Deep Learning for Drug Design

Modeling Molecular Shapes

Miha Skalic

TESI DOCTORAL UPF / year 2019

THESIS SUPERVISOR
Prof. Gianni de Fabritiis
Department de Ciències Experimentals i de la Salut

In order for A.I. systems to work, they need to be trained.
And we, we humans, are their mothers and fathers.
We are their study buddies.
We are the ones these A.I. systems are learning from.

AMY WEBB

Acknowledgments

I would like to thank everyone who contributed to publications presented here and the projects that did not make into publications. Many thanks to family, flatmates and friends who supported me during the PhD period in Barcelona.

Special thanks go to Acellera colleborators who made the developed application available to public as part of *playmolecule.org*.

Finally, I would like to thank my supervisor Gianni De Fabritiis for allowing me to carry out research as part of his lab.

Abstract

Designing novel drugs is a complex process which requires finding molecules in a vast chemical space that bind to a specific biomolecular target and have favorable physio-chemical properties. Machine learning methods can leverage previous data and use it for new predictions helping the processes of selection of molecule candidate without relying exclusively on experiments. Particularly, deep learning can be applied to extract complex patterns from simple representations. In this work we leverage deep learning to extract patterns from three-dimensional representations of molecules. We apply classification and regression models to predict bioactivity and binding affinity, respectively. Furthermore, we show that it is possible to predict ligand properties for a particular protein pocket. Finally, we employ deep generative modeling for compound design. Given a ligand shape we show that we can generate similar compounds, and given a protein pocket we can generate potentially binding compounds.

Resum

El disseny de drogues novells es un procés complex que requereix trobar les molècules adequades, entre un gran ventall de possibilitats, que siguin capaces d'unir-se a la proteïna desitjada amb unes propietats fisicoquímiques favorables. Els mètodes d'aprenentatge automàtic ens serveixen per a aprofitar dades antigues sobre les molècules i utilitzar-les per a noves prediccions, ajudant en el procés de selecció de molècules potencials sense la necessitat exclusiva d'experiments. Particularment, l'aprenentatge profund pot sera plicat per a extreure patrons complexos a partir de representacions simples. En aquesta tesi utilitzem l'aprenentatge profund per a extreure patrons a partir de representacions tridimensionals de molècules. Apliquem models de classificació i regressió per a predir la bioactivitat i l'afinitat d'unió, respectivament. A més, demostrem que podem predir les propietats dels lligands per a una cavitat proteica determinada. Finalment, utilitzem un model generatiu profund per a disseny de compostos. Donada una forma d'un lligand demostrem que podem generar compostos similars i, donada una cavitat proteica, podem generar compostos que potencialment s'hi podràn unir.

Preface

The field of deep learning has exploded. In 2012 Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton showed that neural networks can be trained to predict content of an image in an end-to-end fashion, just by presenting raw images to a network. After that, progress has been made at an unstoppable pace, making it difficult for researchers to keep up with all the latest papers. For example NeurIPS 2018, largest machine learning conference, tickets sold out in just a few minutes, faster than many popular concerts.

On the other hand, pharmaceutical industry is thirsty for disruption. Past promising approaches such as combinatorial chemistry and high-throughput screening did not live up to its expectations. Developing a drug costs billions and can take over decade, in addition to being plagued with high failure rate.

Here we attempt to bridge the gap. We try to leverage the available data and apply it to the process of drug design. Taking inspiration from state-of-the-art methods for image and text analysis, we apply the methods to problems of designing novel drugs. Although these are only the first steps towards new type of data-driven drug discovery, I have a bright outlook on the future. If computers can create images that humans cannot tell apart from real, why can't they design novel pharmaceuticals?

Publications

This section lists publications that were produced during the period of the doctoral study. Publications 1, 2, 3, 4 are published, while publication 5 is in preprint stage. To all publications I contributed to as a first author except for 2nd, to which I contributed as a co-author. The following publications are presented:

1. PlayMolecule BindScope: large scale CNN-based virtual screening on the web. M. Skalic, G. Martínez-Rosell, J. Jiménez and G. De Fabritiis. Bioinformatics 35. 1237-1238 (2018). doi: 10.1093/bioinformatics/bty758.

2. K_{DEEP}: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. J. Jiménez, M. Skalic, G. Martínez-Rosell and G. De Fabritiis. Journal of Chemical Information and Modeling 58. 287-296 (2018). doi: 10.1021/acs.jcim.7b00650.

3. LigVoxel: inpainting binding pockets using 3D-convolutional neural networks. M. Skalic, A. Varela-Rial, J. Jiménez, G. Martínez-Rosell and G. De Fabritiis. Bioinformatics 35. 243-250 (2018). doi: 10.1093/bioinformatics/bty583.

4. Shape-Based Generative Modeling for de-novo Drug Design. M. Skalic, J. Jiménez, D. Sabbadin and G. De Fabritiis. Journal of Chemical Information and Modeling 59. 1205-1214 (2018). doi: 10.1021/acs.jcim.8b00706.

5. From Target to Drug: Generative Modeling for Multimodal Structure-Based Drug Design. M. Skalic, D. Sabbadin, B.Sattarov and G. De Fabritiis. Preprint.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Whether we are aware or not machine learning (ML) technology is well integrated in modern society. Spam filters for our emails, recommendations on e-commerce websites, and auto-completion of our messages are all powered by machine learning algorithms.[1] These algorithms, that learn from data, can also be employed to aid in drug discovery. Herein we present a collection of such applications.

This thesis is structured as follows. Firstly, we introduce concepts of ML with a focus on elements relevant to the carried out research. Description of drug discovery challenges that we tackled follows. In chapter 3 produced publications are presented. Finally, we discuss the research outcome and draw conclusions.

1.1 Machine learning and deep learning

Fundamentally, machine learning is the study of algorithms and statistical models that instead of relying on usage of explicit instructions, is learning from patterns seen in data. After a learning phase the models can be exploited to uncover novel patters or to predict future data.[2] Machine learning algorithms improve their performance as the quality and quantity of data increases. However, the data is often high-dimensional with complex structure and for some algorithms extracting patterns can be hard.

For these tasks deep learning (DL) enabled major breakthroughs. Two most notable examples are image[3] and speech[4] recognition.

DL includes a class of machine learning algorithms that uses artificial neural networks (ANNs). Deep feedforward networks, also known as multilayer perceprots (MLP) are the most common type of networks. MLP can be viewed as an function approximator f^*. For example, regressor $y = f^*(\boldsymbol{x})$ maps input $\boldsymbol{x}$ to prediction y. The mapping is dependent on parameters of the network $\boldsymbol{\theta}$ ($\mathbf{y} = f(\boldsymbol{x}; \boldsymbol{\theta})$) and we want to learn values of these parameters so that we can approximate a target function.[5]

In MLPs the function $f(x)$ is composed of multiple consecutive transformations. For example, a three layer network can be presented in the following way: $f(\boldsymbol{x}) = f^3(f^2(f^1(\boldsymbol{x})))$. Typically, each transformation f^i is defined as $f^i(\boldsymbol{x_i}) = \phi(\boldsymbol{A}\boldsymbol{x_i} + \boldsymbol{b})$, where $\boldsymbol{A}$ and $\boldsymbol{b}$ are learnable parameters of the network and ϕ is an activation function used to introduce non-linear transfromations to the network.

Training a neural network is done by updating weights to minimize a loss function, $\mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y})$, which penalizes the distance between the prediction $\hat{\boldsymbol{y}}$ and the target $\boldsymbol{y}$. Backpropagation algorithm, introduced by Rumelhart et al. [6], is the most commonly used algorithm for this minimization. It uses chain rule to calculate the derivative of the loss function $\mathcal{L}$ with respect to each parameter $\boldsymbol{\theta}$ of the network. The calculated gradients of the weights, $\nabla\boldsymbol{\theta}$, are then used to update the network weights:

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \eta\nabla\boldsymbol{\theta}, \tag{1.1}$$

where η is the learning rate. Nowdays, ANNs are trained with stochastic gradient descent (SGD) using mini-batches. This means that the gradients $\nabla\boldsymbol{\theta}$ are computed and averaged based on the examples in the mini-batch. Over the years several variants of SGD have been developed to accelerate the training process. These include AdaGrad[7], AdaDelta[8], RMSprop and Adam[9].

1.1.1 Convolutional neural networks

Convolutional neural networks[10] (CNNs) are special kind of neural networks for processing data that has a grid-like topology, for example images. Convolution operation is the keys component in these networks and with restrictive connectivity neurons of a convolution can operate only on region of the input, called receptive field. Because the weights, also refereed to as kernels, are shared and used for all receptive fields, same features can be extracted from different locations. Furthermore, by sharing weights CNNs reduce the number of parameters that need to be learned, while at the same time allow for transnational equivariance.

Typically, for two-dimensional input I, such as images, when performing a convolutional function we move receptive fields over both dimensions, i and j, and use a three-dimensional kernel K:

$$S(i,j) = \sum_m \sum_n \sum_c I(i+m, j+n, c) K(m,n,c). \tag{1.2}$$

In this case the receptive field is of size $m \times n$ and the convolution operation with the network will generative single value, $S(i,j)$, for a particular receptive field. The same mathematical operation is depicted in Figure 1.1 top. By applying this operation across the image we generate a feature map. Moreover, because in case of images there are also channels (red, green, blue), kernel needs to be scaled appropriately, taking into the account the extra dimension c. Finally, by applying multiple kernel to the same input we can generate multiple feature maps in the next layer. Commonly in CNNs after the convolutional function a bias term is added to each feature map, followed by an activation function. This sequence of transformations can repeat several, even hundreds, times (Figrure 1.1 bottom). In equation 1.2, a two-dimensional neural network is presented, but convolutions can span arbitrary number of dimensions. For example, one-dimensional convolution can be applied to predict binding DNA motifs[12] and three-dimensional to segmentation of biomedical images.[13] In this work we use three-dimensional convolutions as one of the core processing components.

Arguably, research and usage of convolutional and deep learning in

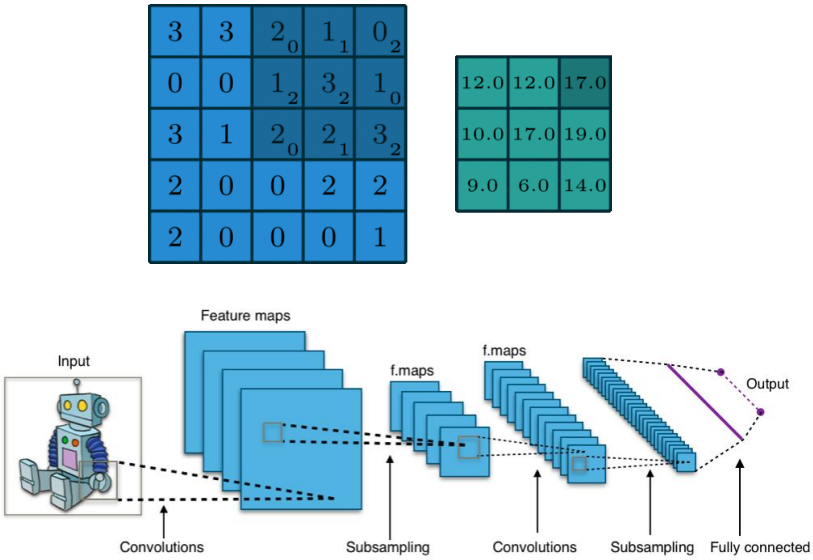

Figure 1.1: Convolutional operation and convolutional neural networks. (top) Example of convolutional operation for two-dimensional input with $|c| = 1$. Input feature map (in blue) is processed by a kernel of size 3×3 (shaded, weights are displayed in bottom right) to produce an output feature map (green). (bottom) Example of a convolutional neural network with two convolutional operations. Image content was taken from [11] and `en.wikipedia.org/wiki/Convolutional_neural_network`

general took off in 2012 with AlexNet[3], winning the ImageNet Large-Scale Image Classification Challenge[14] by a substantial margin. The winning model combined usage of CNNs together with Rectified Linear Units (ReLUs)[15], dropout[14] and pooling. ReLU activation function, defined as $f(x) = \max(0, x)$, allowed better gradient propagation and faster training than previously commonly used $\tanh$ activation function. Dropout, which sets zero to an output of an neuron with certain probability, reduced the overfitting. Finally, pooling layer summarizes the outputs of neighboring groups in the same feature map and thus reduces the size of the feature maps.

Since AlexNet, several improvements have been proposed, further improving the performance of image classification with CNNs. Most notably, the improvements come from changes to architecture, such as increasing the depth (number of layers)[16] and feature map connectivity.[17, 18] It is also worth noting that batch normalization method [19] was a major factor contributing to the ease of training ANNs.

As the methods have been developed, so has hardware evolved. Now it is a norm to train neural networks on graphics processing units (GPUs), enabling orders of magnitude faster training than CPUs.[20] Furthermore, software packages such as Caffe[21], Tensorflow[22] and Pytorch[23] allow fast development and training of models.

1.1.2 Recurrent neural networks

Output of convolutional operation is limited to a receptive field, this can be a problem when dealing with sequential data and long distance dependencies. Recurrent neural networks (RNNs) are connectionist models that can capture the dynamics of sequences via cycles in the network of nodes[24] and have been proposed as a method to deal with sequential data. For image captioning tasks[25], speech synthesis and music generation task they can serve as output producing model. They can be also used for time series prediction e.g. natural language processing, or be applied to interactive tasks such as language translation. For these tasks long short-term memory (LSTM)[26], a variant of RNNs, is the most used

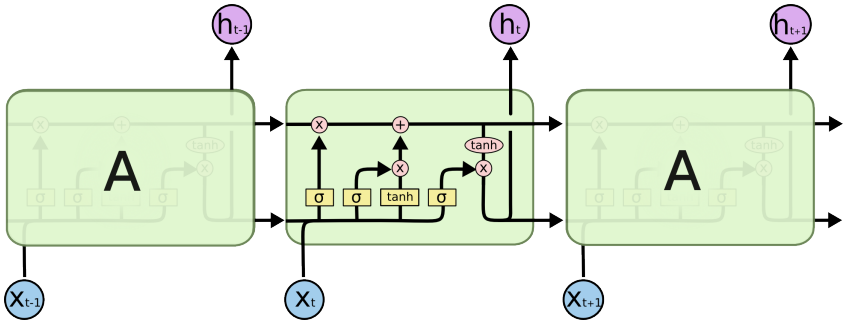

Figure 1.2: Operations in LSTM network. Image take from `https://colah.github.io/`

RNN implementation.

Figure 1.2 shows schematic representation of LSTM network. Given an input $\boldsymbol{x}_t$ at timespet t, output $\boldsymbol{h}_t$ is calculated as a result of operations in a LSTM cell (in green). In total there are four layers (in yellow) in a single cell. in addition to $\boldsymbol{x}_t$ output is also dependent on state of previous cell: memory and hidden state (black arrows from previous cell).

1.1.3 Generative modeling

Complementary to supervised learning, where the goal is to determine mapping of input $\boldsymbol{x}$ to output y, unsupervised learning methods are applied to discover "interesting structure" in the data. Typical applications include dimensionality reduction and clustering. Generative models are also part of unsupervised algorithms and have received a lot of attention in recent years. Generative models try to generate new samples from same probabilistic distribution as a training set. In the following section we describe variational autoencoders (VAE)[27] and generative adversarial networks (GANs)[28], two most commonly neural network-based generative models.

An autoencoder is a pair of networks, an encoder and a decoder. The encoder takes input and compresses it into dense representation (latent

space), which the decoder networks can convert back to the original input (Figure 1.3 top). Typically, the encoder and decoder are trained together through backpropagation. Latent space of autoencoders may not be continuous or easy to interpolate, which might be a problem when using autoencoders as generative models. VAEs overcome this shortcoming by making, by design, the latent space continuous. This is achieved by introducing an encoder that outputs a vector of means $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma}$. Then we sample $\boldsymbol{z}$ from a distribution defined by these two parameters (reparameterization) and pass the sampled values $\boldsymbol{z}$ to the decoder. By doing this certain degree of variability is introduced and the latent space is smoothed. Furthermore, Kullback-Leibler divergence term (KL)[29] is added to loss in addition to the reconstruction loss. KL term ensures that that latent space variables $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ closely resemble a target distribution and thus making the latent space continuous.

Generative adverserial networks (Figure 1.3 bottom), is a special subgroup of generative models, where two types of networks are being trained simultaneously. In the process, a generator (G) is trained to generate sample instances that try to fool a discriminator (D), which on the other hand tries to distinguish between generated and real samples. Essentially, the objective is to find a Nash equilibrium of a value function V for two player min-max problem:

$$\begin{aligned}\min_G \max_D V(D,G) =& \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log(D(x))] + \\ & \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))],\end{aligned} \tag{1.3}$$

where $\mathbf{z}$ is latent variable most commonly drawn from low-dimensional Gaussian or uniform distributions. Over the past few year performance of GANs has drastically improved. [30, 31, 32, 33, 34] Usage of GANs has also been extended to new settings such as image-to-image mapping[35] that can also work on unpaired samples[36] or multimodal one-to-many mapping.[37]

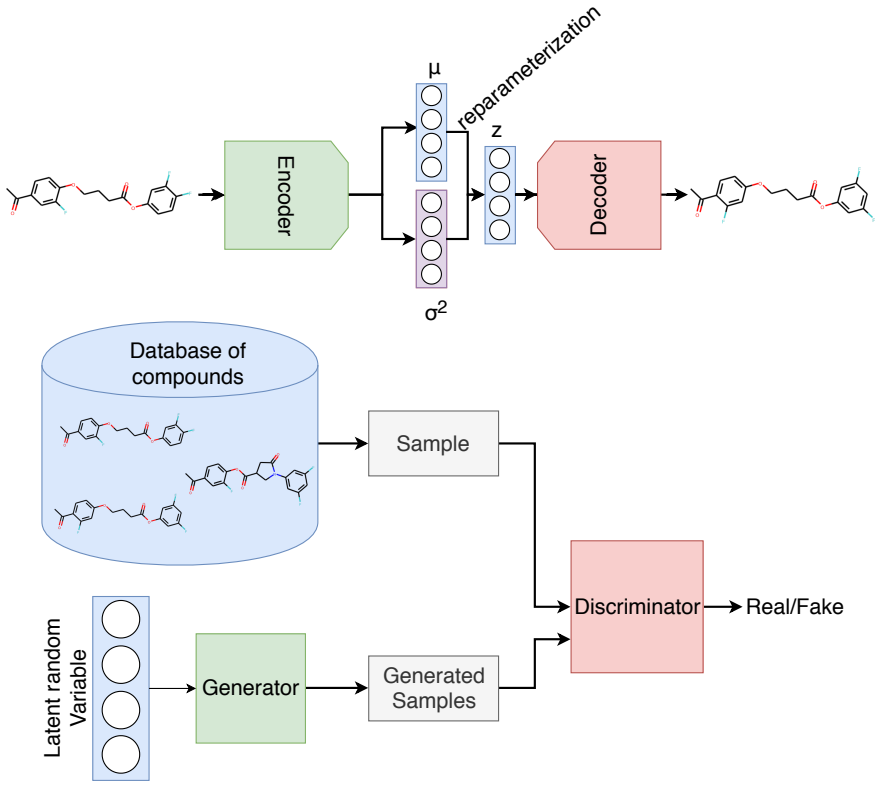

Figure 1.3: Schematic representation of two generative models: variational autoencoder (top) and generative adversarial network (bottom).

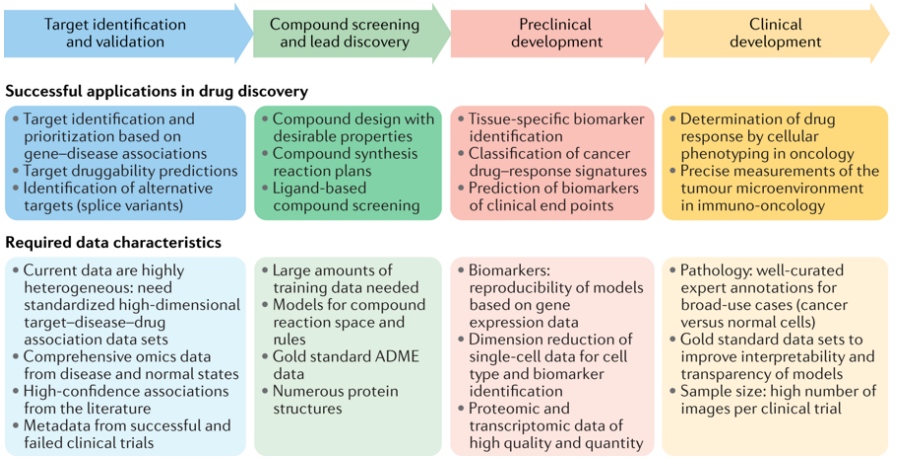

Figure 1.4: Machine learning applications in the drug discovery process and required data characteristics. Figure taken from Vamathevan et. al.[38]

1.2 Challenges of drug design

Developing new drug is a complex process typically taking more than 12 years and costs in excess of one billion dollar.[39] The process can be split into two stages: preclinical stage with R&D and clinical stage. The former stage involves identifying targets, screening of compounds, optimization of promising compounds and testing them in animals. Latter stage includes testing of compounds in humans to evaluate efficacy and safety of the compounds. It is known that in addition to high development costs, there is also a high rate of failure. For example, study by Wong et. al.[40] shows that approval rate of drug approval entering a clinical phase ranges from 3.4% for oncology to 33.4% for infectious disease vaccines. However, throughout the whole process data is being generated and companies, driven by the desire to reduce the costs, want to leverage the obtained data to increase the success rate of future drug development. Hence, for all stages of drug development machine learning algorithms

are being developed and utilized (Figure 1.4). This includes identifying novel targets[41], providing evidence for target-disease associations[42], and predicting success of clinical trial[43], just to name a few.

In this work we focus on the early stages of drug discovery process—lead discovery and lead optimization. for these stages the usage of machine learning methods has a long history. Quantitative structure activity relationship (QSAR), where relationship between target binding ligands and biological activity is modeled, has been used in drug discovery for decades[44, 45], and with advent of deep learning methods in other fields interests data-driven methods for drug discovery has also resurged.[46] However, besides QSAR there are is a variety of tools available to discover novel binding candidates (Figure 1.5).[47] These methods can be divided into structure-based and ligand-based methods[48] and both have potential to be augmented with machine learning approaches.

When designing a drug it is crucial that the ligand binds strongly to a protein target. Binding affinities are usually expressed as equilibrium dissociation constant K_D:

$$K_D = \frac{[R][L]}{[LR]}, \tag{1.4}$$

where $[R]$, $[L]$ and $[LR]$ are concentrations of unbound receptor, unbound ligand and receptor-ligand complex, respectively. The lower the K_D the stronger the protein-ligand binding is. A good starting ligand will have binding affinity in μM range and in the process of lead optimization the goal is to increase the binding affinity to a nM range.

For a successful binding and protein function modulation the ligand must have shape complementary to the protein and form favorable interactions. For example, apolar groups tend to be close, hydrogen bond donors pair with acceptors and charged groups of ligands are frequently neutralized by protein groups of opposite charge.[49] Furthermore, water interactions can play an important role in stabilizing protein-ligand complexes.

During the process of lead optimization it is also important to keep in mind ADME-Tox properties (absorption, distribution, excretion, metabolism,

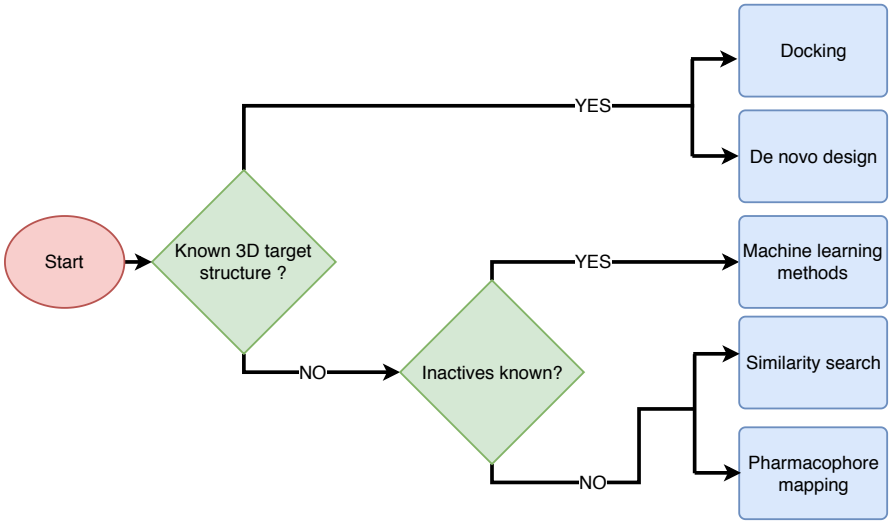

Figure 1.5: Methods available for discovery of novel drug candidates. Which method is employed depends on whether three-dimensional target structure of is known and quality and quantity of data on binding ligands.

and toxicity) so that the final compound will be processed by the body in a desirable way and will have no adverse effects. Prediction and optimization for these properties has also been tackled by deep learning[50, 51], in addition to suggesting chemical syntheses for novel compounds.[52]

1.2.1 Structure-based drug design

Focal point of an early stage drug discovery is identification of lead compounds with pharmacological activity against a target. To this end, experimental screening of large libraries of chemicals against a target has been applied. This process is referred to as high-throughput screening. Through the process compounds that modulate a particular biomolecular pathway can be identified. As High-throughput screening is costly and time-demanding, in silico methods such as structure-based drug design have been proposed as alternative. By utilizing knowledge of the

three-dimensional structure of the biological target it is possible to study protein-ligand molecular interactions using structure-based methods.[53]

Structure-based method can be grouped into two groups: de novo design and virtual screening. For both methods it is critical to obtain a target structure. This is typically done by X-ray crystallography or nuclear magnetic resonance (NMR), although newer methods such as cryogenic electron microscopy[54] are also getting more and more attention. Since 1958 when the first three-dimensional crystal structure of Myoglobin was presented[55], the collection of protein structures is evergrowing[56], making usage data-driven approaches more appealing.

The chemical drug-like space is estimated to consists of between 10^{23} and 10^{60} molecules[57] and only about 10^{8} have ever been synthesized.[58] In order to explore the depth of chemical space structure-based virtual screening projects design virtual libraries that can be orders of magnitude bigger than non-virtual ones. The compounds from these libraries are synthesized only if needed.

Compounds from the virtual libraries are, in a processed called docking[59, 60, 61], fitted into a protein cavity of target protein and evaluated with a scoring function.Typically scoring functions evaluate quality of docked poses and guide the process to to relevant low energy ligand confirmations. Thus design of a good scoring function is critical. Overall, a good scoring function should be able to achieve three things[62]:

- prioritize low energy (experimental) poses
- distinguish active compounds from inactive
- predict absolute binding affinity

Scoring functions can be grouped into three groups[63]: force field-based, knowledge-based and empirical. First group of scoring functions calculates the sum of energy terms from a classical force field, usually considering the interaction energies of the protin-ligand complex and internal ligand energy. Typically, implicit solvent models are used to calculate solvent energy.[64] Knowledge-based scoring functions, on the other hand, are derived from statistical analysis of interacting protein-ligand

atom pairs in known crystal structures.[65, 66] Finally, empirical scoring methods are developed to reproduce experimental affinity data.[67] This group includes machine learning methods.

Although docking and other virtual screening methods, such as molecular dynamics[68], can omit lab experiments, we can still only explore a part of chemical space. De novo drug design approaches, on the other hand, can build up a diverse set of compounds that can accommodate into a protein pocket. These methods can be grouped into two groups: linking and growing. Linking algorithms start from placed fragments in protein pocket and then combine them into compounds. Growing algorithms extend existing fragment and then add, remove or change fragments to improve activity.[69]

1.2.2 Ligand-based drug design

Complementary to structure-based methods computer-aided drug design can leverage information from ligands only. By identify a set of compounds that bind to a particular target, one can extract their structure and search a library for similar compounds or, again, de novo design compounds. Ideally, the goal is to extract from the compounds important physicochemical properties, while discarding extraneous information. The methods can be further divided into machine learning ones (QSAR), similarity search and pharmacophore modeling ones.[70]

Similarity searching is based on theory that structurally similar compounds have similar binding properties.[71] The structure can be further divided grouped into three categories: one-, two- or three-dimensional.[72] One- and two-dimensional methods, such as SMILES strings and structural fingerprints, are efficient at finding close analogs, however they tend to fail at predict activity differences between them.[73] By encoding spatial information it is easier to capture anatomist protein-ligands interactions, even though they might not share similar substrucutre profile. In the following sections we describe some popular shape-based methods.

Ultrafast Shape Recognition (USR)[74] calculates distribution of atomic distances in compounds. To include pharmacophoric feature distribution

of hydrophobic, aromatic, hydrogen bond donor and hydrogen bond acceptor, USR has been extended to USRCAT.[75] The distribution profile can then be used to search for similar compounds.

Gaussian function-based description methods, such as Rapid Overlay of Chemical Structures (ROCS)[76], can be considered a more precise, although computationally demeaning approach. The algorithm searches for optimal alignment of two molecules, query and template, maximizing volume overlap. Other descriptions of shape include surface-based methods such as MSMS[77] and field-based methods[78], that compare how presence of molecules affects other molecules in the space.

Finally, pharmacophore-based modeling tries to model spatial features of molecule that is essential for protein-ligand binding.[79] Pharmacophore models can be build based on structural data[80] or be structure independent. In the latter case different conformations of known ligands are generated, followed by building a consensus model of pharmacophores. Compounds from a database can then be screened against the designed pharmacophore model.

1.2.3 Featurization

One of the core challenges for molecular machine learning is to meaningfully encode molecules so that they can be processed by algorithms, typically this means encoding a molecule into a fixed-length vector. Although SMILES, strings describing structure of chemical species, can be used as molecule representation, most algorithms perform better if informative representation is provided. This holds true especially for linear models, but also for other non-deep learning methods such as support vector machines[81] and random forests[82]. On the other hand, as discussed in first chapter, DL methods can avoid selection of features and make predictions based on complex relations of simple features. Which features in combination with which algorithm performs well is an active area of research. Furthermore, deep learning methods are not limited to a fixed-size vector. Instead, they can accommodate a variety of representations, including variable length SMILES strings or a graph representation.

Thus far several methods of featurization have been proposed and evaluated. Following the work of Wu et. al.[83] (Figure 1.6) this includes:

- Extended-Connectivity Fingerprints.[84] This featurization captures topological characteristics of a molecule by breaking the molecule into substrucutres. Information on substrucutres of the molecule is then hashed into a fixed length fingerprint.
- Coulomb matrix.[85] This matrix holds information on nuclear charges and repulsion between pairs of atoms.
- Grid featurizer.[86, 87] This featurization is designed for protein-ligand complexes. The feature vector consists of ligand fingerprints, fingerprint of protein atoms close to the ligand, fingerprint of atom pairs between protein and ligand atoms and counts of salt bridges, and Hydrogen bonds.
- Symmetry function.[88] This is an atomic coordinates representation, that preserves rotational and permutational symmetry of the system. It includes information on atom distances and angles formed by triplets of atoms.
- Graph convolutions. For each atom a feature vector is computed (e.g. atom type, valance, hybridization) together with a neighbor list. This type or representation is suitable for message passing networks.[89]
- Wave. Similar to wave graph convolutions, but connectivity features contain more information than just pairs of neighbors. Information can include bond properties, distances and ring information.

Complementary to these featurizations, one can also used three-dimensional features such as USR descriptors presented in previous section.

Molecules as images

Here however, we model biological molecules in a way similar to images,

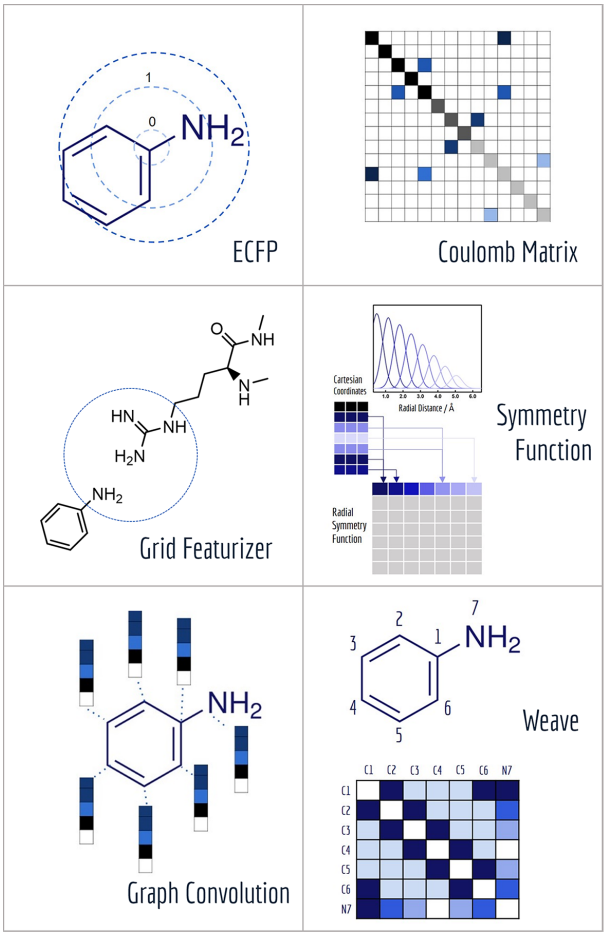

Figure 1.6: Methods for featurization of Molecules. Figure taken from Wu et. al.[83]

but in three dimensions. Biomolecular structures are discretized into a grid, with a fixed resolution (1Å in our case). The fixed size grid (e.g. 24Å for all three sides) is centered around the molecule considered or binding site in case of a protein. Values in the grid are dependent on close by atoms, the distance of an atom from point in grid (r) and van der Waals radius of the considered atom (r_{vdw}), following equation:

$$n(r) = 1 - \exp\left(-\left(r_{\mathrm{vdw}}/r\right)^{12}\right). \tag{1.5}$$

Intuitively, this function interpolates atom densities. If an atom is close to a grid point that point will have a value close to 1 and if it is far the value will be close to 0. Other tackling similar problems have used different representations, mainly binary cutoff[90] or combination of Gaussian and quadratic functions.[91]

Furthermore, as images consist of channels (red, green and blue), so do we use multiple channels to encode atom properties and consequently making training process easier. Atom are assigned to one or multiple channels based on pharmacophore-like properties that are relevant for forming molecular interactions. In the publications we used following channels: hydrophobic, aromatic, hydrogen bond donor, hydrogen bond acceptor and positive or negative partial charge.

1.2.4 Regression and classification for drug design

Once obtaining three-dimensional representations we use convolutional neural networks (CNNs) to process the representation, again, in a way similar to image processing with CNNs. For Publication 3.1[92] and Publication 3.2[93] we downsample feature maps to a vector followed by a single layer perceptron that outputs predicted binding affinity ($-\log K_D$) or probability of bioactivity, respectively. Neural network presented in Publication 3.3[94] does not downsample feature maps, instead the output array has the same dimensions as the input and outputs probabilities

for ligand shape.

Interpreting predictions

Outputs from neural networks undergo multi-layer transformation and thus they are hard to interpret. However, interpretability matters and with obtaining interpretable models we can, according to Selvaraju et.al.[95]: identify model failure modes, establish confidence in users and help us with machine teaching—use models to teach humans make better decisions.

Several methods have been proposed to help visually interpret results. By masking part of the input[96] one can identify salient parts of input. Ragoza et. al.[91] applied this by masking out ligand fragments and they were able to determine which fragments contribute positively or negatively to predicted binding affinity. Other approaches such as guided backpropagation[97], class activation maps (CAM)[98] and grad-CAM[95] have been proven useful in analysis of images.

In Publication 3.1[93] we applied CAM method to highlight for which areas of protein-ligand complex the neural network believes contribute either positively or negatively to bioactivity.

1.2.5 Generative modeling for drug design

Following success of generative models in image, text and music generative models have also been applied to lead generation and optimization. In fact, Elton et. al.[99] counted over 45 publications that use generative models for molecules, all published in recent years. The field took of with work of Gómez-Bombarelli et. al.[100], appearing on arXiv in October 2016. Since then variety of generative models have been proposed. These generative models can be grouped into four categories: recurrent neural networks (RNNs)[101], autoencoders, generative adversarial networks[102] and reinforcement learning methods[103, 104]. Generative models complement genetic algorithms[105] and fragments-based [106] design when building virtual libraries, either diverse or focused.

Recurrent neural networks with SMILES representation are, arguably, the simplest way to generate novel molecules with generative models. By presenting sequence of SMILES tokens as input the network can learn probabilities for next token, conditioned on previous tokens in sequence. For example, Segler et. al.[107] have shown that by training a RNNs on a set of active compounds, one can generate novel compounds, with proven bioactivity.

In publication 3.4[108] we take a slightly different approach. Inspired by image captioning networks[25], we use CNNs to extract a vectorized latent representation of a shape, and feed that into a RNN to generate SMILES strings. by doing string-to-shape encoding, instead of string-to-string, we could capture shape similarity that could be hard to infer from string due to big edit distance. For similar reasons Jin et. al.[109] proposed usage of molecular graphs instead of strings.

Chapter 2

OBJECTIVES

The main objective of this doctorate has been to pioneer the used of deep learning methods to facility drug design process. Specifically, we focused on shape-based modeling where the models are presented with simple 3D representations of molecules. The work can be further divided into two objectives: usage of supervised learning methods, such as regression and classification, and generative modeling. Simplified representation of the carried out work grouped by publications is presented in Figure 2.1.

Developing software was one of the key components and we wanted to ensure accessibility of applications and reproducibility of the results, hence for all publications we made available a web application and for some of them we released the source code.

2.1 Supervised modeling

For supervised predictive modeling we focused on problems of structure-based drug design, where both information from ligands and the target proteins are considered. By presenting spatial information of atoms for both molecules, we want our models to discover arbitrary molecular features that can be either favorable or unfavorable for protein-ligand binding. We apply this to two related problems: classification and regression. In the classification task we train a neural network to distinguish active

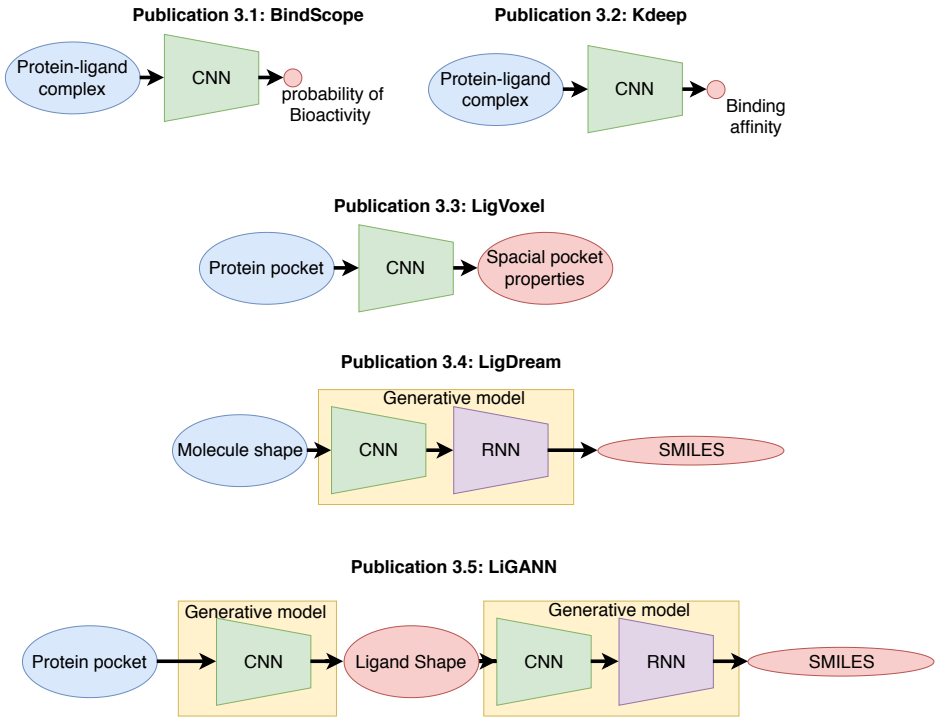

Figure 2.1: Developed predictive and generative models.

ligands from inactives and in regression task our aim is to predict absolute binding affinity of active compounds.

Finally, our objective is also to explore multi-output neural network, where we want to predict spatial properties of ligands for a given protein pocket.

2.2 Generative modeling for de novo compound generation

The second objective was to apply generative modeling to aid in drug design process. We explored the modeling in ligand- and structure-based settings. In the former setting the objective was to, given an input molecule presented as a shape, generate a diverse set of compounds that retain similarity to the starting molecule. For the latter group the objective was to show that it is possible for neural networks to generate ligands with properties complementary to initially presented protein pocket.

To achieve these objectives we drew inspiration from methods develop in the field of image processing and applying them to drug design.

Chapter 3

PUBLICATIONS

3.1 PlayMolecule BindScope: large scale CNN-based virtual screening on the web

M. Skalic, G. Martínez-Rosell, J. Jiménez and G. De Fabritiis. Bioinformatics 35. 1237-1238 (2018).

Summary
Bindscope is a method and a web application for bioactivity prediction in structure-based drug design. Given a protein-ligand complex it leverages 3D convolutional neural networks to discriminate binding compounds from non-binding ones. The network was trained on 101 targets and corresponding known binders and selected decoys from DUD-E database, docked to the target. Compared to previous work we show improved performance by using DenseNet style neural networks, while at the same time reducing complexity of the input representation. The jobs submitted via the web application are processed by a GPU-accelerated computer, thus making it suitable for large-scale screening. Furthermore, the application employs class activation maps to provide an insight into which areas of the complex contribution positively or negatively to the final predicted affinity.

Skalic M, Martínez-Rosell G, Jiménez J, De Fabritiis G. PlayMolecule BindScope: Large scale CNN-based virtual screening on the web. Bioinformatics. 2019 Apr 1;35(7):1237–8. DOI: 10.1093/bioinformatics/bty758

3.2 K_{DEEP}: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks

J. Jiménez, M. Skalic, G. Martínez-Rosell and G. De Fabritiis. Journal of Chemical Information and Modeling 58. 287-296 (2018).

Summary
In this paper we pursue binding affinity prediction, considering $-\log_{10} K_D$ as the target value and leveraging convolutional neural networks in a similar way to previously presented method BindScope. The models were trained and evaluated on PDBbind v.2016 dataset in addition to being evaluated on other datasets such as CSAR and target specific congeneric series. On the standard PDBbind benchmark the method outperformed previously proposed random forrest-based method RF-Score as well as empirical scoring functions. However, for certain targets other methods, such as empirical and machine learning scoring or free energy perturbation, can still outperform the proposed neural network model.

Note: my contribution to this work as second author has been optimization of the procedure, application development as well assisting in preparation of the manuscript.

Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. J Chem Inf Model. 2018 Feb 26;58(2):287–96. DOI: 10.1021/acs.jcim.7b00650

3.3 LigVoxel: inpainting binding pockets using 3D-convolutional neural networks

M. Skalic, A. Varela-Rial, J. Jiménez, G. Martínez-Rosell and G. De Fabritiis. Bioinformatics 35. 243-250 (2018).

Summary
In previous works we have shown that neural networks can be used for regression or classification tasks and here we turn our attention to multi-ouput neural networks. The problem we are trying to solve is inferring ligand shapes from a protein pocket and optionally basic information about the ligand. Given protein structures from scPDB database with well defined pockets we train neural network to map protein shape to ligand shape. We show that the model can capture relevant pharmacophores. Furthermore, we show that the shapes can be used in placing the ligand into the pocket, just by maximizing the overlap between the ligand and the generated shapes.

Skalic M, Varela-Rial A, Jiménez J, Martínez-Rosell G, De Fabritiis G. LigVoxel: Inpainting binding pockets using 3D-convolutional neural networks. Bioinformatics. 2019 Jan 15;35(2):243–50. DOI: 10.1093/bioinformatics/bty583

3.4 Shape-Based Generative Modeling for de-novo Drug Design

M. Skalic, J. Jiménez, D. Sabbadin and G. De Fabritiis. Journal of Chemical Information and Modeling 59. 1205-1214 (2018).

Summary
This work focuses on ligand-based de-novo drug design. The proposed method works similar to image captioning, where a neural network generates sequence of words that describe a given image. In this work, however, we trained networks to generate sequence of SMILES strings, starting from three-dimensional compounds. The networks were trained and evaluated on drug-like ZINC database of compounds. We show that it is possible to generate novel and previously unseen compounds similar to a seed compound and thus explore chemical space of compounds that maintain drug-like characteristics. Furthermore we show in case of several proteins that the model can generate potential binders.

Skalic M, Jiménez J, Sabbadin D, De Fabritiis G. Shape-Based Generative Modeling for de Novo Drug Design. J Chem Inf Model. 2019 Mar 25;59(3):1205–14. DOI: 10.1021/acs.jcim.8b00706

3.5 From Target to Drug: Generative Modeling for Multimodal Structure-Based Drug Design

M. Skalic, D. Sabbadin, B. Sattarov and G. De Fabritiis. Preprint.

Summary
In previous two publications we showed that it is possible to predict ligand properties for a protein pocket and that from ligand shapes we can generate ligand representation as SMILES. In this work we combine these two tasks into an end-to-end pipeline. To produce distinguishable and diverse ligand shapes we apply BiCycleGAN[37] and then use captioning networks to finally generate SMILES strings. We show, using QSAR and docking tools, that there is an enrichment in generating compounds over sampling a virtual library.

Skalic M, Sabbadin D, Sattarov B, Sciabola S, De Fabritiis G. From Target to Drug: Generative Modeling for the Multimodal Structure-Based Ligand Design. Mol Pharm. 2019 Oct 7;16(10):4282–91. DOI: 10.1021/acs.molpharmaceut.9b00634

Chapter 4

DISCUSSION

In this chapter we reflect on the results produced in the publications, describe applicability in drug design pipelines, as well as highlight challenges and future work to be performed.

We have presented a series of applications that use deep learning and can be integrated into a drug design pipeline. In process of lead discovery generative models (Publication 3.4) can be applied to generate a library of compounds, the library can then be docked against a target. The docking can be assisted by property fields generated by method presented in Publication 3.3. By applying method developed in Publication 3.1 compounds in the library can be filtered based on activity and then prioritized by their binding affinity (Publication 3.2). In process of lead optimization, we can again use generative models to find analogs or identify novel scaffolds (Publication 3.4). Alternatively, we have also shown that neural networks can be employed to directly design ligands from protein pocket (Publication 3.5).

Applications BindScope[93] and KDEEP[92] challenge state-of-the art performance on standard evaluation datasets. Although we shown that neural networks can compete with other methods, there are targets on which the methods do not perform well. As these methods are data driven we should be vigilant when applying them. Training set can be skewed and evaluation examples can substantially differ from training, leading

to extrapolation and poorer performance. More recent works describe some of these limitations.[110, 111, 112] As an alternative, physics-based predictive models are being explored, putting less emphasis on training data bias.[113]

We have carried some of the pioneering work of deep learning for drug design, but at the same time a lot of similar research has been done in parallel. AtomNet[90] was the first use case of three-dimensional CNNs for bioactivity classification. Similarly Ragoza et. al.[91] showed usage of CNNs do discriminate between correct and incorrect binding poses. Finally, shortly after publication of Bindscope, Imrie et. al.[114] also proposed CAM-like visualizations.

Arguably, generative models for compound design received even more attention than methods for structure-based virtual screening[99] and a couple of benchmarks have been proposed to evaluate them.[115, 116] However, in drug design at different stages we want to achieve different goals. For example in early stage of lead discovery we want diverse libraries, but in later stage, in lead optimization, we want close analogs to our lead with higher affinity to the target, thus it is hard to evaluate performance of generative models.

In addition to binding affinity, there are three relevant factor for early stage drug discovery that are typically optimized for, but outside the scope of the work present here: synthetic accessibility, pharmacokinetics and off-target activity. Optimization for which have also been tackled by deep learning approaches. For example either by reinforcement learning[117] or Bayesian optimization.[100] Herein proposed generative models take these optimization goals into account only implicitly, e.g. we assume that if a training set consists of synthetically accessible molecules so will the generated ones. However, the proposed shape-based methods can also operate in latent space, and thus optimization through latent space can also be applied to our methods.

When it comes to end-to-end structure-based drug design we have shown that neural networks can achieve this goal and according to QSAR and virtual screening methods generate better compounds than sampling libraries. To the extent of our knowledge this is first attempt at generative

modeling in structure-based setting and with future advances in the field we can expect even more powerful tools and better enrichment.

4.1 Post publication use cases

The present methods are not by any meany finished and wrapped-up projects. All developed applications have been, in collaboration with Accelera labs sl, packed into Singularity[118] containers. With this virtualization the applications can be ran on variety of hardware and operating system combinations. This allows for results reproducibility even on new computational systems.

Furthermore, all applications are available to scientific community as part of `playmolecule.org` platform. Since release over six thousand jobs have been submitted to the web applications (Table 4.1).

Developed methods have also been used on new projects. Testing BindSope on an industrial scale projects reviled that it can challenge commonly used commercial software Glide[60] in in virtual screening campaigns, distinguishing binding compounds from non-binding ones.

KDEEP has was part of best performing solution of D3R challenge, where participants were tasked with ranking compounds based on their binding affinity to BACE protein.[1] The application is also the backbone for a follow-up application that predicts differences in binding affinity for congeneric series.[2]

All in all, drug design still remains a hard problem and we are in early stages in applying deep learning to the problems. However, we do believe that given the complexity of developing novel drugs, deep learning can complement established methods. Work presented is only a small, yet relevant, step towards data-driven solutions.

[1] `https://drugdesigndata.org/php/d3r/gc4/combined/scoringboth/index.php?component=1479&method=combined`

[2] `https://playmolecule.org/DeltaDelta/`

Table 4.1: Developed applications, their access link, number of jobs submitted and publication acceptance date. Number of jobs is extracted at the time of writing (May 27., 2019).

Application	URL	Jobs	Publication
Bindscope	playmolecule.org/BindScope/	1167	29.8.2018
KDEEP	playmolecule.org/Kdeep/	4140	8.1.2018
LigVoxel	playmolecule.org/LigVoxel/	466	6.7.2018
LigDream	playmolecule.org/LigDream/	253	14.2.2019
LiGANN	playmolecule.org/LiGANN/	23	N.A.*

* publication is not accepted yet, web application is undergoing tested.

Chapter 5

CONCLUSIONS

1. Deep learning is shown to be a powerful tool for tackling problems related to drug discovery process. Particularly, three-dimensional shapes have been proven useful as input representation.

2. Convolutional neural networks can be applied to predict bioactivity of small molecules in structure-based settings. Furthermore, regression models can also be applied to predict binding affinity.

3. Convolutions can be used to infer ligand shapes from protein pockets in a purely data-driven setting.

4. Trained captioning-like networks, consisting of convolutional and recurrent neural networks, can be used to variationally decode a compound shape into SMILES strings and thus be applied to explore chemical space focused around a compounds.

5. Generative models have been shown to be a valid option for end-to-end design of ligands in structure-based settings.

Bibliography

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.

[2] Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012.

[3] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. 2012;p. 1097–1105.

[4] Mikolov T, Deoras A, Povey D, Burget L, Černockỳ J. Strategies for training large scale neural network language models. In: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding. IEEE; 2011. p. 196–201.

[5] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016. `http://www.deeplearningbook.org`.

[6] Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science; 1985.

[7] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research. 2011;12(Jul):2121–2159.

[8] Zeiler MD. ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:12125701. 2012;.

[9] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

[10] LeCun Y, Bottou L, Bengio Y, Haffner P, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;86(11):2278–2324.

[11] Dumoulin V, Visin F. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:160307285. 2016;.

[12] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology. 2015;33(8):831.

[13] Roth HR, Shen C, Oda H, Oda M, Hayashi Y, Misawa K, et al. Deep learning and its application to medical image segmentation. Medical Imaging Technology. 2018;36(2):63–71.

[14] Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:12070580. 2012;.

[15] Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. 2010;p. 807–814. Available from: http://dl.acm.org/citation.cfm?id=3104322.3104425.

[16] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014;.

[17] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

[18] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. 2017;p. 4700–4708.

[19] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167. 2015;.

[20] Shi S, Wang Q, Xu P, Chu X. Benchmarking state-of-the-art deep learning software tools. In: 2016 7th International Conference on Cloud Computing and Big Data (CCBD). IEEE; 2016. p. 99–104.

[21] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv preprint arXiv:14085093. 2014;.

[22] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. Software available from tensorflow.org. Available from: `https://www.tensorflow.org/`.

[23] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in pytorch. 2017;.

[24] Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:150600019. 2015;.

[25] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3156–3164.

[26] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9:1735–1780.

[27] Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:13126114. 2013;.

[28] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. Advances

in Neural Information Processing Systems 27. 2014;p. 2672–2680. Available from: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[29] Kullback S, Leibler RA. On information and sufficiency. The annals of mathematical statistics. 1951;22(1):79–86.

[30] Arjovsky M, Chintala S, Bottou L. Wasserstein Generative Adversarial Networks. Proceedings of the 34th International Conference on Machine Learning. 2017 06–11 Aug;70:214–223. Available from: http://proceedings.mlr.press/v70/arjovsky17a.html.

[31] Wei X, Liu Z, Wang L, Gong B. Improving the Improved Training of Wasserstein GANs. International Conference on Learning Representations. 2018;p. 5767–5777. Available from: https://openreview.net/forum?id=SJx9GQb0-.

[32] Karras T, Aila T, Laine S, Lehtinen J. Progressive Growing of GANs for Improved Quality, Stability, and Variation; 2018. Available from: https://openreview.net/forum?id=Hk99zCeAb.

[33] Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral Normalization for Generative Adversarial Networks. International Conference on Learning Representations. 2018;Available from: https://openreview.net/forum?id=B1QRgziT-.

[34] Zhang H, Goodfellow I, Metaxas D, Odena A. Self-attention generative adversarial networks. arXiv preprint arXiv:180508318. 2018;.

[35] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2017;p. 5967–5976.

[36] Zhu JY, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Computer Vision (ICCV), 2017 IEEE International Conference on Computer Vision. 2017;p. 2223–2232.

[37] Zhu JY, Zhang R, Pathak D, Darrell T, Efros AA, Wang O, et al. Toward Multimodal Image-to-Image Translation. Advances in Neural Information Processing Systems 30. 2017;p. 465–476. Available from: http://papers.nips.cc/paper/6650-toward-multimodal-image-to-image-translation.pdf.

[38] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. 2019;p. 1.

[39] Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug discovery. 2010;9(3):203.

[40] Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019;20(2):273–286.

[41] Jeon J, Nim S, Teyra J, Datti A, Wrana JL, Sidhu SS, et al. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. Genome medicine. 2014;6(7):57.

[42] Ferrero E, Dunham I, Sanseau P. In silico prediction of novel therapeutic targets using gene–disease association data. Journal of translational medicine. 2017;15(1):182.

[43] Gayvert KM, Madhukar NS, Elemento O. A data-driven approach to predicting successes and failures of clinical trials. Cell chemical biology. 2016;23(10):1294–1301.

[44] Zupan J, Gasteiger J. Neural networks for chemists: an introduction. John Wiley & Sons, Inc.; 1993.

[45] Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014;57(12):4977–5010.

[46] Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure–activity relationships. Journal of chemical information and modeling. 2015;55(2):263–274.

[47] Leelananda SP, Lindert S. Computational methods in drug discovery. Beilstein journal of organic chemistry. 2016;12(1):2694–2718.

[48] Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. British journal of pharmacology. 2007;152(1):9–20.

[49] Williams MA. Protein–ligand interactions: Fundamentals. In: Protein-Ligand Interactions. Springer; 2013. p. 3–34.

[50] Wenzel J, Matter H, Schmidt F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. Journal of chemical information and modeling. 2019;.

[51] Hop P, Allgood B, Yu J. Geometric deep learning autonomously learns chemical features that outperform those engineered by domain experts. Molecular pharmaceutics. 2018;15(10):4371–4377.

[52] Segler MH, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature. 2018;555(7698):604.

[53] Lionta E, Spyrou G, K Vassilatis D, Cournia Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. Current topics in medicinal chemistry. 2014;14(16):1923–1938.

[54] Cheng Y, Grigorieff N, Penczek PA, Walz T. A primer to single-particle cryo-electron microscopy. Cell. 2015;161(3):438–449.

[55] Kendrew JC, Bodo G, Dintzis HM, Parrish R, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature. 1958;181(4610):662–666.

[56] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic acids research. 2000;28(1):235–242.

[57] Polishchuk PG, Madzhidov TI, Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data. Journal of computer-aided molecular design. 2013;27(8):675–679.

[58] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic acids research. 2015;44(D1):D1202–D1213.

[59] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry. 2010;31(2):455–461.

[60] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Journal of medicinal chemistry. 2004;47(7):1739–1749.

[61] Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. Journal of molecular biology. 1997;267(3):727–748.

[62] Guedes IA, Pereira FS, Dardenne LE. Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. Frontiers in pharmacology. 2018;9:1089.

[63] Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking. Journal of medicinal chemistry. 2003;46(12):2287–2303.

[64] Zou X, Sun Y, Kuntz ID. Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. Journal of the American Chemical Society. 1999;121(35):8033–8043.

[65] Velec HF, Gohlke H, Klebe G. DrugScoreCSD knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. Journal of medicinal chemistry. 2005;48(20):6296–6303.

[66] Muegge I. PMF scoring revisited. Journal of medicinal chemistry. 2006;49(20):5895–5902.

[67] Pason LP, Sotriffer CA. Empirical Scoring Functions for Affinity Prediction of Protein-ligand Complexes. Molecular informatics. 2016;35(11-12):541–548.

[68] Zhao H, Caflisch A. Molecular dynamics in drug design. European journal of medicinal chemistry. 2015;91:4–14.

[69] Sliwoski G, Kothiwale S, Meiler J, Lowe EW. Computational methods in drug discovery. Pharmacological reviews. 2014;66(1):334–395.

[70] Acharya C, Coop A, E Polli J, D MacKerell A. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. Current computer-aided drug design. 2011;7(1):10–22.

[71] Verma J, Khedkar VM, Coutinho EC. 3D-QSAR in drug design-a review. Current topics in medicinal chemistry. 2010;10(1):95–115.

[72] Shin WH, Zhu X, Bures M, Kihara D. Three-dimensional compound comparison methods and their application in drug discovery. Molecules. 2015;20(7):12841–12862.

[73] Mavridis L, Hudson BD, Ritchie DW. Toward high throughput 3D virtual screening using spherical harmonic surface representations. Journal of chemical information and modeling. 2007;47(5):1787–1796.

[74] Ballester PJ, Richards WG. Ultrafast shape recognition to search compound databases for similar molecular shapes. Journal of computational chemistry. 2007;28(10):1711–1723.

[75] Schreyer AM, Blundell T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. Journal of cheminformatics. 2012;4(1):27.

[76] Hawkins PC, Skillman AG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. Journal of medicinal chemistry. 2007;50(1):74–82.

[77] Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers. 1996;38(3):305–320.

[78] Cheeseright TJ, Mackey MD, Melville JL, Vinter JG. FieldScreen: virtual screening using molecular fields. Application to the DUD data set. Journal of chemical information and modeling. 2008;48(11):2108–2117.

[79] Yang SY. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug discovery today. 2010;15(11-12):444–450.

[80] Wolber G, Langer T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. Journal of chemical information and modeling. 2005;45(1):160–169.

[81] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–297.

[82] Barandiaran I. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence. 1998;20(8).

[83] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chemical science. 2018;9(2):513–530.

[84] Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of chemical information and modeling. 2010;50(5):742–754.

[85] Rupp M, Tkatchenko A, Müller KR, Von Lilienfeld OA. Fast and accurate modeling of molecular atomization energies with machine learning. Physical review letters. 2012;108(5):058301.

[86] Durrant JD, McCammon JA. NNScore 2.0: a neural-network receptor–ligand scoring function. Journal of chemical information and modeling. 2011;51(11):2897–2903.

[87] Da C, Kireev D. Structural protein–ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. Journal of chemical information and modeling. 2014;54(9):2555–2561.

[88] Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Physical review letters. 2007;98(14):146401.

[89] Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org; 2017. p. 1263–1272.

[90] Wallach I, Dzamba M, Heifets A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:151002855. 2015;.

[91] Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein–ligand scoring with convolutional neural networks. Journal of chemical information and modeling. 2017;57(4):942–957.

[92] Jiménez J, Skalic M, Martinez-Rosell G, De Fabritiis G. K DEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. Journal of chemical information and modeling. 2018;58(2):287–296.

[93] Skalic M, Martínez-Rosell G, Jiménez J, De Fabritiis G. PlayMolecule BindScope: large scale CNN-based virtual screening on the web. Bioinformatics. 2018;.

[94] Skalic M, Varela-Rial A, Jiménez J, Martínez-Rosell G, De Fabritiis G. LigVoxel: inpainting binding pockets using 3D-convolutional neural networks. Bioinformatics. 2018;35(2):243–250.

[95] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 618–626.

[96] Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection. In: Advances in neural information processing systems; 2013. p. 2553–2561.

[97] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:14126806. 2014;.

[98] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of

the IEEE conference on computer vision and pattern recognition; 2016. p. 2921–2929.

[99] Elton DC, Boukouvalas Z, Fuge MD, Chung PW. Deep learning for molecular generation and optimization-a review of the state of the art. arXiv preprint arXiv:190304388. 2019;.

[100] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS central science. 2018;4(2):268–276.

[101] Gupta A, Müller AT, Huisman BJ, Fuchs JA, Schneider P, Schneider G. Generative recurrent networks for de novo drug design. Molecular informatics. 2018;37(1-2):1700111.

[102] Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. arXiv preprint arXiv:170510843. 2017;.

[103] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Science advances. 2018;4(7):eaap7885.

[104] Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. Journal of cheminformatics. 2017;9(1):48.

[105] Yoshikawa N, Terayama K, Sumita M, Homma T, Oono K, Tsuda K. Population-based de novo molecule generation, using grammatical evolution. Chemistry Letters. 2018;47(11):1431–1434.

[106] Lessel U. Fragment-based design of focused compound libraries. Scaffold Hopping in Medicinal Chemistry. 2014;.

[107] Segler MH, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS central science. 2017;4(1):120–131.

[108] Skalic M, Jiménez Luna J, Sabbadin D, De Fabritiis G. Shape-Based Generative Modeling for de-novo Drug Design. Journal of chemical information and modeling. 2019;.

[109] Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation. arXiv preprint arXiv:180204364. 2018;.

[110] Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. Journal of chemical information and modeling. 2018;58(5):916–932.

[111] Chen L, Cruz A, Ramsey S, Dickson C, Duca JS, Hornak V, et al. Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. 2019;.

[112] Smusz S, Kurczab R, Bojarski AJ. The influence of the inactives subset generation on the performance of machine learning methods. Journal of cheminformatics. 2013;5(1):17.

[113] Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A. Quantum-chemical insights from deep tensor neural networks. Nature communications. 2017;8:13890.

[114] Imrie F, Bradley AR, van der Schaar M, Deane CM. Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data. Journal of chemical information and modeling. 2018;58(11):2319–2330.

[115] Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, et al. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. arXiv preprint arXiv:181112823. 2018;.

[116] Brown N, Fiscato M, Segler MH, Vaucher AC. GuacaMol: Benchmarking Models for de Novo Molecular Design. Journal of chemical information and modeling. 2019;.

[117] Ståhl N, Falkman G, Karlsson A, Mathiason G, Boström J. Deep Reinforcement Learning for Multiparameter Optimization in de novo Drug Design. 2019;.

[118] Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PloS one. 2017;12(5):e0177459.