Communications in Information and Systems
Volume 19, Number 3, 241–277, 2019

Generative network complex (GNC) for drug discovery

Christopher Grow, Kaifu Gao,
Duc Duy Nguyen*, and Guo-Wei Wei†

It remains a challenging task to generate a vast variety of novel compounds with desirable pharmacological properties. In this work, a generative network complex (GNC) is proposed as a new platform for designing novel compounds, predicting their physical and chemical properties, and selecting potential drug candidates that fulfill various druggable criteria such as binding affinity, solubility, partition coefficient, etc. We combine a SMILES string generator, which consists of an encoder, a drug-property controlled or regulated latent space, and a decoder, with verification deep neural networks, a target-specific three-dimensional (3D) pose generator, and mathematical deep learning networks to generate new compounds, predict their drug properties, construct 3D poses associated with target proteins, and reevaluate druggability, respectively. New compounds were generated in the latent space by either randomized output, controlled output, or optimized output. In our demonstration, 2.08 million and 2.8 million novel compounds are generated respectively for Cathepsin S and BACE targets. These new compounds are very different from the seeds and cover a larger chemical space. For potentially active compounds, their 3D poses are generated using a state-of-the-art method. The resulting 3D complexes are further evaluated for druggability by a championing deep learning algorithm based on algebraic topology, differential geometry, and algebraic graph theories. Performed on supercomputers, the whole process took less than one week. Therefore, our GNC is an efficient new paradigm for discovering new drug candidates.

1. Introduction

Drug design and discovery ultimately test our understanding of biological sciences, the status of biotechnology, and the maturity of computational sci-

*Research supported in part by Bristol-Myers Squibb and Pfizer.
†Corresponding author. Research supported in part by NSF Grants DMS-1721024, DMS-1761320, and IIS1900473, NIH grant GM126189, Bristol-Myers Squibb and Pfizer.

ences and mathematics. Technically, drug discovery involves target discovery, lead discovery, lead optimization, preclinical development, three phases of clinical trials, and finally, launching to market only if everything goes well. Among them, lead discovery, lead optimization, and preclinical development disqualify tens of thousands of molecules based on their binding affinities, solubilities, partition coefficients, clearances, permeabilities, toxicities, pharmacokinetics, etc., leaving only about ten compounds for clinical trails. Currently, drug discovery is both expensive and time-consuming. It takes about $2.6 billion dollars and more than ten years, on average, to bring a new drug to the market [1]. Reducing the cost and speeding up the drug discovery process are crucial issues for the pharmaceutical industry. Much effort has been taken to optimize key steps of the drug discovery pipeline. For example, the development of high-throughput screening (HTS) has led to an unprecedented increase in the number of potential targets and leads [2]. HTS is able to quickly conduct millions of tests to rapidly identify active compounds of interest using compound libraries [3].

While there has been an increase in the number of potential targets and leads, the number of new molecular entities generated has remained stable because of a high attrition rate during preclinical development and clinical phases, caused by the selection of leads with inappropriate physicochemical or pharmacological properties [4, 5]. Rational drug design (RDD) approaches are proposed to better identify candidates with the highest probability of success [6]. RDD aims at finding new medications based on the knowledge of biologically druggable targets [1, 7]. Several empirical metrics, such as Lipinski's rule of five (RO5) [8], were established for estimating druglikeness, which describes the druggability of a substance with respect to factors like bioavailability, solubility, toxicity, etc. Generally, the early selection of candidates requires the design of molecules complementary in shape and charge to the target of interest, which leads to a high binding affinity. Additionally, the determination of the nature and rates of physical/chemical/biological processes that are involved in the absorption, distribution, metabolism, and elimination (ADME) of drug candidates are also of primary importance. ADME profiling and prediction are mostly dependent on molecular descriptors such as RO5 [9]. Furthermore, cellular/animal disease models are typically used during lead optimization to measure various pharmacokinetics. Finally, toxicity study is a primary task for preclinical development.

Recently, computer-aided drug design (CADD) has emerged as a useful approach in reducing the cost and period of drug discovery [10]. Computational techniques have been developed for both virtual screening (VS) and optimizing the ADME properties of lead compounds. Essentially, these

methods are designed as in silico filters to eliminate compounds with undesirable properties. These filters are widely applied for the assembly of compound libraries using combinatorial chemistry [11]. The integration of early ADME profiling of lead chemicals has contributed to the speed-up of lead selection for phase-I trials without large amounts of revenue loss [12]. Currently, compounds are added in libraries on the basis of target-focused design or diversity considerations [13]. VS and HTS can screen compound libraries to select a subset of compounds whose properties are in agreement with various criteria [14].

Despite these efforts, the current size of databases of chemical compounds remains small when compared with the chemical space spanned by all possible energetically stable stoichiometric combinations of atoms and topologies in molecules. Considering these factors, it is estimated that there are 10^{60} distinct molecules. Among them, 10^{30} are druglike [3]. As a result, computational techniques are also being developed for the de novo design of druglike molecules [15] and for generating large virtual chemical libraries, which can be more efficiently screened for in silico drug discovery.

Among the computational techniques available, deep neural networks (DNN) have gained much interest for their ability to extract features and learn physical principles from training data. Currently, DNN-based architectures have been successfully developed for applications in a wide variety of fields in the biological and biomedical sciences [16, 17].

More interestingly, several deep generative models based on variational autoencoders (VAEs) [18], adversarial autoencoders (AAEs) [19], recurrent neural networks (RNNs) [20], long short term memory networks (LSTMs) [21] and generative adversarial networks (GANs) [22] have been proposed for exploring the vast druglike chemical space. A policy-based reinforcement learning approach was proposed to tune RNNs for episodic tasks [23, 24]. A VAE was used by Gomez-Bombarelli et al. [25] to encode a molecule in the continuous latent space for exploring associated properties. The usage of these models has been extended to generate molecules with desired properties [26]. Miha Skalic et al. [27] combined a conditional variational autoencoder and a captioning network to generate previously unseen compounds from input voxelized molecular representations. Artur Kadurin et al. [28] built an AAE to generate new compounds. Boris Sattarov et al. [29] combined deep autoencoder RNNs with generative topographic mapping to carry out de novo molecular design.

It is particularly interesting and important to generate potential drug candidates for specific drug targets. To this end, a network complex is required to fulfill various functions, including target-specific molecular generation, target-specific binding affinity ranking, and solubility and partition

coefficient evaluation. In this work, we propose a generative network complex (GNC) to combine drug-property controlled or regulated autoencoder (AE) models and DNN predictors to generate millions of new molecules and select potential drug candidates that have appropriate druggable properties. Our GNC includes the following components:

1) Using known molecules in a target-specific training set as seeds, a SMILES string generator is constructed to generate millions of novel compounds. This generator consists of a CNN-based encoder, a drug-property controlled or regulated latent space, and a LSTM-based decoder.
2) A pre-trained multitask DNN model is constructed to select drug candidates based on druggable properties.
3) A 3D structure generator, MathPose, to convert selected 2D SMILES strings into 3D structures based on target receipt information.
4) A 3D multitask druggable property predictor, mathematical deep learning (MathDL), to further select new drug candidates via various druggable criteria.

Some of these components, namely MathPose and MathDL, have been extensively validated in blind settings [30, 31]. Our GNC can not only generate new molecules, but also construct or pick up the molecules with ideal drug properties. This makes it a very promising method for generating millions of new drug candidates in silico in a very short time period.

2. Methods

2.1. The structure of generative network complex (GNC)

In the proposed GNC, the first component is a generative network including encoder, drug-property regulated latent space, and decoder models. The generative network will take a given SMILES string as input to generate a novel one. The newly generated SMILES strings will be fed into the second component of our GNC, a 2D fingerprint-based deep neural network (2DFP-DNN), so that only ones with desired druggable properties are kept. The next component is the MathPose model which is used to predict the 3D structure information of the compounds selected by 2DFP-DNN. The bioactivities of those compounds are again estimated by the structure-based deep learning model named MathDL. The druggable properties predicted by this last component of our GNC are used as an indicator to select the promising drug candidates. The outline of the GNC is illustrated in Figure 1.

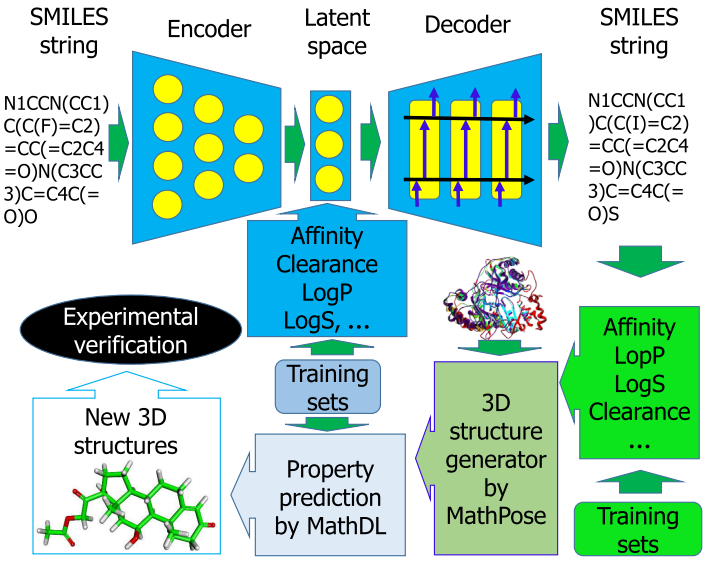

Figure 1: A schematic illustration of a generative network complex. It consists of an autoencoder that takes SMILES strings (SS) into a drug-property regulated latent space, a regulated latent space, a LSTM-based autodecoder, a multitask network for the evaluation of binding affinity, partition coefficient (LogP), solubility (LogS), clearance, etc., a 3D structure generator named MathPose, and MathDL, a refined 3D multitask druggable property predictor based on algebraic topology, differential geometry, and graph theory, to select new drug candidate structures.

2.1.1. Autoencoder An autoencoder is a type of artificial neural network used to encode a set of data into vectors in the latent space. An autoencoder is typically combined with a decoder to transform the encoded vectors back into SMILES strings. In the present work, we propose a latent space technique which controls or regulates various drug properties, such as binding affinity, solubility (LogS), partition coefficient (LogP), clearance, etc.

Encoder The encoder network in the present work is a convolutional neural network (CNN) which takes converts SMILES strings into 3D molecular images before encoding their into the latent space. It consists of five 3D-convolutional layers. The number of output channels for each layer are 32, 32, 64, 64, 32, 32, 32, and 32 with kernel sizes all (1, 1, 1), respectively. For the sake of visualization, the encoder's architecture is outlined in Figure 2.

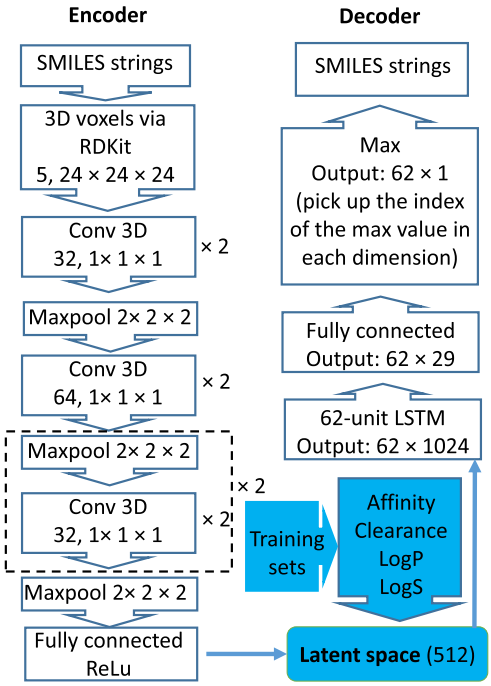

Figure 2: Illustration of an autoencoder, which consists of a CNN-based encoder, a regulated latent space, and a LSTM-based decoder.

In the present encoder, for each SMILES string, 3D conformers were generated via RDKit [32] and optimized using the MMFF94 force field [33] with default settings. Molecule atoms were then voxelized into a discretized 1 Å cubic grid with sides of length 24 Å prior to a random rotation and 2 Å translation of the molecule. The voxelized value is determined by its atom type and the distance $\mathbf{r}$ between neighboring atoms and its center:

$$n(\mathbf{r}) = 1 - \exp[-(r_{\mathrm{vdW}}/\mathbf{r})^{12}], \tag{1}$$

where r_{vdW} is the van der Waals radius.

The voxelized values of five types of properties are calculated: hydrophobic, aromatic, H-bond donors, H-bond acceptors, and heavy atoms, leading to five different channels [27].

Latent space We propose three latent space regulation schemes, i.e., randomized output, controlled output, and optimized output, to construct new

compounds. First, to generate new compounds from seeds, random noise can be added to the latent space. In other words, the encoded latent vectors can be perturbed by standard Gaussian noise, rendering a possible new latent representation. The resulting latent vector will be fed into the decoder network.

Additionally, a more interesting control procedure is to select the latent space output through a druggable property assessment. As shown in Figures 1 and 2, we use the trained encoder to generate latent-space representations of a dataset of interest, such as the BACE dataset. Based on these representations of the dataset and its labels, we train deep learning network models to evaluate and predict various druggable properties, including binding affinities, solubility, partition coefficient, clearance, toxicity, etc. In certain situations, we also build multitask deep learning models to enhance latent-space evaluations. In this approach, each new compound in its latent space representation is evaluated for its druggable properties to determine whether it is to be fed into the decoder.

Finally, a more effective optimization scheme is to actively build new drug candidates in the latent space representation with desirable properties as shown in Figures 1 and 2. With appropriate training datasets, we first construct m latent-space predictive machine learning models as described above for m different properties, such as binding affinities, solubility, partition coefficient, clearance, etc. For each property, we set up a target value, y_{j0}. We then build an L_2 loss function to optimize a given n-component latent-space vector $X \in \mathbb{R}^n$:

$$\min \sum_{j=1}^{m} k_j (\hat{y}_j(X) - y_{j0})^2, \tag{2}$$

where k_j is a preselected weight coefficient for the jth property and $\hat{y}_j(X) : \mathbb{R}^n \to \mathbb{R}$ is the predicted jth property value of the latent-space vector X from latent-space machine learning models. Alternatively, we also use other metrics, such as L_1 or mixed metrics for constructing the loss function. The optimization with a gradient decent algorithm leads to an iterative scheme for regularizing the latent-space vector X. Alternatively, a Monte Carlo procedure can be implemented.

Target values y_{j0} can be chosen to optimize potential drugs. In case of binding affinity (BA), we use a targeted value of $y_{\mathrm{BA}} \leq -9.6$ kcal/mol. For LogP, we set $y_{\mathrm{LogP}} \leq 5$. Note that additionally constraints, such as, similarity, Lipinski's rule of five [8] or their variants for druglikeness can be easily implemented with Eq. (2).

Decoder The decoder network here consists of several LSTMs. LSTMs are variants of RNNs that were proposed to handle language processing problems, which require the network to take into account the relationships between words rather than simply interpreting each word independently. RNNs are designed to pass fixed-size pieces of information from one neuron to others in the network. However, RNNs are not very effective at processing information with long-term dependencies, as the persistence of information within the network is somewhat short-lived. As a result, LSTMs were designed to overcome this problem [21, 34, 35]. The encoder network was trained in the shape encoder framework.

In each LSTM unit, there is a cell consisting of an input gate, an output gate, and a forget gate which are described in the following equations

$$H_t = o_t * \tanh(C_t), \tag{3}$$

where o_t depends on its input X_t and the output of the last layer H_{t-1}:

$$o_t = \sigma(W_o[H_{t-1}, X_t] + b_o). \tag{4}$$

Here, σ is the activation function. Now, the cell state at the ith layer is given by:

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t, \tag{5}$$

where $\hat{C}_t$ is the change of the cell state at the ith layer

$$\hat{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C), \tag{6}$$

and f_t and i_t are given by the following,

$$f_t = \sigma(W_f[H_{t-1}, X_t] + b_f) \tag{7}$$

$$i_t = \sigma(W_i[H_{t-1}, X_t] + b_i). \tag{8}$$

The updated cell state, C_t, is then passed to the next layer along with the output of the current layer and includes accumulated information from all previous layers so that the network can handle long-term dependencies between inputs.

The purpose of LSTMs in our case is to decode molecules from the encoded vectors in the latent space. There is some variation in the decoding process via the use of probabilistic sampling. Due to the LSTM's ability to handle long-term dependencies, it can learn SMILES grammar, and build

SMILES strings by selecting the next token proportionally to its predicted probability [27]. This means that some variation from the seed SMILES string will occur. As a result, even when the input is the same, the output will not always be the same. This causes the generated SMILES strings to be different from their seeds. Our LSTM decoder has 5 layers as the same as the number of layers of the aforementioned CNN encoder. The architecture of the decoder is depicted in Figure 2. The Adam optimizer was applied with the learning rate 0.001 and a batch size of 128 to minimize the following loss

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_j^{(i)} \log p_j^{(i)}, \tag{9}$$

where $y_j^{(i)}$ and $p_j^{(i)}$, respectively, represent the the ground-truth and the predicted probability for component jth in the ith SMILES string. Also, N is the number of samples in each batch and M is the length of the SMILES string.

2.1.2. 2D fingerprint-based binding affinity predictors (2DFP-DNN) The predictors are deep neural networks (DNN) pre-trained on our own training sets. A DNN mimics the learning process of a biological brain by constructing a wide and deep architecture of numerous connected neuron units. A typical DNN often includes multiple hidden layers. In each layer, there are hundreds or even thousands of neurons. During the learning stage, weights on each layer are updated by backpropagation. A deep neural network is able to construct hierarchical features and model complex nonlinear relationships.

The purpose of our DNN predictors is to predict the binding affinities and other properties of the generated compounds and, based on that, screen ideal drug candidates meeting our criteria. Binding affinity assesses a drug's binding strength to its target, which is one of the most important drug properties [36, 37]. The input of predictor networks is 2D molecular fingerprints. In our case, a combination of ECFP [38] and MACCS [39] fingerprints were used, yielding 2214 bits of features (2048 bits from ECFP and 166 bits from MACCS) in total. The output of the network is the drug properties, such as binding affinity, log P, and log S. During the training and prediction processes, the SMILES strings of compounds were first transformed to their 2D fingerprints and then fed into the network. The fingerprint transformation from SMILES strings was conducted by RDKit [32].

With appropriate training data, we can construct multitask DNNs for simultaneous predictions of binding affinity, log P, log S, and toxicity [40, 41].

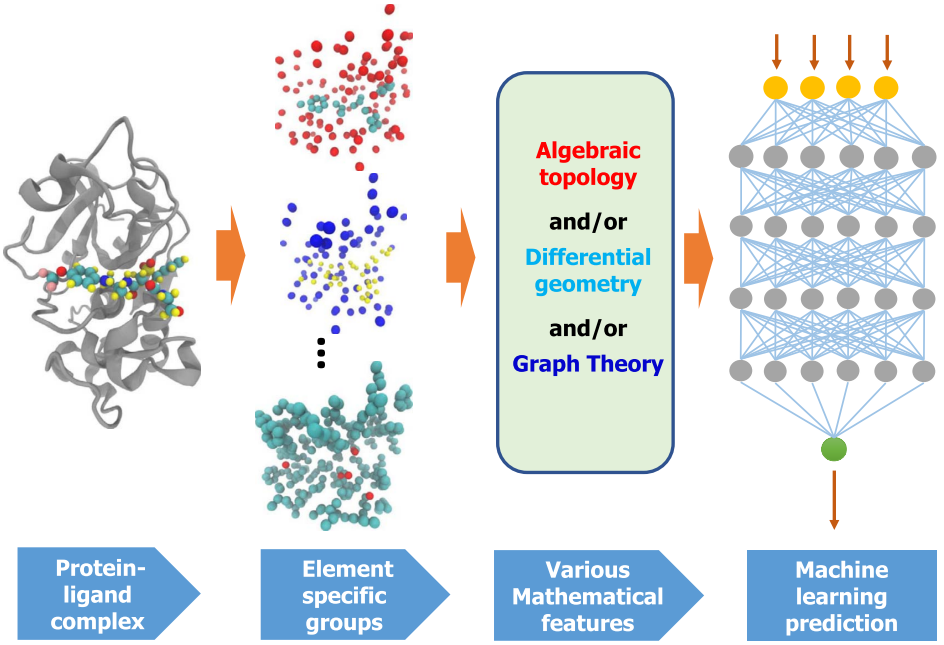

Figure 3: A schematic illustration of the MathDL for binding affinity prediction in which the combination of several advanced mathematical representations is integrated with sophisticated CNN models.

Our DNN predictor networks have 4 layers with 3000, 2000, 1000, and 500 neurons in each hidden layer, respectively. For training, we used stochastic gradient descent with a momentum of 0.5. We trained each network for 2000 epochs with a mini-batch size of 4. We used a learning rate of 0.01 for the first 1000 epochs and reduced it to 0.001 for the last 1000 epochs. Our tests indicate that adding dropout or L_2 decay does not necessarily increase the accuracy of the networks, and as a consequence, we omitted these two techniques. The DNN training and prediction are performed by Pytorch [42].

2.1.3. MathDL for energy prediction Our MathDL is constructed by the integration of mathematical representation features and deep learning networks to generate a powerful binding affinity predictor [30, 31]. Specifically, the MathDL is the blend of intensively validated models based on algebraic topology [43, 44, 45], differential geometry [46], and graph theory [47, 48]. In these methods, algebraic topology model makes use of persistent homology in multi-component and multi-level manners to character-

ize protein-ligand complexes by topological invariants, i.e., Betti numbers counting various dimensional holes. In the 3D space, we have Betti-0, Betti-1, and Bett-2 which receptively counts the numbers of independent components, recognizes numbers of rings, and accounts for the cavity information [49, 50, 51]. Our previous work, we have shown that algebraic topology network has outperformed other state-of-the-art methods in the classifying proteins [43] and active/inactive compounds [44], and the predictions of protein-ligand binding affinity [52, 44], toxicity [41], lop P, and log S [40].

Differential geometry describes how molecules assume complex structures, intricate shapes and convoluted interfaces between different parts [53]. In our differential geometry-based model, essential chemistry, physical, and biological information are encoded into the low-dimensional interactive manifolds which are extracted from high-dimensional data space via a multiscale discrete-to-continuum mapping [54, 46]. Thereby, the molecular structures and atomic interactions can be conveniently represented via interactive curvatures, interactive areas, etc. Numerous numerical validations have shown that the differential geometry model has achieved the state-of-the-art performances on various biological prediction tasks, namely drug toxicity, molecular solvation, and protein-ligand binding affinity [46].

Recently, we have developed a powerful algebraic graph-based scoring function which encodes the important physical and biological properties such as hydrogen bonds, hydrophilicity, hydrophobicity, van der Waals interactions, and electrostatics from the high-dimension space into the low-dimension description via the invariants extracted from Laplacian, its pseudo-inverse, and adjacency matrices [48]. Algebraic graph theory-based models have been widely utilized in the study of physical modeling and molecular analysis such as chemical analysis [55, 56], protein flexibility analysis [57, 58, 59]. Despite its popularity, the graph-based quantitative models typically are not as competitive as other quantitative approaches due to no categorization on element types and the missing crucial non-covalent interactions. This missing information has been encoded in multiscale weighted colored subgraphs in our newly designed algebraic graph-based model, named AGL-Score. Extensive numerical validation on PDBbind benchmarks with various evaluation metrics, namely scoring power, ranking power, docking power, and screening power has shown that our AGL-Score has outperformed other state-of-the-art methods on these evaluations which are the standard criteria for virtual screening in drug discovery [48].

The combination of these three powerful models gives rise to the MathDL model which is expected to be one of the most accurate binding affinity predictors available in the literature. Indeed, the MathDL model achieved

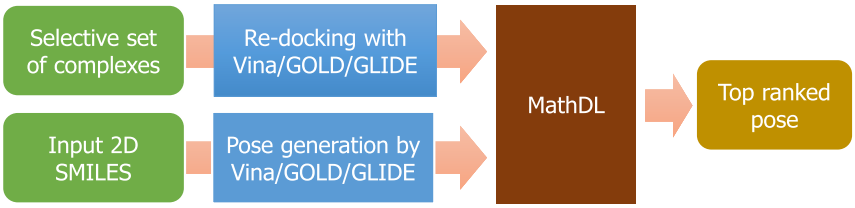

Figure 4: A schematic illustration of the MathPose approach for 3D structure generation from a given input 2D SMILES string.

the top performances on the affinity ranking and free energy prediction for Cathepsin S (CatS) inhibitors in the Drug Design Data Resource (D3R) Grand Challenge 4 (GC4), a worldwide competition series in the computer-aided drug design [31]. Also, MathDL model was the competitive scoring functions on the binding energy predictions for beta-secretase 1 (BACE) compounds in GC4 [31]. The outline of the MathDL model is depicted in Figure 3.

2.1.4. MathPose for 3D structure prediction In our recent work, we have successfully designed an AGL-Score model to achieve the best performances in docking power metrics which validate the scoring function's ability to identify the "native pose" from the computer-generated poses [48]. Specifically, on the CASF-2007 benchmark, [60] our AGL-Score achieves 84% accuracy on the docking power assessment [48]. The second best scoring function on this benchmark is from GOLD software with ASP fitness score (82%) [60]. Our scoring function is still the top performer on docking power test of CASF-2013 benchmark [61] with the accuracy as high as 90% [48], followed by the machine learning based-scoring function $\Delta_{\text{vina}}\text{RF}_{20}$ (87%) [62]. With such promising results, it is expected that the replacement of the single AGL-Score model by intricate MathDL scoring function will certainly improve the quality of pose ranking. This results in the MathPose model whose framework is outlined in Figure 4. In our MathPose, besides the SMILES string of the interested ligand L, we select a set of complexes having similar binding sites to the one the ligand L can bind to. A pool of nearly 1000 poses for the ligand L is generated by several common docking software, namely Autodock Vina [63], GOLD [64], and GLIDE [65]. Additionally, three docking software packages are utilized to re-dock the complexes in the selective data set to form at least 100 decoy complexes per input. Then, our MathDL will be trained on these decoy sets to learn the calculated root mean squared deviation (RMSD) between the decoy and native

structures. The trained MathDL will be applied to pick up the top-ranked pose for the given ligand L.

2.2. The analysis of generated compounds

The 2D similarity analysis between generated compounds and their seeds To investigate how "novel" our generated compounds are from their seeds, a similarity analysis was performed on them. The 2D molecular SMILES strings of the generated molecules were also transformed into 2D molecular fingerprints and then the similarity scores between the fingerprints of the generated molecules and their seeds were calculated. The fingerprints were the same ones used in the DNN predictors, a combination of ECFP and MACCS molecular fingerprints. The criteria used for the similarity scores was the Tanimoto coefficient [66]. The fingerprint transformation was also conducted by RDKit [32].

The k-means clustering analysis of generated compounds Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It has already been widely applied to protein conformation analysis [67, 68, 69]. To present the diversity of our generated active compounds, k-means clustering analysis was performed. The input features were the same molecular fingerprints discussed above, and the k-means clustering was conducted by scikit-learn [70]. For each cluster, the center was extracted to represent the cluster.

3. Results

To examine and validate the performance of our proposed GNC for generating new compounds for drug targets, we consider two specific targets, namely Cathepsin S (CatS) set and Beta-Secretase 1 (BACE). These two targets appeared in the D3R Grand Challenges, worldwide competition series in computer-aided drug design [71, 31], with components addressing pose-prediction, affinity ranking, and free energy calculations.

Both CatS and BACE are potential targets for significant human diseases. CatS constitutes an 11-member family of proteases involved in protein degradation. It is highly expressed in antigen-presenting cells, where it degrades major histocompatibility complex class II (MHC II)-associated invariant chain. CatS is a candidate target for regulating immune hyperresponsiveness, as the inhibition of CatS may limit antigen presentation

[72, 73]. BACE is a transmembrane aspartic-acid protease human protein encoded by the BACE1 gene. It is essential for the generation of beta-amyloid peptide in neural tissue [74], a component of amyloid plaques widely believed to be critical in the development of Alzheimer's, rendering BACE an attractive therapeutic target for this devastating disease [75]. The rest of this section is devoted to the utilization of the proposed GNC on the exploration of new potential drugs for CatS and BACE targets.

3.0.1. Faithful validation of generative network on CatS and ZINC data sets To assess the performance of the autoencoder on the CatS data set, we converted all SMILES strings in the data set into the canonical form using RDKit [32], and kept only the strings with length no more than 60. This was done because the decoder network was designed to produce only SMILES strings of length at most 60. This left us with 1858 of the 2847 molecules in the CatS training set. After feeding these molecules through the network, 1427 (76.8%) yielded valid SMILES strings, with none being identical to the original.

We also tested the performance of the autoencoder on a larger data set of 1 million molecules, randomly chosen from the same subset of the ZINC 15 [76] data set from which the training samples were drawn. The training set produced from the ZINC 15 data set contains 192,813,983 molecules, 26,880,000 of which were previously seen by the autoencoder during training. From these 1 million molecules, 994,219 (99.4%) yielded valid SMILES strings, and 2,724 (0.27%) SMILES strings were reproduced exactly. A high valid molecule generation rate and a low reconstruction rate enable us to generate meaningful compounds with highly diverse chemical properties.

3.1. BACE

3.1.1. Data preparation To enable the proposed GNC to generate meaningful BACE inhibitors, one needs to supply it with seed molecules closely related to the BACE target. To this end, we combine all BACE inhibitors provided in the D3R Grand Challenge 4 (https://drugdesigndata.org/about/grand-challenge-4) with the BACE ligands having the reported binding affinity on the ChemBL database (https://www.ebi.ac.uk/chembl/). That results in a BACE data set of 3916 compounds with binding affinities ranging from −2.84 kcal/mol to −13.22 kcal/mol. If one sets −9.56 kcal/mol as a threshold to label a compound as active, that BACE data set has 1231 active ligands. The distribution of binding affinity in the BACE data set is shown in Figure 5a. That figure reveals that most of the molecules in our

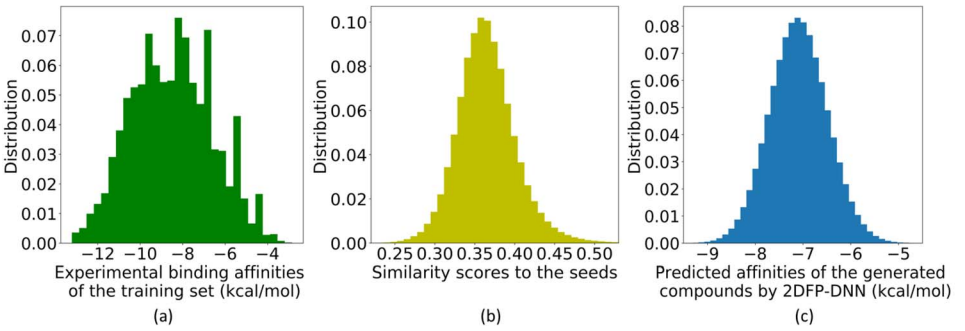

Figure 5: Distributions on the BACE set. (a) The distribution of the experimental binding affinities in the BACE training set; (b) The similarity distribution of new molecules compared with their seeds in the BACE set. (c) The distribution of new BACE molecules' binding affinities predicted by 2D fingerprint network model 2DFP-DNN.

collected data set having affinities between −10 kcal/mol and −7 kcal/mol. Also, there are more BACE inhibitors with binding strength less than −10 kcal/mol than ones having binding affinity higher than −6 kcal/mol.

3.1.2. Structure generation By feeding the BACE data set of 3916 compounds to the generative network, as many as 2.8 million valid compounds were generated by supercomputers in less than one week. To indicate how "novel" our generated compounds are from their seeds, the similarity score between each generated compound and its seed is calculated. The similarity score distribution is illustrated in Figure 5b. It is revealed from the figure that the similarity scores of the generated compounds have a broad range varying from 0.15 to 1.00. This means that our generated compounds cover a very large chemical space. A similarity score being 1 indicates the generated compound is exactly the same as the seed. Fortunately, this is very rare, happening only 9 times in all 2,727,379 generated compounds. In most cases, the similarity scores are very low with an average value of 0.34, implying the wide range of diversity among the generated samples.

To further verify that the generated compounds are really unique from the seeds, a seed molecule and several generated compounds are shown in Figure 6. In which, Figure 6a depicts a seed, Figure 6b illustrates the most similar compound generated from the seed, Figure 6c plots a compound with a medium similarity score of 0.34, and Figure 6d presents the most different one. One can realize that even the most similar one with a similarity

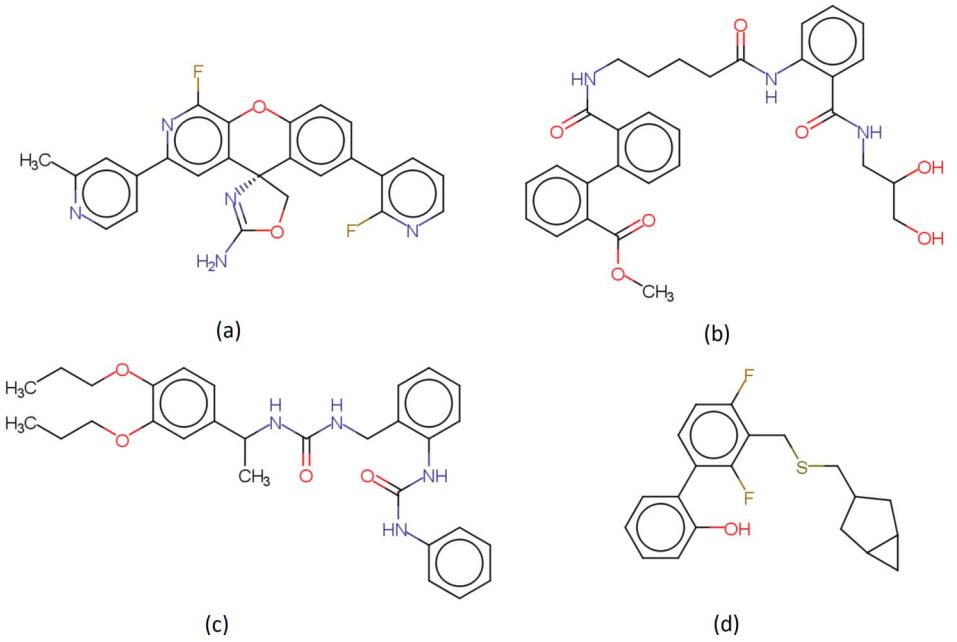

Figure 6: The illustration of similarity between a seed molecule in the BACE set and some generated compounds: (a) The seed; (b) The most similar compound generated from the seed (similarity score=0.50); (c) A compound with a medium similarity score of 0.34; (d) The most different one from the seed (similarity score=0.23).

score as high as 0.50, chemical structures are still quite different due to the replacement of the fused ring by a carbon chain (see Figures 6a and 6b).

3.1.3. Binding affinity screening by 2DFP-DNN To efficiently select the potential drug candidates, we carry out the 2D fingerprint DNN model discussed in Section 2.1.2 to predict the binding affinities of more than 2.7 millions compounds. Figure 5 illustrates those predicted energies. From Figure 5, one can notice that the predicted affinities of the generated BACE compounds are distributed in a Gaussian manner. This result is probably due to the Gaussian-like distribution of similarity scores between generated ones and their corresponding seeds depicted in Figure 5b.

The range of their binding affinities of predicted molecules is widely spread from −3.89 kcal/mol to −10.20 kcal/mol, confirming that large chemical space is covered. The peak is at −7.1 kcal/mol, which means about half

of the generated compounds have binding affinity smaller than −7.1 kcal/mol. Among this first half with the binding affinity smaller than −7.1 kcal/mol, 5 compounds have predicted binding affinity smaller than −10 kcal/mol which indicates they are promising drug candidates. Moreover, there are 2130 compounds with binding affinity smaller than −9 kcal/mol, and 178250 compounds with the binding affinities smaller than −8 kcal/mol. In this work, we use a common binding affinity threshold, i.e. −9.56 kcal/mol, to screen out high-likely less active compounds. As a result, we are left with 99 generated inhibitors having the lowest binding energy in term of kcal/mol.

It is noticed that the 2D fingerprint DNN model for binding affinity prediction only relies on the ligand information without the involvement of target proteins. Therefore, its accuracy is not as high as its 3D counterparts (e.g. MathDL) [44, 31] in which the interactions between the target binding site and the interesting compounds are fully incorporated. However, it is projected to be time consuming when carrying out those 3D-based binding affinity predictor models on a large pool of molecules. Thus, in this work, we make use of the advantage of simple calculations in the 2D-based models to filter out a large number of compounds with highly predicted affinities.

3.1.4. Clustering analysis of selected compounds To illustrate how diverse our generated active compounds are, clustering analysis was performed on the 99 generated compounds with the most highly predicted binding affinities discussed in Section 3.1.3. By carrying out k-means clustering method, one can find 6 clusters in our generated set, and the center of each cluster is shown in Figure 7.

Statistically, the sizes of these 6 clusters are 7, 38, 10, 7, 12, and 25, respectively. Inside these 6 clusters, the average similarity scores to the centers are 0.69, 0.58, 0.62, 0.66, 0.63, and 0.67, respectively, which indicates the compounds in the same cluster are relatively similar. By contrast, the similarity scores between different clusters are much lower. Specifically, the similarity score between these 6 cluster centers are only around 0.40; thereby, implying the high diversity in our generated compounds.

Among these 6 clusters, cluster 2 is the biggest one with 38 compounds. Moreover, it contains the largest numbers of the highly predicted binding affinities. Particularly, cluster 2 has 5 compounds with predicted binding affinities smaller than −10 kcal/mol. Since the compounds in the same cluster are similar, it suggests that other compounds in cluster 2 may also have a high potential to become drugs. SMILES strings of all 99 compounds in 6 clusters are included in Table S1 in Supporting Information.

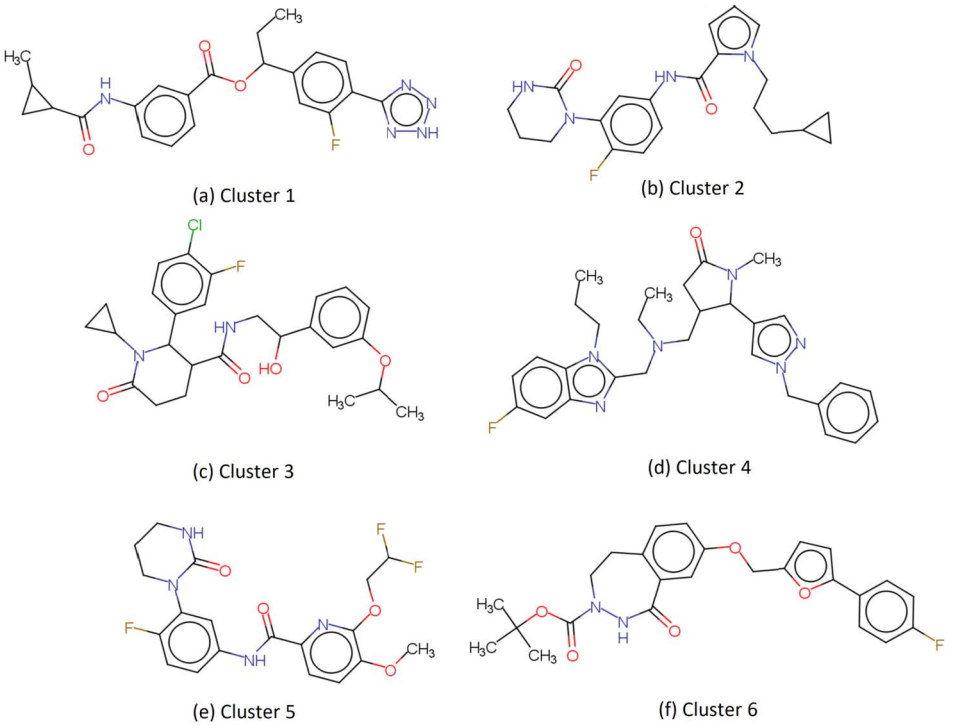

Figure 7: The center of the 6 clusters found in our BACE generated set.

3.1.5. Binding affinity screening by MathDL The DNN model using the 2D fingerprint features, as discussed in Section 3.1.3, only relies on the ligand information and lacks the receptor environment. As a result its reliability is not guaranteed when identifying the most promising drug candidates. It has been shown that structure-based models often outperform the ligand-based models in diverse datasets [44, 30, 31]. Therefore, our MathDL model, discussed in Section 2.1.3, is utilized to re-rank the compounds picked out by 2DFP-DNN models. The MathDL model is trained on the BACE data set of 3916 compounds whose 3D structures are generated by MathPose mentioned in Section 2.1.4.

The Kendall's Tau coefficient (τ) and Pearson correlation coefficient (R_p) of the cross-validation on the training data are 0.608 and 0.797, respectively. These accuracy evaluations guaranteed a well-trained MathDL model on that specific training set. A generated compound set of 99 molecules are fed into MathPose to obtain 3D structures provided in File S1 in Supporting Information. All of them were docked to the protein extracted from a

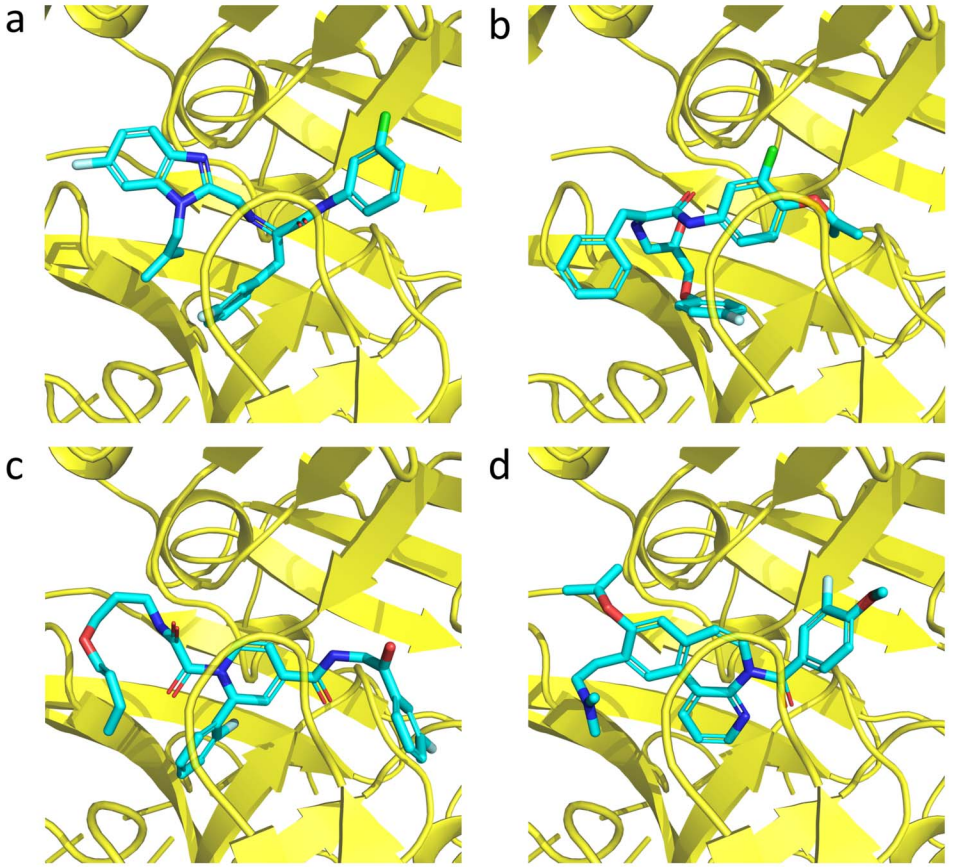

Figure 8: Top 4 generated BACE compounds having the lowest binding affinities predicted by MathDL model. Their 3D structures were constructed by MathPose. Their IDs under our naming system and the predicted energies are, respectively (a) BACE_gen_35 (−8.263 kcal/mol); (b) BACE_gen_66 (−8.258 kcal/mol); (c) BACE_gen_29 (−8.202 kcal/mol); and (d) BACE_gen_25 (−8.20 kcal/mol). Their SMILES strings are provided in Table S1, and their corresponding 3D structures are included in File S1. All of those molecules were docked to protein extracted from the complex with PDB ID 3dv5.

complex with PDB ID 3dv5. Their binding affinities are, then, predicted by the aforementioned trained MathDL model. It is noticed that binding affinity values of 99 generated molecules predicted by MathDL are higher than ones estimated by the 2D fingerprint DNN approach in term of kcal/mol.

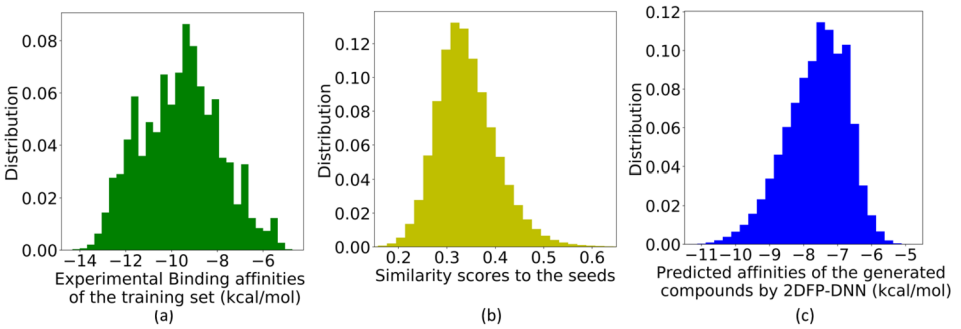

Figure 9: The three distributions about the CatS set. (a) The distribution of experimental binding affinity in the CatS data set; (b) The distribution of similarity scores to their seeds in the CatS generated set. (c) he distribution of the CatS generated set's binding affinities predicted by 2D fingerprint network model 2DFP-DNN.

Specifically, based on MathDL predictor, the lowest binding energy is −8.263 kcal/mol, the highest energy is −5.972 kcal/mol, and the averaged energy over 99 compounds is −7.33 kcal/mol. Figure 8 illustrates the binding poses of top four ligands, namely BACE_gen_35, BACE_gen_66, BACE_gen_29, and BACE_gen_25, in term of affinity. The predicted energies of those top 4 molecules are −8.263 kcal/mol, −8.258 kcal/mol, −8.202 kcal/mol, and −8.20 kcal/mol, respectively.

Despite having nearly the same values of predicted affinities among those top 4 compounds, they are quite different molecules judged by their 2D similarity scores. Specifically, among those 4 compounds, BACE_gen_35 and BACE_gen_66 are the most similar structures but their similarity score is as low as 0.265. In addition, BACE_gen_29 and BACE_gen_25 are the most dissimilar compounds with 2D singularity score being 0.11. Generating very low binding affinity compounds with diverse chemical formulas is an important goal for the pre-clinical stage since that will enhance the chance of selecting promising drug candidates with low risk of having a side effect. Obtaining top and disparate molecules demonstrates the capacity of our proposed GNC in capturing the wide range of chemical space.

3.2. CatS

3.2.1. Data preparation Similar to the BACE target, CatS inhibitors were presented in the D3R grand challenges. Thus, these compounds (n =

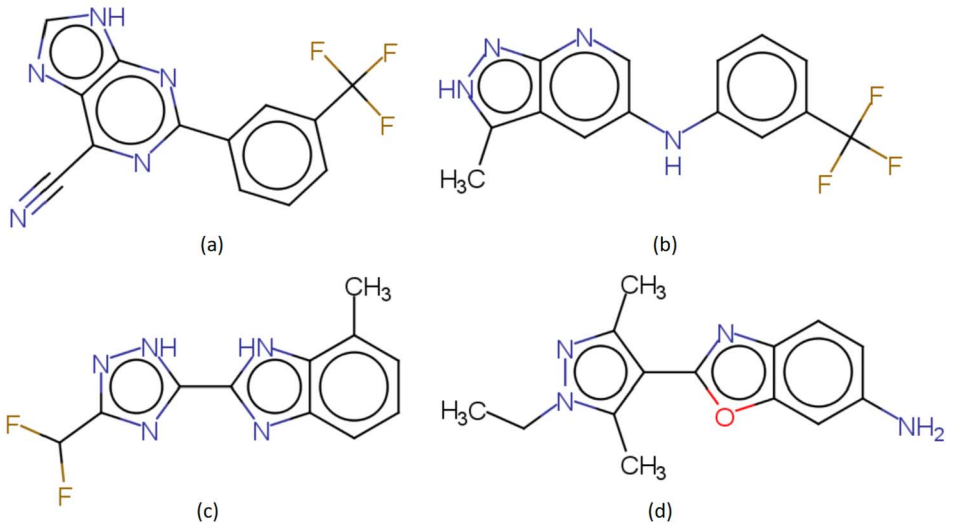

Figure 10: Illustration of similarity between a seed in the CatS set and some generated compounds: (a) The seed; (b) The most similar compound generated from the seed (similarity score=0.45); (c) A compound with a medium similarity score of 0.31; (d) The most different one from the seed (similarity score=0.18).

593) are used as seeds to produce new CatS molecules. Other CatS compounds reported in the ChemBL database are also included in our seeds. In total, we collected a data set of 2847 compounds. The binding affinity of these molecules ranges from −4.72 to −14.33 kcal/mol. As with the BACE data set, we chose −9.56 kcal/mol as the threshold for the active compound selection. With this threshold, 1461 of the 2847 compounds in our seeds are active. The distribution of binding affinity in our collected CatS data set is shown in Figure 9a.

3.2.2. Structure generation Using the 2847 compounds in the CatS collected set as seeds and feeding them into the generator network, we generated 1000 distinct compounds for each seed, for a total of 2,847,000 generated compounds. However, there was some duplication among the compounds generated by different seeds, resulting in only 2,080,566 distinct compounds being generated. To determine the novelty of our generated network, the similarity score between each generated compound and its seed is evaluated and depicted in Figure 9b.

Similar to the results from the BACE data set, the similarity scores of the generated compounds have a broad range from 0.06 and 1.00. A similarity score of 1.00 was obtained only 12 times in all 2,847,000 generated molecules. In most cases, the similarity scores are very low with an average value of 0.34, indicating there is a lot of diversity among the generated samples. To further verify that the generated compounds are really different from the seeds, a seed molecule and several generated compounds are shown in Figure 10. In which, Figure 10a is one seed, Figure 10b is the most similar compound generated from the seed with a similarity score of 0.45, Figure 10c is the compound with a medium similarity score of 0.31, and Figure 10d is the most different one with a similarity score of 0.18. Obtaining low similarity scores between generated compounds and feeding target is one of the desired features in our GNC model in which the novelty of computer-generated molecules is emphasized.

3.2.3. Binding affinity screening by 2DFP-DNN Here, we carry out the 2DFP-DDN model to filter out the "bad" generated CatS molecules by the binding affinity criterion. Similar to the BACE compound screening conditions, we use an affinity threshold at −9.56 kcal/mol. Specifically, any molecules with predicted energy higher than that threshold are left out. As a result, we selected 61,571 potentially "good" compounds.

Furthermore, we are interested in the overall distribution of the binding affinity of the generated compounds. Figure 9c depicts the distribution of the predicted affinity for all 2,080,566 molecules. The distribution is fairly close to a Gaussian distribution. Consistent with the similarity score distribution above, the range of their binding affinity prediction is very large, from −4.61 kcal/mol to −12.12 kcal/mol, confirming that large chemical space is covered. The mean binding affinity is −7.62 kcal/mol. Among the compounds with the smallest predicted binding affinity, 21,283 compounds have binding affinity smaller than −10 kcal/mol, 510 compounds have binding affinity smaller than −11 kcal/mol, and 1 compound has a binding affinity smaller than −12 kcal/mol. These are potentially very highly active compounds. However, as discussed in Section 3.1.3, there is no free lunch in the development of binding prediction models. 2DFP-DNN predictor is extremely fast in training millions of molecules. However, its accuracy is less competitive in comparison to 3D-based models such as MathDL. Thus, we still utilize the MathDL scoring function to select the most promising drug candidates.

3.2.4. Clustering analysis of selected compounds To illustrate how diverse our generated compounds are, clustering analysis were performed

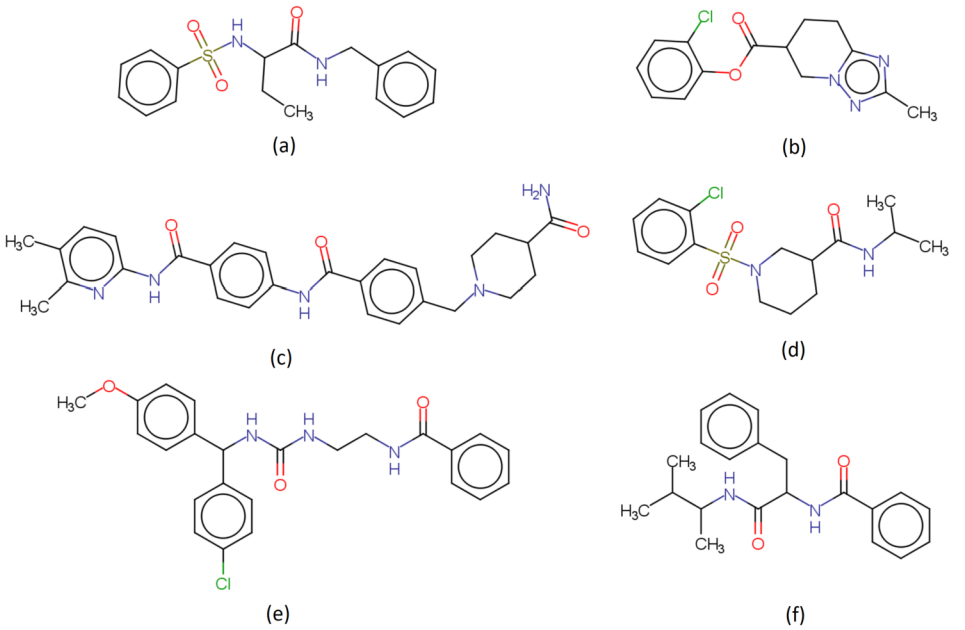

Figure 11: The center of the 6 clusters found in our CatS generated set.

to the 61,571 selected compounds generated by our model. The 6 clusters are found in our generated set, and the center of each cluster is shown in Figure 11. The sizes of these 6 clusters are 7077, 13059, 15048, 9221, 6884, and 10282 respectively. Inside these 6 clusters, the average similarity scores to the centers are 0.37, 0.34, 0.34, 0.39, 0.41, and 0.36 respectively, which indicates that there is a significant variation among compounds in each cluster. In addition, the average binding affinity of each cluster is -10.01, -9.89, -9.91, -9.98, -9.92, and -9.93 kcal/mol respectively. Unlike the BACE data set, there is not much difference in the average energies between different clusters. Therefore, it is expected to obtain highly potential drug candidates with dissimilar physical and biological chemical properties.

3.2.5. Binding affinity screening by MathDL MathDL here was trained with 2847 seeds used in the generator network. The Pearson's correlation coefficient and Kendall's Tau coefficient on the 10-fold cross-validation (CV) of the training set was found to be $R_p = 0.746$ and $\tau = 0.577$, respectively. The promising CV performance ensures a well-trained machine learning model. Furthermore, the reliability of the MathDL models on the

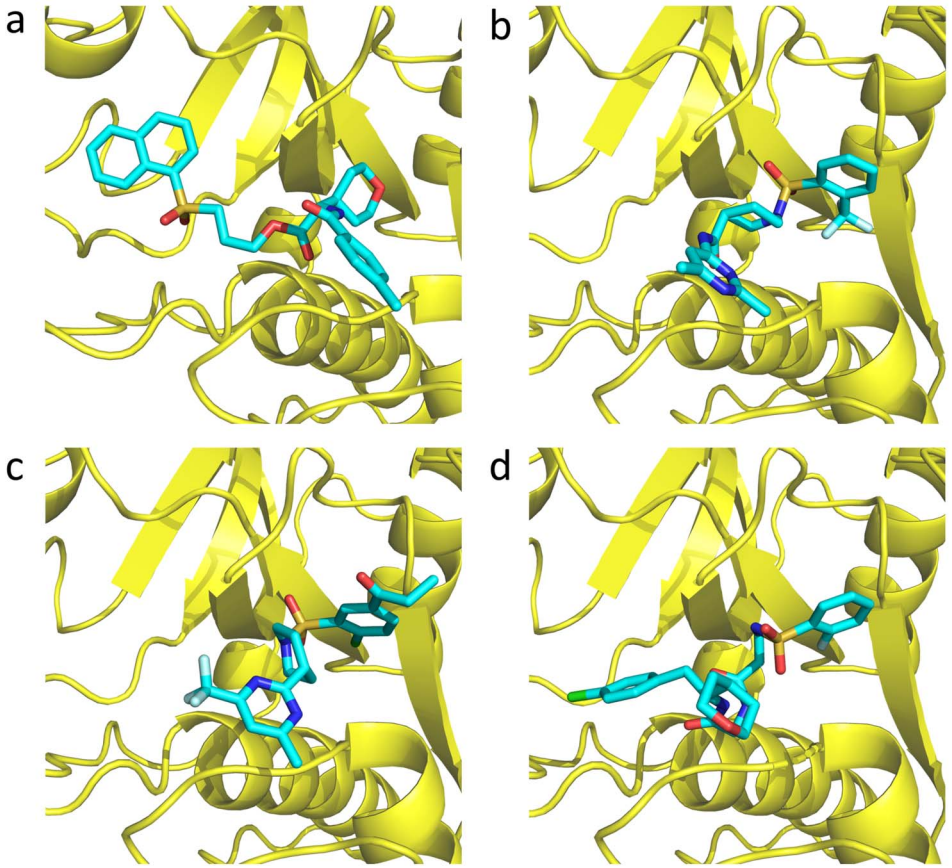

Figure 12: Four generated CatS compounds having the lowest binding affinities predicted by MathDL model. Their 3D structures were predicted by MathPose. Their IDs under our naming system and the predicted energies are, respectively (a) CatS_gen_195 (−11.681 kcal/mol); (b) CatS_gen_968 (−11.608 kcal/mol); (c) CatS_gen_902 (−11.540 kcal/mol); and (d) CatS_gen_228 (−11.536 kcal/mol). Their SMILES strings are provided in Table S2, and their corresponding 3D structures are included in File S2.

affinity ranking of the CatS inhibitors has been shown in the Grand Challenges 3 [30] and 4 [31] where our models were ranked 1st place among more than 50 teams from over the world.

To further validate the generated molecules, the top 1050 compounds in term of energy indicated by 2DFP-DNN network are re-ranked by the

MathDL model. To get the input ready for the structure-based model, the 3D poses of those 1050 molecules are predicted by our MathPose. Under the 2DFP-DNN predictor, the average binding affinity of those 1050 compounds is −11.05 kcal/mol and their affinities range from −10.693 kcal/mol to −12.159 kcal/mol with standard deviation being −0.213 kcal/mol. On the other hand, by utilizing the MathDL model, the average binding affinity of the selected molecules is −9.27 kcal/mol with a range between −7.008 kcal/m and −11.681 kcal/mol, and the standard deviation is found to be −0.736 kcal/mol. The Pearson's correlation coefficient on the energy prediction for the generated compounds by 2DFP-DNN and MathDL is as low as 0.112 which indicates the disagreement between those two models. That discrepancy was also observed when predicting the affinity ranking of the CatS molecules in the Grand Challenges 4 where the structure-based model MathDL outperformed its ligand-based counterpart. Therefore, MathDL's predicted energies are chosen to select the promising drug candidates among the computer-generated compounds.

The 3D structures of top 4 compounds in term of affinity, namely CatS_gen_195, CatS_gen_968, CatS_gen_902, and CatS_gen_228, are plotted in Figure 12. Their reported affinities are, respectively, −11.681 kcal/mol, −11.608 kcal/mol, −11.540 kcal/mol, and −11.536 kcal/mol. Despite similarly predicted affinities, their structures are quite dissimilar from each other. Specifically, the highest similarity score is 0.297 obtained between CatS_gen_968 and CatS_gen_902 molecules. While the lowest similarity score is 0.11 evaluated between CatS_gen_902 and CatS_gen_228. The statistical information again confirms the ability of our proposed GNC to cover large chemical space.

4. Discussions

Since the chemical space is huge, there is a need to generate a wide variety of novel compounds for all kinds of properties. This work introduces the GNC to generate novel molecules, predict their druggable properties, and finally pick up the drug candidates that fulfill the threshold for drug properties such as binding affinity. We discuss a number of issues concerning generative networks.

4.1. Latent space design of new compounds

Latent space information can be effectively modified by a variety of methods. In the current work, we propose three approaches, including 1) randomized

output, 2) controlled output, and 3) optimized output. The first approach can certainly create new molecules. We note that some of the new latent space configurations cannot be interpreted by the decoder.

The second approach is designed to discriminate potentially drug-like molecules from potentially inactive ones. Currently, we found that machine learning models built from latent space representations are highly accurate. Therefore, the proposed approach is potentially very useful. Nonetheless, the performance of this method depends crucially on the quality of training datasets. Additionally, for this approach to effectively control the druggability of the generated compounds, the decoder must be intensively trained with tens of millions of molecules and have a near-perfect reconstruction rate. Achieving a high reconstruction rate for a diverse class of test compounds is a challenging issue in the design of molecular autoencoders. This issue is under our consideration.

The third approach is introduced to create new compounds with desirable druggable properties. Similarly to the last method, the success of this approach depends on the quality of training datasets and machine models and the reconstruction rate of the decoder. Additionally, reference selection for each drug property is another important issue. It depends on our current understanding and criteria of drug-like molecules. However, this approach is very promising and will be an important direction for future studies.

Finally, it is noted that the third approach does not depend on the seed configuration. Therefore, its initial latent space distribution can be chosen randomly. As such, this method can be very fast and efficient.

4.2. Generator efficiency

One challenge traditional pharmaceutical industry faces is that designing new drug candidates is very time-consuming. This low efficiency obviously can not tackle a variety of health crises human being currently encounters, such as drug-resistant infections and fast mutation of viruses, which requires lots of new drugs in a very short time.

Computers are typically faster than human beings. Therefore, generating new drugs by computers is a potential solution. Such as in our case, just using one K20 Nvidia CUDA GPU card, our generator network can generate 2.08 million and 2.8 million novel compounds for the CatS and BACE targets in less than one week, such task is far beyond human power. Moreover, such process is fully automatic and even does not need human supervision. So, such automatic generators can provide us a huge drug-candidate database rather than some sporadic ones. What is more, just we already showed in

this work, combining this generator with reliable automatic DNN predictors, the bulk of drug candidates can be further screened based on the properties predicted by automatic predictors. This whole automatic workflow should be a promising future of the pharmaceuticals industry.

4.3. Chemical spaces generated by generators

Our generated compounds are originated from their seeds, some known ligands binding to CatS and BACE from PDBbind database. Thanks to the magic of the autoencoder including some random source, these generated compounds are truly novel and quite far from their seeds: no matter for CatS and BACE, the average similarity scores to the seeds are just around 0.3. This means the two sources of random works and the generator creates novel compounds rather than just playing some "copy" games.

More importantly, these generated compounds spread in huge chemical space, this means our generated compounds cover a large range of chemical properties, so it is more possible to hit potential drug candidates. First, the similarity scores to the seeds have a large range, for BACE it is 0.2 to 0.6, for CatS, it is 0.15 to 0.65. Second, our predicted binding affinities also have a wide range, from -5 kcal/mol to -10 kcal/mol or even to -11 kcal/mol. All in all, the generator is powerful, originating from seeds and but cover a huge chemical space far away from the seeds.

4.4. Faith vs novelty

The random noise regulated latent space we designed here can generate lots of novel compounds far from their seeds, this is due to its design: random sources are included in the model. However, in other words, this generator is not faithful, since the output is quite different from the input. Such architecture is good for our purpose – what we want is to create broad new compounds from the seeds rather than faithful ones.

However, in another scenario, faith is highly needed. Griffiths et al. [77], Jin et al. [78], Kusner et al. [79] and Dai et al. [80] perform Bayesian optimization in the latent space to obtain compounds with desired properties. We have also designed controlled latent space and optimized latent space in the present work. In these cases, outputs should faithfully reflect the latent space. Otherwise, optimization in the latent space could not be faithfully passed to the output. The reconstructing accuracy is a very critical evaluation. Much effort has already put to reinforce reconstructing accuracies, such as grammar VAE [79], syntax-directed VAE [80] and junction tree VAE

[78], to achieve a reconstructing accuracy as high as 0.76. In comparison, the VAE we applied only has a reconstructing accuracy of 0.20. It means that our outputs are always new compounds. However, our new gated recurrent unit (GRU)-based autoencoder can achieve a 99% reconstructing accuracy, which enables us to carry out desirable design in the latent space. The detail of this work will be published elsewhere.

4.5. Chemical spaces of predicted high binding affinity compounds

Using our predictors, high binding affinity compounds can be screened. So one concern is whether these high binding affinity compounds spread in a large chemical space or they are similar to each other and in a small range. According to our results, even the numbers of high binding affinity compounds are only 1050 and 99 for CatS and BACE respectively, they are still quite different. In our clustering analysis, these high binding affinity compounds are classified into 6 clusters, the similarities between clusters are only around 0.4.

The large chemical space covered by the high binding affinity compounds is beneficial to drug design. First, a good drug not only depends on binding affinity but also depends on other properties such as toxicity, log P, log S, clearance, etc. A large chemical space means these high binding affinity compounds have different other properties, so there is more chance for them to pass the screenings based on other properties. Additionally, more types of related drugs are easier to tackle the fast mutation of viruses.

5. Conclusion

In our work, a generative network complex (GNC) is introduced. We propose three latent-space techniques, including randomized output, controlled output, and optimized output to generate novel and potential compounds. Additionally, their physical and chemical properties are predicted by a two-dimensional (2D) fingerprint-based deep learning predictor, and potential drug candidates are preliminarily screened by predicted properties. Moreover, for promising drug candidates, their 3D poses associated with specific protein targets are predicted by our MathPose, one of the most accurate pose prediction schemes according to D3R Grand Challenges, a worldwide competition series in computer-aided drug design [31]. Finally, more accurate property estimations based on the 3D poses are performed by our MathDL, a advanced mathematics-based deep learning network, leading to new drug

candidates with the desirable drug properties. This automated platform has been used to generate 2.08 million new drug candidates for Cathepsin S and 2.8 million novel compounds for BACE. For 1050 potential drug candidates for CatS and 99 potential drug candidates for BACE, 3D poses associated with their target proteins have been created to further evaluate their druggable properties. Our framework is designed to create new drugs in silico, so as to save time and reduce cost in drug discovery. Designing gated recurrent unit (GRU)-based autoencoders with near perfect reconstruction accuracies is under our consideration to achieve robust latent space drug design.

Supplementary materials

Supplementary materials are available upon request for potential drug candidates for Cathepsin S and BACE targets.

TableS1.csv A list of SMILES strings and predicted binding affinities of 99 potentially active compounds for the BACE target.

TableS2.csv A list of SMILES strings and predicted binding affinities of 1050 potentially active compounds for the CatS target.

Additional supplementary materials are available upon request for potential drug candidates for Cathepsin S and BACE targets (near 70 gigabytes in size).

FileS1.zip Zip file of 3D structure information of 99 selectively generated BACE compounds and their receptors.

FileS2.zip Zip file of 3D structure information of 1050 selectively generated CatS compounds and their receptors.

References

[1] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of Health Economics*, 47:20–33, 2016.

[2] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, 2011.

[3] Ricardo Macarron, Martyn N Banks, Dejan Bojanic, David J Burns, Dragan A Cirovic, Tina Garyantes, Darren VS Green, Robert P Hertzberg, William P Janzen, Jeff W Paslay, et al. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10(3):188, 2011.

[4] Ann I Graul, Laura Revel, Esmeralda Rosa, and Elisabet Cruces. Overcoming the obstacles in the pharma/biotech industry: 2008 update. *Drug News Perspect*, 22(1):39, 2009.

[5] Mahmud Tareq Hassan Khan. Predictions of the admet properties of candidate drug molecules utilizing different qsar/qspr modelling approaches. *Current Drug Metabolism*, 11(4):285–295, 2010.

[6] Michael J Waring, John Arrowsmith, Andrew R Leach, Paul D Leeson, Sam Mandrell, Robert M Owen, Garry Pairaudeau, William D Pennie, Stephen D Pickett, Jibo Wang, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, 14(7):475, 2015.

[7] Kristian Strømgaard, Povl Krogsgaard-Larsen, and Ulf Madsen. *Textbook of drug design and discovery*. CRC Press, 2017.

[8] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1-3):3–25, 1997.

[9] Katya Tsaioun, Michel Bottlaender, and Aloise Mabondzo. Addme – avoiding drug development mistakes early: central nervous system drug discovery perspective. In *BMC neurology*, volume 9, page S1. BioMed Central, 2009.

[10] Saeed Alqahtani. In silico adme-tox modeling: progress and prospects. *Expert Opinion on Drug Metabolism & Toxicology*, 13(11):1147–1158, 2017.

[11] KV Balakin, YA Ivanenkov, and NP Savchuk. Compound library design for target families. In *Chemogenomics*, pages 21–46. Springer, 2009.

[12] IM Kapetanovic. Computer-aided drug discovery and development (caddd): in silico-chemico-biological approach. *Chemico-Biological Interactions*, 171(2):165–176, 2008.

[13] Renjie Huang and Ivanhoe Leung. Protein-directed dynamic combinatorial chemistry: a guide to protein ligand and inhibitor discovery. *Molecules*, 21(7):910, 2016.

[14] Paweł Szymański, Magdalena Markowicz, and Elżbieta Mikiciuk-Olasik. Adaptation of high-throughput screening in drug discovery-toxicological screening tests. *International Journal of Molecular Sciences*, 13(1):427–452, 2012.

[15] Gisbert Schneider and Uli Fechner. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4(8):649, 2005.

[16] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18(5):851–869, 2017.

[17] Polina Mamoshina, Armando Vieira, Evgeny Putin, and Alex Zhavoronkov. Applications of deep learning in biomedicine. *Molecular Pharmaceutics*, 13(5):1445–1454, 2016.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[20] Danilo P Mandic and Jonathon Chambers. *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. John Wiley & Sons, Inc., 2001.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[22] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[23] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017.

[24] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science Advances*, 4(7):eaap7885, 2018.

[25] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

[26] Seokho Kang and Kyunghyun Cho. Conditional molecular design with deep generative models. *Journal of Chemical Information and Modeling*, 59(1):43–52, 2018.

[27] Miha Skalic, José Jiménez, Davide Sabbadin, and Gianni De Fabritiis. Shape-based generative modeling for de novo drug design. *Journal of Chemical Information and Modeling*, 59(3):1205–1214, 2019.

[28] Artur Kadurin, Sergey Nikolenko, Kuzma Khrabrov, Alex Aliper, and Alex Zhavoronkov. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular Pharmaceutics*, 14(9):3098–3104, 2017.

[29] Boris Sattarov, Igor I Baskin, Dragos Horvath, Gilles Marcou, Esben Jannik Bjerrum, and Alexandre Varnek. De novo molecular design by combining deep autoencoder recurrent neural networks with generative topographic mapping. *Journal of Chemical Information and Modeling*, 59(3):1182–1196, 2019.

[30] Duc Duy Nguyen, Zixuan Cang, Kedi Wu, Menglun Wang, Yin Cao, and Guo-Wei Wei. Mathematical deep learning for pose and binding affinity prediction and ranking in d3r grand challenges. *Journal of Computer-Aided Molecular Design*, 33(1):71–82, 2019.

[31] Duc Duy Nguyen, Kaifu Gao, Menglun Wang, and Guo-Wei Wei. Mathdl: Mathematical deep learning for d3r grand challenge 4. *Journal of Computer Aided Molecular Design*, in press, 2019. arXiv preprint arXiv:1909.07784.

[32] Greg Landrum et al. Rdkit: Open-source cheminformatics, 2006.

[33] Thomas A Halgren. Merck molecular force field. I. basis, form, scope, parameterization, and performance of mmff94. *Journal of Computational Chemistry*, 17(5-6):490–519, 1996.

[34] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[35] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[36] Marco Daniele Parenti and Giulio Rastelli. Advances and applications of binding affinity prediction methods in drug discovery. *Biotechnology Advances*, 30(1):244–250, 2012.

[37] Kaifu Gao, Jian Yin, Niel M Henriksen, Andrew T Fenley, and Michael K Gilson. Binding enthalpy calculations for a neutral host–guest pair yield widely divergent salt effects across water models. *Journal of Chemical Theory and Computation*, 11(10):4555–4564, 2015.

[38] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[39] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 2002.

[40] Kedi Wu, Zhixiong Zhao, Renxiao Wang, and Guo-Wei Wei. TopP-S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *Journal of Computational Chemistry*, 39(20):1444–1454, 2018.

[41] Kedi Wu and Guo-Wei Wei. Quantitative toxicity prediction using topology based multitask deep neural networks. *Journal of Chemical Information and Modeling*, 58(2):520–531, 2018.

[42] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch: Tensors and dynamic neural networks in python with strong GPU acceleration. *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, 6, 2017.

[43] Zixuan Cang, Lin Mu, Kedi Wu, Kristopher Opron, Kelin Xia, and Guo-Wei Wei. A topological approach for protein classification. *Computational and Mathematical Biophysics*, 3(1), 2015.

[44] Zixuan Cang, Lin Mu, and Guo-Wei Wei. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Computational Biology*, 14(1):e1005929, 2018.

[45] Zixuan Cang and Guo-Wei Wei. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International Journal for Numerical Methods in Biomedical Engineering*, 34(2):e2914, 2018.

[46] Duc Duy Nguyen and Guo-Wei Wei. Dg-gl: Differential geometry-based geometric learning of molecular datasets. *International Journal for Numerical Methods in Biomedical Engineering*, 35(3):e3179, 2019.

[47] Duc D Nguyen, Tian Xiao, Menglun Wang, and Guo-Wei Wei. Rigidity strengthening: A mechanism for protein–ligand binding. *Journal of Chemical Information and Modeling*, 57(7):1715–1721, 2017.

[48] Duc Nguyen and Guo-Wei Wei. Agl-score: Algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. *Journal of Chemical Information and Modeling*, 59(7):3291–3304, 2019.

[49] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005. MR2121296

[50] Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering*, 30(8):814–844, 2014. MR3247713

[51] Kelin Xia, Xin Feng, Yiying Tong, and Guo Wei Wei. Persistent homology for the quantitative prediction of fullerene stability. *Journal of Computational Chemistry*, 36(6):408–422, 2015.

[52] Zixuan Cang and Guo-Wei Wei. Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Computational Biology*, 13(7):e1005690, 2017.

[53] Guo-Wei Wei. Differential geometry based multiscale models. *Bulletin of Mathematical Biology*, 72(6):1562–1622, 2010. MR2671586

[54] Kelin Xia and Guo-Wei Wei. Multidimensional persistence in biomolecular data. *Journal of Computational Chemistry*, 36(20):1502–1520, 2015.

[55] Dejan Plavšić, Sonja Nikolić, Nenad Trinajstić, and Douglas J Klein. Relation between the wiener index and the Schultz index for several classes of chemical graphs. *Croatica Chemica Acta*, 66(2):345–353, 1993. MR1169298

[56] Dusanka Janezic, Ante Milicevic, Sonja Nikolic, and Nenad Trinajstic. *Graph-theoretical matrices in chemistry*. CRC Press, 2015. MR3381136

[57] Nobuhiro Go, Tosiyuki Noguti, and Testuo Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Sciences*, 80(12):3696–3700, 1983.

[58] Ali Rana Atilgan, SR Durell, Robert L Jernigan, Melik C Demirel, O Keskin, and Ivet Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1):505–515, 2001.

[59] Ivet Bahar, Timothy R Lezon, Lee-Wei Yang, and Eran Eyal. Global dynamics of proteins: bridging between structure and function. *Annual Review of Biophysics*, 39:23–42, 2010.

[60] Tiejun Cheng, Xun Li, Yan Li, Zhihai Liu, and Renxiao Wang. Comparative assessment of scoring functions on a diverse test set. *Journal of Chemical Information and Modeling*, 49(4):1079–1093, 2009.

[61] Yan Li, Li Han, Zhihai Liu, and Renxiao Wang. Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *Journal of Chemical Information and Modeling*, 54(6):1717–1736, 2014.

[62] Cheng Wang and Yingkai Zhang. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *Journal of Computational Chemistry*, 38(3):169–177, 2017.

[63] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010.

[64] Gareth Jones, Peter Willett, Robert C Glen, Andrew R Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267(3):727–748, 1997.

[65] Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, 2004.

[66] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20, 2015.

[67] Kaifu Gao, Hongqing He, Minghui Yang, and Honggao Yan. Molecular dynamics simulations of the escherichia coli hppk apo-enzyme reveal a network of conformational transitions. *Biochemistry*, 54(44):6734–6742, 2015.

[68] Kaifu Gao, Ya Jia, and Minghui Yang. A network of conformational transitions revealed by molecular dynamics simulations of the binary complex of escherichia coli 6-hydroxymethyl-7, 8-dihydropterin pyrophosphokinase with mgatp. *Biochemistry*, 55(49):6931–6939, 2016.

[69] Kaifu Gao and Yunjie Zhao. A network of conformational transitions in the apo form of ndm-1 enzyme revealed by md simulation and a Markov state model. *The Journal of Physical Chemistry B*, 121(14):2952–2960, 2017.

[70] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011. MR2854348

[71] Zied Gaieb, Conor D Parks, Michael Chiu, Huanwang Yang, Chenghua Shao, W Patrick Walters, Millard H Lambert, Neysa Nevins, Scott D Bembenek, Michael K Ameriks, et al. D3r grand challenge 3: blind prediction of protein–ligand poses and affinity rankings. *Journal of Computer-Aided Molecular Design*, 33(1):1–18, 2019.

[72] Michael K Ameriks, Scott D Bembenek, Matthew T Burdett, Ingrid C Choong, James P Edwards, Damara Gebauer, Yin Gu, Lars Karlsson, Hans E Purkey, Bart L Staker, et al. Diazinones as p2 replacements for pyrazole-based cathepsin s inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 20(14):4060–4064, 2010.

[73] Danielle K Wiener, Alice Lee-Dutra, Scott Bembenek, Steven Nguyen, Robin L Thurmond, Siquan Sun, Lars Karlsson, Cheryl A Grice, Todd K Jones, and James P Edwards. Thioether acetamides as p3 binding elements for tetrahydropyrido-pyrazole cathepsin s inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 20(7):2379–2382, 2010.

[74] Robert Vassar, Dora M Kovacs, Riqiang Yan, and Philip C Wong. The β-secretase enzyme bace in health and Alzheimer's disease: regulation, cell biology, function, and therapeutic potential. *Journal of Neuroscience*, 29(41):12787–12794, 2009.

[75] Federica Prati, Giovanni Bottegoni, Maria Laura Bolognesi, and Andrea Cavalli. Bace-1 inhibitors: From recent single-target molecules to multitarget compounds for Alzheimer's disease: Miniperspective. *Journal of Medicinal Chemistry*, 61(3):619–637, 2017.

[76] Teague Sterling and John J. Irwin. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015. http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00559.

[77] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained

bayesian optimization for automatic chemical design. *arXiv preprint arXiv:1709.05501*, 2017.

[78] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. *arXiv preprint arXiv:1802.04364*, 2018.

[79] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning, Volume 70*, pages 1945–1954. JMLR.org, 2017.

[80] Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.

Christopher Grow
Department of Mathematics
Michigan State University
East Lansing, MI 48824
USA
E-mail address: growchri@msu.edu

Kaifu Gao
Department of Mathematics
Michigan State University
East Lansing, MI 48824
USA
E-mail address: gaokaifu@msu.edu

Duc Duy Nguyen
Department of Mathematics
Michigan State University
East Lansing, MI 48824
USA
E-mail address: ddnguyen@msu.edu

Guo-Wei Wei
Department of Mathematics
Michigan State University
East Lansing, MI 48824
USA
E-mail address: weig@msu.edu