity and number of chiral centers [28]. More recent open-source software for retrosynthetic planning includes AiZynthFinder [29], LillyMol [30] and ASKCOS [15]. Comparison of such methods has been limited and would require synthesis of compounds using proposed routes obtained with each method [29].

We now describe a generative model which is flexible to address the needs of many drug discovery projects, as well as a prototype Pipeline Pilot protocols for automated lead expansion, filtration of analogs, and selection of a representative set that is user-accessible. Because the molecules are generated in an automated fashion, some of the molecules may be difficult or impractical to pursue from a synthetic chemistry perspective. Thus, we also created an automated tool to predict the relative difficulty of synthesis for targeted analog molecules utilizing automated retrosynthetic analysis coupled with a fragment analysis to score molecule on their synthetic feasibility. We have evaluated these tools using a set of FDA approved drugs as well as a recently published set of natural products [31]. We also provide several test cases to recapitulate a recently described known analog of ibogaine [32] and develop analogs of lapatinib with improved

predicted properties [33] using MegaSyn, which shows that it can generate synthetically feasible compounds with desired properties.

METHODS

Activity models for MegaSyn

All activity models consisted of naïve bayes trained models using the scikit-learn package in python. Datasets were acquired for each target (e.g., HER1, HER2) from ChEMBL target activities. All activities were binarized according to an activity threshold, with 1 indicating active and 0 indicating inactive. Each model was trained and calibrated using isotonic regression with 3-fold cross-validation from which all statistics were generated from. As input into MegaSyn multi-objective scores, the calibrated prediction score was used as input.

Evaluation of Variational Autoencoder, Generative Adversarial Networks, and Recurrent Neural Networks

As input, molecules are represented as tokenized SMILES strings. Briefly, each SMILES is tokenized, and each character is represented in a vocabulary (e.g. “c” [nH]”, “1”, =”). Each token in the vocabulary has a corresponding numerical representation (e.g., all “c” are represented by 1, all “=” are represented by the number 2, etc). SMILES are encoded by their integer vocabulary representation, and padded to the longest sequence length with zeroes which were masked during training. Beyond this, several differences exist between models during training.

Variational Autoencoder (VAE): The variational autoencoder utilizes an encoder-decoder architecture to map chemical space into a latent vector [34]. The encoder is composed of 3 LSTM layers of 512 units each followed by a linear layer of 64 units (the latent space). Our decoder is comprised of 3 LSTM layers of 512 units each with dropout of 0.2 between all layers. We used KL-divergence as our loss term with an Adam optimizer = 0.0001, patience = 10, 200 epochs, and batch size of 64.

Generative Adversarial Networks (GAN): We implemented a latentGAN [35] architecture for our generative GAN model. Wasserstein GAN with gradient penalty was utilized for the GAN model. The heteroencoder was comprised of 3 LSTM layers of 512 units each with a final linear layer of 64 units (the latent space), while the decoder was comprised of 3 LSTM layers of 512 units each followed by a linear layer with softmax activation to return the probability of each character in the vocabulary. The autoencoder was trained for 100 epochs with a batch size = 128 and an Adam optimizer with a learning rate = 0.0001 using teacher forcing. The discriminator of the GAN was formed by 3 linear layers of 256 hidden units each with ReLU activation between each layer (except for the last layer). The generator consists of 5 linear layers of 512 hidden units each with batch normalization = 0.9 and leaky ReLU activation between each layer.

The autoencoder was pre-trained using the ChEMBL dataset followed by training of the full GAN model.

Recurrent Neural Networks (RNN) [5]: Each LSTM-based model is composed of an embedding layer, three LSTM layers (512 hidden units), followed by a linear layer with softmax activation the size of a vocabulary generated from the training data.

MegaSyn design

Each LSTM-based model is composed of an embedding layer, three LSTM layers (512 hidden units), followed by a linear layer with softmax activation the size of a vocabulary generated from the training data. As input for all models, molecules are represented as tokenized SMILES strings. MegaSyn is composed of three distinct model types: The initial pre-trained model, a set of primed models, and finally a set of exploratory models.

Initial model

The initial model is trained on ChEMBL 28's ~2 million compounds [13]. The loss function for a sequence of encoded SMILES is the Negative Log-Likelihood. The model uses an Adam optimizer with learning rate = 0.002. Teacher-forcing is used to expedite training of the generative model.

Primed models

For each set of primed models, the initial model is trained for *n* epochs, with a new agent model saved every 2 epochs. The target molecule(s) of interest are broken down into substructures based on RECAP rules. Simplified carbon-only versions of these substructures and the original molecule are also generated. The initial model is trained on this set of structures and substructures alone, using the same parameters as the initial model (described above) using teacher forcing. Every *i* epoch, the model is saved, until a set of *n* primed models have been created. We find that 16 total epochs with a model saved after every 2 epochs represents the gradient of general to diverse reasonably well for a number of target molecules.

Exploration models

For each primed model, *de novo* molecules are generated. The generated molecules are then ranked based on a composite score from any number of criteria. The composite score is represented as below:

For each criteria *i* in the composite score (i.e., predicted target activity or drug-likeness (QED) score [36]), the composite score is defined as

$$\sum_{i=1}^{n} ln(\frac{x_i}{y_i})$$

where x_i is the ith score for molecule and y_i is the ith desired score. This usually includes QED, activity against target (target model), and any other desired scores. As long as a score can be assigned to a compound, it can be included in the final composite score, given a large potential to the tasks the generative model can be applied to. The top 10% of ranked compounds are kept and fed back into the model for training using NLL and teacher forcing, a training concept called hill-climb maximum likelihood estimation (MLE). A new set of molecules is generated after training, and the cycle continues. Importantly, the top 10% of compounds are kept from one epoch to the next; only if a newly generated compound has a score higher than one in the current top 10% list does it replace one in the set. Eventually the model will find a substructure minima and is then capable of generating analogs of this specific substructure. Often, based on the initial seed molecules of the very first iteration, the model will converge to one local minima. At least four models are trained and generated from each primed model node, to obtain models

that focus on different substructures of the original target molecule. The top scoring 10% of compounds found over the training loop for each model are kept.

Automated Analog Generation

Lead Expansion/Enumeration

We have developed a Pipeline Pilot (Biovia, San Diego, version 19.1.0.1964) [37] protocol for automated lead expansion, filtration of analogs, and selection of a representative set. For lead expansion, we encoded several different medicinal chemistry strategies to generate potential analogs. Included in these strategies are classical bioisosteric replacement and similarity "bioisosteres" (for which Pipeline Pilot components already exist) [38-43]. The classical bioisosteres include the replacement of several common functional groups with sterically similar functional groups believed to have similar physicochemical effects in a biological environment. Similarity bioisosteres locates fragments within molecules and replaces them with similar fragments, based on a user specified similarity measure (e.g. FCFP_6 and PHFC_2). Another strategy involves the enumeration of heteroatomic regioisomers. Heteroatoms are identified and relocate them to every possible position around within the molecule [44]. Finally, a large number of molecular transformations (37 aromatic/phenyl replacement; 2 conformational restriction/ expansion, 92 Topliss, 8 Magic methyl) have been encoded to identify modification sites on molecules and automate the enumeration of analogs using common medicinal chemistry approaches [45-47]. These approaches include, Topliss, Magic Methyl,

conformational restriction/relaxation, and ring expansion/contraction. Users can select or deselect the different transformation categories as desired.

Tagging and scoring

Using these techniques, 10s to 1000s of analogs are generated for a typical lead molecule, depending on its complexity. These molecules are then examined for any undesirable functional groups such as reactive functional groups and toxicophores [48]. Molecules with any of these features are tagged (and can be removed later as desired). The molecules are then scored for synthetic feasibility, using a newly developed algorithm (see below). The molecules are then clustered using FCFP_4 fingerprints, so that a diverse set can be selected if desired. The canonical tautomer is generated for each molecule and duplicate molecules are removed.

Selection

After the analogs are enumerated, tagged, and scored, the resulting analogs are displayed in graphical and tabular format. Categorical and numeric charts, such as pie charts and histograms are then generated along with a tabular output in PipelinePilot. The charts and tabular output are linked together such that the user can select subsets of molecules and export them readily.

Automated Retrosynthetic Analysis and Synthetic Feasibility Prediction

Three primary methodologies were used to evaluate synthetic feasibility. The first method involves the fragmentation of known (synthesized) molecules and the relative presence

of those fragments in targets. The second method couples automated retrosynthesis with the first method. In addition to using automated retrosynthesis to rate synthetic feasibility, a separate application for retrosynthetic analysis was created. Finally, a weighting mechanism was added to penalize molecular elements that are undesirable from a synthetic perspective.

Fragmentation

To create the fragments used in the first method, two molecular sources were used. These included eMolecules [49] consisting of 26,400,125 molecules (at the time of download) and ChEMBL version 24 consisting of 1,820,035 molecules [50]. Separately, these sources were subjected to fragmentation using Pipeline Pilot using the Generate Fragments component. Specifically, ring assemblies (contiguous ring systems), BridgeAssemblies (contiguous ring systems that share two or more bonds), and Chains (Contiguous atoms not in rings) and BemisMurcko assemblies were generated [51]. Canonical SMILES were generated for each fragment. Fragments containing less than 2 atoms were filtered. Fragments that occurred more than 10 times over the entire molecule set were retained, along with their frequency (occurrence count). Each unique fragment and its frequency were saved in comma separated files for each source.

Fragmentation Scoring

Molecules evaluated for synthetic feasibility are fragmented in the same way as the source sets. A baseline score is created by the ratio of fragments of the incoming molecules [52]. The baseline score is calculated as follows:

Matches = incoming molecule fragments that are also present in source set.

Misses = incoming molecule fragments that are not present in the source set.

Baseline Score = Matches/Matches+Misses

The score is then weighted using an algorithm that takes the size of the fragment (number of atoms) and the frequency of occurrence in the source set.

Retrosynthetic analysis

This was carried out by applying a set of transformations that apply known reactions in reverse [30]. Our solution has 2 primary sources of these transformations. The primary source is a set of reactions extracted from patents by a group at Eli Lilly (Lilly) [30]. The secondary source is a set of reactions detailed in by a group at Astra Zeneca (AZ) [53]. All of the reactions were reversed so that they could be applied that way. In the case of the Lilly reactions [30], a set of 1,929,251 reactions in a format similar to SMIRKS were culled to a set of 8,040 reactions simply by looking at the number of characters in the text for each reaction. The idea here was that smaller reactions were more likely to represent the core or substructures of reactants and products, and therefore would be applicable to a larger number of molecules. The reactions were then reversed by swapping the products with the reactants and converted from SMIRKS format to RXN format in Pipeline Pilot. Approximately 10,000 druglike molecules were tested by running each of the 8040 reactions on them. Of the 8040 reactions, 2632 unique reactions were used at least once. This set of reactions was used as the final Lilly reaction set. A much smaller set of ~45 common reactions were derived from the AZ group [53]. These reactions were hand-written SMIRKS that represented common transformations used in organic synthesis. These SMIRKS were reversed by hand. Some were removed due to their promiscuous nature when applied in reverse (e.g. carbon-carbon bond formation reactions).

Once the core set of retrosynthetic reactions were selected and curated, the retrosynthetic analysis tool was developed and subjected to numerous rounds of testing (using experienced medicinal chemists) and enhancement, where various rules were imposed to encourage better outcomes. It was arbitrarily determined that up to 5 rounds of retrosynthetic reactions should be applied to each molecule. In the first round, each unique set of reaction products is retained. The “size” of each product molecule was determined by the number of non-hydrogen atoms. Most retrosynthetic reactions produce more than one product. For each set of products that are created by an individual reaction, the largest product (selected product) is retained. In rounds 2-5, an additional restraint is imposed. Only the 5 smallest of the selected products are allowed into to the next round. In rounds 4 and 5, another additional restraint is imposed. Selected products must be smaller than the smallest selected product in all other rounds to be moved to the next round or to be reported. Results are reported for each round that is executed with all precursor molecules from each round.

Fragmentation and Retro Combined Scoring

The retrosynthetic analysis tool was combined with the fragmentation score to enhance the synthetic feasibility score. For the enhanced scoring, the selected product from the last 3 executed rounds that were executed (if at least 3 rounds were executed) are scored using the fragmentation scoring system. The highest score is then selected as the consensus score.

Weighting Mechanisms

After reviewing results with our experienced synthetic chemists, it was clear that a certain key weighting mechanism was required to be added for certain features that are difficult to synthesize. The presence of one or more absolute chiral center is one example of a penalizing feature. The presence of one or more spiro atom is another example. For each of these elements that is present in the molecule, the score is reduced by a certain relative ratio.

Software testing

A set of 'best-selling 25 small molecule drugs' were selected as an example of well-known molecules in order to test the automated retrosynthetic analysis software (Supplemental Table S1 and Figure S4). A set of 346 natural products (Canvass) was used to compare with a library of 201 FDA approved drugs. [31]

Visualization of FDA approved drugs and natural products

The molecular property space of FDA approved drugs and the Canvass dataset [31] were compared using a t-SNE plot.

Data analysis

To determine if FDA approved drugs were considered more synthetically feasible than the Canvass natural products library, Bootstrap hypothesis testing was performed on the two datasets [54]. Briefly, both datasets (FDA library and Canvass) are combined into one dataset. Two datasets of size *n* and *m* (the size of the FDA library and Canvass library, respectively) are randomly sampled from the combined dataset. The mean and

standard deviation are calculated. A p-value is calculated by determining the likelihood of the true mean occurring from the bootstrapped sample means.

t-SNE plot generation

All t-distributed stochastic neighbor embedding (t-SNE) plots were generated using the sklearn package in python with default parameters (number of components = 2, perplexity = 30.0, early exaggeration = 12, learning rate = 200, number of iterations = 1000, number of iterations without progress = 300, minimum gradient norm = 1e-07, metric = Euclidean).

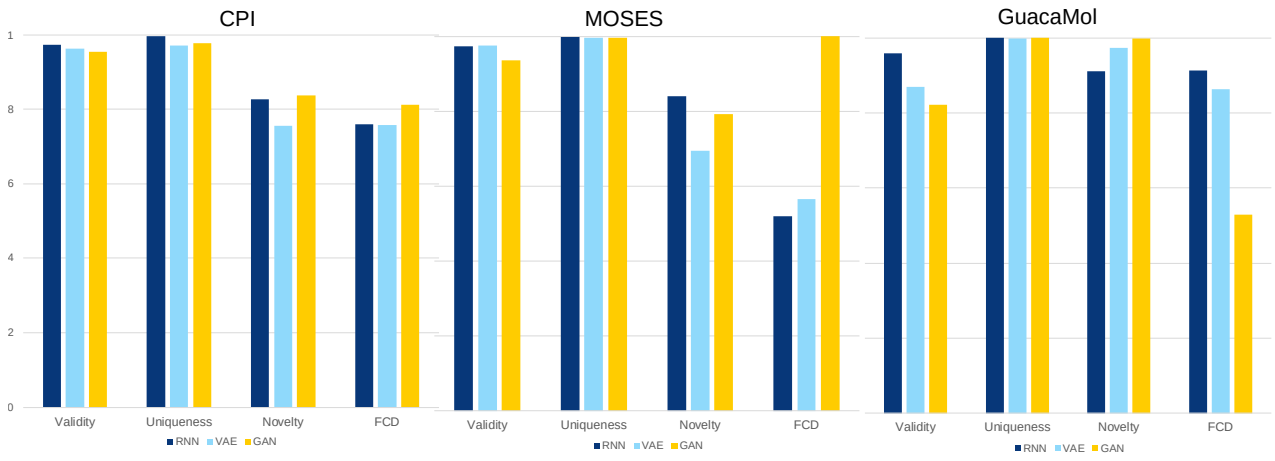
RESULTS

Evaluation of different generative approaches

First we evaluated several different generative model architectures (to compare with published benchmark resources MOSES [55] and GuacaMol [56]) which had been introduced in the literature in recent years: Recurrent Neural Networks (RNNs) [5], Generative Adversarial Networks (GANs) [35], and Variational Autoencoders (VAEs) [7]. In order to assess the capabilities of each architecture, we decided to use a number of metrics proposed in the literature, including Validity: whether the compounds generated are theoretically realistic molecules; Uniqueness: the fractions of molecules which are unique; Novelty: the fraction of molecules generated not in the training set; and finally, the Fréchet ChemNet Distance: (FCD) [57] a measure of how close distributions of generated data are to the molecules in the training set. As comparing architectures is difficult given the ability of different hyperparameter tuning to alter results, we chose

Hyperparameters based off their initial implementation. We then trained each architecture (RNN, VAE, and GAN) on 1.2 million ChEMBL compounds and filtered to between 10-50 heavy atoms. We employed early stopping to reduce the length of time to train each model. Finally, we generated 100,000 compounds per architecture. We found that all three architectures performed similarly, and were all capable of generating valid, unique, and novel compounds with a good FCD score (Figure 1) [55, 56]. These scores were comparable to those reported in the literature with other benchmarking studies (Figure 1) and suggested that the choice of generative model architecture was not a significant factor for improving generative model capabilities.

Figure 1: comparison of different model architectures for generative models using our models (CPI) in comparison to values reported from two other published benchmark resources (MOSES [55] and GuacaMol [56]).



MegaSyn design

At its core, MegaSyn uses long-short term memory (LSTM)-based generative models to learn the proper structure of SMILES strings [5]. As input, molecules are represented as tokenized SMILES strings. MegaSyn is composed of three distinct model types: The initial pre-trained model, a set of primed models, and finally a set of exploratory models (Figure 2).

Initial model

The initial model is trained on ChEMBL28's ~2 million compounds. The purpose of training this model is to teach it how to create drug-like molecules. Once trained, The initial model "knows" how to put together drug-like molecules, and can be queried to generate compounds that fall within ChEMBL's chemical-space. This represents the prior knowledge of chemistry: valid chemical structures and how they are put together, atom-by-atom, is learned in this initial model. This large chemical information will be transfer-learned in the subsequent model. This initial model takes the largest amount of time to train; however, once trained, it can be re-used for many projects as the prior model, and the overall training time of MegaSyn is small in comparison to a full retraining of a typical generative model starting from training on the entire ChEMBL database.

Primed models

After the initial model is trained, a set of “primed” models are trained (Figure 2). The initial model is first presented molecule(s) of interest. The molecule(s) of interest are broken down into substructures based on RECAP rules [58]. Simplified carbon-only versions of these substructures and the original molecule are also generated. Model 1 is trained on this list of structures and substructures for several epochs using teacher forcing. Every *i* epochs, the model is saved, until a set of *n* primed models have been created. Each of these primed models represent generic exploration of chemical space (early primed models) to enumeration of the target molecule(s) (late primed models). How many epochs the model is trained on is critical; if too little, the primed models explore a very wide chemical space around the target molecule. If too many epochs are trained, the model learns to focus only on the specific structure and substructures of the target(s) of interest themselves. We find that 16 total epochs with a model saved after every 2 epochs represents the gradient of general to diverse reasonably well for a number of target molecules. Due to the few targets trained at a time, primed models can be quickly generated.

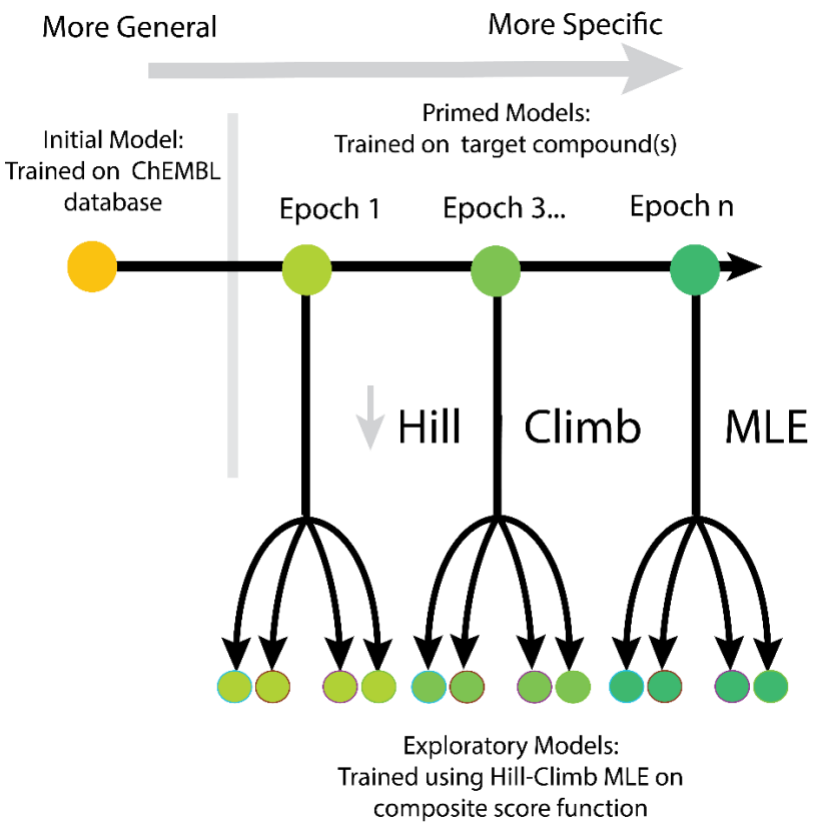
Exploration models

Primed models represent nodes along a singular branch. To explore more diverse chemical space around each of these nodes, a final set of exploration models are branched off of each primed model node. For each primed model, *de novo* molecules are generated (~2,000-10,000 appear sufficient to cover a broad chemical space). The generated molecules are then ranked based on a composite score from a number of criteria. This usually includes QED [36], activity against the target (target model), and any

other desired scores. Notably, as long as a score can be assigned to a compound, it can be included in the final composite score, this provides flexibility to the tasks the generative model can be applied to. We can weight each objective according to their importance, on a scale from 0 to 1, with 1 being extremely important and 0 representing no importance. After the generated set of compounds are scored, the top 10% of ranked compounds are kept the model is trained on these top compounds, a training concept called hill-climb MLE [59]. A new set of molecules is generated after training, and the cycle continues. Importantly, the top 10% of compounds are kept from one epoch to the next; only if a newly generated compound has a score higher than one in the current top 10% list does it replace one in the set. Eventually the model will find a substructure minima and is then capable of generating analogs of this specific substructure. Often, based on the initial seed molecules of the very first iteration, the model will converge to one local minima. At least four models are trained and generated from each primed model node to obtain models that focus on different substructures of the original target molecule. The top scoring 10% of compounds found over the entire training loop for each model are kept. The collection of models are indexed to give flexibility on what regions of chemical space the user could explore. Instead of sampling from a single generative model, MegaSyn randomly samples from a collection of *t* total models (initial model + (*i/n) * 4*) in parallel. It should be noted that training multiple models from the initial model takes a limited amount of time, only requiring 6 hours on a single 1080 GPU to generate 32 models, the number of models generated per MegaSyn case study in this paper. The desired “focus” of the model can be driven by a generative specificity parameter, which weights the chance of

a model to be sampled from, either closer molecules to the training target(s) or driving away from the targets to generate novel compounds.

Figure 2: MegaSyn architecture. First, an initial model is trained on a drug database (i.e., ChEMBL). Next, a set of Primed Models are generated by training on a target compound(s). Finally, exploratory models are generated from each Primed Model node, completing a set of generative models that range from general, drug-like molecules to analogs of the target compound(s).

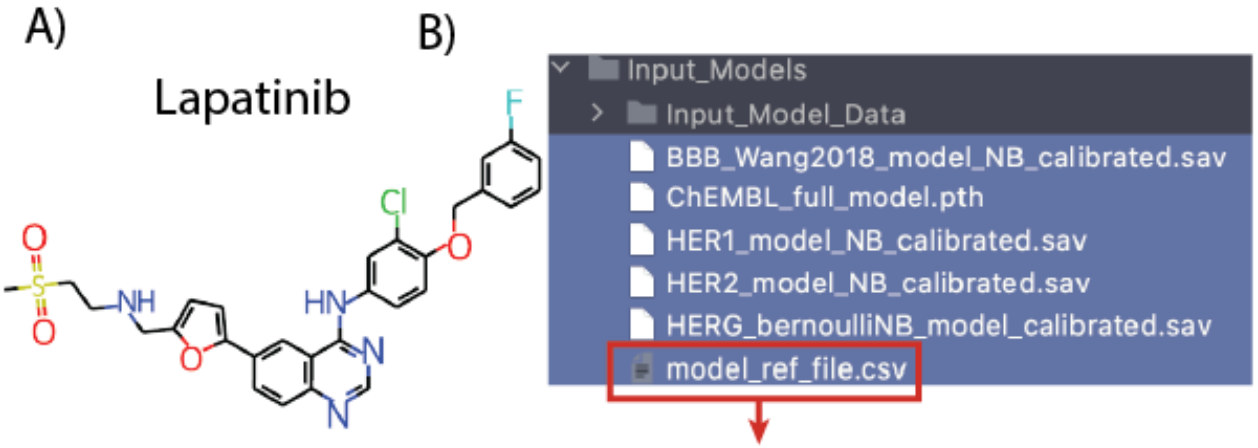


Evaluation of *de novo* molecules generated from MegaSyn

Case study 1 – Lapatinib analogs

We decided to evaluate the capability of MegaSyn to generate valid, novel molecules with desired properties by employing several case studies. As an example of our generative approach we chose to optimize Lapatinib, an orally active drug for breast cancer and other tumors (Figure 3A). Lapatinib inhibits EGFR (HER1) and HER2 kinases, and thus is commonly used in combination therapy for HER2-positive breast cancer [60]. Lapatinib, however, is relatively poor at crossing the blood brain barrier (BBB), with highly variable metastasis uptake and is not detected in normal brain tissue [61]. We used MegaSyn to design analogs that simultaneously optimizes for HER1 and HER2 activity with an improved ability to cross the BBB. All activity models were built using naïve bayes (Table S2; see methods). For inputs to the scoring function, we considered QED score > 0.6, Similarity to lapatinib or lapatanib fragments (Tanimoto similarity >0.6), and prediction scores from machine learning models we constructed for crossing the BBB, HER1 inhibition, HER2 inhibition, and finally a HERG model to ensure the molecules avoid this ion channel (Figure 3B). Figure 3B shows our selected weighting scheme for the Lapatinib MegaSyn model. We ran MegaSyn for 16 total epochs, saving a Primed model node every 2 epochs, and generating 4 exploratory models per primed model node for a total of 32 RNN-based models. 10,000 molecules were generated from each of the 32 RNN-based models.

Figure 3: Case study 1. A) Structure of Lapatinib, the target molecule of interest. B) Model reference file and predictive models used for MegaSyn.



model_name	wanted_activity	importance
BBB_Wang2018_model_NB_calibrated.sav	1	0.8
HER1_model_NB_calibrated.sav	1	0.8
HER2_model_NB_calibrated.sav	1	0.8
HERG_bernoulliNB_model_calibrated.sav	0	0.6

MegaSyn explores diverse chemical space

t-SNE plots of the top 200,000 scored molecules shows that MegaSyn explores a rich chemical space around Lapatinib (Figure 4A), ranging from Tanimoto similarity scores of 0.1 to 0.97. To contrast to other generative models, we also used a single LSTM-based generative model pre-trained on ChEMBL and used the exact same loss function and multi-parameter optimization score to drive the generative model. We then sampled 5000 compounds from the LSTM-based generative model to compare against MegaSyn, from which we also sampled 5000 compounds. MegaSyn had significantly higher multi-parameter optimization scores, suggesting it is capable of finding better composite score maxima (Figure 5). When we limit our number of molecules down to the

top 2000 scored generated molecules, molecular diversity is still common, suggesting that MegaSyn is not just enumerating on a common core structure alone, but exploring diverse options to meet the criteria used in the scoring function (Figure 4B). In contrast to the Tanimoto similarity score, the region in the t-SNE plot with the highest multi-optimization score is distinct from the location of Lapatinib, suggesting MegaSyn is capable of finding novel chemical space with better molecular properties than Lapatinib (Figure 6A). While the majority of the top 2000 compounds are predicted to cross the BBB (Figure 6B), there is a clear structure-activity relationship between activity relationship with HER1 activity and especially HER2 activity, which shows higher selectivity amongst the top compounds (Figure 6C, 6D). We evaluated the atomic contribution to model prediction for Lapatinib and two of the top-scoring generated compounds (Figure S1). While the BBB model suggests that the smaller generated compounds have no distinct atom-specific prediction differences (Figure S1), the HER1 model suggests that the core atomic contribution to predicted activity is retained, with a new strong atomic contributor (the carbon atom highlighted in the first top-generated molecule under HER1) in addition (Figure S1). For HER2, however, the strongest atomic contributor is not retained from Lapatinib in the top-scoring generated compounds, and instead novel atomic contributors are highlighted, suggesting the optimization of the generated molecules can “find” distinct properties that allow the generated molecules to still be active against the target (Figure S1). We next evaluated the synthetic feasibility of the top 2000 of compounds by using our newly built retrosynthetic analysis tool.

Figure 4: t-SNE plots of structural diversity of MegaSyn generated compounds. A) t-SNE plot based on ECFP6 for 200,000 top-scoring generated molecules colored by Tanimoto similarity to Lapatinib. B) t-SNE plot based on ECFP6 for 2,000 top-scoring generated molecules colored by Tanimoto similarity to Lapatinib. Blue dot represents Lapatinib.

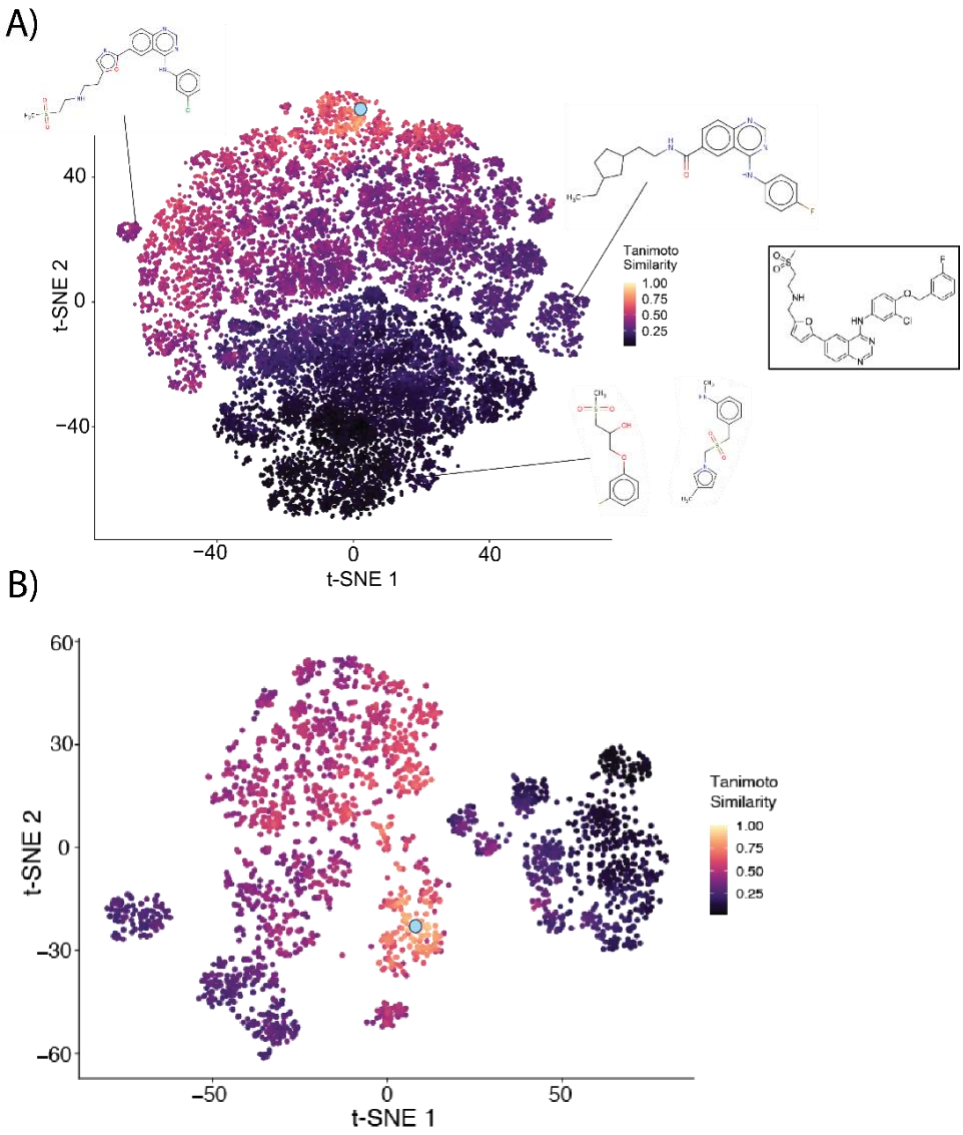


Figure 5: Comparison of MegaSyn vs. a single LSTM model multi-optimization score using the same training setup with the same ChEMBL pre-trained model for set up. Boxplot showing the multi-parameter optimization score for the same generated compounds.

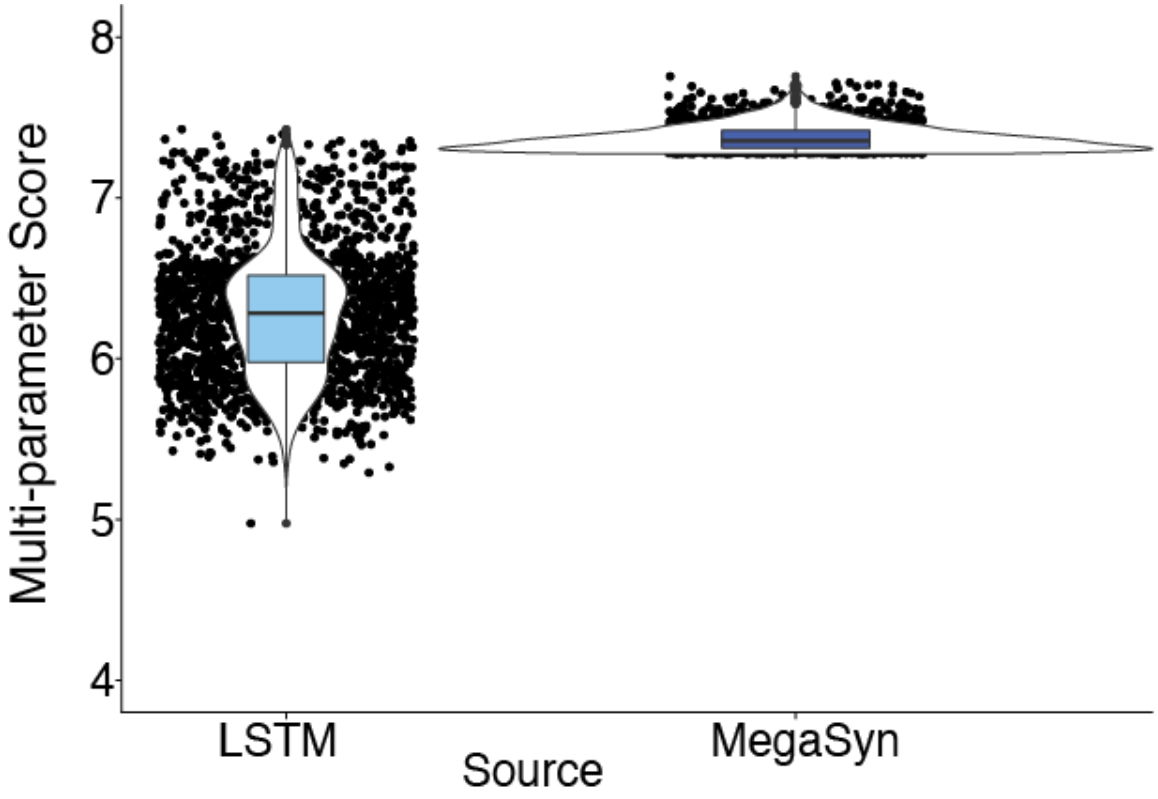
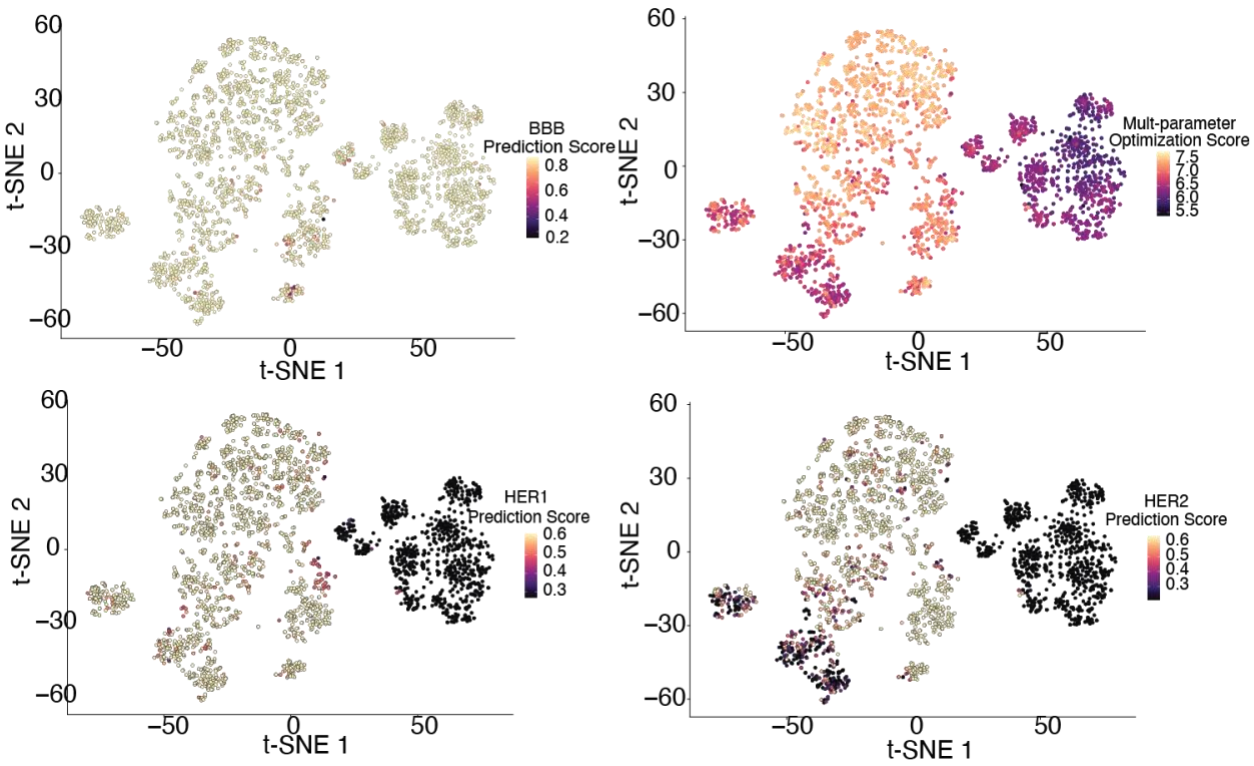


Figure 6: t-SNE plots based on ECFP6 of the top 2,000 scoring compounds generated by MegaSyn colored by A) mult-objective optimization score, B) predicted ability to cross the BBB, C) predicted HER1 inhibition, or D) predicted HER2 inhibition.



Case Studies for Retrosynthetic Analysis

Before scoring the retrosynthetic feasibility of MegaSyn generated compounds, we first evaluated test cases to show the utility of the retrosynthetic analysis tool. Initially, the retrosynthetic analysis tool was tested on several examples to illustrate potential utility. As an example of application of this software, Sorenson *et al*., recently described a 3-step synthesis for the antiviral drug tilorone [62]. Our software suggests several approaches to deliver tilorone (Figure S2). Another molecule tested in this way was the kinase inhibitor axitinib [63]. The retrosynthetic analysis results were comparable with known synthesis route (Figure S3). We have also generated a much larger evaluation for the top 25 selling small molecule drugs (Table S1). This resulted in a similar number of alternative synthetic routes for these drugs (Figure S4). 15 out of 25 were 'retro-synthesized' completely to

commercially available reactants (eMolecules was checked for commercial availability). 2 of the drugs only required 1 step, 4 required 2 steps, 8 required 3 steps and 1 required 5 steps to break down into commercially available reactants. In many cases, the retrosynthesis went further than required to reach commercially available reactants (Table S1).

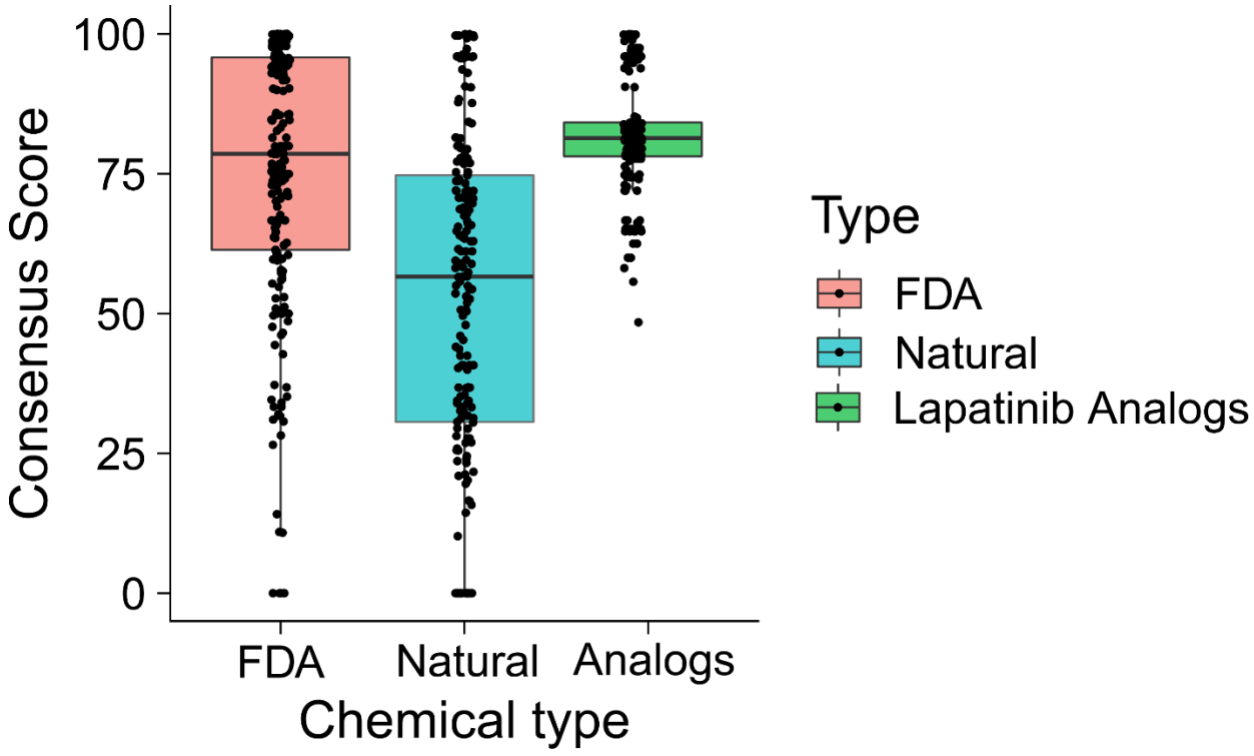
Synthetic feasibility prediction

An example of using tilorone for synthetic feasibility prediction is shown in Figure S5, which illustrates the analysis of results as a whole and also the scoring. In addition, we have compared the synthetic feasibility consensus scores of an FDA approved drug library versus 346 natural products in the Canvass dataset [31] (Figure 7). This analysis shows a good separation of drugs from natural products using this score. We decided to use reference points of a synthetic feasibility score of <60 to indicate synthetic feasibility, and a score of > 90 to indicate a compound that is easily synthesizable. The FDA dataset and Canvass were statistically significantly different ($p=0.00318$), suggesting that the synthetic feasibility tool is easily capable of discerning difficult to synthesize molecules (natural products) from generally simpler molecules like drugs. Visualization of the chemistry space of these approved drugs and the natural products further demonstrate that they cover different chemical areas with drugs generally focused in the center of the plot while natural products are on the periphery (Figure S6).

Synthetic Feasibility of MegaSyn generated compounds

After validating our synthetic feasibility tool earlier, we used the consensus model to score the top 200 MegaSyn generated lapatinib analogs ranked by MPO score and (Figure 7). The majority of compounds (97.5%) were scored as synthetically feasible, nearly a quarter (23%) being considered easily synthesizable (Figure 7). This suggests that MegaSyn can generate valid, drug-like, easily synthesizable compounds with desired predicted physicochemical and bioactivity properties.

Figure 7. Boxplot comparing the consensus synthetic feasibility score for an FDA approved library versus 346 natural products in the Canvass dataset and 200 of the top-scoring MegaSyn generated lapatinib analog compounds. 195/200 megaSyn compounds had a score > 60 and 46/200 compounds had a score >90, indicating the compounds were synthetic feasibility and easily synthesizable, respectively.



Case Study 2. Ibogaine analogs

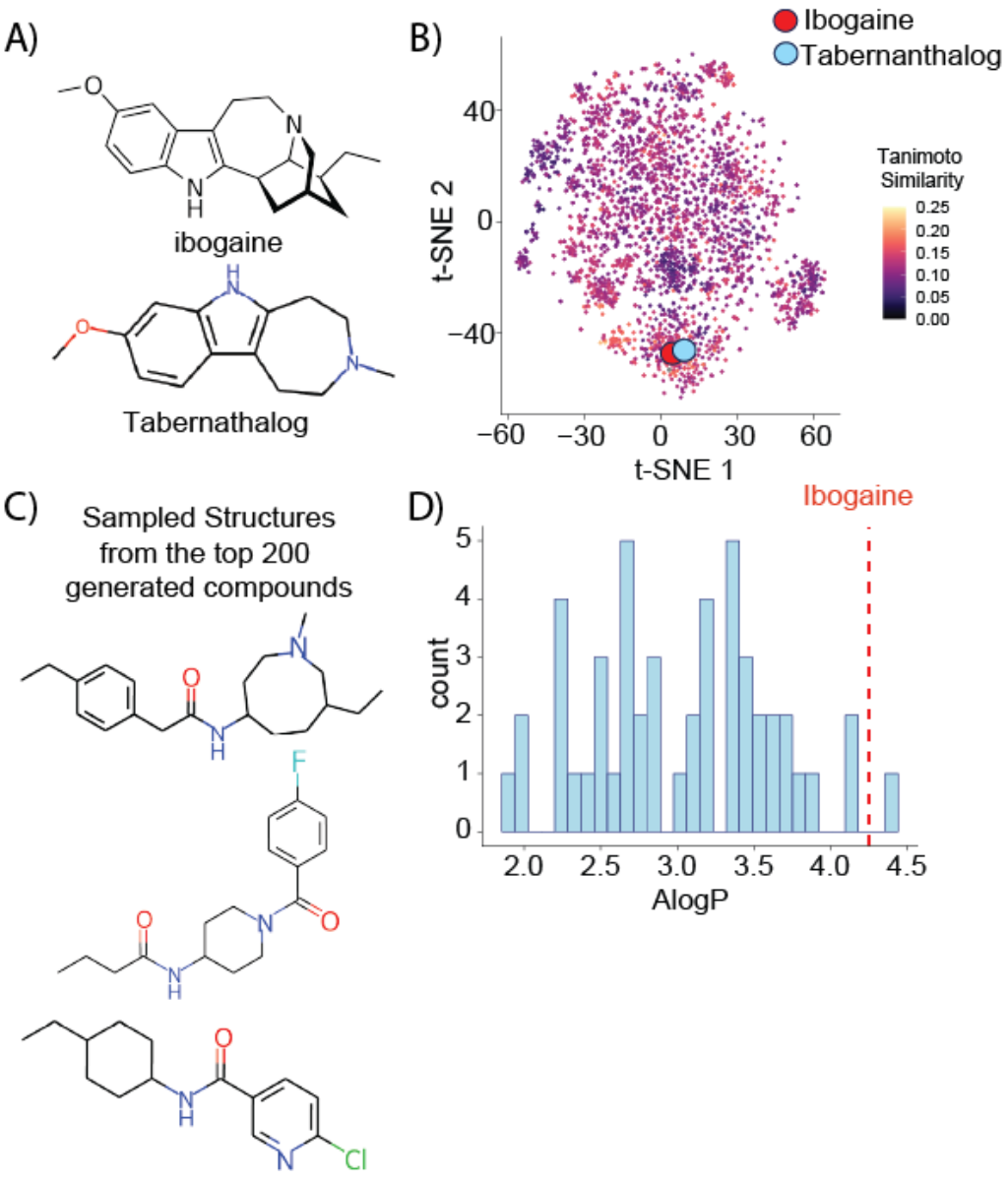
As a second more challenging case study we chose to potentially improve upon a natural product, ibogaine. Ibogaine is a natural product derived from *tabernanth iboga* (Figure 8A). Recent research has shown that psychedelics such as ibogaine may have therapeutic potential as anti-addictive agents. However, ibogaine has several undesirable properties, including inhibition of the hERG channel and the induction of a psychedelic experience. In a recent publication, Cameron *et al,* proposed, synthesized, and tested new ibogaine analogs with the following targeted properties in mind: that it does not inhibit the hERG channel, it maintains specificity to the 5-HT_{2A}, which is thought to be necessary for the therapeutic action; and it does not induce a psychedelic experience [32]. Ultimately the authors discovered tabernanthalog, an ibogaine derivative with these desired properties [32].

We have used this paper as a test case and challenged MegaSyn to find tabernanthalog, using the following criteria: activity against 5-HT_{2A}, inactivity against hERG, 5-HT_{1A}, 5-HT_{1F}, 5-HT_{2C}, similarity to ibogaine and it's substructures (Tanimoto > 0.6), and lower cLogP than ibogaine. We ran MegaSyn for 16 total epochs, saving a Primed model node every 2 epochs, and generating 4 exploratory models per primed model node for a total of 32 LSTM-based models.

We built machine learning models against 5-HT_{2A}, hERG, 5-HT_{1A}, 5-HT_{1F}, and 5-HT2C to include in the multi-objective scoring function to drive MegaSyn. All activity models were built using naïve bayes (Table S2; see Methods). We then generated 100,000 compounds and took the top 50 highest multi-objective scoring compounds. tabernanthalog was included in the top 50 highest scoring compounds. In addition,

MegaSyn captured a wide variety of other related structures, including dissimilar scaffolds to ibogaine (Figure 8B, C). The majority of the top-50 compounds had a lower AlogP than ibogaine, suggesting that MegaSyn was capable of finding molecules with improved predicted drug-like properties (Figure 8D). In addition, the top 10 generated compounds had an MPO score comparable or better than tabernanthalog and all had a higher MPO score than ibogaine, suggesting that several of the novel MegaSyn generated compounds have a higher chance of crossing the BBB (Table S3).

Figure 8: MegaSyn generation of new molecules based on ibogaine. A) The structures of ibogaine and tabernanthalog. B) t-SNE plots of the top 2,000 generated molecules based on ECFP6 fingerprints colored by Tanimoto similarity. C) structures of three randomly sampled molecules from the top 200 compounds. D) Histogram of the AlogP of the top 50 generated compounds. The AlogP of ibogaine is indicated by the red dashed line.



Automated analog generation

In addition to *de novo* design of molecules with MegaSyn, we have also developed an easy-to-use web interface using Pipeline Pilot for running an automated analog generation protocol which can be used for lead expansion. We encoded several different medicinal chemistry strategies to generate potential analogs. A file of molecules to

generate analogs for is uploaded and and the output consists of a pie-chart summarizing the make-up of the molecule analogs and bar charts of their properties (Figure S7). The charts and tabular output are linked together such that the user can select subsets of molecules and readily export them. This tool can also be used with the retrosynthetic analysis described earlier to score the likely synthetic feasibility.

DISCUSSION

The goal in this study was to generate a complementary suite of accessible tools for generative molecular design, computer assisted synthesis, retrosynthesis and synthetic viability in order to propose new analogs or additional as the next steps after identification of a potential hit. We aimed to make use of existing data and algorithms where possible in order to deliver this additional functionality to provide meaningful synthesis suggestions for each molecule. We have now delivered methods for automated lead expansion, filtration of analogs, and selection of a representative set of molecules that is user-accessible. This collection of capabilities can also be combined with other software or machine learning tools to score proposed analogs with models of interest.

Over the past few years, new discoveries in the field of *de novo* drug design has renewed interest in generating new molecules using machine learning [5]. RNNs have been used for generating libraries for HTS, hit to lead optimization and fragment-based hit discovery [64-68]. A feature of these generative models is the ability to optimize multiple parameters such as physicochemical properties or biological activity. While these new approaches are promising, a critical gap in knowledge is that limited experimental

validation data was generated by synthesizing compounds and testing for activity for any of the aforementioned studies with only a few groups validating their approach by making and testing compounds [69]. Default or 'vanilla' generative models, while capable of generating novel compounds often do not end up in the desired chemical space. In our conversations with numerous drug discovery experts at various companies, the major complaints regarding generative models are that they either end up enumerating on the same initial target molecules, essentially rediscovering what their medicinal chemists have already proposed (the model is too focused), or end up far outside of "realistic" drug designs, proposing molecules which are well outside the realm of synthesizability. We believe that a single model is not sufficient to cover all the possible tasks requested of a generative model, so we tried to circumvent these issues by creating a large enumeration of models, from very general (little information is taken into account about the desired molecular space) to the specific (models which generate only analogs of the desired target molecules).

MegaSyn initial model choice

The current MegaSyn models are all initially based off a single pre-trained model on the ChEMBL database. This serves two purposes: First, the model has already learned how to compose correct molecule structures from SMILES strings. Second, despite the large number of learned molecules, ChEMBL has the additional bonus of being comprised almost entirely of drug-like molecules. This works to the advantage of MegaSyn due to the unique training strategy of using the hill-climb MLE algorithm. The use of a hill-climb MLE means that only molecules that the initial generative model is capable of generating

can be used for training, creating a feedback loop of only drug-like molecules being generated and trained on, and prevents undesirable properties from being generated. This is further re-enforced through use of a QED score to prevent molecules from straying too far into non-druglike space. The initial choice of database to train the initial model, then, is critical to the success as a generative model. Exploration of other databases to train the initial-model can be used to change the desired outcome.

Composite score function

The core driver of MegaSyn is the composition of the composite score function, which often includes a score for drug-likeness (such as QED) and similarity to target molecule(s) (i.e., Tanimoto similarity) in addition to the primary activity scoring models for potential drug-targets. The accuracy and choice of scores, then, is also critical for the success of MegaSyn. The number of possible scores to include are unbounded, and only require that a molecule can be scored and ranked numerically. For example, including machine learning models on toxicity (HERG, drug induced liver toxicity, CYPs) can be combined with on-target (i.e., a 5-HT receptor) and off-target (other 5-HT receptors) to create a composite score of dozens of scoring functions. We included a weighting value, from 0-1, which allows flexibility in score inclusion; instead of only the important score functions, several “nice-to-have” scores may also be introduced with a lower weight value than the more critical score functions.

Case study results pros and cons

In the absence of prospective validations of the approaches the use of case studies are a promising way to explore the possible application and limitations of generative *de novo* design software as we have demonstrated. We illustrated that MegaSyn, even when faced with a natural product (ibogaine) is capable of discovering the same molecular analog as proposed by medicinal chemists (using 'traditional' medicinal chemistry approaches to design), suggesting it is capable of supplementing medicinal chemistry exploration. In addition, a number of the top-scoring compounds in this case study had molecular scaffolds distinct from ibogaine, highlighting that ibogaine is not considered a 'drug-like' molecule. The proposed top-scored compounds for lapatinib, were similar to lapatinib, but with improved predicted molecular properties.

The downside of case studies, however, is that the interpretation of success is only as good as the accuracy of the composite scoring function. While we can judge generated molecules as being reasonable from a chemistry point of view, it remains to be seen whether the other top-scoring compounds are in fact active, non-toxic, and selective without making them and testing them. This is a critical point that has yet to be fully investigated by any generative model proposed to date (to our knowledge), and we do not know whether the bias of using machine learning models to drive generative models also affects the probability of the top-scoring generated compounds to be truly active, non-toxic, or selective. We would argue, however, that these same machine learning models would be used to drive drug-discovery projects irregardless of the origin of the proposed molecules, therefore suggesting that generative models may provide a promising route to finding new molecules to test, especially when combined with retrosynthetic analysis.

Retrosynthetic analysis and analog designer for pipeline pilot

While some of the tools described are less sophisticated than the approaches described earlier for computer assisted synthesis [20-22], retrosynthesis [15, 23-27] and synthetic viability tools (e.g. AutoGrow 3 [70], chemical stability [71] and others [28] in order to eliminate invalid options) they can be readily implemented in PipelinePilot which is a widely used and commercially available product. Similarly, this approach and software could be readily reimplemented in open-source tools such as KNIME [37, 72].

In conclusion, we have demonstrated that MegaSyn can propose synthesizable analogs for molecules based on the integration of various software components (open source and commercial). We have also demonstrated that we can recapitulate synthetic approaches for approved drugs in our case studies and that our synthetic feasibility score can reliably differentiate between approved drugs that are likely to be more synthetically feasible than more complex natural products. While these represent essentially retrospective evaluations of the software in line with what has been demonstrated with several more sophisticated tools described earlier, the next step is using this MegaSyn suite of tools to propose analogs, define how to make them, rank their synthetic feasibility before ultimately selecting molecules to synthesize and test *in vitro*. This tool is currently being applied to do just this on various internal research projects.

Acknowledgments

We kindly acknowledge NIH funding: R44GM122196-02A1, 2R44GM122196-04A1 from NIGMS, 3R43AT010585-01S1 from NCCAM, and 1R43ES031038-01 from NIEHS. “Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number R43ES031038. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.” Biovia are kindly acknowledged for providing Discovery Studio and PipelinePilot. We thank Dr. Ian Watson and Dr. Alex M. Clark for early discussions.

Author contributions

F.U. performed machine learning model building, data analysis and coded MegaSyn. C.T.L. and J.C.C. curated data for and developed automated analog designer and synthetic viability prediction tools in PipelinePilot as well as performed data analysis. S.E. led the project and designed experiments. All authors contributed to writing the manuscript.

Competing interests

SE is founder and owner and FU is an employee of Collaborations Pharmaceuticals, Inc. CL and JCC consulted for Collaborations Pharmaceuticals, Inc.

Abbreviations

absorption, distribution, metabolism, excretion and toxicity (ADME/tox); extended connectivity fingerprints (ECFP4); long short term memory (LSTM); recurrent neural network (RNN); Quantitative estimate of drug likeness (QED); maximum likelihood

estimation (MLE); Simplified molecular-input lin-entry system (SMILES); Blood brain barrier (BBB); Fréchet ChemNet Distance (FCD)

Data and Software Availability

Datasets used for machine learning model building are available upon request (Table S2). Output molecules from MegaSyn test cases are proprietary. Pipeline Pilot software is licensed from Biovia. The MegaSyn software and our Pipeline Pilot protocols are available for licensing.

SUPPORTING INFORMATION

Supporting further details on the models, structures of public molecules and computational models are available. This material is available free of charge via the Internet at http://pubs.acs.org.

References

1. Vignaux, P.; Minerali, E.; Foil, D. H.; Puhl, A. C.; Ekins, S., Machine Learning for Discovery of GSK3β Inhibitors. *ACS Omega* **2020**, 5, 26551-26561.

2. Vignaux, P. A.; Minerali, E.; Lane, T. R.; Foil, D. H.; Madrid, P. B.; Puhl, A. C.; Ekins, S., The Antiviral Drug Tilorone Is a Potent and Selective Inhibitor of Acetylcholinesterase. *Chem Res Toxicol* **2021**, 34, 1296-1307.

3. Klein, J. J.; Baker, N.; Foil, D. H.; Zorn, K. M.; Urbina, F.; Puhl, A. C.; Ekins, S., Using Bibliometric Analysis and Machine Learning to Identify Compounds binding to Sialidase-1. *ACS Omega* **2021**, 6, 3186-3193.

4. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H., Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics* **2017**, 9, 48.

5. 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* **2018**, 4, 120-131.

6. Hochreiter, S.; Schmidhuber, J., Long Short-Term Memory. *Neural Computation* **1997**, 9, 1735-1780.

7. Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H., Application of Generative Autoencoder in De Novo Molecular Design. *Mol Inform* **2018**, 37.

8. Sanchez-Lengeling, B.; Outeiral, C.; Guimaraes, G. L.; Aspuru-Guzik, A. Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). https://chemrxiv.org/engage/chemrxiv/article-details/60c73d91702a9beea7189bc2

9. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, 28, 31-36.

10. Holliday, J. D.; Hu, C. Y.; Willett, P., Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb Chem High Throughput Screen* **2002**, 5, 155-66.

11. Thanh-Tung, H.; Tran, T., On Catastrophic Forgetting and Mode Collapse in Generative Adversarial Networks. *arXiv* **2018**, 1807.04015.

12. Breiman, L., Random Forests. *Machine Learning* **2001**, 45, 5-32.

13. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., The ChEMBL database in 2017. *Nucleic Acids Res* **2017**, 45, D945-D954.

14. Christ, C. D.; Zentgraf, M.; Kriegl, J. M., Mining electronic laboratory notebooks: analysis, retrosynthesis, and reaction based enumeration. *Journal of chemical information and modeling* **2012**, 52, 1745-56.

15. Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F., Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent Sci* **2017**, 3, 434-443.

16. Proudfoot, J. R., Molecular Complexity and Retrosynthesis. *J Org Chem* **2017**, 82, 6968-6971.

17. Cadeddu, A.; Wylie, E. K.; Jurczak, J.; Wampler-Doty, M.; Grzybowski, B. A., Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. *Angewandte Chemie* **2014**, 53, 8108-12.

18. Todd, M. H., Computer-aided organic synthesis. *Chemical Society reviews* **2005**, 34, 247-66.

19. Nair, V. H.; Schwaller, P.; Laino, T., Data-driven Chemical Reaction Prediction and Retrosynthesis. *Chimia (Aarau)* **2019**, 73, 997-1000.

20. Warr, W. A., A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol Inform* **2014**, 33, 469-76.

21. Szymkuc, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A., Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew Chem Int Ed Engl* **2016**, 55, 5904-37.

22. Badowski, T.; Molga, K.; Grzybowski, B. A., Selection of cost-effective yet chemically diverse pathways from the networks of computer-generated retrosynthetic plans. *Chem Sci* **2019**, 10, 4640-4651.

23. Segler, M. H. S.; Waller, M. P., Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chemistry* **2017**, 23, 5966-5971.

24. Shibukawa, R.; Ishida, S.; Yoshizoe, K.; Wasa, K.; Takasu, K.; Okuno, Y.; Terayama, K.; Tsuda, K., CompRet: a comprehensive recommendation framework for chemical synthesis planning with algorithmic enumeration. *J Cheminform* **2020**, 12, 52.

25. Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y., Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J Chem Inf Model* **2020**, 60, 47-55.

26. Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J. L.; Butler, C. R., Molecular Transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chem Commun (Camb)* **2019**, 55, 12152-12155.

27. Bai, R.; Zhang, C.; Wang, L.; Yao, C.; Ge, J.; Duan, H., Transfer Learning: Making Retrosynthetic Predictions Based on a Small Chemical Reaction Dataset Scale to a New Level. *Molecules* **2020**, 25.

28. Fukunishi, Y.; Kurosawa, T.; Mikami, Y.; Nakamura, H., Prediction of synthetic accessibility based on commercially available compound databases. *J Chem Inf Model* **2014**, 54, 3259-67.

29. Genheden, S.; Thakkar, A.; Chadimova, V.; Reymond, J. L.; Engkvist, O.; Bjerrum, E., AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J Cheminform* **2020**, 12, 70.

30. Watson, I. A.; Wang, J.; Nicolaou, C. A., A retrosynthetic analysis algorithm implementation. *J Cheminform* **2019**, 11, 1.

31. Kearney, S. E.; Zahoranszky-Kohalmi, G.; Brimacombe, K. R.; Henderson, M. J.; Lynch, C.; Zhao, T.; Wan, K. K.; Itkin, Z.; Dillon, C.; Shen, M.; Cheff, D. M.; Lee, T. D.; Bougie, D.; Cheng, K.; Coussens, N. P.; Dorjsuren, D.; Eastman, R. T.; Huang, R.; Iannotti, M. J.; Karavadhi, S.; Klumpp-Thomas, C.; Roth, J. S.; Sakamuru, S.; Sun, W.; Titus, S. A.; Yasgar, A.; Zhang, Y. Q.; Zhao, J.; Andrade, R. B.; Brown, M. K.; Burns, N. Z.; Cha, J. K.; Mevers, E. E.; Clardy, J.; Clement, J. A.; Crooks, P. A.; Cuny, G. D.; Ganor, J.; Moreno, J.; Morrill, L. A.; Picazo, E.; Susick, R. B.; Garg, N. K.; Goess, B. C.;

Grossman, R. B.; Hughes, C. C.; Johnston, J. N.; Joullie, M. M.; Kinghorn, A. D.; Kingston, D. G. I.; Krische, M. J.; Kwon, O.; Maimone, T. J.; Majumdar, S.; Maloney, K. N.; Mohamed, E.; Murphy, B. T.; Nagorny, P.; Olson, D. E.; Overman, L. E.; Brown, L. E.; Snyder, J. K.; Porco, J. A., Jr.; Rivas, F.; Ross, S. A.; Sarpong, R.; Sharma, I.; Shaw, J. T.; Xu, Z.; Shen, B.; Shi, W.; Stephenson, C. R. J.; Verano, A. L.; Tan, D. S.; Tang, Y.; Taylor, R. E.; Thomson, R. J.; Vosburg, D. A.; Wu, J.; Wuest, W. M.; Zakarian, A.; Zhang, Y.; Ren, T.; Zuo, Z.; Inglese, J.; Michael, S.; Simeonov, A.; Zheng, W.; Shinn, P.; Jadhav, A.; Boxer, M. B.; Hall, M. D.; Xia, M.; Guha, R.; Rohde, J. M., Canvass: A Crowd-Sourced, Natural-Product Screening Library for Exploring Biological Space. *ACS Cent Sci* **2018**, 4, 1727-1741.

32. Cameron, L. P.; Tombari, R. J.; Lu, J.; Pell, A. J.; Hurley, Z. Q.; Ehinger, Y.; Vargas, M. V.; McCarroll, M. N.; Taylor, J. C.; Myers-Turnbull, D.; Liu, T.; Yaghoobi, B.; Laskowski, L. J.; Anderson, E. I.; Zhang, G.; Viswanathan, J.; Brown, B. M.; Tjia, M.; Dunlap, L. E.; Rabow, Z. T.; Fiehn, O.; Wulff, H.; McCorvy, J. D.; Lein, P. J.; Kokel, D.; Ron, D.; Peters, J.; Zuo, Y.; Olson, D. E., A non-hallucinogenic psychedelic analogue with therapeutic potential. *Nature* **2021**, 589, 474-479.

33. Rusnak, D. W.; Lackey, K.; Affleck, K.; Wood, E. R.; Alligood, K. J.; Rhodes, N.; Keith, B. R.; Murray, D. M.; Knight, W. B.; Mullin, R. J.; Gilmer, T. M., The effects of the novel, reversible epidermal growth factor receptor/ErbB-2 tyrosine kinase inhibitor, GW2016, on the growth of human normal and tumor-derived cell lines in vitro and in vivo. *Mol Cancer Ther* **2001**, 1, 85-94.

34. Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* **2018**, 4, 268-276.

35. Prykhodko, O.; Johansson, S. V.; Kotsias, P. C.; Arus-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H., A de novo molecular generation method using latent vector based generative adversarial network. *J Cheminform* **2019**, 11, 74.

36. Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L., Quantifying the chemical beauty of drugs. *Nat Chem* **2012**, 4, 90-8.

37. Warr, W. A., Scientific workflow systems: Pipeline Pilot and KNIME. *J Comput Aided Mol Des* **2012**, 26, 801-4.

38. Silverman, R., *The Organic Chemistry of Drug Design and Drug Action*. Elsevier: 2004.

39. Brown, N., *Bioisosteres in Medicinal Chemistry*. Wiley-VCH Verlag GmbH & Co. KGaA: 2012.

40. Langmuir, I., ISOMORPHISM, ISOSTERISM AND COVALENCE. *Journal of the American Chemical Society* **1919**, 41, 1543-1559.

41. Erlenmeyer, H.; Willi, E., Zusammenhänge zwischen Konstitution und Wirkung bei Pyrazolonderivaten. *Helvetica Chimica Acta* **1935**, 18, 740-743.

42. Erlenmeyer, H.; Leo, M., Über Pseudoatome. *Helvetica Chimica Acta* **1932**, 15, 1171-1186.

43. Doble, M.; Kruthiventi, A. K.; Gajanan, V., *Biotransformations and Bioprocesses.* . CRC Press: 2004.

44. Tyagarajan, S.; Lowden, C. T.; Peng, Z.; Dykstra, K. D.; Sherer, E. C.; Krska, S. W., Heterocyclic Regioisomer Enumeration (HREMS): A Cheminformatics Design Tool. *J Chem Inf Model* **2015**, 55, 1130-5.

45. Topliss, J. G., Utilization of operational schemes for analog synthesis in drug design. *Journal of Medicinal Chemistry* **1972**, 15, 1006-1011.

46. Schönherr, H.; Cernak, T., Profound Methyl Effects in Drug Discovery and a Call for New C□H Methylation Reactions. *Angewandte Chemie International Edition* **2013**, 52, 12256-12267.

47. Pinheiro, P. S. M.; Rodrigues, A. D.; Maia, R. C.; Thota, S.; Fraga, C. A. M., The Use of Conformational Restriction in Medicinal Chemistry. *Curr Top Med Chem* **2019**, 19, 1712-1733.

48. Liu, D.; Jiang, H.; Chen, K.; Ji, R., A New Approach to Design Virtual Combinatorial Library with Genetic Algorithm Based on 3D Grid Property. *Journal of Chemical Information and Computer Sciences* **1998**, 38, 233-242.

49. Anon eMolecules. https://www.emolecules.com/info/plus/download-database

50. Anon ChEMBL. https://chembl.gitbook.io/chembl-interface-documentation/downloads

51. Bemis, G. W.; Murcko, M. A., The properties of known drugs. 1. Molecular frameworks. *J Med Chem* **1996**, 39, 2887-93.

52. Sheridan, R. P.; Hunt, P.; Culberson, J. C., Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *Journal of Chemical Information and Modeling* **2006**, 46, 180-192.

53. Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S., A collection of robust organic synthesis reactions for in silico molecule design. *J Chem Inf Model* **2011**, 51, 3093-8.

54. Stine, R., An Introduction to Bootstrap Methods: Examples and Ideas. *Sociological Methods and Research* **1989**, 18, 243-291.

55. Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A., Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front Pharmacol* **2020**, 11, 565644.

56. Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C., GuacaMol: Benchmarking Models for de Novo Molecular Design. *J Chem Inf Model* **2019**, 59, 1096-1108.

57. Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G., Frechet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J Chem Inf Model* **2018**, 58, 1736-1741.

58. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M., RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *Journal of Chemical Information and Computer Sciences* **1998**, 38, 511-522.

59. Neil, D.; Segler, M. H. S.; Guasch, L.; Ahmed, M.; Plumbley, D.; Sellwood, M.; Brown, N. Exploring Deep Recurrent Models with Reinforcement Learning for Molecule Design In ICLR 2018 Conference, 2018; 2018.

60. Higa, G. M.; Abraham, J., Lapatinib in the treatment of breast cancer. *Expert Rev Anticancer Ther* **2007**, 7, 1183-92.

61. Saleem, A.; Searle, G. E.; Kenny, L. M.; Huiban, M.; Kozlowski, K.; Waldman, A. D.; Woodley, L.; Palmieri, C.; Lowdell, C.; Kaneko, T.; Murphy, P. S.; Lau, M. R.; Aboagye, E. O.; Coombes, R. C., Lapatinib access into normal brain and brain metastases in patients with Her-2 overexpressing breast cancer. *EJNMMI Research* **2015**, 5, 30.

62. Chen, X.-Y.; Ozturk, S.; Sorensen, E. J., Synthesis of Fluorenones from Benzaldehydes and Aryl Iodides: Dual C–H Functionalizations Using a Transient Directing Group. *Organic Letters* **2017**, 19, 1140-1143.

63. Singer, R. A. Commercial Development of Axitinib (AG-013736): Optimization of a Convergent Pd-Catalyzed Coupling Assembly and Solid Form Challenges. In *Transition Metal-Catalyzed Couplings in Process Chemistry*; 2003, pp 165-180.

64. Gupta, A.; Muller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G., Erratum: Generative Recurrent Networks for De Novo Drug Design. *Mol Inform* **2018**, 37.

65. Bjerrum, E. J.; Threlfall, R., Molecular Generation with Recurrent Neural Networks (RNNs). *arXiv* **2017**, 1705.04612.

66. Domenico, A.; Nicola, G.; Daniela, T.; Fulvio, C.; Nicola, A.; Orazio, N., De Novo Drug Design of Targeted Chemical Libraries Based on Artificial Intelligence and Pair-Based Multiobjective Optimization. *J Chem Inf Model* **2020**, 60, 4582-4593.

67. Maziarka, L.; Pocha, A.; Kaczmarczyk, J.; Rataj, K.; Danel, T.; Warchol, M., Mol-CycleGAN: a generative model for molecular optimization. *J Cheminform* **2020**, 12, 2.

68. Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noe, F.; Clevert, D. A., Efficient multi-objective molecular optimization in a continuous latent space. *Chem Sci* **2019**, 10, 8016-8024.

69. Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A., Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol* **2019**, 37, 1038-1040.

70. Durrant, J. D.; Lindert, S.; McCammon, J. A., AutoGrow 3.0: an improved algorithm for chemically tractable, semi-automated protein inhibitor design. *Journal of molecular graphics & modelling* **2013**, 44, 104-12.

71. Clark, A. M.; Dole, K.; Coulon-Spector, A.; McNutt, A.; Grass, G.; Freundlich, J. S.; Reynolds, R. C.; Ekins, S., Open source bayesian models: 1. Application to ADME/Tox and drug discovery datasets. *Journal of chemical information and modeling* **2015**, 55, 1231-1245.

72. Saubern, S.; Guha, R.; Baell, J. B., KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Mol Inform* **2011**, 30, 847-50.