bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Transformer neural network for protein specific de novo drug generation as machine translation problem

Daria Grechishnikova[1]*

[1] Physics Department, Lomonosov Moscow State University, Moscow, 119991, Russia

E-mail: daria.grechishnikova@gmail.com

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Abstract

Drug discovery for the protein target is a very laborious, long and costly process. Machine learning approaches, and deep generative networks in particular, can substantially reduce development time and costs. However, the majority of methods imply prior knowledge of protein binders, their physicochemical characteristics or three-dimensional structure of the protein. The method proposed in this work generates novel molecules with predicted ability to bind target protein relying on its amino acid sequence only. We consider target specific de novo drug design as a translational problem between amino acid "language" and SMILE (Simplified Molecular Input Line Entry System) representation of the molecule. To tackle this problem, we apply Transformer neural network architecture, the state-of-the-art approach in sequence transduction tasks. The Transformer is based on a self-attention technique which allows capturing long-range dependencies between items in sequence. The model generates realistic diverse compounds with structural novelty. The computed physicochemical properties and common metrics used in drug discovery fall within the plausible drug-like range of values.

Introduction

Drug development is a multi-stage process that requires many resources. Bringing drug to market may take up to 20 years [1]. The total cost may vary within from US$0.5 billion to US$2.6 billion [2]. The estimated amount of drug-like molecule space is 10^{60} while the number of synthesized compounds is on the order of 10^{8} [3]. Therefore, the search for a promising molecule that may bind to a target protein is a challenging task for chemists. A high-throughput screening technique allows testing millions of molecules in vitro to determine compounds that may act on the protein of interest[4]. However, this method is expensive and time-consuming. Virtual screening is used to search libraries of billions of molecules in silico [5]. This method requires information about compounds active against the protein or knowledge of protein three-dimensional structure and operates on already known molecules which span only small part of synthetically accessible molecule space. In de novo drug design one have to create a molecule active toward the desired biological target from scratch. Existing computational methods often generate molecules that are hard to synthesize or restrict accessible chemical space via coded rules [6]. Despite all efforts, targeted generation of molecules remains a challenging task. Recently machine learning methods were proposed to tackle this problem [7].

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Most of the deep learning models for molecule generation are based on recurrent neural network (RNN). RNN is commonly used for modeling sequence data. The main feature of RNN allowing it to work with sequential data is the ability to make use of information from preceding steps. RNN can reveal links between distant elements of a sequence [8]. Unfortunately, RNNs suffer from the problem of vanishing gradients which significantly limits the ability to work with long sequences. long short-term memory and gated recurrent units partially solve this issue [9]. Recently, several works introduced recurrent neural networks based on the long short-term memory for de novo molecule generation [10-12]. They use Simplified Molecular-Input Line-Entry (SMILE) strings as input. Fine-tuning on a smaller dataset with compounds known to be active against biological target force the models to generate focused molecule libraries with the desired activity toward the same target. Several research groups applied a reinforcement learning approach to bias the generator to produce molecules with desired properties [13-20]. In the reinforcement learning paradigm the agent (generator in de novo drug generation problem) takes some action (choosing next character during new SMILE string generation) to maximize reward (function computed after SMILE string completion). Olivecrona et al. fine-tuned the RNNs for generating compounds binding Dopamine Receptor Type 2 (DRD2). To predict molecule activity they built a Support Vector Machine (SVM) classifier with a Gaussian kernel and trained it on the DRD2 activity dataset [13]. The output of this model was used to formulate the reward function. Popova et al. [14] suggested training separately two neural network – generative and predictive- then use them jointly to generate novel libraries of compounds with desired properties e.g. targeted toward Janus kinase 2. Several research groups applied generative adversarial network concept to design compounds with optimized properties but they did not consider activity against any biological target [17,18].

Another fundamental approach to de novo compound generation is based on autoencoder architecture [21-28]. Autoencoder consists of encoder and decoder networks [8]. Former one convert the input data into a latent representation (vector of fixed dimension), the second one reconstructs the initial object from the latent code. The hidden layer forming the latent representation vector is an informational bottleneck which induces the model to capture the most important features of the input object [8]. Variational and adversarial autoencoders are two types of autoencoders widely used to generate molecules. In variational autoencoders a prior distribution, usually normal, is imposed on latent space to make it smooth and suitable for sampling [29]. In adversarial autoencoders the discriminator neural network is introduced into architecture to force the distribution of latent codes to follow arbitrary prior distribution [30]. Gómez-Bombarelli et al. [21] suggested variational autoencoder extended by attaching a multilayer perceptron to the latent layer for property prediction. Joint training of this enlarged model forces the latent space to organize by property values. On top of this model, authors trained the Gaussian process to predict target

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

compound properties using the latent space representation as input. In recent publication [22] authors have compared several autoencoder architectures including variational and adversarial ones. The adversarial autoencoder provides the highest fraction of valid SMILE strings. Authors trained the SVM classifier to predict activity against DRD2. They use this probability as the objective function and maximized it during the latent space Bayesian optimization. Also, autoencoder can be used for a conditional generation [31-33]. In these studies properties were directly imposed on latent space during the training. Polykovskiy et al. introduced a conditional adversarial autoencoder to design compounds with specified properties [33]. After training on a set of Janus kinase 2 (JAK2) and Janus kinase 3 (JAK3) inhibitors and conditioning on the selective activity against JAK2 the model generated compound which turned out to be active toward JAK2 during in vitro tests. Recently, Zhavoronkov et al. developed discoidin domain receptor 1 (DDR1) inhibitor in 21 days using variational autoencoder fine-tuned with reinforcement learning approach [25]. One molecule successfully passed experiments in mice.

However, all these methods imply prior knowledge of protein binders and their physicochemical characteristics. Structure-based drug design approaches require three-dimensional structure of the protein. In this work we introduce an approach to targeted drug generation which uses only protein amino acid sequence as input. We consider target specific de novo drug generation problem as a translation from amino acid "language" to SMILE representation of the molecule. Recently the Transformer based models demonstrated state-of-the-art results on neural machine translation tasks [34,35]. We adopt the Transformer to generate molecules. The network takes amino acid sequence as input and generates molecules with predicted ability to bind the protein target. The model outputs valid structures with plausible values of computed physicochemical characteristics, drug-likeness metric, synthetic accessibility score.

The main contributions of our work are as follows:

1. We formulate targeted drug generation problem as translational task and applied the Transformer architecture. This allows molecule generation based on protein amino acid sequence only.
2. Our approach requires neither prior knowledge of protein binders nor preparing libraries of ligands active against the target.
3. The proposed model is based on a self-attention technique which allows better capturing of long-range dependencies than recurrent neural networks.

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Methods

Data

We retrieved data from BindingDB [36]. BindingDB contains a measured binding affinity of interactions between proteins and drug-like molecules. The full database version was downloaded. The raw dataset contained over 1.5 million data records. We selected records from the raw dataset using the following criteria:

1. The field "Target Source Organism According to Curator or DataSource" equals to "Homo sapiens".
2. The record has IC50 value less than 100 nm; if IC50 is missing then Kd is less than 100 nm; if both are missing then EC50 is less than 100 nm.
3. The record has a chemical identifier (PubChem CID).
4. The record has SMILE representation.
5. The molecular weight is less than 1,000 Da.
6. The record has a protein identifier (Uniprot ID).
7. Protein amino acid sequence length is greater than 80 and lower than 2050.

This yielded a result dataset containing 214 749 records. There were 1128 unique amino acid sequences and 140575 unique ligand SMILE strings. All SMILE strings used in this work were canonicalized using RDKit. We randomly selected 100 unique protein sequences from dataset five times (i.e. roughly 10% of unique proteins). It gave five train/test dataset pairs. Proteins from training and test sets did not intersect.

Data representation

We considered each symbol in amino acid sequence or in SMILE string as token. The vocabulary was determined by the dataset and contained 71 symbol. Each token was converted into vector using trainable embedding in the first layer of encoder.

Model

We adopted the Transformer model for targeted drug generation using the original implementation described in [35]. The Transformer has an encoder-decoder structure. The encoder maps a protein amino acid sequence $(a_1,..,a_n)$ to a sequence of continuous representations $z=(z_1,\ldots,z_n)$. Then the decoder takes z as input and generates SMILE string in auto-regressive manner. At every step of generation, the decoder may attend to any elements of z due to the attention mechanism. The latent code z may be considered as a "protein context" used by decoder to generate a molecule structure. The model yields a probability distribution over each element in the vocabulary for each position in the output sequence. The Transformer is based on attentional mechanism only. It lacks any kind of convolutional or recurrent neural network components. Transformer uses self-attention to compute

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

the representations of input and output sequences. Self-attention refers to different components of a single sequence in relation to other components to compute sequence representation. Each layer of the encoder is composed of multi-head self-attention sub-layer and feed-forward sub-layer. In addition to them, each layer of the decoder has multi-head attention layer attending encoder output. The self-attention mechanism successfully copes with long-range dependencies while being faster than recurrent layers. The attention layer at first calculates three vectors from each "word" of a "sentence' – key, query and value. To process all words in a sentence simultaneously key vectors are packed together into matrix *K*, queries and values produce matrices *Q* and *V*. In our task definition "words" are amino acid residuals or characters in SMILE strings. The attention is computed as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

, where d_k is a scaling factor.

The multi-head attention mechanism produces *h* different representations of *Q*, *K*, *V* values, and computes attention function for each representation:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

The outputs are concatenated and projected one more time yielding final values:

$$Multihead(Q, K, V) = (head_1, \dots, head_h)W^O$$

, where W_i^Q, W_i^K, W_i^V are matrices of learned weights.

Since the model lacks any recurrent component it has no information about the order of tokens in a sequence. To address this, the model adds position-dependent signals to the input embedding. There are many possible choices for signal functions. In Transformer sine and cosine functions are used:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right)$$

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

, where *pos* is the position, i is the dimension and d_{model} is the size of embedding.

We use beam search to decode SMILE strings. While constructing a sequence the beam search evaluates all possible next steps and keeps top n candidates with the highest probability, where n is the user-specified parameter referred to as beam size. If beam size is equal to one the beam search becomes the greedy search. If beam size is greater than one the output sequences differ only slightly from each other. It might be beneficial if generated molecule is good enough and small structure optimizations are needed. But in a process of target specific de novo drug generation it would be better to have more diverse variants per certain protein. There are several ways to potential improvement. We discuss them in the "Result and Discussion" section. For each protein we ran beam search with beam size of 4 and 10. In the first case we left only one SMILE string with the highest probability (one per one mode). In the second case we left all ten generated molecules for subsequent analysis (ten per one mode).

All work was performed in Google Colaboratory. We use open-source tensor2tensor library for building, training and testing the model [37]. We experimented with different numbers of attentional heads, layers, and their sizes. The optimal proportion between the amount of valid and unique SMILE strings gives the model containing four layers of size 128 and four attention heads. We use Adam optimizer and learning rate decay scheme proposed in [35], the batch size was set to 4096 tokens. The training runs for 600K epoch using one GPU.

To test the model, we perform Monte-Carlo cross validation. We randomly selected 100 unique proteins to create test dataset. Then we trained the model and tested it on selected proteins. This procedure was repeated five times.

Model evaluation

We used RDKit [38] to check chemical validity, calculate properties, compute similarity scores and produce SMILE canonical representation of molecule structures. Molecules known to be active against given target protein and generated ones were docked in binding sites using SMINA [39]. Protein PDB structures were downloaded from the Protein Data Bank [40]. We followed the standard procedure to prepare protein for docking, heteroatoms were removed, hydrogens were added via PyMol [41]. We utilized OpenBabel [42] to generate three dimensional conformers.

Results and Discussion

Chemical feasibility

This section demonstrates the effectiveness of the proposed approach for the generation of valid realistic molecules. We created five different divisions of the initial dataset to train and test parts. For each division, we performed training of the model followed by validation on the corresponding test set. Each test set contained 100 unique amino acid sequences. At first we ran the model in one

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

per one mode (see Methods). For each protein in test datasets the model generated a molecule. Thus, one hundred SMILE strings were produced for each proteins test set. We checked the chemical validity of molecules with RDKit software, analyzed uniqueness and searched the ZINC15 database for generated compounds. All characteristics were averaged across five test sets. About 93% of generated molecules were valid and almost 94% were unique (Table 1). About 35% of compounds were found in ZINC15 database.

	Dataset 1	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Dataset 5**	**Average**
Total number of generated SMILE strings (one per one target protein)	100	100	100	100	100	100
Valid (%)	95,0	94,0	87,0	97,0	92,0	93,0
Unique (%)	99,0	91,0	96,0	91,0	92,0	93,8
Match with ZINC15 database (%)	30,0	40,0	37,0	38,0	30,0	35,0

Table 1. Percentages of valid, unique and found in ZINC15 database SMILES strings generated by the model in one per one mode.

In case of generating one ligand per protein, the outputted compound might be considered as valid starting points for a subsequent modification during the drug discovery process. Nevertheless, it would be useful to obtain more drug candidates for the given target protein. To achieve this, we expanded beam size to ten allowing the model to output ten most probable variants of the compound for inputted protein (ten per one mode). In this mode the model generated over 87% valid SMILE strings and 88% unique on average across five datasets (Table 2). Over 24% of novel compounds matched the ZINC15 database.

The number of valid and unique SMILE strings is lower in ten per one mode. We assume that this is caused by the problem of performance degradation in beam search. Recently proposed method may possibly increase the performance [43]. However, this improvement is outside the scope of our work.

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Average
Total number of generated SMILE strings (ten per one target protein)	1000	1000	1000	1000	1000	1000
Valid	86,2	92,8	83,4	87,2	89,1	87,7
Unique	94,9	85,1	88,5	85,8	86,5	88,2
Match with ZINC15 database (%)	21,9	27,9	22,7	22,8	25,6	24,2

Table 2. Percentages of valid, unique and found in ZINC15 database SMILES strings generated by the model in ten per one mode.

Testing the binding affinity between generated compounds and target proteins

In this section, our goal is to test binding affinity of generated molecules to the active site of the target protein. At first, we randomly selected a number of proteins from test datasets. We manually inspected a PDB file of each protein, eliminated those containing complexes between the protein of interest and other ones, removed ligands. This procedure resulted in 12 proteins. Our model generated 10 candidate ligands for each of them. For comparison, we collected known binders of these proteins from BindingDB. The binding energies between ligands and target proteins active sites were computed using SMINA. In molecular docking the binding energy indicates how well a ligand can fit in protein binding pocket. A more negative value corresponds to stronger protein-ligand interaction. Figure 1 shows the comparison of binding energies (reported in kcal/mol) between proteins and generated ligands or ligands from the database. Vertical lines represent mean values while box borders depict standard deviations. We see that in 8 cases of 12 the generated ligands outperform known binders in energy values. In two cases the energy ranges overlap. For the two remaining proteins, the ligands from the database exhibit stronger binding energies, however, the differences are not very large (0.8 and 1.2 kcal/mol). The docking results indicate the ability of the proposed model to generate ligands with favorable binding affinity to target protein.

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

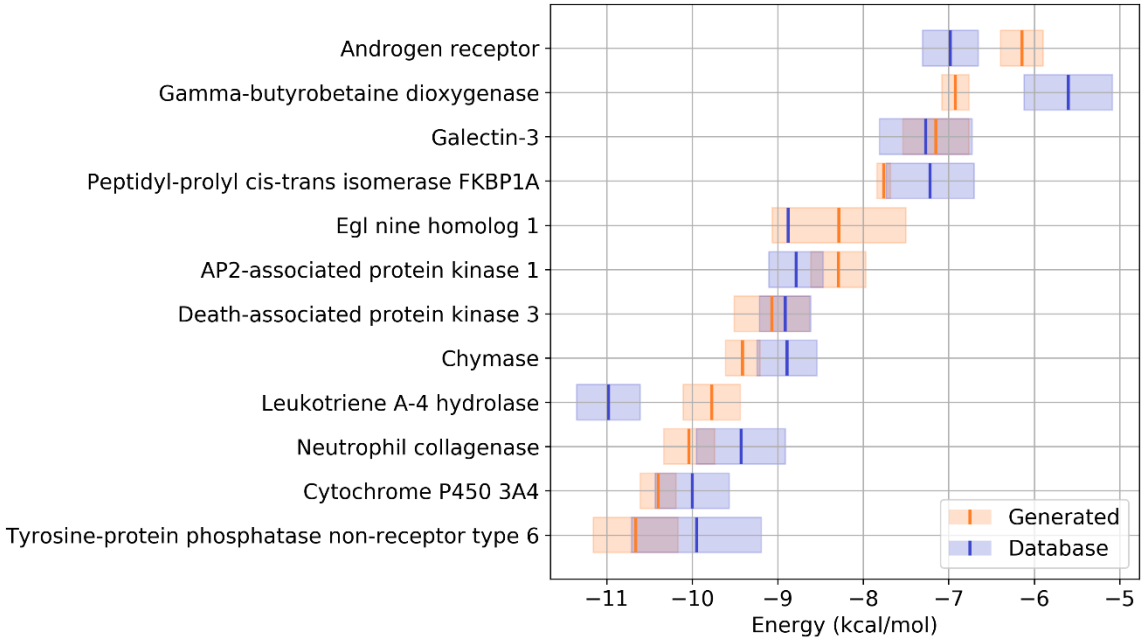

Figure 1. Predicted binding energies of interactions between proteins (randomly selected from test dataset) and generated ligands, known binders from database.

We visualized three randomly chosen complexes of target proteins and generated ligands using PyMol software. Figure 2 shows the docking poses of generated molecules in the active site of the corresponding targets. The first protein is neutrophil collagenase (MMP8). MMP8 is involved in the degradation of extracellular matrix components and plays a key role in tissue remodeling [44]. Thus excessive protein activity may lead to the development of destructive processes in connective tissue. MMP8 is known to contribute to the pathophysiology of many diseases including periodontitis, psoriasis, multiple sclerosis, osteoarthritis, rheumatoid arthritis, osteoporosis, and Alzheimer's Disease [44,45]. The second one is death-associated protein kinase 3 (DAPK3). The two main biological roles of DAPK3 are regulation of apoptosis and smooth muscle contraction [46]. DAPK3 is involved in various pathogenic processes including regulation of tumor suppression or progression, vascular inflammation, smooth muscle cell proliferation, migration and hypertension. Hereby DAPK3 turns out to be a therapeutic target for many diseases [47]. The third protein is tyrosine-protein phosphatase non-receptor type 11 (Shp2). Shp2 is involved in signal transduction processes important for proliferation, differentiation, metabolism and others [48]. The protein is related to adult acute myelogenous leukemia and human solid tumors. Recent studies suggest Shp2 involvement in other human cancers, including breast cancer, liver cancer, gastric cancer, oral cancer, non–small cell lung cancer and thyroid cancer [49]. It makes Shp2 a valuable target for drug development.

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

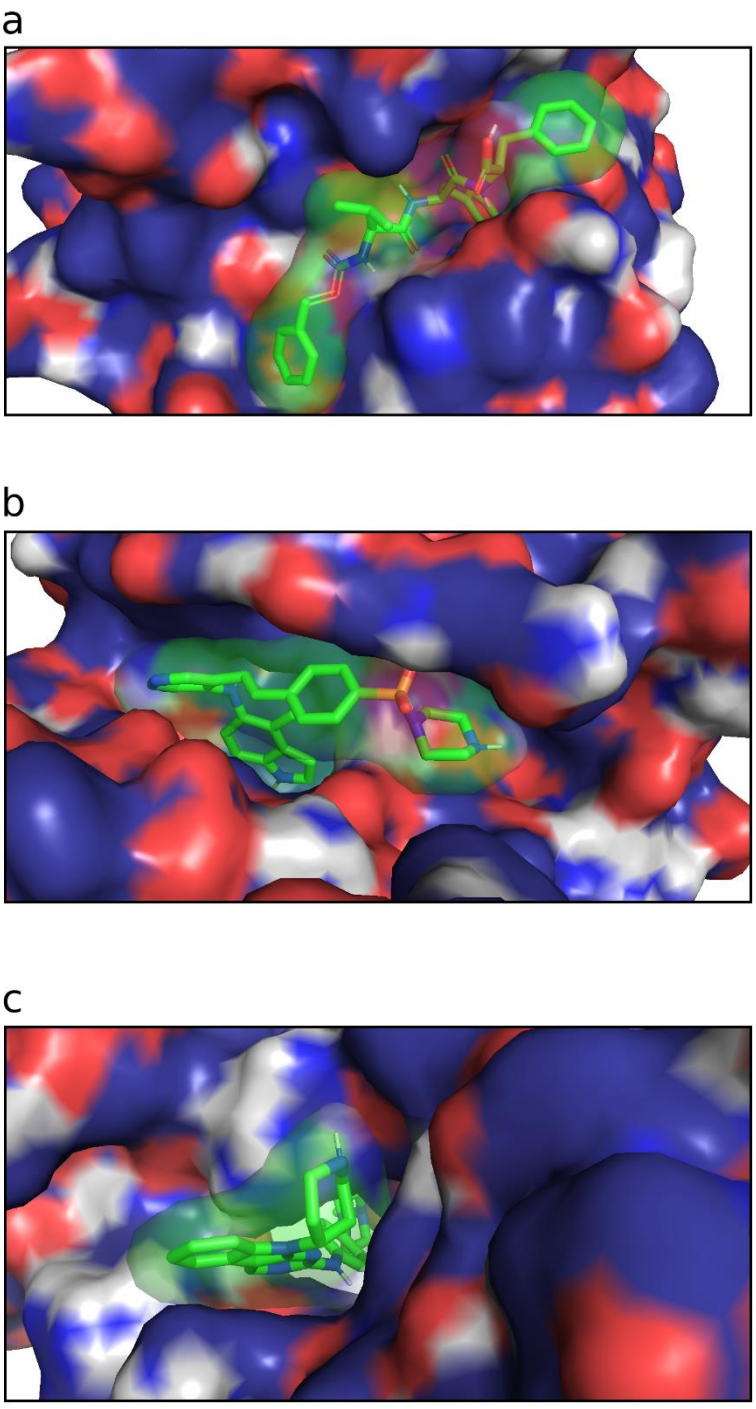

Figure 2. Positions of the generated molecules in the active sites of the following proteins: a) Neutrophil collagenase, b) Death-associated protein kinase 3, c) Tyrosine-protein phosphatase non-receptor type 6.

Physicochemical properties and metrics

It is not enough for the model to output chemically valid molecules active against a certain target. The model should also take care of parameters crucial for a molecule to be a potential drug. We computed several important metrics and physicochemical properties for generated compounds and

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

compared them with corresponding characteristics of the molecules from the training dataset. The goal was to access the ability of the model to generate compounds satisfying typical drug-likeness metrics. According to the famous Lipinski's rule of five, the water-octanol partition coefficient (logP) of a potential orally active drug should not exceed five. Molecules with molecular weight less than 500 show better permeability and solubility. Numbers of the hydrogen donors, acceptors and rotatable bonds have to be no more than 5,10 and 10 respectively [50,51]. Although the Lipinski's rule was developed for oral medications, it gives a good reference point for evaluating the properties of the generated molecules.

Property name	**Constraints**	**Structures satisfying the constraints (%)**	
		Generated molecules (one per one) **	**Generated molecules (ten per one) ****
logP	<5	76.9	76.5
Molecular weight (Da)	<500	84.4	79.5
Number of hydrogen donors	<5	94.0	92.0
Number of hydrogen acceptors	<10	88.0	87.2
Number of rotational bonds	<10	92.3	89.0
Topological polar surface area ($Å^2$)	<140	94.0	92.0
Quantitative Estimate of Drug-likeness (QED) *		0.60 ± 0.22	0.55 ± 0.22
Synthetic accessibility score	<6	99.8	100.0

* There is no common threshold for QED. QED vary in range [0,1]. The higher QED value is, the better. The columns show mean values and standard deviations.

**averaged across five cross validational datasets

Table 3. Percentage of generated molecules falling within plausible for drug-like molecules ranges of values.

The Topological Polar Surface Area (TPSA) is another important characteristic of a drug candidate. Chemists assume that molecules having a topological polar surface area greater than 140$Å^2$ are absorbed poorly [51]. To overcome blood-brain barrier the molecule should have TPSA less than 90 $Å^2$ [52]. Quantitative Estimate of Drug-likeness (QED) is based on the desirability functions for molecular properties widely used to select appropriate compounds during the early stages of drug discovery. In other words, QED is the measure of drug-likeness [53]. It ranges from zero to one where zero indicates totally unsuitable molecule while one corresponds to molecules with favorable

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

characteristics. The Synthetic Accessibility Score (SAS) are of great importance as many computational approaches often yield molecules tending to be difficult to synthesize (SAS > 6) [54]. Table 3 summarizes data about the compliance of the generated molecules with the rules mentioned above across five datasets used for Monte-Carlo cross-validation. For each constraint, the majority of generated compounds lie in acceptable for drug-like molecule boundaries. Figure 3 shows the distributions of logP, the number of H-donors, H-acceptors, and rotatable bonds, QED, SAS, TPSA, molecular weight and length for the first test dataset. The distributions for four remaining test datasets are almost identical. We analyzed computed characteristics of molecules from three datasets: structures generated in one per one mode, ten per one mode and training set. For each parameter, the histograms display almost complete overlapping between datasets. It indicates that the model reproduces the properties distribution of molecules in training set very well.

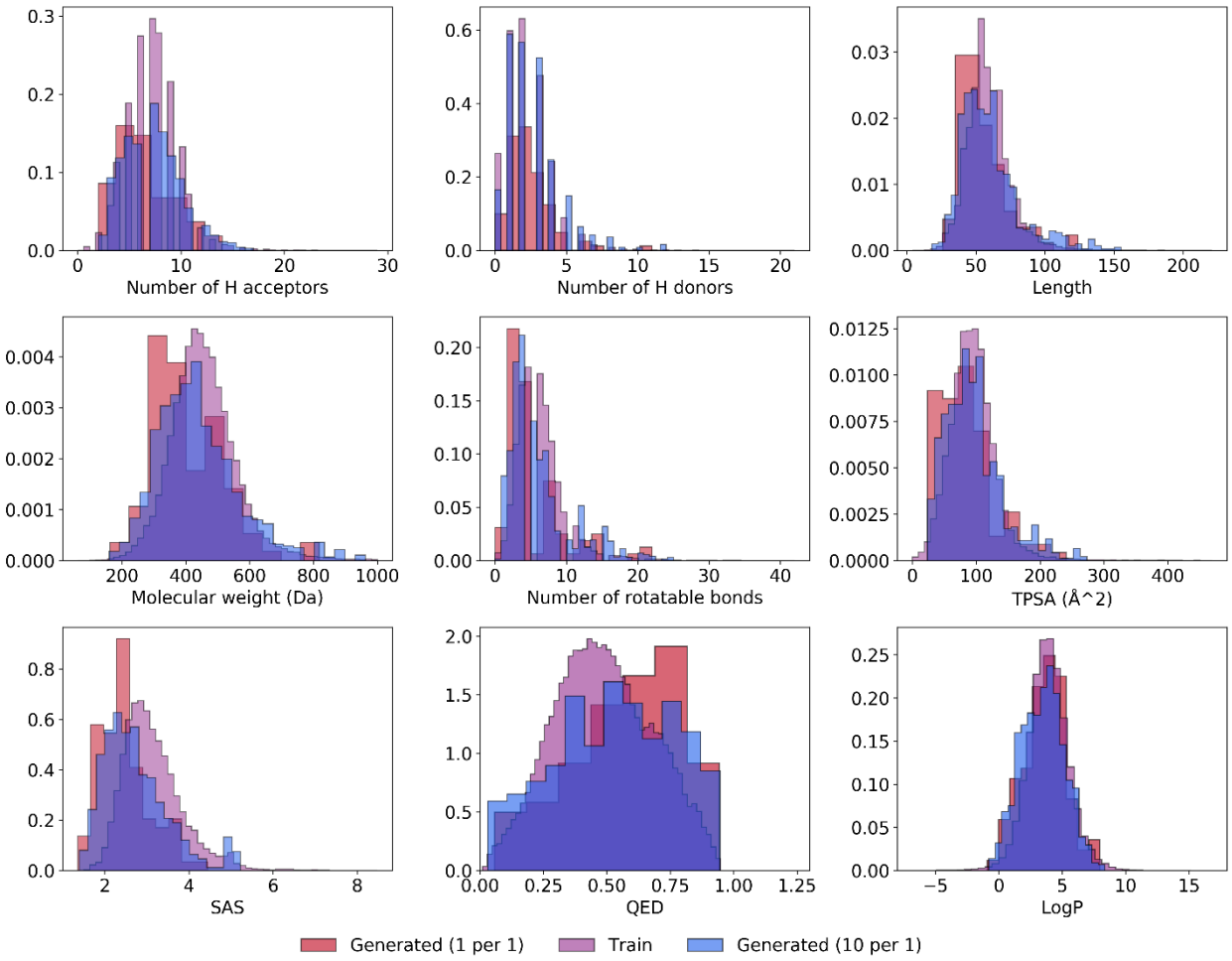

Figure 3. Distribution of properties for the generated molecules. Properties include: water-octanol partition coefficient (logP), the number of H-donors, the number of H-acceptors, the number of rotatable bonds, Quantitative Estimation of Drug-likeness (QED), the synthetic accessibility score (SAS), total polar surface area, molecular weight and length. Vertical lines represent mean values while box borders depict standard deviations.

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The favorable values of these parameters do not necessarily indicate that generated compound will become a drug. It can be checked only in an experiment. Nevertheless, we can conclude that generated molecules may be considered as starting points in developing novel drugs with activity against given protein targets.

We assessed the structural diversity between generated molecules and molecules from the training dataset by calculating the Tanimoto similarity score implemented in RDKit. Figure 4 shows the distributions of the nearest neighbor Tanimoto coefficients over all pairs of molecules. Only 20% of all generated structures have Tanimoto score above similarity threshold (Tanimoto score > 0.85) and can be considered similar to structures from the training dataset.

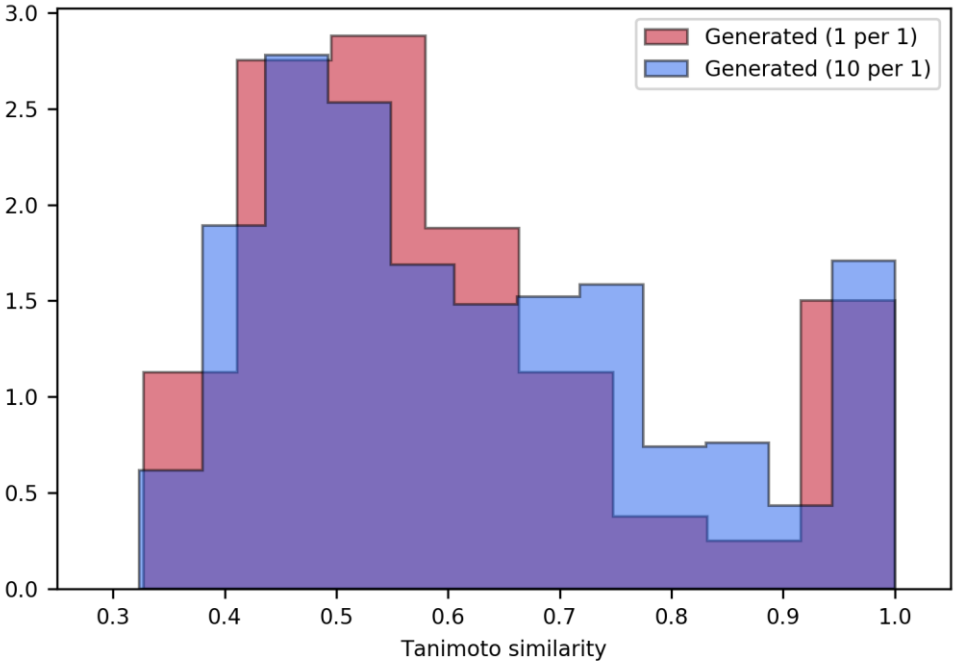

Figure 4. Tanimoto similarity of generated molecules to the nearest neighbor in training dataset.

We also analyzed Bemis-Murcko scaffolds of molecules in training and generated datasets. Bemis and Murcko proposed a molecule structure classification system based on a decomposition of compounds into the ring, linker, framework, and side-chain atoms [55]. The framework includes a ring system and linkers in molecule and is referred to as “Murcko scaffold”. Our analysis shows that only 12% of scaffolds in the generated dataset could be found among compounds used for training. These results indicate that the model does not just copy items from the training dataset, but generates molecules with structural novelty.

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The Transformer applicability to drug generation task

Usually deep learning methods need library of molecules with known activity against a certain protein to generate ligand binding with the target. The specific library is used to fine-tune the model or to train predictive network that assigns reward to generator output in a reinforcement learning approach (e.g., [10,14,22]). In several research works authors used seed molecule to generates structures with desired activity (e.g., [27,28]). In other words, these approaches demand some prior information about compounds which are active against given target. The method proposed in this work does not imply knowledge of active ligands or any kind of chemical descriptors of the molecule. At the same time, the method does not rely on information about the three-dimensional structure of the protein of interest. Usually, protein three-dimensional structure determination is not an easy task. Also, it may be quite costly. Therefore, the usage of an amino acid sequence as input may substantially simplify one of initial stages of drug discovery – search of the lead compound and can be very fruitful in case of targeting protein with limited or no information about inhibitors and three-dimensional structure.

To the best of our knowledge, this paper is the first attempt to present de novo drug generation problem as a translational task between protein sequence and SMILE representation of the molecule. The method has benefited from the recent progress in the neural machine translation field where the Transformer architecture demonstrated state-of-the-art results [34]. Recently the Transformer also exhibited very promising results in predicting the products of chemical reaction and retrosynthesis [56,57]. One of the key features of the Transformer is self-attention layers. They reduce the length of the paths the signal should travel during deep network learning. It allows the model to maintain long-range dependencies in sequence much better than in recurrent neural networks. The self-attention in Transformer architecture operates on both – input amino acid sequence and already generated part of SMILE string giving access to any part of them at any time. Intuitively, self-attention is a good choice for translation between protein and molecule. Firstly, protein sequence may be quite long - dozens of times longer than SMILE string. Secondly, three-dimensional structural feature of the protein may be formed by amino acid residues located far from each other in the sequence representation. That is why it is so important for the algorithm to reference elements coming long before the current one. The multi-head self-attention mechanism allows the model to jointly attend to different aspects of positions being important in relation to proceeding elements. In language translation tasks it means that the Transformer may capture for example both semantic and grammatical meaning of the particular word. Intuitively, it appears that this ability may be helpful in capturing 3D features of protein binding pocket. For example, a model may consider a certain

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

residue simultaneously in two aspects: forming the pocket and interacting directly with the drug. This is just our assumption requiring additional checking.

Currently, the vast amount of deep learning approaches to the drug generation task uses the similarity of organic chemistry structure and natural human language. Chemists understand molecule structure much like a human understands words. Segler et al. introduced encoder-decoder RNN architecture for the construction of a chemical language model i.e. the probability distribution over sequence of characters in SMILE notation [10]. Others implemented variational and adversarial autoencoders to create a continuous latent representation of chemical spaces (e.g. [22]). It allows easy sampling of latent codes and decoding them to SMILE strings corresponding to novel molecules. Reinforcement learning technique and fine-tuning on specific datasets were proposed to bias probability distribution toward desired properties (e.g. [13]). In all these approaches the source "language" and the target "language" should ideally have the same distribution and deep learning methods are used to construct the best fitting between them. Unlike previous studies, in our approach we attempt to tackle the problem where source language and target language have different distributions. It allows creation of a molecule with intended binding affinity using minimum information about the target i.e. amino acid sequence only. As a proof of concept, we investigated several types of end points: chemical feasibility, physical properties, predicted biological activity and achieved promising results in each of them. However, the method can be improved in several directions. One of them is generation of more diverse valid variants per protein. The Diverse Beam Search may be beneficial in this respect as it optimizes objective containing dissimilarity term [58]. However, a more fundamental approach is to couple The Transformer with variational or adversarial autoencoder. These networks can be trained on large datasets of molecule structure to produce a latent continuous representation of chemical space. Joint training The Transformer with such an additional component will allow usage of benefits from both approaches: sampling from continuous representation and conditioning on the target protein. Another important improvement is an increase in the model interpretability. A visualizable interpretation may provide valuable biological insights and substantially improve understanding of the protein-ligand interaction.

Conclusion

In this work we introduced the deep neural network based on the Transformer architecture for protein specific de novo molecule design. Computational experiments demonstrated the efficiency of the proposed method in terms of predicted binding affinity of generated ligands to the target protein, percentages of valid diverse structure, drug-likeness metrics and synthetic accessibility. Our model is based solely on protein sequence. This may be beneficial in the early stage of drug discovery i.e. during identification of lead compound for protein target. The proposed method may

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

be useful if information about 3D protein structure is inaccessible due to difficulties in protein expression, purifying and crystallizing. However, our approach can be extended to yield more interpretable model. We will address this improvement in our future studies.

Data Availability

The code and data are available at
https://github.com/dariagrechishnikova/molecule_structure_generation

References

1 Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature reviews. Drug discovery* **9**, 203-214, doi:10.1038/nrd3078 (2010).

2 Avorn, J. The $2.6 billion pill--methodologic and policy considerations. *The New England journal of medicine* **372**, 1877-1879, doi:10.1056/NEJMp1500848 (2015).

3 Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of computer-aided molecular design* **27**, 675-679, doi:10.1007/s10822-013-9672-4 (2013).

4 High-throughput screening. *Nature biotechnology* **19**, 689-689, doi:10.1038/90317 (2001).

5 Stumpfe, D. & Bajorath, J. Similarity searching. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **1**, 260-282, doi:10.1002/wcms.23 (2011).

6 Hartenfeller, M. *et al.* A collection of robust organic synthesis reactions for in silico molecule design. *Journal of chemical information and modeling* **51**, 3093-3098, doi:10.1021/ci200379p (2011).

7 Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering* **4**, 828-849, doi:10.1039/c9me00039a (2019).

8 Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. (2016).

9 Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., & Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *In Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

10 Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science* **4**, 120-131, doi:10.1021/acscentsci.7b00512 (2018).

11 Gupta, A. *et al.* Generative recurrent networks for de novo drug design. *Molecular informatics* **37**, doi:10.1002/minf.201700111 (2018).

12 Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Molecular informatics* **37**, doi:10.1002/minf.201700153 (2018).

13 Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics* **9**, 48, doi:10.1186/s13321-017-0235-x (2017).

14 Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci Adv* **4**, eaap7885, doi:10.1126/sciadv.aap7885 (2018).

15 Jaques, N., Gu, S., Turner, R. E. & Eck, D. Tuning Recurrent Neural Networks with Reinforcement Learning. *Preprint at https://arxiv.org/abs/1611.02796* (2016).

16 Benjamin, S.-L., Carlos, O., GuimaraesGabriel, L. & Alan, A. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (organic). *Preprint at ChemRxiv https://doi.org/10.26434/chemrxiv.5309668.v3* (2017).

17 Putin, E. *et al.* Reinforced adversarial neural computer for de novo molecular design. *Journal of chemical information and modeling* **58**, 1194-1204, doi:10.1021/acs.jcim.7b00690 (2018).

18 Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C. & Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *Preprint at https://arxiv.org/abs/1705.10843* (2017).

19 Putin, E. *et al.* Adversarial threshold neural computer for molecular de novo design. *Molecular pharmaceutics* **15**, 4386-4397, doi:10.1021/acs.molpharmaceut.7b01137 (2018).

20 Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of molecules via deep reinforcement learning. *Scientific reports* **9**, 10752, doi:10.1038/s41598-019-47148-x (2019).

21 Gomez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **4**, 268-276, doi:10.1021/acscentsci.7b00572 (2018).

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

22 Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. & Chen, H. Application of generative autoencoder in de novo molecular design. *Molecular informatics* **37**, doi:10.1002/minf.201700123 (2018).

23 Kadurin, A. *et al.* The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* **8**, 10883–10890, doi:10.18632/oncotarget.14073 (2017).

24 Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. druGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular pharmaceutics* **14**, 3098-3104, doi:10.1021/acs.molpharmaceut.7b00346 (2017).

25 Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology* **37**, 1038-1040, doi:10.1038/s41587-019-0224-x (2019).

26 Sattarov, B. *et al.* De novo molecular design by combining deep autoencoder recurrent neural networks with generative topographic mapping. *Journal of chemical information and modeling* **59**, 1182-1196, doi:10.1021/acs.jcim.8b00751 (2019).

27 Harel, S. & Radinsky, K. Prototype-based compound discovery using deep generative models. *Molecular pharmaceutics* **15**, 4406-4416, doi:10.1021/acs.molpharmaceut.8b00474 (2018).

28 Skalic, M., Jimenez, J., Sabbadin, D. & De Fabritiis, G. Shape-based generative modeling for de novo drug design. *Journal of chemical information and modeling* **59**, 1205-1214, doi:10.1021/acs.jcim.8b00706 (2019).

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

30 Makhzani, A. S., J. & Jaitly, N. & Goodfellow, I. Adversarial autoencoders. (2015).

31 Lim, J., Ryu, S., Kim, J. W. & Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of cheminformatics* **10**, 31, doi:10.1186/s13321-018-0286-7 (2018).

32 Kang, S. & Cho, K. Conditional molecular design with deep generative models. *Journal of chemical information and modeling* **59**, 43-52, doi:10.1021/acs.jcim.8b00263 (2019).

33 Polykovskiy, D. *et al.* Entangled conditional adversarial autoencoder for de novo drug discovery. *Molecular pharmaceutics* **15**, 4398-4405, doi:10.1021/acs.molpharmaceut.8b00839 (2018).

34 Edunov, S., Ott, M., Auli, M. & Grangier, D. Understanding back-translation at scale. *Preprint at https://arxiv.org/abs/1808.09381*, 489-500, doi:10.18653/v1/D18-1045 (2018).

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

35 Vaswani, A. *et al.* Attention Is All You Need. *Preprint at https://arxiv.org/abs/1706.03762* (2017).

36 Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* **44**, D1045-1053, doi:10.1093/nar/gkv1072 (2016).

37 Vaswani, A. *et al.* Tensor2Tensor for Neural Machine Translation. *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers). Preprint at https://arxiv.org/abs/1803.07416* (2018).

38 RDKit: Open-source cheminformatics *http://www.rdkit.org*.

39 Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of chemical information and modeling* **53**, 1893-1904, doi:10.1021/ci300604z (2013).

40 Rcsb Protein Data Bank. *https://www.rcsb.org/*.

41 PyMOL. *https://pymol.org/2/*.

42 O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of cheminformatics* **3**, 33, doi:10.1186/1758-2946-3-33 (2011).

43 Cohen, E. & Beck, C. Empirical Analysis of Beam Search Performance Degradation in Neural Sequence Models. *Proceedings of the 36th International Conference on Machine Learning* **97**, 1290--1299 (2019).

44 Pochetti, G. *et al.* Structural insight into the stereoselective inhibition of MMP-8 by enantiomeric sulfonamide phosphonates. *Journal of medicinal chemistry* **49**, 923-931, doi:10.1021/jm050787+ (2006).

45 Raisanen, I. T. *et al.* Active matrix metalloproteinase-8 point-of-care (poc)/chairside mouthrinse test vs. Bleeding on probing in diagnosing subclinical periodontitis in adolescents. *Diagnostics* **9**, doi:10.3390/diagnostics9010034 (2019).

46 Usui, T., Okada, M. & Yamawaki, H. Zipper interacting protein kinase (ZIPK): function and signaling. *Apoptosis : an international journal on programmed cell death* **19**, 387-391, doi:10.1007/s10495-013-0934-3 (2014).

47 Elbadawy, M., Usui, T., Yamawaki, H. & Sasaki, K. Novel functions of death-associated protein kinases through mitogen-activated protein kinase-related signals. *International journal of molecular sciences* **19**, doi:10.3390/ijms19103031 (2018).

48 Chen, C. *et al.* Discovery of a novel inhibitor of the protein tyrosine phosphatase shp2. *Scientific reports* **5**, 17626, doi:10.1038/srep17626 (2015).

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

49 Hill, K. S. *et al.* PTPN11 plays oncogenic roles and is a therapeutic target for braf wild-type melanomas. *Molecular cancer research : MCR* **17**, 583-593, doi:10.1158/1541-7786.MCR-18-0777 (2019).

50 Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **46**, 3-26 (2001).

51 Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem.* **45**, 2615-2623, doi:10.1021/jm020017n (2002).

52 Hitchcock, S. A. & Pennington, L. D. Structure-brain exposure relationships. *Journal of medicinal chemistry* **49**, 7559-7583, doi:10.1021/jm060642i (2006).

53 Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature chemistry* **4**, 90-98, doi:10.1038/nchem.1243 (2012).

54 Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics* **1**, 8, doi:10.1186/1758-2946-1-8 (2009).

55 Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry* **39**, 2887-2893, doi:10.1021/jm9602928 (1996).

56 Schwaller, P. *et al.* Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **5**, 1572-1583, doi:10.1021/acscentsci.9b00576 (2019).

57 Karpov, P., Godin, G. & Tetko, I. V. A transformer model for retrosynthesis. *Preprint at ChemRxiv https://doi.org/10.26434/chemrxiv.8058464.v1* **11731**, 817-830, doi:10.1007/978-3-030-30493-5_78 (2019).

58 Vijayakumar, A. K. *et al.* Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *Preprint at https://arxiv.org/abs/1610.02424* (2016).

Acknowledgements

The author would like to thank Leonid Grechishnikov for his expertise, help in code creation, in the visualization of the figures and support. The author thanks Irina Bannikova, Irina Volodina, Sergey Pavlishchev for fruitful discussions and support.

bioRxiv preprint first posted online Dec. 3, 2019; doi: http://dx.doi.org/10.1101/863415. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Author contributions statement

DG conceived the presented idea, performed the computations, analyzed the results and wrote the manuscript.

Additional Information

Competing Interests: The author declares no competing interests.