Machine learning in structural biology and chemoinformatics

Driving drug discovery one epoch at a time

José Jiménez Luna

TESI DOCTORAL UPF / year 2019

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

A learning experience is one of those things that says:
"You know that thing you just did? Don't do that."

–DOUGLAS ADAMS

Real stupidity beats artificial intelligence every time.

—TERRY PRATCHETT - Hogfather (Discworld # 20, 1996)

Sometimes it seems as though each new step towards artificial intelligence, rather than producing something which everyone agrees is real intelligence, merely reveals what real intelligence is not.

–DOUGLAS HOFSTADTER - Gödel, Escher, Bach: An Eternal Golden Braid (1979)

Acknowledgements

Well, this was quite the trip. And yet, it all seems so short now: I came to meet the people in this lab back in the late 2015, with the only hope of finding a part-time job that would help me pay the bills while I was finishing my master's degree. Me, a statistician, a guy who was used to writing proofs and reporting p-values, and without barely any biology or chemistry knowledge, was strolling into a computational biophysics laboratory like it was the most normal thing in the world. Suddenly, the next thing I knew is that I was doing a PhD. And boy, was I scared: back then it all sounded Greek to me. But I *liked* my project, so I got out of my comfort zone and started bothering my supervisor and peers with what probably were at the time extremely stupid questions. Their patience shall probably never be fully repaid.

I owe it all to you. To Gianni, for entrusting me with such an opportunity and his support. To my colleagues and friends in the lab and Acellera, with whom I have had the luxury of sharing countless hours and laughs. The environment we together had created made me not even realize I was doing work. I will cherish these memories for the rest of my life.

Abstract

Deep learning approaches have become increasingly popular in the last years thanks to their state-of-the-art performance in fields such as computer vision and natural language understanding. The first goal of this thesis was to adapt such approaches, and particularly those used in image recognition, to the domains of structural biology and chemoinformatics. We do so by the development of a novel three-dimensional biomolecular representation that can be used in conjunction with 3D-convolutional neural networks for a variety of tasks. We test the applicability of such methods in several relevant problems in the early drug discovery pipeline, such as protein binding site prediction, protein-ligand binding affinity prediction, drug selectivity elucidation and molecular generative models. The second goal of this thesis was to facilitate the use and accessibility of such tools by their implementation and deployment in an easy-to-use web application.

Resum

Els mètodes d'aprenentatge profund han guanyat molta popularitat en els últims anys gràcies al seu rendiment en camps com la visió per ordinador o l'aprenentatge del llenguatge natural. El primer objectiu de la tesi va ser adaptar aquests mètodes, particularment els utilitzats en el reconeixement d'imatges, als camps de la biologia estructural i la quimioinformàtica. L'adaptació s'ha fet mitjançant el desenvolupament d'una representació biomolecular tridimensional que pot ser utilitzada en conjunt amb xarxes neuronals convolucionals tridimensionals en diverses tasques. Hem testat l'aplicabilitat d'aquests mètodes en varis problemes rellevants per als primers estadis de desenvolupament de drogues, com la predicció de la zona d'unió de proteïnes, l'afinitat d'unió entre proteïna i lligand, la elucidació de selectivitat de drogues i models generatius de molècules. El segon objectiu de la tesi ha sigut el facilitar la utilització i l'accessibilitat d'aquestes eines mitjançant la seva implementació i desplegament en una aplicació web de fàcil ús.

Preface

Research on deep learning approaches such as convolutional and recurrent neural networks drew a lot of attention in the early 2010s, and such is not undeserved: they provided a significant performance leap in areas such as in computer vision, natural language understanding or the development of the self-driving car.

Back in 2016 when I first started my PhD studies, I remember being astonished by these advances, and by how fast these were being deployed and affecting our everyday lives. Moreover, my supervisor and I worryingly suspected that it would be only a matter of time until these modern machine-learning approaches were used in other research fields, therefore opening a lot of opportunities, but only for those able to seize them quickly enough. Overall, time has proven that we were certainly not wrong: the past five years have experienced an explosion of deep-learning applications in most of the related areas to bioinformatics and chemoinformatics.

Given that both my background in statistics and the timing were fitting, we decided that it was a good idea to go with the flow: in the research presented here I tried my best to bridge the gap between deep-learning models and several problems relevant in drug discovery, such as protein binding site prediction, protein-ligand affinity prediction, compound selectivity elucidation and generative models of drugs. Most of the research carried out during these studies has involved a lot of frustration and failures, but also a good deal of very lucky successes. Only the latter are presented here, while the former I have certainly learned from.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 The machine learning context

We are generating data at an unprecedented rate in human history. For instance, more than 300 hours of video are uploaded to YouTube every minute [1], Amazon handles more than 400 orders per second in holiday season [2] or more than 100k whole human genomes have been fully sequenced [3]. This vast amount of information therefore called for the development of machine learning models: their use nowadays ubiquitous in modern society as we enter an era of big data.

Machine learning is the field occupied with the development of general-purpose approaches that directly learn patterns from data without explicit functional pre-specification and their use in future prediction and smart decision making [1]. Arguably, these methodologies can be classified into several paradigms. In **supervised learning** we are interested in fit-

[1] http://www.everysecond.io/youtube

[2] https://www.theverge.com/2013/12/26/5245008/amazon-sees-prime-spike-in-2013-holiday-season

[3] https://www.broadinstitute.org/news/broad-institute-sequences-its-100000th-whole-human-genome-national-dna-day

ting a function $f : \mathcal{X} \to \mathcal{Y}$ using n training data points from a set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Typical applications include regression and classification tasks. In **unsupervised learning**, our data consists of only inputs $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{n}$, and our goal is to find some notion of internal structure. Clustering, anomaly detection and dimensionality reduction techniques fall into this category. **Reinforcement learning**, on the other hand, deals with the task of an finding optimal policy for an agent on a environment, given some notion of reward.

For the work presented in this thesis we mostly are concerned with the supervised and unsupervised paradigms: particularly on deep-learning-based models and their application at several stages of the drug discovery pipeline. The thesis is structured as follows: first we provide motivation for deep learning methods, to then explain their potential use in relevant drug-discovery problems. Finally we discuss in detail all the particular challenges studied and corresponding machine-learning applications developed throughout this thesis.

1.1.1 A brief history of artificial neural networks

Among the plethora of machine-learning approaches, deep learning methods, those based on artificial neural networks (ANNs) [2–4], have become increasingly popular over the last few years. The impact these models are having in the present world is unquantifiable: they have become de-facto a common tool in many scientific fields, such as computer vision [5–9], natural language understanding [10–15], speech recognition [16–19], recommender systems [20, 21] and the development of self-driving vehicles [22, 23], in many cases surpassing human levels of performance.

One may think that the aforementioned approaches are novel, but in fact, the earliest artificial neural network, the perceptron, was developed by Rosenblatt back in the 1950s [24]. Moreover, early drafts of back-propagation, a very popular technique through which neural networks are trained, were derived in the context of control theory in the early

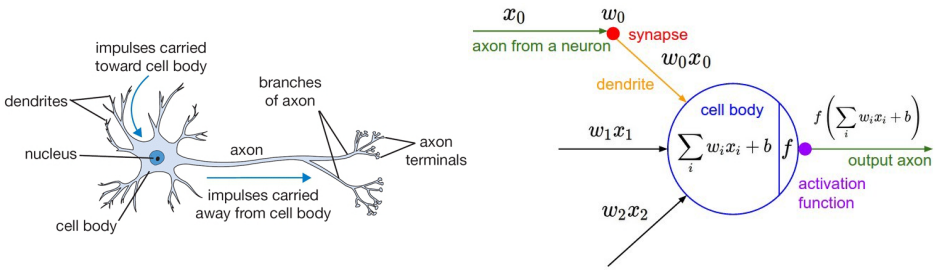


Figure 1.1: A comparison between biological (left) and artificial (right) neural networks. Neurons transmit information via dendrites, and if activated will further transmit its non-linear transformation to other neighboring neurons.

1960s [25], although their significance was not properly recognised until the works of Hinton in the 1980s [26, 27]. It would not be until the early 2010s, however, when the potential of artificial neural networks was finally unhindered.

1.1.2 Deep learning and representations

Anatomy of an ANN

An artificial neural network is a machine learning model that takes inspiration from their biological counterpart (Figure 1.1, taken from Stanford's CS231 notes [4]). A neuron ϕ receives a linear combination of learnable weights $\boldsymbol{w}$ and bias b and the output of preceding neurons $\boldsymbol{x}_i$ and applies a (typically non-linear) activation function f. This result is then passed to subsequent neurons:

$$\phi = f\left(\sum_i w_i x_i + b\right). \tag{1.1}$$

[4] http://cs231n.github.io/neural-networks-1/

In regular feed-forward neural networks, neurons are sequentially organized in layers, information passed from the previous to the next until an output of a desired dimensionality is obtained. Layers that are either not the input or the output ones are named *hidden*, and a neural network with more than a single hidden layer is said to be *deep*. A *loss function* $\mathcal{L}$ is used at the output of the network and compared with the real targets y so as to quantify how close these are to each other. In order to adjust the learnable parameters at each layer, we use a gradient-descent-based approach named backprogatation [28, 29], that iteratively updates parameters through the chain rule. As a motivating example, assume a regression setup with a single layer for examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, learnable parameters $\boldsymbol{w}$ and the mean squared error as loss function $\mathcal{L}$:

$$\mathcal{L}\left(\boldsymbol{x}_i, \boldsymbol{w}\right) = \frac{1}{2}\left\|N(\boldsymbol{x}_i, \boldsymbol{w}) - y_i\right\|^2, \tag{1.2}$$

where $\mathcal{L}\left(\boldsymbol{x}_i, \boldsymbol{w}\right)$ is the loss for example i between the output of the network $N(\boldsymbol{x}_i, \boldsymbol{w}) = f\left(\boldsymbol{x}_i^t \boldsymbol{w}\right)$ and its corresponding targets. The expression can be averaged over a batch of examples:

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(\boldsymbol{x}_i, \boldsymbol{w}\right). \tag{1.3}$$

We are interested in finding the partial derivative of the loss w.r.t. learnable parameters, so that they can be updated using standard gradient descent:

$$\boldsymbol{w}^{\text{new}} = \boldsymbol{w} - \eta \frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}}, \tag{1.4}$$

where $\eta \in (0, 1)$ is the learning rate. Furthermore, for the gradient to be computable, it is necessary that the loss and activation functions in the network are differentiable. Up to recently, the gradients needed to be symbolically derived and then coded for each particular neural network model, but thanks to the advent to automatic differentiation packages [30, 31] such as Tensorflow or PyTorch [32, 33] this is no longer necessary, tremendously facilitating the design of complex architectures.

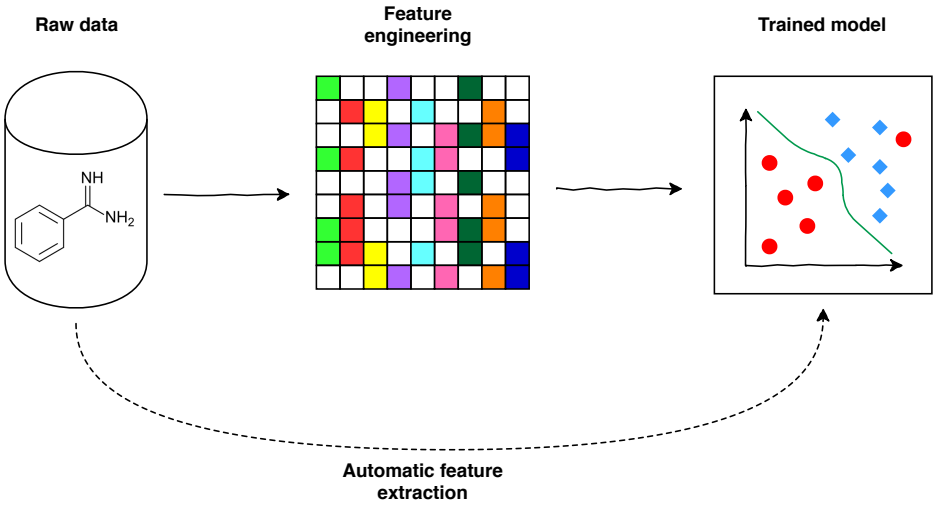


Figure 1.2: Contrary to other machine learning approaches, which require users to engineer their own sets of features, deep learning procedures are able to extract valuable features from a closer representation to the raw data itself.

Automatic feature extraction: going deep

There were several reasons why artificial neural networks fell out of favour against other popular supervised machine or statistical learning algorithms, such as random forests [34] or support vector machines [35,36] in the last 50 years. Mostly these come from the inability to train deep models, those that feature more than a single hidden layer, which was partially caused by what is known as the exploding or vanishing gradient problem [37]: when backpropating errors in deep architectures, partial derivatives have a tendency to be numerically unstable, rendering training impossible. This issue is nowadays partially solved by modern activation functions, such as ReLU (i.e. $r(x) = \max(0, x)$), and architectures such as ResNet [38]. The other reason is computational efficiency: training deep models is very expensive compared to other well-performing yet simpler alterna-

tives, and it would not be until the advent of graphical processing units (GPUs) that this issue would be mitigated.

The training of neural networks was confined to shallow models up to recently, but there is good reason why one would be interested in deep architectures. The main reason why that deep learning approaches have revolutionized many research fields is because of their ability to perform automatic feature extraction [39–41] (Figure 1.2). A typical machine-learning workflow involved the computation of handcrafted descriptors on which a model is trained later, their choice being completely problem-specific, and the performance of such model heavily depending on such representation. Deep learning models however, learn features hierarchically from a closer representation of data itself, higher layers representing more abstract concepts. For instance, in the case of computer vision, convolutional neural networks [6] work directly at the pixel level of data, extracting the most relevant features in each picture so as to maximize predictive performance, with earlier layers serving the purpose of edge detectors, and the higher representing more complex concepts (such as the nose of a dog). Alternatively, in the case of natural language processing, recurrent neural networks [42, 43] can work directly with text data to predict, for instance, which is the most probable word to follow an incomplete sentence. This ability to work with a closer representation of the real modelled data, instead of requiring the practitioner to manually extract its own set of features, resulted in deep learning approaches achieving state-of-the-art performance in many problems.

1.2 The promise of deep learning in drug discovery

1.2.1 A mentality shift

Machine learning (ML) is barely a newcomer to the related subfields of drug discovery, such as chemoinformatics or structural biology. A long

tradition stems from the early quantitative structure activity relationship models first reported in the early 1960s [44], which became commonplace in a computational chemist's toolbox [45–49]. Traditionally, machine learning and classical statistical approaches have never been an easy subject in the aforementioned fields, as they involve the description of complex entities, such as molecules, through a one-dimensional vector that can later be used for modeling [50]. In fact, hundreds of descriptors have been developed in the context of molecular property prediction alone [51–58].

Given the success of early deep learning approaches in other fields, researchers did not wait to explore their applicability in all the stages of drug discovery, and in fact, such is the case in the thesis presented here. While machine learning models can be deployed at all stages of the drug discovery pipeline (Figure 1.3), a lot of effort is currently being spent in the earlier stages, those dedicated to target identification as well as molecular generation and property prediction.

The success of ML-based models in the field has gained momentum in the last years, encouraged by several early promising results. In 2013, deep neural networks models were the top performing ones in the Merck molecular activity challenge [59] and in 2015 similar results were obtained in the Tox21 toxicity data challenge [60]. Furthermore, deep learning approaches implied that models were no longer restricted to traditional data types, such as compound fingerprints, but could also extend to the structure of proteins, images or transcriptomics. Driven by these successes, many of the major pharmaceutical companies already have began to explore machine learning initiatives [5], in some cases with the collaboration of IT giants such as Google [6].

[5] https://emerj.com/ai-sector-overviews/ai-in-pharma-and-biomedicine/

[6] https://www.slashgear.com/google-and-pharma-company-sanofi-team-up-for-big-data-processing-18580858/

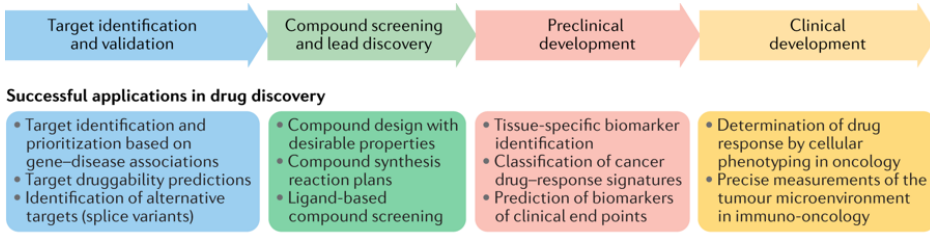


Figure 1.3: Drug discovery pipelines and corresponding machine learning applications at each of the stages. Taken and adapted from Vamathevan *et al.* [65]

1.2.2 A modern deep-learning representation for biomolecular complexes

Moving beyond feature engineering

Traditional descriptor sets for structural biology have traditionally fallen into the feature engineering mindset. That means that researchers tried to come up with one-dimensional descriptions or representations of what in reality is a three-dimensional physical object. Examples of such descriptions include, among many others, atom type or aminoacid counts at different distance thresholds [61, 62], protein-ligand interaction fingerprints [63] or pharmacophoric descriptors [64].

Naturally, these representations represent a simplification of the problem in order to accomodate the standard supervised learning framework of predicting a scalar variable using a vector of features [66]. While these strategies were certainly useful, and many researchers developed creative solutions to adapt features to said paradigm, the modeling of three-dimensional structures, such in the case of proteins remained far from ideal. Drawing parallels with the computer vision field, it was also common practice to engineer features to the standard machine learning paradigm, using descriptors such the scale-invariant feature transform [67], speeded up robust features [68] or the histogram of oriented

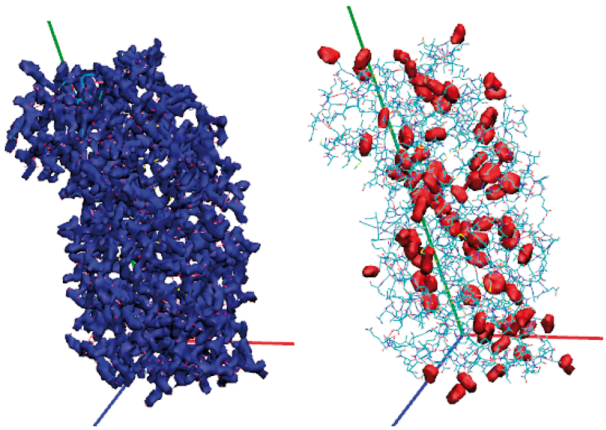


Figure 1.4: Example of descriptor computation output for the hydrophobic and aromatic channels, respectively for PDB Id 4NIE. Taken from Pub. 1 [77]

gradients [69]. With the popularization of convolutional neural networks and their clear victory in the well-known image classification challenge ImageNet [7], their usage became common both in academia and industry. Furthermore, the flexibility of these gradient-based approaches allowed the training of models that go beyond simple image classification, but also segmentation [70, 71], colorization [72], stereo conversion [73] and high-dimensional generative modeling [74–76], among many others.

Drawing inspiration from these advances in computer vision, we started adapting similar models for biomolecular complexes. On one hand, images are typically represented as three two-dimensional matrices (or *channels*), each of them representing the color intensity at each pixel location (i.e. RGB color space). Proteins on the other hand are represented by a collection of atoms and their coordinates in three-dimensional space, so the first question became how to adapt current computer vision know-how in this context

A natural idea is to first discretize the space into equally spaced $k\text{Å}^3$

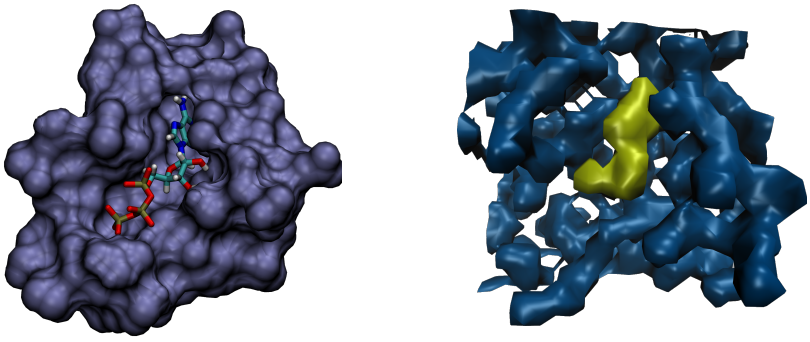
voxels. Then, a compendium of properties (the aforementioned channels) can be computed for each of these voxels, predefined by a simple atom typing. For most the work presented in this thesis we have used pharmacophoric-like features (hydrophobic, aromatic, hydrogen-bond donor or acceptor, positive or negatible ionizable and metals). Direcly translating atoms into volumetric space results in a very sparse representation, so an atomic influence to each voxel location is computed following a pair correlation function which depends on their euclidean distance r:

$$n(r) = 1 - \exp\left(-\left(\frac{r_{\text{vdw}}}{r}\right)^{12}\right), \tag{1.5}$$

where r_{vdw} is the Van der Waals radius of the atom in question. This description can also be seen as a distance-based interpolation. A simple atomic occupancy channel, is also typically included to account for explicit geometric information of the complex. In terms of the choice of atom typing it is been common in the literature to see the use of those provided by either the AutoDock 4 [78] or Smina [79] software packages. Other authors have chosen a different distance-based functional forms for the interpolation function [80]. An example of the proposed representation for a protein can be checked in Figure 1.4. The latter is general for biomolecular complexes and therefore is able to describe other entities such as small compounds, a fact that can be exploited to model, for instance, protein-ligand interactions. An example of the voxelization of an interaction between a small compound and a protein pocket can be seen in Figure 1.5.

3D-convolutional neural networks

Once a representation similar to the one images has been developed, the deep-learning techniques commonly used in computer vision, such as three dimensional convolutional neural networks become readily applicable. These are specifically designed to work with spatial inputs (i.e. voxels), by trying to emulate the response of an individual neuron to vi-



(a) (b)

Figure 1.5: (a) PDB Id 2HMU pocket and bound ligand ATP. (b) Voxel representation of the hydrophobic channel for both protein (blue) and ligand (yellow). Taken from Pub. 2 [81]

sual stimuli [82]. This fact allows us to encode certain properties into the architecture that cannot be assumed in regular, fully connected networks. Concretely, the layers of a three-dimensional neural network are arranged in 4 dimensions: height, width, depth and number of channels, with each neuron only locally connected to a localized region of the preceding layer. In the case of voxels, the output of a neuron, a feature map ϕ, is a three dimensional tensor, obtained through discrete convolution of a filter W_i over an input feature map $z_i(x, y, z)$:

$$\phi = f\left(\sum_i W_i * z_i(x, y, z) + b\right), \tag{1.6}$$

where $*$ represents a three-dimensional discrete convolution operation [83]:

$$w * f(x, y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} \sum_{l=-c}^{c} w(s, t, l) f(x - s, y - t, z - l). \tag{1.7}$$

In Figure 1.6 we show an example of the the multiplication of a

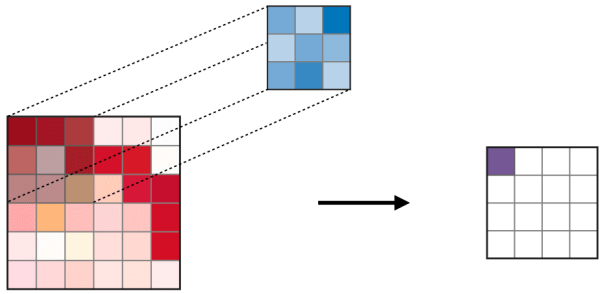


Figure 1.6: Example of a single convolution operation over an input feature map z, which is element-wise multiplied by a learnable filter W to obtain an output of arbitrary size.

single filter with an image patch[7]. The connectivity of a feature map is controlled by a parameter named kernel or filter size, and such is only defined across the spatial dimensions, while full connectivity is applied to all feature channels. Parameters in a convolutional neural network are also typically shared in the channel dimension: if one feature is useful for a particular position in the image, it should also be for a different one. This simplification results in a significant reduction of learnable parameters, which in turn makes current implementations more computationally approachable at a scale.

Apart from convolution, other layers are commonly used in the development of this type of neural networks. For instance, pooling layers apply a non-learnable transformation to an input, typically reducing its size in order to simplify further calculations and reduce the number of parameters in the network, operating independently on each channel. Normalization layers, such as batch normalization [84], ensure the weight distribution remains similar across batches, while dropout layers [85] randomly drop neuron connections with the intent of avoiding overfitting.

[7]Taken from `https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks`

The size of the three-dimensional filters and the output of the network can be freely modified depending on the task. In typical classification or regression problems, for instance, the size of the feature maps is reduced enough so that it can be flattened out to a vector, and regular fully connected layers are then applied to ensure an output of size one.

Limitations

The voxelization of proteins represent a natural representation of their three dimensional structure in space, but it suffers from several issues. In particular, three dimensional arrays take significantly more space in memory than images in computer vision applications. Furthermore, voxelizing only a part of the protein, such as a binding site, entails choosing a window [81, 86, 87], since the shape of the arrays needs to be fixed for their use in CNN architectures, while protein pockets and ligands can significantly differ in size [88, 89]. Finally, the representation is neither rotationally nor translationally invariant [90, 91], properties which are desirable when modelling biomolecular complexes with standard convolutional neural networks.

1.3 Applications studied

In this section I summarize the areas I have studied during my PhD, mostly with a focus towards deep learning techniques and their application in well-known problems of the drug-discovery pipeline, comparing them with other current state-of-the-art approaches whenever possible. In particular, I have focused on protein binding site prediction, protein-ligand binding affinity prediction, compound selectivity elucidation via their association with pathways and ligand generation.

1.3.1 Protein binding site prediction

Proteins produce the building blocks of cells, performing functions that are critical for life. They do not perform such actions alone, but by interacting with other molecules [92], and such are mediated by only a few aminoacids. Therefore, identifying a protein's binding sites a priori can substantially clarify and help understand their underlying mechanisms of function. Furthermore, in structure-based virtual screening applications [93, 94], where one is interested in ranking a set of ligands in terms of activity against a particular target, an accurate identification of the binding pocket is crucial.

Classical approaches for predicting protein binding sites roughly fall into two categories. The first one include sequence-based approaches, which mostly rely on the conservation information extracted from multiple sequence alignments [95]. The success of these approaches has been known to greatly depend with the type of functional residue sought [96]. The second type encompass structure-based approaches, which seek to identify regions on the protein surface which are likely to bind to a ligand, by using geometric information and other types of descriptors [97–100].

In **Pub. 1** [77] we provide, to the best of our knowledge, the first fully machine-learning-based approach towards the identification of protein ligand binding sites. The algorithm is completely learned from examples, and is based on the previously described 3D-convolutional-neural-network paradigm. Furthermore, we test its performance to find it is comparable to that of the state-of-the-art methods for pocket detection.

1.3.2 Protein-ligand binding affinity prediction

Drug discovery is inherently a multiobjective optimization problem [101] as several variables need to be taken into account when considering a compound, e.g. solubility [102], toxicity [103], selectivity [104] or kinetics [105, 106]. Among these, perhaps the most important is potency,

which measures how strongly a small molecule binds to its protein target to produce a desired effect or inhibition. Chemists typically study this variable through experimental affinity measurements (e.g. K_i, K_d, IC_{50}) in different types of assays in the laboratory, such as phenotypic or cell-based ones.

Experimentally determining binding affinities is a long and costly process, approaches to predict these quantities *in silico* were consequently developed in order to prioritize testing of compounds. In fact, quantitative structure activity relationship (QSAR) approaches, based on simple linear or empirical models of molecules have found their way into a computational chemist's toolbox in the last 30 years. With the advent of increasing available affinity data coming from compound databases such as ChEMBL [107] and protein-ligand ones, such as PDBbind [108] and cheaper computational resources [109], opportunities to explore more data hungry machine learning approaches have become prevalent in the last decade. In the case of structure-based approaches, these typically allowed more flexibility by not requiring an explicit mathematical relationship of the protein-ligand complex [110] and their affinity, which in practice resulted in greatly improved performances compared to classical QSAR approaches.

Scoring functions can arguably be classified into three different categories depending on the nature of their modelling [111]: potential-, simulation- and data-based. The first type model binding affinities as the sum of statistical potentials between protein and ligand atoms [112–116]. Simulation-based methods make use of available force fields such as AMBER or CHARMM [117, 118] to model protein-ligand interactions, resulting in approaches like free energy perturbation methods [119–121]. The last category uses experimental data to fit statistical or machine-learning regression models to predict potency [45, 49, 61].

In **Pub. 2** [81] we developed a scoring function based on 3D-convolutional neural networks, named K_{DEEP} and extensively tested it in

several public datasets. We found that its performance is state-of-the-art in most benchmarks. In fact, the performance of K_{DEEP} was validated by several industrial partners and in the 4th D3R Grand Challenge [122,123], where it came out first place in two affinity subchallenges.

The congeneric series case

K_{DEEP} was developed using the v.2016 iteration of the PDBbind database, which is composed of a diverse set of protein-ligand complexes, with the intent of it being as general as possible. After its development we started testing its performance in the lead optimization scenario, that is, in the congeneric series case, which are sets of closely related molecules that are typically modified by medicinal chemists with the intent of improving several molecular properties (Figure 1.7). Early tests suggested the generalization capabilities of our model, trained on PDBbind, were limited when predicting the small differences expected in congeneric series.

Given the lack of publicly available congeneric series data, most of it available in the BindingDB database [125], we contacted several industrial partners (Janssen, Pfizer, Biogen) to collaborate in both the training and testing of machine learning models, again based on 3D-convolutional neural networks, on several congeneric series related to different targets and diseases. This resulted in **Pub. 3**, where we retrospectively show that such models can achieve superior performance even to the most sophisticated simulation-based free energy perturbation methods with very few training ligands. Furthermore, these avoid most of the issues related with simulation-based approaches, such as the treatment of waters or ligand parameterization [126], at a fraction of the computational cost.

1.3.3 Molecular pathway association

Drug discovery is a very expensive process than can easily span more than ten years since the inception of a project until the release of a

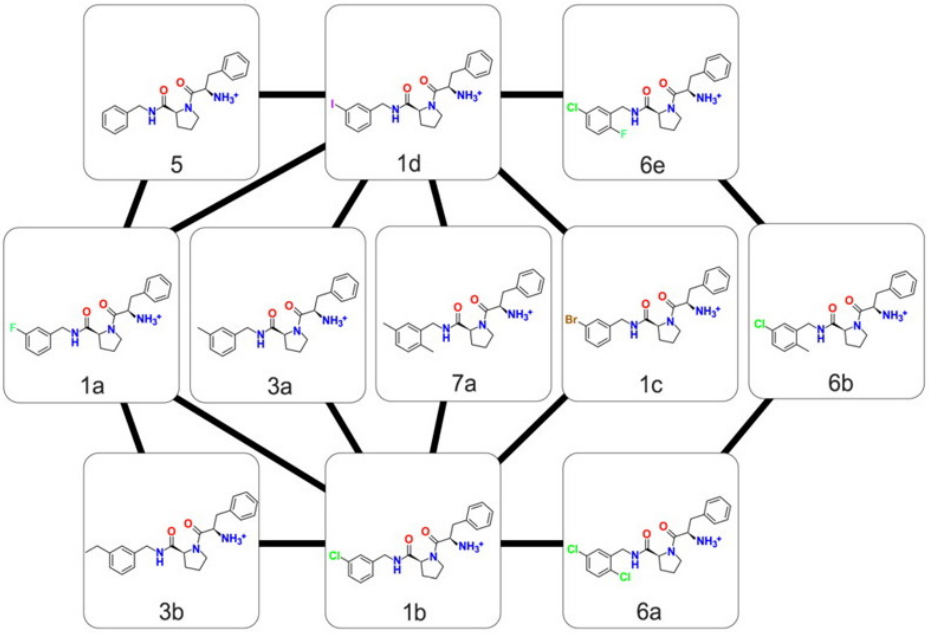


Figure 1.7: A set of ligands belonging to the same congeneric series that bind to thrombin, commonly used as benchmark in free energy perturbation studies. Taken from Wang *et al.* [124]

compound to the market [127]. In particular, it suffers from a heavy attrition problem [128]: drugs initially deemed as promising candidates can later show ineffectiveness, toxicity, or promiscuity due to an unclear mechanism of action. In the context of high-throughput and phenotypic screening, compounds marked active against a target are obscure by a large number of molecules that act through unknown or undesired mechanisms of action. Therefore, early attention in lead discovery is crucial to minimize costs caused by these reasons at later stages.

A computational tool that predicts all pathways a particular drug intervenes in could allow: (i) a fast depriorization of compounds with unwanted mechanisms of action, (ii) the identification of chemicals that

target disease-relevant biological pathways, and (iii) the identification of compounds that with yet-to-be known mechanisms of action. The latter compounds were named by researchers as dark chemical matter [129], and evidence suggests that these molecules could have a unique activity profile against new proteins, and correspondingly have a better chance of a safer activity profile.

In **Pub. 4** [130] we developed a model based on self-normalizing neural networks that given a particular compound, is capable of predicting which pathways it interferes with. We used several large pathway databases such as KEGG [131], Reactome [132]. Ligands were extracted from one of the latest versions of ChEMBL, and associated with their corresponding pathways through Uniprot [133]. To the best of our knowledge, this is the most extensive study on molecular pathway association reported so far, both in terms of the amount of data processed and ability to deal with multifunction compounds. Furthermore, we validated our models using both publicly available and industrial data, thanks to a collaboration with Novartis.

1.3.4 Generative modeling for drug design

The main step in a typical drug discovery campaign for the formulation of new hypothesis is a well-motivated lead compound [134], which sometimes is extracted from vast synthetically feasible libraries. Medicinal chemists modify these lead compounds, their design hypothesis typically biased towards preferred chemistry [135]. Since drug-like molecule space was estimated to range between 10^{30} and 10^{60} compounds, and to avoid such prohibitive sampling, the process of automating the de-novo design of compounds with a desired set of properties has become an active field of research in the last 15 years [136, 137].

While QSAR-like models have been extensively used in the last 40 years, with the arrival of novel generative machine learning models, such as variational autoencoders [139], generative adversarial networks [140],

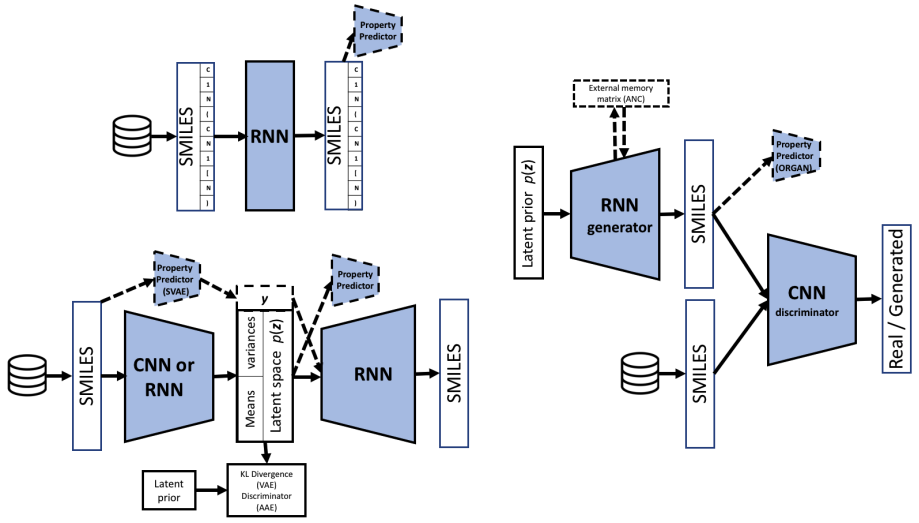


Figure 1.8: Example schemes of different deep learning architectures used in drug design, with SMILES as input representation. A recurrent neural network architecture (top left), a variational autoencoder architecture that can either use 1d-convolution or recurrent layers (bottom left), and a generative adversarial network approach (right). Figure taken from Elton *et al.* [138]

recurrent neural networks [42] and adversarial autoencoders [141], significant attention has been drawn towards its inverse problem (i.e. predicting structures given a set of properties). Early approaches have shown promising results [142, 143], and in fact, this has become a very popular topic of research, with over 45 papers published only in the last two years [138]. Sadly, validation of these methods has lacked standardization up until recently [144], severely limiting their benchmarking and reproducibility.

Different approaches towards this problem use different input representations, the most common being the conjunction of SMILES with recurrent neural network architectures, although recent approaches have

seemed to adapt a graph-like representation of molecules [145, 146]. One limitation of such models is that their generated structures are biased towards small chemical modifications of a provided seed molecule [147]. On the other hand, shape-based tools [148] allow the exploration of much larger chemical space.

Concurrently in computer vision, research in image captioning models (i.e. those that are trained to predict a written description of a given picture) were also gathering significant attention due to their impressive results [149, 150]. Given that we had developed a 3D representation that could be used in conjunction with such deep learning approaches, in **Pub. 5** we explored shape-based generative models in drug design and compared them against other ligand-based methods.

1.4 Deployment of models

A main focus of my PhD was to develop machine-learning applications that could help accelerate and provide better decision making in the early phases of the drug discovery pipeline. As a result of my research, a lot of code was produced. It was clear for us from the beginning that in order to obtain higher visibility and ease the job of computational and medicinal chemists as much as possible that these tools needed to be freely available in some form.

A typical problem of professionals working in the bio and chemoinformatics fields is ensuring their software works under the most general conditions as possible. It is very common to find the implementation of particular publication in the version of a language that is no longer supported, with unmet dependencies, using non-existing web resources, or simply not developed with a software maintainability mindset. In order for us to achieve this, all of our models are fully containerized using Singularity [8] and uploaded to the PlayMolecule.org repository of applica-

[8] `https://www.sylabs.io/singularity/`

Table 1.1: Applications developed and deployed in the PlayMolecule.org repository of applications in the duration of the thesis

Application name	**Number of jobs** [a]	**Publication date**
DeepSite	8621	31/May/2017
KDeep	6567	8/Jan/2018
DeltaDelta	21	(submitted)
PathwayMap	231	26/Dic/2018
LigDream	289	14/Feb/2019

[a] As of 21st June 2019.

tions, where users can freely submit their own jobs.

Chapter 2

OBJECTIVES

The objectives of the thesis presented here were threefold: the first was the exploration of modern representations for biomolecular complexes towards their use in modern deep learning architectures, such as in the case of voxelization and convolutional neural networks. The second was to apply such models in projects relevant in drug discovery pipelines, comparing their performance to existing approaches whenever possible. Finally, the last goal was to deploy such models in the PlayMolecule.org repository of applications so as to facilitate and promote their use to computational and medicinal chemists.

2.1 Development of a novel deep-learning representation for biomolecular complexes

One of the main advantages of modern deep learning approaches such as convolutional or recurrent neural networks is their ability to work directly (or closely) on input data itself, whether it is images for the former or text corpora for the latter. In this thesis we introduced a density representation for biomolecular complexes, which uses a distance-based interpolation of atoms and their corresponding user-defined properties and provides the possibility to use it in conjunction with 3D-convolutional

neural networks, the de-facto state-of-the-art model in computer vision applications.

2.2 Deep learning in drug discovery applications

Deep-learning-based approaches have atracted a considerable amount of attention in the past years, and consequently their exploration in structural biology and chemoinformatics was only a matter of time. Most of the work presented in this thesis therefore responds to this goal: the application and benchmarking of deep learning approaches in relevant drug discovery problems. Among the applications explored in this thesis we can find:

- DeepSite: A protein binding site prediction tool that uses 3D-convolutional neural networks.
- K_{DEEP}: A 3D-convolutional neural network scoring function for protein-ligand binding affinity prediction.
- PathwayMap: Selectivity elucidation of compounds via their association with pathways using self-normalizing neural networks.
- LigDream: Shape-based generative modeling of compounds via conditional variational autoencoders and captioning networks.

Chapter 3

PUBLICATIONS

3.1 Journal articles

3.1.1 DeepSite: protein-binding site predictor using 3D-convolutional neural networks

Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., & De Fabritiis, G. (2017). Bioinformatics, 33(19), 3036-3042.
https://doi.org/10.1093/bioinformatics/btx350

Summary
In this paper we presented DeepSite, a protein binding site predictor based on 3D-convolutional neural networks and the novel biomolecular representation of compounds presented in this thesis. Contrary to other structure-based approaches, that used algorithmic approaches via the clever exploitation of geometric, evolutionary or chemical features to detect druggable protein cavities, our approach is entirely data-based. We used the scPDB (v.2013) database, and more than 7000 proteins in order to train and validate our approach, showing state-of-the-art performance when comparing with other geometric-based approaches such as fPocket [99] and Concavity [151].

Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017 Oct 1;33(19):3036–42. DOI: 10.1093/bioinformatics/btx350

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

Jiménez, J., Skalic, M., Martinez-Rosell, G., & De Fabritiis, G. (2018). Journal of Chemical Information and Modeling, 58(2), 287-296.
https://doi.org/10.1021/acs.jcim.7b00650

Summary
We continued exploring our 3D dimensional representation of biomolecular complexes for a different task: the prediction of protein-ligand binding affinities. The development of machine-learning-based scoring functions is a well-studied field, with high-quality curated databases such as PDBbind available. In this paper we developed K_{DEEP}, our own approach based on 3D-convolutional networks and compared its performance to other existing machine-learning and empirical scoring functions, to find its performance at least competitive with the current state of the art. Furthermore, we investigate the utility of the developed approach in the lead optimization scenario, to find that results greatly vary depending on the congeneric series studied.

Note: The supplementary information of this publication has been modified in the thesis presented here due to space restrictions.

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

3.1.3 DeltaDelta Neural Networks for Lead Optimization of Small Molecule Potency

Jiménez, J., Pérez-Benito, L., Martínez-Rosell, G., Sciabola, S., Torella, R., Tresadern G. & De Fabritiis, G. (submitted)

Summary
Given the limitations of the previous presented method K_{DEEP} at ranking close compounds, such as the ones found in a congeneric series, in this paper we developed similar approaches in the lead optimization scenario. We trained and tested our models in public databases such as the BindingDB protein-ligand validation and Schrödinger free energy perturbation benchmark sets. However, given the very little freely available data, in order to further validate our method we collaborated with several pharmaceutical companies to blindly train and validate our models. We found that 3D-convolutional neural network models are competitive to other docking and simulation-based approaches, such as Glide and free energy perturbation, respectively, with very few training ligands. We also provided a retrospective simulation scenario where the model is tasked with choosing the best available compound out of a pool, to find that in most cases the model is able to find it before the recorded experimental order.

Jiménez-Luna J, Pérez-Benito L, Martínez-Rosell G, Sciabola S, Torella R, Tresadern G, et al. DeltaDelta neural networks for lead optimization of small molecule potency. Chem Sci. 2019. DOI: 10.1039/C9SC04606B

3.1.4 PathwayMap: Molecular Pathway Association with Self-Normalizing Neural Networks

Jiménez, J., Sabbadin, D., Cuzzolin, A., Martínez-Rosell, G., Gora, J., Manchester, J., Duca, J., & De Fabritiis G.. Journal of Chemical Information and Modeling 2019 59 (3), 1172-1181
https://doi.org/10.1021/acs.jcim.8b00711

Summary
Attrition is a serious problem in drug discovery: compounds initially deemed as promising might act through obscure or unknown mechanisms of action, or hit unwanted targets (i.e. the selectivity problem). In this paper we propose a model, based on multi-label self-normalizing neural networks, that is able to associate compounds and the pathways it intervenes in. We used public compound databases such as ChEMBL and pathway databases such as KEGG and Reactome and associated both via Uniprot. The models and evaluation provided here are (to the best of our knowledge) the most extensive provided up to date for this task and can naturally tackle multifunction compounds. In the paper, an applicability scenario is exemplified by the identification of dark chemical matter (i.e. those identified not to bind to more than a predefined number of assays in another study).

Note: The supplementary information of this publication has been modified in the thesis presented here due to space restrictions.

Jiménez J, Sabbadin D, Cuzzolin A, Martínez-Rosell G, Gora J, Manchester J, et al. PathwayMap: Molecular Pathway Association with Self-Normalizing Neural Networks. J Chem Inf Model. 2019 Mar 25;59(3):1172–81. DOI: 10.1021/acs.jcim.8b00711

3.1.5 Shape-Based Generative Modeling for de Novo Drug Design

Skalic, M., Jiménez, J., Sabbadin, D., & De Fabritiis, G.. Journal of Chemical Information and Modeling 2019 59 (3), 1205-1214
https://doi.org/10.1021/acs.jcim.8b00706

Summary

The past few years have witnessed an explosion of generative-based approaches in *de-novo* drug design. Common to most of these is way ligands are featurized, most models opting for a SMILES or a graph representation of compounds, that allow for small modifications from a seed compound. In this work we focus on such models but through the use of the 3D-dimensional representation presented in our previous works. Our approach also takes great inspiration from other computer vision tasks, such as image captioning: given a 3D-representation of a ligand, we are able to generate SMILES of arbitrarily similar compounds through a conditional variational autoencoder and a captioning network.

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

3.2 Book contributions

3.2.1 Predicting protein-ligand binding affinities

Jiménez, J. & De Fabritiis, G. Part of the Royal Society of Chemistry's upcoming book *Artificial Intelligence in Drug Discovery*, edited by Dr. Nathan Brown.

Summary

In this chapter we were tasked with summarizing the current state-of-the-art approaches for binding affinity prediction from a structure-based viewpoint. We first overviewed other classical methods, ranging from those that are empirical or simulation-based to later explain in detail the advances in structure-based machine learning models: namely 3D-convolutional and graph-based neural networks. Particular attention is given to relevant topics such as model interpretability, benchmarking and available databases for development.

4 | Predicting Protein-Ligand Binding Affinities

José Jiménez-Luna and Gianni De Fabritiis

Abstract

Accurate in silico protein-ligand binding affinity prediction can substantially accelerate drug discovery pipelines by prioritizing compounds for experimental testing, a typically lengthy and costly process. Given the success of machine-learning and artificial intelligence approaches in areas such as computer vision and natural language processing in the last few years, there have been significant developments towards their application in structure-based potency prediction. In this chapter we summarize recent progress in this field, and we provide readers with a thorough introduction of the basic aspects to take into account when developing such models.

Contents

4.1 Introduction

Drug discovery is inherently a multiobjective optimization problem [1,2] as several variables need to be taken into account, e.g. solubility [3], toxicity [4,5], selectivity [6,7] or kinetics [8,9]. Among these, perhaps the most important is potency, which measures how strongly a small molecule binds to its protein target to produce a desired effect or inhibition. Chemists typically study this variable through experimental affinity measurements (e.g. K_i, K_d, IC_{50}) in different types of assays in the laboratory, such as phenotypic or cell-based ones.

Experimentally determining binding affinities is a costly process, and therefore computational approaches to predict these quantities *in silico* were consequently developed in order to prioritize testing of compounds. In fact, quantitative structure activity relationship (QSAR) approaches, based on fitted linear or empirical models of molecules have been common among those in a computational chemist's toolbox for the last 30 years. With the advent of increasing available affinity data coming from compound databases such as ChEMBL [10] and protein-ligand ones, such as PDBbind [11] and cheaper computational resources [12], opportunities to explore more data hungry machine learning approaches have become prevalent in the last decade. In the case of structure-based approaches, these typically allowed more flexibility by not requiring an explicit mathematical relationship of the protein-ligand complex [13] and their affinity, which in practice resulted in greatly improved performances compared to classical QSAR approaches.

In this chapter we provide readers with an introduction to the field of structure-based potency prediction via machine learning. Though the work here does not intend to be an exhaustive review of proposed approaches, which at the time of writing continues to grow at a fast pace, we believe a disciplined introduction on the basics regarding this area, namely classification and scope of models, descriptor generation and evaluation standards to be beneficial for the community. Finally an overview of the most important techniques in the last few years, to the best of our knowledge, is provided.

4.2 A brief background on classical methodologies

A scoring function $f : \mathcal{X} \to \mathbb{R}$ maps a ligand, or a protein-ligand complex $x \in \mathcal{X}$, to a quantity which is either its binding affinity or a proportional proxy. Over the years, many types have been proposed, most claiming advantage of some form over already existing approaches. They can arguably be classified into three different categories depending on the nature of their modelling [14], potential-, simulation- and data-based. In this section we provide a small background on previous work focusing on the subject of binding affinity prediction, based on such classification, before describing the more recent machine-learning approaches.

4.2.1 Potential-based

Methods in this category model binding affinities as the sum of statistical potentials between protein and ligand atoms:

$$\Delta G = \sum_{i \in A_l} \sum_{j \in A_p} \omega_{ij}(r), \tag{4.1}$$

where A_t and A_p are the sets of atoms in the ligand and protein respectively and

$$\omega_{ij}(r) = -k_B T \log\left(\frac{\rho_{ij}(r)}{\rho^*_{ij}}\right), \tag{4.2}$$

where $\rho_{ij}(r)$ is the number density of atom pair ij at distance r, ρ^*_{ij} is the same quantity at a reference state with no interatomic interactions, k_B is the Boltzmann constant and T a temperature. The reasoning behind this approach is probabilistic: if a certain interatomic contact appears more frequently than expected of its reference state, it is energetically favorable and viceversa. Potential-based approaches have been widely used mainly thanks to the simplicity in their construction. Some popular implementations of these approaches are SMoG [15], Muegge's PMF [16], DrugScore [17], IT-Score [18] and KECSA [19], among many others [20–24].

4.2.2 Simulation-based

Molecular mechanics force fields such as AMBER [25] or CHARMM [26] are regularly used for predicting protein-ligand interactions [27–33]. Often, approximate solutions such as the Poisson-Boltzmann or Poisson Generalized Born models are used, where van der Waals, electrostatic, and desolvation terms are taken into account:

$$\Delta G = \Delta E_{\text{vdw}} + \Delta E_{\text{electrostatic}} + \Delta E_{\text{H}-\text{Bond}} + \Delta G_{\text{desolvation}}. \tag{4.3}$$

Full simulation based approaches such as free energy perturbation (FEP) [34–40], or thermodynamic integration (TI) [41, 42] have shown to provide excellent performance despite being computationally demanding. These methods can also take advantage of the advances of modern force fields, quantum mechanic methods and solvation models. However, recent evaluations [43] have shown that their performance is very sensitive to starting parameters such as force-field selection or treatment of waters, limiting their applicability in prospective scenarios. Other related approaches, such as linear interaction energy (LIE) [44], linear response approximation (LRA) [45] and MM-PBSA/GBSA [46] methods have shown alternate successes and failures [47]. These are typically named end-point approximation methods as they only consider both protein and ligand in their bound and unbound states.

4.2.3 Data-based

In this category we find classical scoring functions that use statistical methods such as linear regression or partial least squares (PLS) [48] to adjust the contribution of several physico-chemical terms (or other descriptors) towards an affinity prediction. Therefore a set of known protein-ligand complexes with affinity data is needed to find the aforementioned coefficients for an optimal fit. For instance, X-Score [49] adopts the following functional form:

$$\Delta G_{\text{bind}} = \beta_0 + \beta_1 \Delta G_{\text{vdw}} + \beta_2 \Delta G_{\text{H-bond}} + \beta_3 \Delta G_{\text{deformation}} + \beta_4 \Delta G_{\text{hydrophobic}} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.4}$$

where ΔG_{vdw} accounts for the van der Waals interaction between the protein and ligand, $\Delta G_{\text{H-bond}}$ for hydrogen bonding , $\Delta G_{\text{deformation}}$ for the deformation effect and $\Delta G_{\text{hydrophobic}}$ for hydrophobic interactions. An estimate of these coefficients $\hat{\boldsymbol{\beta}}$ is determined through a least squares fit. Many other empirical scoring functions have been developed over the years, such as cyScore [50], ChemScore [51], GlideScore-SP [52], LudiScore [53], among many others [54, 55].

At first glance, one may be right to think that modern machine-learning scoring functions belong to this category, as they use a training set and several fit parameters to perform predictions. The main difference is that classical empirical scoring functions assume a fixed, pre-defined mathematical relationship among handcrafted features to model the target affinity. Such is not the case for modern machine-learning scoring functions, which automatically extract the most important features from a closer representation from the real data, in practice allowing them to be considerably more flexible and consequently provide better performance. Recent advances in modern machine-learning scoring functions are summarized in the subsequent section.

4.3 Modern machine-learning scoring functions

Research on machine-learning based scoring functions is currently a very active field, with dozens of proposed approaches only in the last ten years. In this section, we first describe the domain applicability of the proposed models, depending on the nature of their training data. We then make a thorough summary on the types of features (or predictors) used for the training of models in the literature, to later discuss recent developments on specific structure-based machine-learning algorithms. Finally we describe several approaches towards model interpretability and discuss about the availability of the proposed algorithms at the time of writing.

4.3.1 Domain applicability

It is common in the development of scoring functions to distinguish between two scenarios, depending on the nature of the data at hand. If the goal itself is to predict binding affinity or a proportional proxy, one needs continuous binding constants, and typically tackles the problem as a regression task. Other scenarios may feature binary data (i.e. active/inactive), which is better suited for a more classical virtual screening scenario, where the goal is to select as many active molecules (also called binders), from a considerably big database, against a target as possible. Since only two classes are considered in this scenario, binary classification models are commonly used to tackle this problem instead. Note that any regression-based scoring function can be used as a binary classification one given an appropriate threshold.

It is also worth mentioning that in this chapter we focus on the structure-based case, that is, we are interested in a model that generalizes not only in chemical space, but also across different targets, which is common at earlier phases of drug discovery, such as target identification. If the study at hand only takes into account a single protein target, ligand-based models are significantly more popular, the archetypal application in this case being lead optimization.

4.3.2 Descriptors

In this section we describe different sets of descriptors studies used in the development of machine-learning models for the prediction of both continuous and binary binding affinity estimates. It was the norm until recently for researchers not only to design their own set of features, but to perform descriptor selection as well. This process of handcrafted feature creation and filtering was subsequently abandoned, when it was shown that modern deep learning architectures could perform automatic feature extraction from a closer representation of the original data. This allowed models to design a more diverse, non-linear, latent set of features that in practice obtains better performance when enough data is available. Because of the mentioned mentality swift that deep-learning architectures brought to feature space, we make the distinction when describing them here as well.

Handcrafted features

Since the inception of the (arguably) first structure-based machine-learning-based scoring functions in 2004, researchers have used an increasing number of feature sets in order to find out the better performing ones. Although we do not provide an exhaustive list, among the most popular we can find:

- Occurrences of each protein-ligand atom pair at different distance thresholds [56–60]
- Electronegativities of ligand and protein atom types [61]
- Atom and group interactions such as van der Waals, electrostatics, hydrogen-bonding, π-system, aromatic and metals [62, 63].
- Energy terms representing desolvation and entropic losses [62].
- Geometrical description of the binding, such as shape and surface property matching [62].
- Property-encoded shape distributions [64].
- Protein-ligand interaction fingerprints [65–67]
- Intermolecular interaction terms extracted from AutoDock Vina [68] and BINANA [69,70].
- Distance-based fuzzy membership functions accounting for attraction and repulsion terms between atoms [71]
- Knowledge-based potentials combining SYBYL atom types [72].
- Other structural features regarding β-contacts, crystallographic-normalized B factors and polar and hydrophobic contact surfaces [73].
- Multiscale weighted labeled algebraic subgraphs [74]

Interestingly, it was found that a more detailed description of the protein-ligand interaction landscape did not necessarily result in better performance in the standard PDBbind benchmark [75], despite of some studies considering more than 100 feature subsets.

Automatic feature extraction

One of the reasons for the success of deep learning [76] architectures in fields like computer vision [77] and natural language processing [78, 79] was the fact that modern neural networks perform automatic feature extraction. Among the first structure-based approaches that adopt this strategy we acknowledge the one taken by DeepVS [80]. In particular, in this approach, the local context of an atom in a protein-ligand complex is embedded into several learned fixed sized vectors using several basic features such as atom types, partial charges and distances to the closest ligand neighbours and aminoacids. The definition of the atom context, however, is user dependent, since a number of neighboring ligand and protein atoms needs to be pre-specified. For each type of basic feature, a column lookup operation is performed in a predefined learnable embedding matrix W_t for each possible discretized value of a feature. The embedding of an atom z_i is then constructed by the concatenation of the column vectors:

$$z_i = \{z_{\text{atom}}, z_{\text{distance}}, z_{\text{charge}}, z_{\text{aminoacid}}\} \tag{4.5}$$

Since the number of atoms varies with the particular system, this representation needs then to be summarized in a single latent vector v, representing an embedding for the entire complex, which we describe in the next section. This set of descriptors suffer mainly from two problems: the first being the pre-specification of context, and the second the need to discretize continuous atom features into bins to extract its corresponding learned latent space from W_t

Other approaches have more closely followed computer vision architectures. Using the example of image classification, a convolutional neural network (CNN) learns which picture patches are the most informative in order to arrive to a correct classification, only using pixel information. A similar argument can be made for proteins and small chemical compounds: they are structures in three-dimensional space with different physiochemical values as properties. Identically to a two-dimensional image, which can be represented using three different two-dimensional arrays, or channels, representing its colors (i.e. RGB color space), a protein or a ligand could potentially also be represented with k three-dimensional arrays, and modern computer vision architectures would be readily applicable. In fact, this is a promising direction that several researchers have taken in the last few years: the same way that a pixel in an image holds three values for its colors, a voxel in a three dimensional image (in our case, a protein or ligand) can feature different values representing different molecular properties. Directly translating atomic positions into volumetric space can result in a very sparse representation, and therefore most works use a distance-based interpolation over pre-defined atom types. For instance, Ragoza *et al.* [81] use the following functional:

$$A_1(d,r) = \begin{cases} \exp\left(-\dfrac{2d^2}{r^2}\right) & \text{if } 0 \leq d < r \\ \dfrac{4d^2}{e^2r^2} - \dfrac{12d}{e^2r} + \dfrac{9}{e^2} & \text{if } r \leq d < \dfrac{3r}{2} \\ 0 & \text{if } d \geq \dfrac{3r}{2} \end{cases}, \tag{4.6}$$

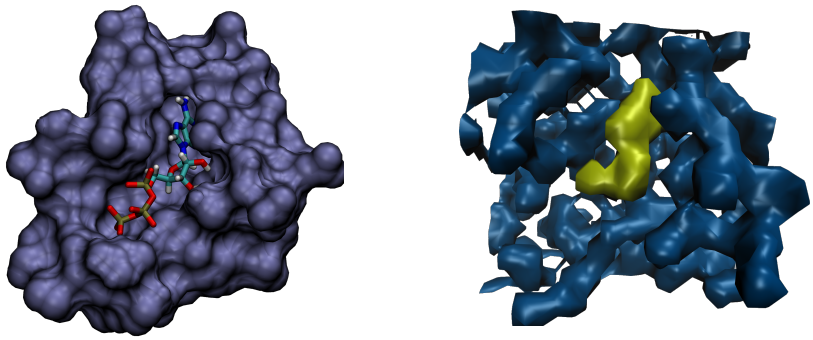
where d is the distance of each atom to a particular voxel, and r is the atom's van der Waals radius. Jiménez *et al.* [82], on the other hand, use the following functional in the K_{DEEP} protein-ligand affinity predictor:

$$A_2(d,r) = 1 - \exp\left(-\left(\frac{r}{d}\right)^{12}\right) \tag{4.7}$$

In terms of channel selection, the vast majority of approaches use some predefined notion of atom typing available in other applications, such as the ones defined in smina [83] or in the AutoDock PDBQT format (hydrophobic, aromatic, hydrogen-bond acceptor/donor, positive/negative ionizable and metals) [84]. A general occupancy channel is also typically included to include explicit geometrical information of the molecular object (Figure 4.1 [1]). This particular representation of biomolecular complexes has not only been used for protein-ligand binding affinity prediction, but for protein binding site prediction [85], pharmacophore elucidation [86] and de-novo generation of molecules [87], with varying success.

These sets of descriptors represent the three dimensional structure of proteins in space, but they also suffer from several issues. In particular, three dimensional arrays take significantly more space in memory than images in computer vision applications. Furthermore, voxelizing

[1] Reprinted with permission from Journal of Chemical Information and Modeling, 58 (2), Jiménez *et al.*, KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks, 287-296. Copyright (2018) American Chemical Society.



(a) (b)

Figure 4.1: (a) PDB Id 2HMU pocket and bound ligand ATP. (b) Voxel representation of the hydrophobic channel for both protein (blue) and ligand (yellow).

the binding site entails choosing a window, since the shape of the arrays needs to be fixed for their use in CNN architectures, while protein pockets and ligands can significantly differ in size. Finally, the representation is neither rotationally nor translationally invariant, properties which are desirable when modelling biomolecular complexes with the aforementioned models.

Given the success of graph convolution architectures in ligand-based approaches [88], some attention has been given into adapting the same framework for structure-based approaches. In particular, PotentialNet [89] uses the concept of adjacency from an atomic distance matrix $R \in \mathbb{R}^{N\times N}$, where N is the number of atoms in a predefined environment of the binding site of the system. While ligand-based graph-convolution architectures use the notion of bonds to represent adjacency, in a co-crystal it can encompass a wider range of chemical interactions among neighbors, such as π-π stacking, hydrogen bonds or hydrophobic contact. In fact, a simple distance-based threshold may serve to construct an adjacency matrix $A^{N\times N\times c}$, where c represents the number of edge types. Ordering the rows of the adjacency matrix by the membership of each atom to their corresponding protein or ligand complex, A can be seen as a block matrix, where the diagonal blocks are interactions inside the same complex, while the off-block elements represent interactions between the protein and ligand atoms:

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N1} & A_{N2} & \dots & A_{NN} \end{bmatrix} = \begin{bmatrix} A_{L:L} & A_{L:P} \\ A_{P:L} & A_{P:P} \end{bmatrix}, \tag{4.8}$$

where $A_{ij} = 1$ for an occurring interaction and 0 in any other case. We will explain how this input is used in a graph-based convolutional neural network in the next subsection.

4.3.3 Models

In this section we will discuss recent advances in the modelling of protein-ligand binding affinities with modern structure-based machine-learning architectures. Mostly we focus here on those

models that perform automatic feature extraction (see previous section) out of a representation closer to its source. In particular, we briefly describe custom embedding approaches, discuss three-dimensional convolutional neural networks whose corresponding input is the atom-type voxelization, and finally graph-convolution neural networks, which in turn use a distance-based representation between atoms and their features in a system.

Custom embedding approaches

Here we describe the approach of DeepVS [80]. Once atom feature vectors z_i have been computed, they are mapped to:

$$u_i = \tanh(W^e z_i + b^e), \tag{4.9}$$

where W^e, b^e are a set of learnable weights and biases, respectively, which are shared for all embeddings z_i. A maximum-over-columns operation is then applied to obtain a fixed-size vector representation for the entire complex, which is used by subsequent fully connected layers to obtain the desired scalar output.

3D-convolutional neural networks

The output ϕ of a neuron in a regular fully connected layer is obtained by multiplying an input $\boldsymbol{x}$ with some learnable weights $\boldsymbol{w}$, adding a bias b and applying a non-linearity f:

$$\phi = f\left(\sum_i w_i x_i + b\right). \tag{4.10}$$

Convolutional neural networks are specifically designed to work with spatial inputs, such as images (or voxels, in our case), trying to emulate the response of an individual neuron to visual stimuli. This fact allows us to encode certain properties into the architecture that cannot be assumed in fully connected ones. Concretely, the layers of a three-dimensional neural network are arranged in 4 dimensions: height, width, depth and number of channels, with each neuron only locally connected to a localized region of the preceding layer, since it is impractical to connect to all of the previous neurons [90]. In the case of voxels, the output of a neuron is called a feature map ϕ, which is a three dimensional tensor, obtained through discrete convolution of a filter W_i over an input feature map $z_i(x, y, z)$:

$$\phi = f\left(\sum_i W_i * z_i(x, y, z) + b\right), \tag{4.11}$$

where $*$ represents a three-dimensional discrete convolution operation (i.e $w * f(x, y, z) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} \sum_{l=-c}^{c} w(s, t, l) f(x - s, y - t, z - l)$) [91]. The connectivity of a neuron is controlled by a parameter named kernel or filter size, and its locality is only defined across the spatial dimensions, while full connectivity is applied to all feature channels. Parameters in a convolutional neural network are also typically shared in the channel dimension: if one feature is useful for a particular position in the image, it should also be for a different one. This simplification results in a significant reduction of learnable parameters, which in turn makes current implementations more computationally approachable at a scale.

Apart from convolution, other layers are commonly used in the development of this type of neural networks. For instance, pooling layers apply a non-learnable transformation to an input, typically reducing its size in order to simplify further calculations and reduce the number of

parameters in the network, operating independently on each channel. Normalization layers, such as batch normalization [92], ensure the input distribution remains similar across batches, while dropout layers randomly drop neuron connections with the hope of avoiding overfitting. Once the size of the three-dimensional filters has been reduced enough, these are typically flattened out to a vector, so that regular fully connected layers can be applied afterwards to ensure a one-dimensional output corresponding to the predicted binding affinity.

Approaches based on three-dimensional CNNs, such as K_{DEEP} [82] have been shown to work particularly well in practice, scoring first in several targets of the D3R Grand Challenge 4 [93]. Other approaches to tackle lead optimization of congeneric series, such as DeltaDelta [94], have been recently developed.

Graph-based models

In a regular convolutional neural network layer, the output of each layer is composed by the convolution of the previous one by the use of linear kernels and non-linearities, effectively gathering information from neighbouring pixels. Similarly, a graph has an inherent structure that can be efficiently exploited: each node (or atom) v_i can have a vector of features $\boldsymbol{x}_i$, and a set of neighbors based on an adjacency matrix A, as described in the previous section. Each node also features a latent representation h_i, which is iteratively updated by several functions (Figure 4.2):

$$h_i^{(t+1)} = U\left(h_i^{(t)}, \sum_{v_j \in N(v_i)} m^{(t)}\left(h_j^{(t)}\right)\right), \tag{4.12}$$

where:

- U is a differentiable *update* function that updates the latent representation of a node depending on the one from its neighbors.
- m is a differentiable *message* function sending a transformation of the hidden states from nodes v_j to v_i.
- $N(v_i)$ is the set of neighbours of node v_i.

In order to obtain a single representation for the entire graph, a node-order invariant *readout* function R (also known as graph gather) is typically applied. In general, update, message and readout functions can be fully parameterized by neural networks, and such is the case in most applications [95–101]. While this is the general scheme [88] for most architectures, we here focus on gated graph neural networks (GGNNs) [102], which uses a gated recurrent unit (GRU) [103] module as its update function and independent linear message functions for each edge type:

$$h_i^{(t+1)} = \text{GRU}\left(h_i^{(t)}, \sum_e W^{(e)} A^{(e)} h^{(t)}\right), \tag{4.13}$$

where $A^{(e)}$ and $W^{(e)}$ are the adjacency and learnable weight matrices for edge type e, respectively. A simple readout function which sums over the final node embeddings is applied to obtain the desired output size:

$$h^{(0)} = \sum_{r=1}^{N} \left(\sigma\left(i\left(h^{(K)}, x\right) \odot j\left(h^{(K)}\right)\right)\right)_r, \tag{4.14}$$

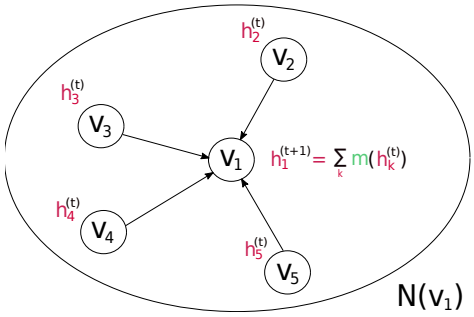


Figure 4.2: In graph convolution, the hidden state h of each node v is iteratively updated with its neighboring ones via a differentiable message function m.

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function, i, j are arbitrary learnable functions and $\odot$ is the element-wise multiplication. Once a single vector of the desired size is obtained per graph, standard fully connected layers can be used for classification or regression tasks. A structure-based generalization of this approach is the one proposed by PotentialNet, which introduces nonlinearity in the message function:

$$h_i^{(t)} = \text{GRU}\left(h_i^{(t-1)}, \sum_e^{N_{\text{et}}} \sum_{j \in N^{(e)}(v_i)} \text{NN}^{(e)}\left(h_j^{(t-1)}\right)\right), \tag{4.15}$$

where $\text{NN}^{(e)}$ represents a standard feed-forward neural network for edge type e and $N^{(e)}(v_i)$ are the neighbours for node i with edge type e. Two different *stages* are defined in the PotentialNet architecture, depending on where message functions are applied. In stage 1, named covalent propagation, only graph convolutions over ligand bonds are applied, similarly to the ligand-based counterpart. In stage 2 both bond-based and spatial-based graph convolutions are implemented, effectively propagating information between ligand and protein atoms. This stage is known as dual non-covalent and covalent propagation. Finally in stage 3, a ligand-based readout function is applied in order to obtain a fixed-size feature vector. (Figure 4.3 [2]).

4.3.4 Interpretability

The ability to interpret predictions is one of the most important features when it comes to convince computational and medicinal chemists of the usefulness of a particular model. When we speak about interpretability we refer to that of its parameters, that is, the influence of each input towards its predicted affinity value. When simpler data-based scoring functions are considered, such as those using a linear model, interpretation is straightforward since each corresponding coefficient β_i is interpreted as the individual contribution of each of the input variables towards the prediction.

[2] Reprinted with permission from ACS Central Science, 4(11), Feinberg *et al.*, PotentialNet for Molecular Property Prediction, 1520-1530. Copyright (2018) American Chemical Society. (https://pubs.acs.org/doi/full/10.1021/acscentsci.8b00507)

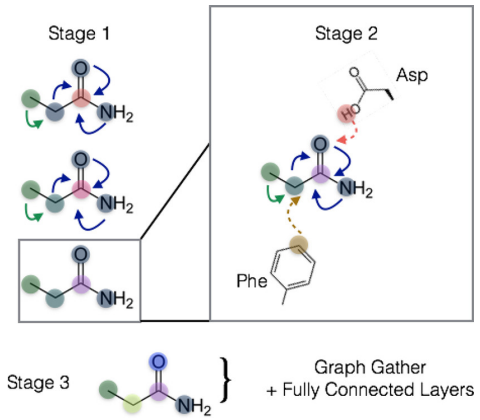


Figure 4.3: Depiction of the several stages defined by the PotentialNet approach. Stage 1 only takes into account updates over ligand bonds, while stage 2 also considers updates over neighbouring protein atoms, based on a distance threshold.

When it comes to modelling data, two different and opposing approaches arose in the statistics literature [104]. The first one assumes that the relationship between a target variable and a certain set of predictors follows a particular, but known, functional form f, which typically depends on certain parameters θ fit using available data. On the other hand, modern machine-learning approaches do not assume a particular functional form f, and instead choose non-linear algorithmic approaches that approximate the target variable as well as possible. In practice, the latter kind of approaches have been shown to provide superior predictive performance. Earlier data-driven approaches are direct descendants of the first category, while more modern scoring functions are closer in nature to the second. A negative consequence of this is the fact is that the latter do not satisfactorily address model interpretability. In this sense, there has been some recent effort in the community to satisfy the need for interpretable machine-learning models [105, 106]. In this section we summarize several presented approaches towards that end.

Masking

Masking is a popular yet simple approach applied in computer vision to find out which parts of the input a trained convolutional neural network finds important in its corresponding prediction. In masking, parts of the input, in this case the ligand or the protein are sequentially removed and re-scored, the difference between the non-altered image and the altered one representing the importance of that particular part. For the ligand, one can remove either individual atoms, or choose to remove entire molecular subgraphs (fragments). Similarly, for proteins, individual residues can be removed to find their individual contribution. Masking, however, is computationally demanding due to the number of evaluations needed, as they grow polynomially with ligand subgraph generation.

Atomic gradient

When training 3D-based convolutional neural network based models, one typically minimizes a loss function by optimizing a set of learnable parameters. This process requires numerically

computing the gradients of the loss with respect to them, which can then be naturally extended to computing the corresponding gradient with respect to the input representation. The negative of the aforementioned gradient can be interpreted as the directions in three-dimensional space which maximize the network's predicted binding affinity. The functions mapping a particular atom with a type to a voxel density are differentiable with respect to distance d and the gradient of the neural network scoring function f with respect to atom coordinates $\boldsymbol{a}$ is found via the chain rule and aggregating all grid points $\boldsymbol{G}_a$ overlapping each atom with a particular type:

$$\frac{\partial f}{\partial \boldsymbol{a}} = \sum_{A \in \boldsymbol{G}_a} \frac{\partial f}{\partial A} \frac{\partial A}{\partial d} \frac{\partial d}{\partial \boldsymbol{a}} \tag{4.16}$$

Layer-wise relevance propagation

Layer-wise relevance propagation [107] defines a measure R_d over the voxels x_d of a volumetric input which decomposes the output of a neural network binary classifier into a sum of relevances:

$$f(x) \approx \sum_{d=1}^{V} R_d, \tag{4.17}$$

with the qualitative interpretation that $R_d < 0$ contributes negative evidence for a classification, and R_d contributes positively. The goal is to find a separate relevance per layer of the classifier, with the constraint that their sums are as close as possible:

$$f(x) = \cdots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \cdots = \sum_{d} R_d^{(l)}. \tag{4.18}$$

Iterating Eq. 4.18 from the last classification layer to the first one yields the desired heatmap over voxels defined by Eq. 4.17. However, a decomposition satisfying said constraint is not unique, and one popular alternative is to propagate relevances according to their corresponding neuron activations $z_{ij} = x_i w_{ij}$. The relevance of node i at layer l is defined as the sum of the ones from of its following nodes j, weighted by z_{ij}:

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{ij} z_{ij}} R_j^{(l+1)}. \tag{4.19}$$

Layer-wise relevance propagation distributes the output value of the network as an explanation for the reason a particular input generated it. Similarly to the atomic gradient approach, this method only requires a single backwards pass through the network to obtain the desired results. However, there are issues with this proposed methodology, especially when propagating through nodes whose activation is zero. Several solutions to this problematic have been proposed in the literature, such as the alpha-beta decomposition or the conserved layer-wise relevance propagation [108]. Examples of the previously mentioned interpretability techniques can be checked in Figure 4.4 [3].

Class activation maps

Class activation maps [109] use the fact that the filters of convolutional neural networks behave as object detectors without explicit supervision in classification tasks (such as in virtual screening). A technique named global average pooling [110] is used to output the spatial average of the

[3] Reprinted from Journal of Molecular Graphics and Modelling, 84, Hochuli *et al.*, Visualizing convolutional neural network protein-ligand scoring, 96-108, (2018), with permission from Elsevier

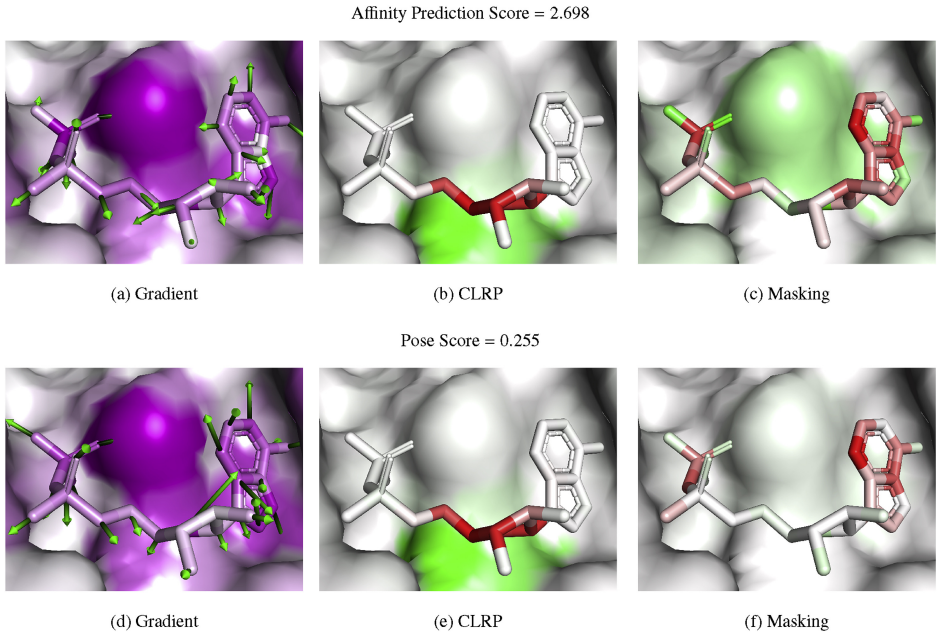


Figure 4.4: Masking, atomic gradient and layer-wise relevance propagation interpretability techniques for PDB. Id. 1o0h. For both masking and layer-wise relevance propagation approaches green and red represent a positive and negative contribution to binding, respectively, while for the atomic gradient technique its norm is represented in a purple scale.

feature map of each unit at the last convolutional layer, whose weighted sum is used to generate the final output. In a similar way, a weighted sum of the feature maps of the last convolutional layer is used to obtain class activation maps. Formally, let $f_k(x, y, z)$ represent the activation of unit k in the last convolutional layer at location (x, y, z). Then the result of applying global average pooling for unit k is $F_k = \sum_{x,y,z} f_k(x, y, z)$, and therefore for a given class c, the input of the softmax classification layer is $S_c = \sum_k w_k^c F_k = \sum_{x,y,z} \sum_k w_k^c f_k(x, y, z)$, where w_k^c is the weight that corresponds to class c for unit k. The class activation map M_c at location (x, y, z) is then defined by:

$$M_c(x, y, z) = \sum_k w_k^c f_k(x, y, z). \tag{4.20}$$

Therefore $S_c = \sum_{x,y,z} M_c(x, y, z)$, and thus M_c indicates activation importance at (x, y, z). Upsampling is performed in order to map these features to the original input size, motivated by the fact that each unit is activated by some visual pattern in its receptive field [110] (Figure 4.5).

More modern techniques, such as gradient class activation maps [112] remove the restriction of a particular model architecture for producing activation maps by letting the gradient information flow into the last available convolutional layer. That is, in order to obtain the localization map $L^c_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v \times o}$ of width u, height v and depth o, we compute the gradient of the score for

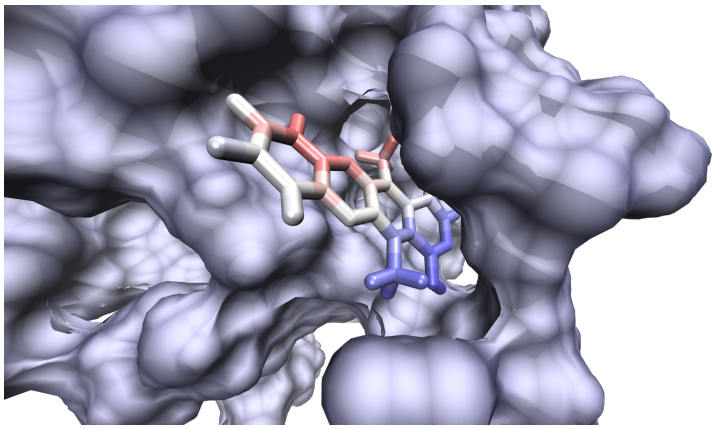


Figure 4.5: Class activation map representation taken from a public example submission to Bindscope [111], available through the PlayMolecule.org repository of applications. Influence over a positive prediction is depicted on a blue to red scale.

class c, y^c with respect to A^k, the feature maps of the convolutional layer, (i.e. $\frac{\partial y^c}{\partial A^k}$). These then are average-pooled in order to obtain neuron importance weights a_k^c:

$$a_k^c \propto \sum_{i,j,l} \frac{\partial y^c}{\partial A_{ijl}^k}, \tag{4.21}$$

where a_k^c represents the contribution of feature map k towards c. We then apply these in a weighted sum of activation maps and use a non-linearity:

$$L_{\text{Grad-CAM}}^c = g\left(\sum_k a_k^c A^k\right). \tag{4.22}$$

Typically g is a ReLU activation function since we are interested only in those features with a positive influence towards a particular classification. A combination of up-sampling with bi-linear interpolation and guided backpropagation [113] is then used to reshape filters to the original input size.

4.3.5 Implementation and availability

Reproducibility is a known issue in the machine learning community [114, 115], although most of the models cited here provide either source code or a web-based service. In particular, in this chapter we have mainly focused on three approaches: DeepVS [80] provides code implemented in the R programming language for training and application in virtual screening [4], where no license of use is specified. Another software that follows the same approach is gnina [81] [5], with Python code for performing molecular docking and virtual screening available under an Apache license. K_{DEEP} [82] and Deltadelta [94] are available through the PlayMolecule.org repository of

[4] https://github.com/Rietaros/deepvs
[5] https://github.com/gnina/gnina

applications, where users can freely submit their own protein-ligand complexes, although backend code is unavailable. Finally, PotentialNet [89] does not provide an online service nor available code.

4.4 Available data and evaluation

In this section we focus mainly on providing basic guidance towards the training and evaluation of a new machine-learning-based scoring function. In particular we first detail the most popular databases with available binding affinity data, to later discuss the most common evaluation procedures when comparing it to other approaches, both in terms of metrics and cross-validation procedures.

4.4.1 Scope and databases

In terms of available data to train structure-based models, there are several databases that can be accessed, depending on the nature of the drug discovery project at hand. In the earlier phases of a drug discovery project one may be interested in targeting a particular protein or phenotypic assay and therefore exploring a wide variety of ligands from a library of compounds. In this sense, one is concerned with training models that explore as wide chemical space as possible, so as to know which scaffolds interact more strongly with the target of interest. To train these models several databases of diverse compounds and targets have been developed over the years. Such an example is PDBbind, which extracts and curates ligand-binding affinities from the literature for most types of biomolecular complexes deposited in the Protein Data Bank (PDB). It releases yearly, with the latest (as of the time of writing) being the 2018 release, featuring 19588 manually curated protein-ligand complexes and their corresponding affinities. A *refined* set is selected out of all the available compounds, following filters regarding the quality of the data, excluding complexes with a resolution higher than 2.5 Å, an R-factor higher than 0.25, ligands bound through covalent bonds, ternary complexes or steric clashes, affinity not reported either in K_d or K_i, falling out of a desired range ($K_d < 1$pM) among other criteria. Finally, a high-quality *core set* is extracted out of the refined one, with the intent of validating scoring functions and therefore providing a standard benchmark. It is common in the development of scoring functions to train models on the difference between the refined and core sets and testing only on the latter [116–118].

The number of protein-ligand complexes available in the PDBbind database has substantially grown since its inception in 2002 (Fig. 4.6), although the number of compounds in each set follow different trends. In particular, the general set size has increased more than ten fold in this period, while the refined set, having stricter inclusion requirements has grown only five fold since its birth. The core set, on the other hand, has remained relatively stable for benchmarking, with sizes ranging from 195 to 290 compounds. Overall, the PDBbind database is one if not the most extensive collection of quality protein-ligand complexes and affinity data available today, making it the *de facto* choice for developing novel structure-based scoring functions. BindingMOAD is another well known structure-based affinity database [119, 120].

When it comes to binary affinity data, resources like the Database of Useful Decoys DUD [121] and its enhanced version DUD-E, are commonly used. The latest release, as of the time of writing, features 22886 active compounds drawn from ChEMBL and their binary activity label against 102 protein targets, an average of 224 compounds per each, as well as 50

decoys drawn from ZINC [122] per binder with similar physico-chemical properties but different two-dimensional topologies.

Once a hit, or posteriorly a lead, has been identified we focus on *lead optimization*. In this phase of drug discovery, the chemical structure of such lead is typically modified by a medicinal chemist team with the intent of improving its potency, selectivity, and many other pharmacokinetic and toxicological parameters. These modifications result in congeneric series, a set of ligands with few atom changes between them, usually around a unique or small number of different scaffolds for which there are experimental structures of the complex with the target protein. The scope of this scenario is completely different from the previous one, since a diverse set does most likely not contain enough information so as to distinguish similar molecules whose potency against a target in practice can differ in less than 1 Kcal/mol. In this sense, several simulation-based approaches (which we introduced in the first section) have been developed to estimate the relative binding affinity between a pair of analogues, with relative success. Despite this, these methods suffer from several issues, such as system preparation, treatment of waters, force-field selection, protein flexibility and computational cost, making their prospective application difficult in practice. Due to this, machine learning approaches based on convolutional neural networks have recently been developed for this task [94], showing promising results.

In terms of available databases for congeneric series, the BindingDB [123] protein-ligand validation sets provide (as of the time of writing) 645 congeneric series, which can serve as a base for prototyping models. However, it is likely that the congeneric series of the project of

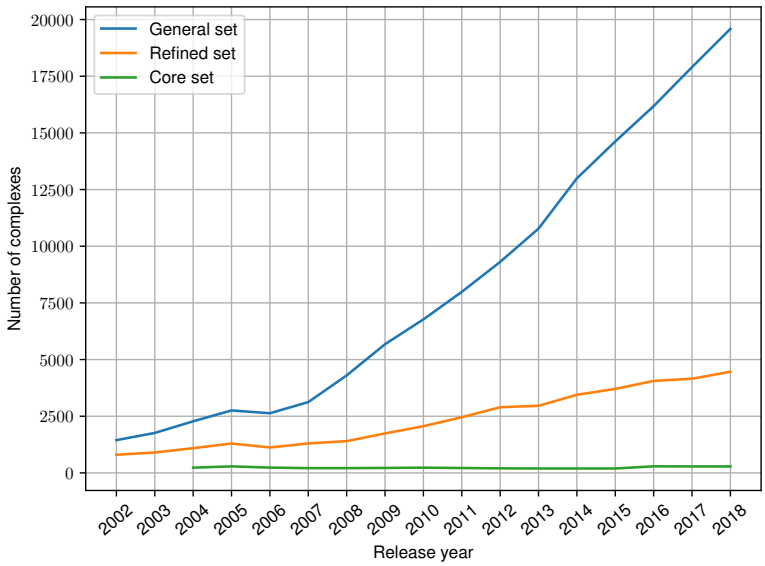


Figure 4.6: Evolution of the number of protein-ligand complexes available in the PDBbind database from 2002 to 2018.

interest has no relation with freely available ones, so that new models have to be built from scratch to take into account its particularities. In this sense, simulation-based approaches have an advantage over data-driven alternatives, such as machine learning models, as they not require prior knowledge of affinity data [124]. However, the latter can always be incrementally trained, increasing its accuracy when more data is available and taking into account particularities of the congeneric series of interest without relying on a physical model.

4.4.2 Evaluation

We divide this subsection in two brief parts. In the first one we discuss several commonly used metrics in the evaluation of scoring functions, while in the second we focus on different data splitting procedures, which can significantly condition results.

Metrics

Scoring functions are evaluated via several metrics, depending on the setup of the study and its goals. In the standard regression setup, some of them are such as the root mean squared error (RMSE), Pearson's correlation coefficient (R) :

$$\mathrm{RMSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4.23}$$

$$R\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}}, \tag{4.24}$$

where y_i and $\hat{y}_i$ represent experimental and predicted affinity i. Spearman's ρ is also a commonly used metric, computed as the Pearson correlation coefficient of the ranked experimental and predicted variables. In a retrospective lead optimization scenario, when one has the chronological experimental order of the series, a simulation-based approach is also appropriate for data-driven scoring functions. This evaluation represents a paradigm where the model "chooses" the ligands to synthesize next, in each iteration increasing its training pool. The evaluation ends when the model picks the ligand with the highest affinity in the series, and its order is then compared with the experimental one achieved by medicinal chemists. If the model was able to retrieve the mentioned compound faster, it taken as an indication that such models are able to accelerate the lead optimization process.

In terms of binary classification (active/inactive) scoring functions, one is instead interesting in retrieving as many active compounds as possible from a given library, therefore metrics such as the enrichment factor (EF) are commonly used:

$$\mathrm{EF}_{k\%} = \frac{\frac{N_a(k\%)}{N(k\%)}}{\frac{N_a}{N}}, \tag{4.25}$$

where $N_a(k\%)$ and $N(k\%)$ are the number of active and total molecules in the top $k\%$ ranked molecules in the library according to the scoring function, and N_a and N are the number of actives and total number of molecules in the entire library. More classical classification metrics are also commonly used, such as the area under the receiver operating curve (AUC) or BEDROC [125].

Splitting procedures

When evaluating scoring functions, if a standard benchmark to compare against is not set beforehand it is important to check model performance under different scenarios. In particular, the most common type of split is some variation of k-fold cross-validation, where we evaluate our model several times in k different splits, using the remaining ones as training data. How the splits are performed can significantly vary results: the most common strategy being randomly assigning each protein-ligand pair. It is well known that this type of split can produce overoptimistic results, particularly in cases when the test set is far (in a statistical distribution sense) from our the data we have previously used to train our model (e.g. different chemical spaces). Several authors have then proposed several alternatives to provide more realistic performance measurements and minimize the possibility of bias, such as scaffold-based splits [126]. In situations where a time-stamp is available, such as in lead discovery, a temporal-based split can also be appropriate [127]. Finally, when developing structure-based approaches, it is also common to consider the performance of the model both in a intra-target and inter-target sense, that is, within and between proteins [128], the latter kind of split typically performed via means of sequence information.

4.5 Discussion

In this chapter our intention was to provide readers with an overview of modern machine-learning structure-based approaches for the prediction of protein-ligand affinities. The success and attention of deep learning methods in this task is unquestionable, with modern models focusing on a representation of the system that is closer to reality, changing the classical feature handcrafting paradigm. In this sense, the most popular approaches are those focusing on either a voxelized representation of the protein, to later use readily available computer vision techniques such as three-dimensional CNNs. The other family of approaches, on the other hand, is a direct descendent of recent ligand-based models, extending the concept of graph convolution to include the protein structure.

While deep learning approaches are attractive for their automatic feature extraction capabilities, among other things, their use in structure-based approaches is necessarily limited. It is a well known fact that deep learning approaches are very data-hungry, and significantly more so than other simpler models if satisfactory performance is to be achieved. For instance, modern CNN architectures for computer vision are trained on the ImageNet database of images [129], which contains more than 14m labeled instances. In contrast, the most extensive structure-based affinity database, namely PDBbind, only features around 20k protein-ligand pairs. This scarcity necessarily forces researchers to design more data-efficient models than the current state of the art.

Interpretability of deep-learning-based models is also one of the topics that has grabbed significant attention in the last years, as we have explained in the corresponding subsection of this chapter. While these tend to be significantly more accurate than other alternatives, they are commonly treated as black boxes, and this comes with two costs: first they are harder to debug, meaning it would be difficult to identify whether a model is learning an inherent bias in the data or the real signal. Secondly, all proposed models have a domain of applicability, in most cases being the prioritization of compound testing. If the decision a model makes is hard to understand for a team of medicinal or computational chemists, justifying its usefulness becomes as difficult.

Bibliography

[1] Ismail Kola and John Landis. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nature Reviews Drug discovery*, 3(8):711, 2004.

[2] Christos A Nicolaou and Nathan Brown. Multi-objective Optimization Methods in Drug Design. *Drug Discovery Today: Technologies*, 10(3):e427–e435, 2013.

[3] Li Di, Paul V Fish, and Takashi Mano. Bridging Solubility Between Drug Discovery and Development. *Drug Discovery Today*, 17(9-10):486–495, 2012.

[4] Jing Lin, Diana C Sahakian, SM De Morais, Jinghai J Xu, Robert J Polzer, and Steven M Winter. The Role of Absorption, Distribution, Metabolism, Excretion and Toxicity in Drug Discovery. *Current Topics in Medicinal Chemistry*, 3(10):1125–1154, 2003.

[5] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. DeepTox: Toxicity Prediction Using Deep Learning. *Frontiers in Environmental Science*, 3:80, 2016.

[6] David J Huggins, Woody Sherman, and Bruce Tidor. Rational Approaches to Improving Selectivity in Drug Design. *Journal of Medicinal Chemistry*, 55(4):1424–1444, 2012.

[7] José Jiménez, Davide Sabbadin, Alberto Cuzzolin, Gerard Martínez-Rosell, Jacob Gora, John Manchester, Jose Duca, and Gianni De Fabritiis. PathwayMap: Molecular Pathway Association with Self-Normalizing Neural Networks. *Journal of Chemical Information and Modeling*, 2018.

[8] David C Swinney. The Role of Binding Kinetics in Therapeutically Useful Drug Action. *Current Opinion in Drug Discovery & Development*, 12(1):31–39, 2009.

[9] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. VAMPnets for Deep Learning of Molecular Kinetics. *Nature Communications*, 9(1):5, 2018.

[10] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 2011.

[11] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The PDBbind Database: Collection of Binding Affinities for Protein-ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980, 2004.

[12] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the Decline in Pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11(3):191, 2012.

[13] Qurrat Ul Ain, Antoniya Aleksandrova, Florian D. Roessler, and Pedro J. Ballester. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 5(6):405–424, 2015.

[14] Jie Liu and Renxiao Wang. Classification of Current Scoring Functions. *Journal of Chemical Information and Modeling*, 55(3):475–482, 2015.

[15] Robert S DeWitte and Eugene I Shakhnovich. SMoG: De Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *Journal of the American Chemical Society*, 118(47):11733–11744, 1996.

[16] Ingo Muegge and Yvonne C Martin. A General and Fast Scoring Function for Protein-ligand Interactions: a Simplified Potential Approach. *Journal of Medicinal Chemistry*, 42(5):791–804, 1999.

[17] Holger Gohlke, Manfred Hendlich, and Gerhard Klebe. Knowledge-based Scoring Function to Predict Protein-ligand Interactions. *Journal of Molecular Biology*, 2000.

[18] Sheng-You Huang and Xiaoqin Zou. An Iterative Knowledge-based Scoring Function to Predict Protein-ligand Interactions: I. Derivation of Interaction Potentials. *Journal of Computational Chemistry*, 27(15):1866–1875, 2006.

[19] Zheng Zheng and Kenneth M Merz Jr. Development of the Knowledge-based and Empirical Combined Scoring Algorithm (KECSA) to Score Protein-ligand Interactions. *Journal of Chemical Information and Modeling*, 53(5):1073–1083, 2013.

[20] Manfred Hendlich, Peter Lackner, Sabine Weitckus, Hannes Floeckner, Rosina Froschauer, Karl Gottsbacher, Georg Casari, and Manfred J Sippl. Identification of Native Protein Folds Amongst a Large Number of Incorrect Models: The Calculation of Low Energy Conformations from Potentials of Mean Force. *Journal of Molecular Biology*, 216(1):167–180, 1990.

[21] Manfred J Sippl. Calculation of Conformational Ensembles from Potentials of Mean Force: An Approach to the Knowledge-based Prediction of Local Structures in Globular Proteins. *Journal of Molecular Biology*, 213(4):859–883, 1990.

[22] Paul D Thomas and Ken A Dill. An Iterative Method for Extracting Energy-like Quantities from Protein Structures. *Proceedings of the National Academy of Sciences*, 93(21):11628–11633, 1996.

[23] Paul D Thomas and Ken A Dill. Statistical Potentials Extracted from Protein Structures: How Accurate are They? *Journal of Molecular Biology*, 257(2):457–469, 1996.

[24] Hui Lu and Jeffrey Skolnick. A Distance-dependent Atomic Knowledge-based Potential for Improved Protein Structure Selection. *Proteins: Structure, Function, and Bioinformatics*, 44(3):223–232, 2001.

[25] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and Testing of a General AMBER Force Field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[26] Kenno Vanommeslaeghe, Elizabeth Hatcher, Chayan Acharya, Sibsankar Kundu, Shijun Zhong, Jihyun Shim, Eva Darian, Olgun Guvench, P Lopes, Igor Vorobyov, et al. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-atom Additive Biological Force Fields. *Journal of Computational Chemistry*, 31(4):671–690, 2010.

[27] Stefan Doerr and Gianni De Fabritiis. On-the-fly Learning and Sampling of Ligand Binding by High-throughput Molecular Simulations. *Journal of Chemical Theory and Computation*, 10(5):2064–2069, 2014.

[28] Gerard Martinez-Rosell, Matt J Harvey, and Gianni De Fabritiis. Molecular-simulation-driven Fragment Screening for the Discovery of New CXCL12 Inhibitors. *Journal of Chemical Information and Modeling*, 58(3):683–691, 2018.

[29] Ignasi Buch, Toni Giorgino, and Gianni De Fabritiis. Complete Reconstruction of an Enzyme-inhibitor Binding Process by Molecular Dynamics Simulations. *Proceedings of the National Academy of Sciences*, 108(25):10184–10189, 2011.

[30] Noelia Ferruz, Stefan Doerr, Michelle A Vanase-Frawley, Yaozhong Zou, Xiaomin Chen, Eric S Marr, Robin T Nelson, Bethany L Kormos, Travis T Wager, Xinjun Hou, et al. Dopamine D3 Receptor Antagonist Reveals a Cryptic Pocket in Aminergic GPCRs. *Scientific Reports*, 8(1):897, 2018.

[31] Noelia Ferruz, Gary Tresadern, Antonio Pineda-Lucena, and Gianni De Fabritiis. Multibody Cofactor and Substrate Molecular Recognition in the Myo-inositol Monophosphatase Enzyme. *Scientific Reports*, 6:30275, 2016.

[32] Vittorio Limongelli, Massimiliano Bonomi, and Michele Parrinello. Funnel Metadynamics as Accurate Binding Free-energy Method. *Proceedings of the National Academy of Sciences*, 110(16):6358–6363, 2013.

[33] Jagdish Suresh Patel, Anna Berteotti, Simone Ronsisvalle, Walter Rocchia, and Andrea Cavalli. Steered Molecular Dynamics Simulations for Studying Protein-ligand Interaction in Cyclin-dependent Kinase 5. *Journal of Chemical Information and Modeling*, 54(2):470–480, 2014.

[34] L. Wang, B. J. Berne, and R. A. Friesner. On Achieving High Accuracy and Reliability in the Calculation of Relative Protein-Ligand Binding Affinities. *Proceedings of the National Academy of Sciences*, 109(6):1937–1942, 2012.

[35] Eelke B. Lenselink, Julien Louvel, Anna F. Forti, Jacobus P. D. van Veldhoven, Henk de Vries, Thea Mulder-Krieger, Fiona M. McRobb, Ana Negri, Joseph Goose, Robert Abel, Herman W. T. van Vlijmen, Lingle Wang, Edward Harder, Woody Sherman, Adriaan P. IJzerman, and Thijs Beuming. Predicting Binding Affinities for GPCR Ligands Using Free-Energy Perturbation. *ACS Omega*, 1(2):293–304, 2016.

[36] Dahlia A. Goldfeld, Robert Murphy, Byungchan Kim, Lingle Wang, Thijs Beuming, Robert Abel, and Richard A. Friesner. Docking and Free Energy Perturbation Studies of Ligand Binding in the Kappa Opioid Receptor. *Journal of Physical Chemistry B*, 119(3):824–835, 2015.

[37] Laura Pérez-Benito, Henrik Keränen, Herman van Vlijmen, and Gary Tresadern. Predicting Binding Free Energies of PDE2 Inhibitors. The Difficulties of Protein Conformation. *Scientific Reports*, 8(1):4883, 2018.

[38] Myriam Ciordia, Laura Pérez-Benito, Francisca Delgado, Andrés A Trabanco, and Gary Tresadern. Application of Free Energy Perturbation for the Design of BACE1 Inhibitors. *Journal of Chemical Information and Modeling*, 56(9):1856–1871, 2016.

[39] Christina Schindler, Friedrich Rippmann, and Daniel Kuhn. Relative Binding Affinity Prediction of Farnesoid X Receptor in the D3R Grand Challenge 2 Using FEP+. *Journal of Computer-Aided Molecular Design*, 32(1):1–8, 2017.

[40] Henrik Keränen, Laura Pérez-Benito, Myriam Ciordia, Francisca Delgado, Thomas B Steinbrecher, Daniel Oehlrich, Herman W T van Vlijmen, Andrés A Trabanco, and Gary Tresadern. Acylguanidine Beta Secretase 1 Inhibitors: A Combined Experimental and Free Energy Perturbation Study. *Journal of Chemical Theory and Computation*, 13(3):1439–1453, 2017.

[41] Shunzhou Wan, Agastya P. Bhati, Sarah Skerratt, Kiyoyuki Omoto, Veerabahu Shanmugasundaram, Sharan K. Bagal, and Peter V. Coveney. Evaluation and Characterization of Trk Kinase Inhibitors for the Treatment of Pain: Reliable Binding Affinity Predictions from Theory and Computation. *Journal of Chemical Information and Modeling*, 57(4):897–909, 2017.

[42] Anita de Ruiter, Stefan Boresch, and Chris Oostenbrink. Comparison of Thermodynamic Integration and Bennett Acceptance Ratio for Calculating Relative Protein-ligand Binding Free Energies. *Journal of Computational Chemistry*, 34(12):1024–1034, 2013.

[43] Zoe Cournia, Bryce Allen, and Woody Sherman. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of Chemical Information and Modeling*, 57(12):2911–2937, 2017.

[44] Johan Åqvist and John Marelius. The Linear Interaction Energy Method for Predicting Ligand Binding Free Energies. *Combinatorial Chemistry & High Throughput Screening*, 4(8):613–626, 2001.

[45] Tomas Hansson, John Marelius, and Johan Åqvist. Ligand Binding Affinity Prediction by Linear Interaction Energy Methods. *Journal of Computer-aided Molecular Design*, 12(1):27–35, 1998.

[46] Samuel Genheden and Ulf Ryde. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-binding Affinities. *Expert Opinion on Drug Discovery*, 10(5):449–461, 2015.

[47] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling*, 51(1):69–82, 2010.

[48] Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: A Basic Tool of Chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[49] Renxiao Wang, Luhua Lai, and Shaomeng Wang. Further Development and Validation of Empirical Scoring Functions for Structure-based Binding Affinity Prediction. *Journal of Computer-aided Molecular Design*, 16(1):11–26, 2002.

[50] Yang Cao and Lei Li. Improved Protein-ligand Binding Affinity Prediction by Using a Curvature-dependent Surface-area Model. *Bioinformatics*, 30(12):1674–1680, 2014.

[51] Marcel L Verdonk, Jason C Cole, Michael J Hartshorn, Christopher W Murray, and Richard D Taylor. Improved Protein-ligand Docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, 2003.

[52] Matthew P Repasky, Mee Shelley, and Richard A Friesner. Flexible Ligand Docking with Glide. *Current Protocols in Bioinformatics*, 18(1):8–12, 2007.

[53] Hans-Joachim Böhm. Prediction of Binding Constants of Protein Ligands: a Fast Method for the Prioritization of Hits Obtained from De Novo Design or 3D Database Search Programs. *Journal of Computer-aided Molecular Design*, 12(4):309–309, 1998.

[54] Matthias Rarey, Bernd Kramer, Thomas Lengauer, and Gerhard Klebe. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *Journal of Molecular Biology*, 261(3):470–489, 1996.

[55] Renxiao Wang, Liang Liu, Luhua Lai, and Youqi Tang. SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-ligand Complex. *Molecular Modeling Annual*, 4(12):379–394, 1998.

[56] Wei Deng, Curt Breneman, and Mark J Embrechts. Predicting Protein-ligand Binding Affinities Using Novel Geometrical Descriptors and Machine-learning Methods. *Journal of Chemical Information and Computer Sciences*, 44(2):699–703, 2004.

[57] Natalia Artemenko. Distance Dependent Scoring Function for Describing Protein-ligand Intermolecular Interactions. *Journal of Chemical Information and Modeling*, 48(3):569–574, 2008.

[58] Pedro J Ballester and John BO Mitchell. A Machine Learning Approach to Predicting Protein-ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics*, 26(9):1169–1175, 2010.

[59] Kun-Yi Hsin, Samik Ghosh, and Hiroaki Kitano. Combining Machine Learning Systems and Multiple Docking Simulation Packages to Improve Docking Prediction Reliability for Network Pharmacology. *PloS ONE*, 8(12):e83922, 2013.

[60] David Zilian and Christoph A Sotriffer. SFCscore RF: A Random Forest-based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling*, 53(8):1923–1933, 2013.

[61] Shuxing Zhang, Alexander Golbraikh, and Alexander Tropsha. Development of Quantitative Structure- Binding Affinity Relationship Models Based on Novel Geometrical Chemical Descriptors of the Protein- Ligand Interfaces. *Journal of Medicinal Chemistry*, 49(9):2713–2724, 2006.

[62] Guo-Bo Li, Ling-Ling Yang, Wen-Jing Wang, Lin-Li Li, and Sheng-Yong Yang. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein-ligand Interactions. *Journal of Chemical Information and Modeling*, 53(3):592–600, 2013.

[63] Tomohiro Sato, Teruki Honma, and Shigeyuki Yokoyama. Combining Machine Learning and Pharmacophore-based Interaction Fingerprint for In Silico Screening. *Journal of Chemical Information and Modeling*, 50(1):170–185, 2009.

[64] Sourav Das, Michael P Krein, and Curt M Breneman. Binding Affinity Prediction with Property-encoded Shape Distribution Signatures. *Journal of Chemical Information and Modeling*, 50(2):298–308, 2010.

[65] Vladimir Chupakhin, Gilles Marcou, Igor Baskin, Alexandre Varnek, and Didier Rognan. Predicting Ligand Binding Modes from Neural Networks Trained on Protein-ligand Interaction Fingerprints. *Journal of Chemical Information and Modeling*, 53(4):763–772, 2013.

[66] Zhan Deng, Claudio Chuaqui, and Juswinder Singh. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-dimensional Protein-ligand Binding Interactions. *Journal of Medicinal Chemistry*, 47(2):337–344, 2004.

[67] Chris de Graaf, Chantal Rein, David Piwnica, Fabrizio Giordanetto, and Didier Rognan. Structure-based Discovery of Allosteric Modulators of Two Related Class BG-protein-coupled Receptors. *ChemMedChem*, 6(12):2159–2169, 2011.

[68] 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.

[69] Jacob D Durrant and J Andrew McCammon. BINANA: a Novel Algorithm for Ligand-binding Characterization. *Journal of Molecular Graphics and Modelling*, 29(6):888–893, 2011.

[70] Jacob D Durrant and J Andrew McCammon. NNScore: A Neural-network-based Scoring Function For the Characterization of Protein-ligand Complexes. *Journal of Chemical Information and Modeling*, 50(10):1865–1871, 2010.

[71] Xuchang Ouyang, Stephanus Daniel Handoko, and Chee Keong Kwoh. Cscore: A Simple Yet Effective Scoring Function for Protein-ligand Binding Affinity Prediction Using Modified Cmac Learning Architecture. *Journal of Bioinformatics and Computational Biology*, 9:1–14, 2011.

[72] Liwei Li, Bo Wang, and Samy O Meroueh. Support Vector Regression Scoring of Receptor-ligand Complexes for Rank-ordering and Virtual Screening of Chemical Libraries. *Journal of Chemical Information and Modeling*, 51(9):2132–2138, 2011.

[73] Qian Liu, Chee Keong Kwoh, and Jinyan Li. Binding Affinity Prediction for Protein-ligand Complexes Based on β Contacts and B Factor. *Journal of Chemical Information and Modeling*, 53(11):3076–3085, 2013.

[74] Duc Duy Nguyen and Guo-Wei Wei. Algebraic Graph Learning of Protein-ligand Binding Affinity. *arXiv preprint arXiv:1812.08328*, 2018.

[75] Pedro J Ballester, Adrian Schreyer, and Tom L Blundell. Does a More Precise Chemical Description of Protein-ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *Journal of Chemical Information and Modeling*, 54(3):944–955, 2014.

[76] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.

[77] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[78] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[79] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[80] Janaina Cruz Pereira, Ernesto Raul Caffarena, and Cicero Nogueira dos Santos. Boosting Docking-based Virtual Screening with Deep Learning. *Journal of Chemical Information and Modeling*, 56(12):2495–2506, 2016.

[81] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein-ligand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 57(4):942–957, 2017.

[82] José Jiménez, Miha Skalic, Gerard Martínez-Rosell, and Gianni De Fabritiis. K DEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 58(2):287–296, 2018.

[83] David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 2013.

[84] Garrett M Morris, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, 2009.

[85] José Jiménez, Stefan Doerr, Gerard Martínez-Rosell, Alexander S Rose, and Gianni De Fabritiis. DeepSite: Protein-binding Site Predictor Using 3D-convolutional Neural Networks. *Bioinformatics*, 33(19):3036–3042, 2017.

[86] Miha Skalic, Alejandro Varela-Rial, José Jiménez, Gerard Martínez-Rosell, and Gianni De Fabritiis. LigVoxel: Inpainting Binding Pockets Using 3D-convolutional Neural Networks. *Bioinformatics*, 35(2):243–250, 2018.

[87] 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*, 2019.

[88] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272. JMLR. org, 2017.

[89] Evan N Feinberg, Debnil Sur, Zhenqin Wu, Brooke E Husic, Huanghao Mai, Yang Li, Saisai Sun, Jianyi Yang, Bharath Ramsundar, and Vijay S Pande. PotentialNet for Molecular Property Prediction. *ACS Central Science*, 4(11):1520–1530, 2018.

[90] Hamed Habibi Aghdam and Elnaz Jahani Heravi. Guide to Convolutional Neural Networks. *New York, NY: Springer. doi*, 10:978–3, 2017.

[91] Dan E Dudgeon. Multidimensional digital signal processing. *Engewood Cliffs*, 1983.

[92] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167*, 2015.

[93] 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.

[94] José Jiménez-Luna, Laura Pérez-Benito, Gerard Martínez-Rosell, Simone Sciabola, Rubben Torella, Gary Tresadern, and Gianni De Fabritiis. DeltaDelta Neural Networks for Lead Optimization of Small Molecule Potency. *(submitted, under review)*, 2019.

[95] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems*, pages 2224–2232, 2015.

[96] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction Networks for Learning About Objects, Relations and Physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016.

[97] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular Graph Convolutions: Moving Beyond Fingerprints. *Journal of Computer-aided Molecular Design*, 30(8):595–608, 2016.

[98] Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical Insights from Deep Tensor Neural Networks. *Nature Communications*, 8:13890, 2017.

[99] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[100] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852, 2016.

[101] Thomas N Kipf and Max Welling. Semi-supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*, 2016.

[102] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated Graph Sequence Neural Networks. *arXiv preprint arXiv:1511.05493*, 2015.

[103] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[104] Leo Breiman. Statistical Modeling: The Two Cultures). *Statistical Science*, 16(3):199–231, 2001.

[105] Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, 2017.

[106] Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making Machine Learning Models Interpretable. In *ESANN*, volume 12, pages 163–172. Citeseer, 2012.

[107] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-wise Explanations for Non-linear Classifier Decisions by Layer-wise Relevance Propagation. *PloS ONE*, 10(7):e0130140, 2015.

[108] Joshua Hochuli, Alec Helbling, Tamar Skaist, Matthew Ragoza, and David Ryan Koes. Visualizing Convolutional Neural Network Protein-ligand Scoring. *Journal of Molecular Graphics and Modelling*, 84:96–108, 2018.

[109] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[110] Min Lin, Qiang Chen, and Shuicheng Yan. Network in Network. *arXiv preprint arXiv:1312.4400*, 2013.

[111] Miha Skalic, Gerard Martínez-Rosell, José Jiménez, and Gianni De Fabritiis. PlayMolecule BindScope: Large-scale CNN-based Virtual Screening on the Web. *Bioinformatics*, 2018.

[112] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[113] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net. *arXiv preprint arXiv:1412.6806*, 2014.

[114] Sören Sonnenburg, Mikio L Braun, Cheng Soon Ong, Samy Bengio, Leon Bottou, Geoffrey Holmes, Yann LeCun, Klaus-Robert MÃžller, Fernando Pereira, Carl Edward Rasmussen, et al. The Need for Open Source Software in Machine Learning. *Journal of Machine Learning Research*, 8(Oct):2443–2466, 2007.

[115] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep Reinforcement Learning That Matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[116] Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling*, 2018.

[117] 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.

[118] 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.

[119] Mark L Benson, Richard D Smith, Nickolay A Khazanov, Brandon Dimcheff, John Beaver, Peter Dresslar, Jason Nerothin, and Heather A Carlson. Binding MOAD, a High-Quality Protein-ligand Database. *Nucleic Acids Research*, 36:D674–D678, 2007.

[120] Aqeel Ahmed, Richard D Smith, Jordan J Clark, James B Dunbar Jr, and Heather A Carlson. Recent Improvements to Binding MOAD: A Resource for Protein-ligand Binding Affinities and Structures. *Nucleic Acids Research*, 43(D1):D465–D469, 2014.

[121] Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 2012.

[122] John J Irwin and Brian K Shoichet. ZINC, a Free Database of Commercially Available Compounds for Virtual Screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005.

[123] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. BindingDB: A Web-accessible Database of Experimentally Determined Protein-ligand Binding Affinities. *Nucleic Acids Research*, 35:D198–D201, 2006.

[124] Adrià Pérez, Gerard Martínez-Rosell, and Gianni De Fabritiis. Simulations Meet Machine Learning in Structural Biology. *Current Opinion in Structural Biology*, 49:139–144, 2018.

[125] Jean-François Truchon and Christopher I Bayly. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *Journal of Chemical Information and Modeling*, 47(2):488–508, 2007.

[126] Christian Kramer and Peter Gedeck. Leave-cluster-out Cross-validation is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *Journal of Chemical Information and Modeling*, 50(11):1961–1969, 2010.

[127] Robert P Sheridan. Time-split Cross-validation as a Method for Estimating the Goodness of Prospective Prediction. *Journal of Chemical Information and Modeling*, 53(4):783–790, 2013.

[128] Jochen Sieg, Florian Flachsenberg, and Matthias Rarey. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling*, 2019.

[129] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Chapter 4

DISCUSSION

The last few years have witnessed an explosion of machine learning applications in many scientific fields. In line with that current trend of research, in this thesis we have explored the potential of deep learning approaches, such as the ones used in computer vision and natural language processing, in drug discovery. This allowed the automation of many tasks that were previously tackled with clever hand-crafted, empirical or algorithmic approaches. Here we discuss the implications and future prospects of the type of models proposed here.

The development of a three-dimensional representation for biomolecular complexes suitable for use in conjunction with convolutional neural networks allowed an unprecedented flexibility in the modeling of relevant drug discovery problems. However, they are not free of limitations, the first one being merely computational, as deep learning architectures are significantly more expensive than other algorithmic approaches, their speed only on par through the use of graphical processing units (GPUs), which severely limits their applicability in many scenarios. While it is also true that deep learning models, and particularly convolutional neural networks, work with a representation closer to the raw data itself, another of the limitations remains their lack of rotational invariance, and the need to choose a resolution. Furthermore, except for standard models such as

ResNets [38] or VGGNets [152], that have been shown to work well in computer vision applications, architecture and hyperparameter search to find the optimal neural network model in drug-discovery-related projects remains an active topic of study.

In our first publication we presented this new paradigm (**Pub. 1**), by tackling the prediction of druggable binding sites we showed that the proposed method had a performance superior or comparable to that of the state of the art for the task. However, posterior works [153] showed that such performance was very sensitive when testing in datasets whose proteins significantly differed (e.g. overall size or number of pockets) from the ones used for model development (namely the scPDB database), suggesting the need for further training data or to account for these cases. Furthermore, the lack of common standardized protocols and benchmarks for this task to fairly evaluate different methodologies remains an open problem.

We then explored in our second publication (**Pub. 2**) the applicability of such models for the well-known task of predicting protein-ligand binding affinities, and developed a model that we named K_{DEEP}. While the model quickly challenged the state-of-art status of other data-driven approaches, performance remained unequal with respect to the target and ligands considered. In particular, we show that while K_{DEEP} and other approaches perform relatively well in the PDBbind binding affinity benchmark, they can perform very poorly when predicting close differences, such as in the congeneric series case of lead optimization, suggesting that the applicability scenarios of the models trained in the aforementioned database is severely restricted. Towards that end we started exploring and developing approaches to predict relative binding affinities in chemically close series (**Pub. 3**), showing good performance with only a few training ligands.

In **Pub. 4** we explored multitask self-normalizing feed-forward neural networks for the task of predicting which metabolic or signaling

pathways a particular molecule can interfere in. Towards that goal we mined the entire ChEMBL database and analized via Uniprot which ligands were active in a specific pathway filtering by an activity threshold. With that in mind, it was only possible to mark compounds as active only if explicit activity information was available, which implied that both non-active compounds towards a target and those with no explicit information towards the same fell in the same ”negative” category. A better treatment of this subtlety remains a further topic of study.

We explored modern molecular generative approaches such as variational autoenconders in **Pub. 5**, to show that our shape-based representation is competitive with other approaches based on SMILES or graph representation of compounds. These approaches have gathered a significant amount of attention in the last years, and can have different goals: from generating compounds similar to a seed, or models that generate molecules to satisfy a predefined goal (e.g. solubility). Validation of these methods has been widely inconsistent, since their usefulness greatly depends on the set objective at a given stage of the drug-discovery pipeline. Consequently, and while some recent effort has been put in standardizing benchmarking protocols [144], there is need for wider adoption of such procedures.

Overall, interpretability also remains one of the main problems of deep learning approaches. Compared to other simpler alternatives such as linear or tree-based models, the mechanism through which a neural network produces a specific answer remains elusive. Despite neural networks having been shown to work particularly well in several fields, they are still considered overparametrized black boxes. There has been some recent effort in improving the explainability of such models, and in particular convolutional neural networks (e.g. atomic gradients [154], class activation maps [155] or layer-wise relevance propagation [156]), but it is still in very early stages. Further effort in this area would foster further collaboration in industrial settings and solidify the utility of machine learning approaches within professionals with a medicinal

chemistry background.

Last but not least, there is need for better standards of collaboration within the field if we want deep learning methods to truly succeed. The latter are notoriously famous for needing a significantly larger amount of data compared to other simpler machine learning algorithms, yet the amount of structural data is insignificant (e.g. the PDBbind refined set, at the time of writing consists of around 4k protein-ligand cocrystals) if we compare it to the millions of pictures in the ImageNet dataset. Pharmaceutical companies have the potential to exploit and share larger amounts of information with the community to drive innovation in deep-learning approaches. In that sense, we are hopeful about the future, with some initial attempts towards the development of a platform for federated and privacy-preserving machine learning in drug discovery [1] [2].

[1] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/imi2-2018-14-03

[2] https://www.ft.com/content/ef7be832-86d0-11e9-a028-86cea8523dc2

Chapter 5

CONCLUSIONS

1. Volumetric representations of biomolecular complexes are a novel and flexible way of modeling shapes and solving different structural biology and chemoinformatics tasks.

2. Pipelines featuring 3D-convolutional neural networks can outperform complex hand-crafted geometric algorithms for the detection of druggable binding pockets, given enough curated training data.

3. Similar approaches that featurize the binding pocket as well as the pose of a compound have been shown to be state of the art in protein-ligand affinity prediction, compared to other scoring functions of diverse nature. However, the performance for ranking chemically close compounds in lead optimization is not consistent, therefore requiring further training in the congeneric series at hand. Such models show promise by outperforming simulation and docking-based approaches with very few examples.

4. Multilabel neural networks are a fast and efficient model that can be used in the association of compounds to the pathways they intervene in, at an unprecedented scale. An open issue remains the equal treatment of negative and unknown activity ligands towards a target.

5. Generative models such as variational autoencoders and captioning networks can be used in conjunction with volumetric representations to generate novel compounds with desirable characteristics while retaining similarity to a seed molecule.

Chapter 6

LIST OF COMMUNICATIONS

This chapter summarizes talks and posters during my PhD

Talks

- Predicting protein-binding affinities with PyTorch. Bioinformatics Open Days 2019. Universidade do Minho (PT), Feb. 2019.

Posters

- Jiménez, J. , Skalic, M. & De Fabritiis, G. K_{DEEP}: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. 1st DCEXS PhD symposium. Barcelona (SP) Nov. 2017.

- Jiménez, J. & De Fabritiis. Relative Protein-ligand Binding Affinity Prediction with 3D-convolutional Neural Networks. 2018 Workshop on Free Energy Methods, Kinetics and Markov State Models in Drug Design. Boston (MA), May. 2018.

- Jiménez, J. & De Fabritiis. Lead Optimization of Congeneric Series via Convolutional Neural Networks. 1st RSC AI in Chemistry Symposium. London (UK), Jun. 2018.

Chapter 7

APPENDIX: OTHER PUBLICATIONS

This section includes publications in which I have contributed a minor part during my PhD.

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

Skalic, M., Varela-Rial, A., Jiménez, J., Martínez-Rosell, G. & De Fabritiis, G. (2018). Bioinformatics, 35(2), 243-250.
https://doi.org/10.1093/bioinformatics/bty583

Abstract

Structure-based drug discovery methods exploit protein structural information to design small molecules binding to given protein pockets. This work proposes a purely data driven, structure-based approach for imaging ligands as spatial fields in target protein pockets. We use an end-to-end deep learning framework trained on experimental protein–ligand complexes with the intention of mimicking a chemist's intuition at manu-

ally placing atoms when designing a new compound. We show that these models can generate spatial images of ligand chemical properties like occupancy, aromaticity and donor-acceptor matching the protein pocket. The predicted fields considerably overlap with those of unseen ligands bound to the target pocket. Maximization of the overlap between the predicted fields and a given ligand on the Astex diverse set recovers the original ligand crystal poses in 70 out of 85 cases within a threshold of 2Å RMSD. We expect that these models can be used for guiding structure-based drug discovery approaches. LigVoxel is available as part of the PlayMolecule.org molecular web application suite.

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

Skalic, M., Martínez-Rosell, G., Jiménez, J. & De Fabritiis, G. (2018). Bioinformatics, 35(7), 1237–1238
https://doi.org/10.1093/bioinformatics/bty758

Abstract

Virtual screening pipelines are one of the most popular used tools in structure-based drug discovery, since they can can reduce both time and cost associated with experimental assays. Recent advances in deep learning methodologies have shown that these outperform classical scoring functions at discriminating binder protein-ligand complexes. Here, we present BindScope, a web application for large-scale active-inactive classification of compounds based on deep convolutional neural networks. Performance is on a pair with current state-of-the-art pipelines. Users can screen on the order of hundreds of compounds at once and interactively visualize the results.

Bibliography

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

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

[3] Schmidhuber J. Deep learning in neural networks: An overview. Neural networks. 2015;61:85–117.

[4] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.

[5] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural computation. 1989;1(4):541–551.

[6] 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.

[7] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.

[8] Cireşan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. arXiv preprint arXiv:12022745. 2012;.

[9] Lawrence S, Giles CL, Tsoi AC, Back AD. Face recognition: A convolutional neural-network approach. IEEE transactions on neural networks. 1997;8(1):98–113.

[10] Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. Journal of machine learning research. 2003;3(Feb):1137–1155.

[11] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

[12] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning. ACM; 2008. p. 160–167.

[13] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473. 2014;.

[14] Socher R, Huval B, Manning CD, Ng AY. Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics; 2012. p. 1201–1211.

[15] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems; 2014. p. 3104–3112.

[16] Graves A, Mohamed Ar, Hinton G. Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE; 2013. p. 6645–6649.

[17] Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. ACM; 2006. p. 369–376.

[18] Saon G, Kurata G, Sercu T, Audhkhasi K, Thomas S, Dimitriadis D, et al. English conversational telephone speech recognition by humans and machines. arXiv preprint arXiv:170302136. 2017;.

[19] Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, et al. Achieving human parity in conversational speech recognition. arXiv preprint arXiv:161005256. 2016;.

[20] Wang H, Wang N, Yeung DY. Collaborative deep learning for recommender systems. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2015. p. 1235–1244.

[21] Cheng HT, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, et al. Wide & deep learning for recommender systems. In: Proceedings of the 1st workshop on deep learning for recommender systems. ACM; 2016. p. 7–10.

[22] Bojarski M, Del Testa D, Dworakowski D, Firner B, Flepp B, Goyal P, et al. End to end learning for self-driving cars. arXiv preprint arXiv:160407316. 2016;.

[23] Huval B, Wang T, Tandon S, Kiske J, Song W, Pazhayampallil J, et al. An empirical evaluation of deep learning on highway driving. arXiv preprint arXiv:150401716. 2015;.

[24] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review. 1958;65(6):386.

[25] Kelley HJ. Gradient theory of optimal flight paths. Ars Journal. 1960;30(10):947–954.

[26] Ackley DH, Hinton GE, Sejnowski TJ. A learning algorithm for Boltzmann machines. Cognitive science. 1985;9(1):147–169.

[27] Rumelhart DE, Hinton GE, Williams RJ, et al. Learning representations by back-propagating errors. Cognitive modeling. 1988;5(3):1.

[28] Dreyfus SE. Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. Journal of guidance, control, and dynamics. 1990;13(5):926–928.

[29] Griewank A. Who invented the reverse mode of differentiation. Documenta Mathematica, Extra Volume ISMP. 2012;p. 389–400.

[30] Baydin AG, Pearlmutter BA, Radul AA, Siskind JM. Automatic differentiation in machine learning: a survey. Journal of Marchine Learning Research. 2018;18:1–43.

[31] Bartholomew-Biggs M, Brown S, Christianson B, Dixon L. Automatic differentiation of algorithms. Journal of Computational and Applied Mathematics. 2000;124(1-2):171–190.

[32] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16); 2016. p. 265–283.

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

[34] Breiman L. Random forests. Machine learning. 2001;45(1):5–32.

[35] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011;2(3):27.

[36] Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V. Support vector regression machines. In: Advances in neural information processing systems; 1997. p. 155–161.

[37] Hochreiter S, Bengio Y, Frasconi P, Schmidhuber J, et al.. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. A field guide to dynamical recurrent neural networks. IEEE Press; 2001.

[38] 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.

[39] Hamel P, Eck D. Learning features from music audio with deep belief networks. In: ISMIR. vol. 10. Utrecht, The Netherlands; 2010. p. 339–344.

[40] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer; 2014. p. 818–833.

[41] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Object detectors emerge in deep scene CNNs. arXiv preprint arXiv:14126856. 2014;.

[42] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–1780.

[43] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014;.

[44] Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature. 1962;194(4824):178.

[45] Cao Y, Li L. Improved Protein-ligand Binding Affinity Prediction by Using a Curvature-dependent Surface-area Model. Bioinformatics. 2014;30(12):1674–1680.

[46] Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved Protein-ligand Docking using GOLD. Proteins: Structure, Function, and Bioinformatics. 2003;52(4):609–623.

[47] Repasky MP, Shelley M, Friesner RA. Flexible Ligand Docking with Glide. Current Protocols in Bioinformatics. 2007;18(1):8–12.

[48] Wang R, Liu L, Lai L, Tang Y. SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-ligand Complex. Molecular Modeling Annual. 1998;4(12):379–394.

[49] Wang R, Lai L, Wang S. Further Development and Validation of Empirical Scoring Functions for Structure-based Binding Affinity Prediction. Journal of Computer-aided Molecular Design. 2002;16(1):11–26.

[50] Hochreiter S, Klambauer G, Rarey M. Machine Learning in Drug Discovery. ACS Publications; 2018.

[51] Deng W, Breneman C, Embrechts MJ. Predicting Protein-ligand Binding Affinities Using Novel Geometrical Descriptors and Machine-learning Methods. Journal of Chemical Information and Computer Sciences. 2004;44(2):699–703.

[52] Zhang S, Golbraikh A, Tropsha A. Development of Quantitative Structure- Binding Affinity Relationship Models Based on Novel Geometrical Chemical Descriptors of the Protein- Ligand Interfaces. Journal of Medicinal Chemistry. 2006;49(9):2713–2724.

[53] Li GB, Yang LL, Wang WJ, Li LL, Yang SY. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein-ligand Interactions. Journal of Chemical Information and Modeling. 2013;53(3):592–600.

[54] Das S, Krein MP, Breneman CM. Binding Affinity Prediction with Property-encoded Shape Distribution Signatures. Journal of Chemical Information and Modeling. 2010;50(2):298–308.

[55] Chupakhin V, Marcou G, Baskin I, Varnek A, Rognan D. Predicting Ligand Binding Modes from Neural Networks Trained on Protein-ligand Interaction Fingerprints. Journal of Chemical Information and Modeling. 2013;53(4):763–772.

[56] Durrant JD, McCammon JA. BINANA: a Novel Algorithm for Ligand-binding Characterization. Journal of Molecular Graphics and Modelling. 2011;29(6):888–893.

[57] Ouyang X, Handoko SD, Kwoh CK. Cscore: A Simple Yet Effective Scoring Function for Protein-ligand Binding Affinity Prediction Using Modified Cmac Learning Architecture. Journal of Bioinformatics and Computational Biology. 2011;9:1–14.

[58] Nguyen DD, Wei GW. Algebraic Graph Learning of Protein-ligand Binding Affinity. arXiv preprint arXiv:181208328. 2018;.

[59] 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.

[60] Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. Frontiers in Environmental Science. 2016;3:80.

[61] Ballester PJ, Mitchell JB. A Machine Learning Approach to Predicting Protein-ligand Binding Affinity with Applications to Molecular Docking. Bioinformatics. 2010;26(9):1169–1175.

[62] Zilian D, Sotriffer CA. SFCscore RF: A Random Forest-based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. Journal of Chemical Information and Modeling. 2013;53(8):1923–1933.

[63] de Graaf C, Rein C, Piwnica D, Giordanetto F, Rognan D. Structure-based Discovery of Allosteric Modulators of Two Related Class BG-protein-coupled Receptors. ChemMedChem. 2011;6(12):2159–2169.

[64] Sato T, Honma T, Yokoyama S. Combining Machine Learning and Pharmacophore-based Interaction Fingerprint for In Silico Screening. Journal of Chemical Information and Modeling. 2009;50(1):170–185.

[65] 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.

[66] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics New York; 2001.

[67] Lowe DG, et al. Object recognition from local scale-invariant features. In: iccv. vol. 99; 1999. p. 1150–1157.

[68] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features. In: European conference on computer vision. Springer; 2006. p. 404–417.

[69] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: international Conference on computer vision & Pattern Recognition (CVPR'05). vol. 1. IEEE Computer Society; 2005. p. 886–893.

[70] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3431–3440.

[71] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–241.

[72] Iizuka S, Simo-Serra E, Ishikawa H. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Transactions on Graphics (TOG). 2016;35(4):110.

[73] Xie J, Girshick R, Farhadi A. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In: European Conference on Computer Vision. Springer; 2016. p. 842–857.

[74] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:181204948. 2018;.

[75] Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in neural information processing systems; 2016. p. 2172–2180.

[76] Zhu JY, Zhang R, Pathak D, Darrell T, Efros AA, Wang O, et al. Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems; 2017. p. 465–476.

[77] Jiménez J, Doerr S, Martínez-Rosell G, Rose A, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017;33(19):3036–3042.

[78] Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. Journal of Computational Chemistry. 2009;30(16):2785–2791.

[79] Koes DR, Baumgartner MP, Camacho CJ. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. Journal of Chemical Information and Modeling. 2013;53(8):1893–1904.

[80] 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.

[81] 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.

[82] Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. The Journal of physiology. 1968;195(1):215–243.

[83] Dudgeon DE. Multidimensional digital signal processing. Engewood Cliffs. 1983;.

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

[85] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 2014;15(1):1929–1958.

[86] 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.

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

[88] Liang J, Woodward C, Edelsbrunner H. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein science. 1998;7(9):1884–1897.

[89] Kahraman A, Morris RJ, Laskowski RA, Thornton JM. Shape variation in protein binding pockets and their ligands. Journal of molecular biology. 2007;368(1):283–301.

[90] Cohen TS, Geiger M, Köhler J, Welling M. Spherical CNNs. arXiv preprint arXiv:180110130. 2018;.

[91] Cohen T, Welling M. Group equivariant convolutional networks. In: International conference on machine learning; 2016. p. 2990–2999.

[92] Karp G. Cell and molecular biology: concepts and experiments. John Wiley & Sons; 2009.

[93] Lyne PD. Structure-based virtual screening: an overview. Drug discovery today. 2002;7(20):1047–1055.

[94] Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-based virtual screening for drug discovery: a problem-centric review. The AAPS journal. 2012;14(1):133–141.

[95] Valdar WS. Scoring residue conservation. Proteins: structure, function, and bioinformatics. 2002;48(2):227–241.

[96] Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007;23(15):1875–1882.

[97] Huang B, Schroeder M. LIGSITE csc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC structural biology. 2006;6(1):19.

[98] Weisel M, Proschak E, Schneider G. PocketPicker: analysis of ligand binding-sites with shape descriptors. Chemistry Central Journal. 2007;1(1):7.

[99] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC bioinformatics. 2009;10:168.

[100] Xie L, Bourne PE. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. In: BMC bioinformatics. vol. 8. BioMed Central; 2007. p. S9.

[101] Nicolaou CA, Brown N. Multi-objective Optimization Methods in Drug Design. Drug Discovery Today: Technologies. 2013;10(3):e427–e435.

[102] Di L, Fish PV, Mano T. Bridging Solubility Between Drug Discovery and Development. Drug Discovery Today. 2012;17(9-10):486–495.

[103] Lin J, Sahakian DC, De Morais S, Xu JJ, Polzer RJ, Winter SM. The Role of Absorption, Distribution, Metabolism, Excretion and Toxicity in Drug Discovery. Current Topics in Medicinal Chemistry. 2003;3(10):1125–1154.

[104] Huggins DJ, Sherman W, Tidor B. Rational Approaches to Improving Selectivity in Drug Design. Journal of Medicinal Chemistry. 2012;55(4):1424–1444.

[105] Swinney DC. The Role of Binding Kinetics in Therapeutically Useful Drug Action. Current Opinion in Drug Discovery & Development. 2009;12(1):31–39.

[106] Mardt A, Pasquali L, Wu H, Noé F. VAMPnets for Deep Learning of Molecular Kinetics. Nature Communications. 2018;9(1):5.

[107] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research. 2011;40(D1):D1100–D1107.

[108] Wang R, Fang X, Lu Y, Wang S. The PDBbind Database: Collection of Binding Affinities for Protein-ligand Complexes with Known Three-Dimensional Structures. Journal of Medicinal Chemistry. 2004;47(12):2977–2980.

[109] Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the Decline in Pharmaceutical R&D efficiency. Nature Reviews Drug Discovery. 2012;11(3):191.

[110] Ain QU, Aleksandrova A, Roessler FD, Ballester PJ. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2015;5(6):405–424.

[111] Liu J, Wang R. Classification of Current Scoring Functions. Journal of Chemical Information and Modeling. 2015;55(3):475–482.

[112] DeWitte RS, Shakhnovich EI. SMoG: De Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. Journal of the American Chemical Society. 1996;118(47):11733–11744.

[113] Muegge I, Martin YC. A General and Fast Scoring Function for Protein-ligand Interactions: a Simplified Potential Approach. Journal of Medicinal Chemistry. 1999;42(5):791–804.

[114] Gohlke H, Hendlich M, Klebe G. Knowledge-based Scoring Function to Predict Protein-ligand Interactions. Journal of Molecular Biology. 2000;.

[115] Huang SY, Zou X. An Iterative Knowledge-based Scoring Function to Predict Protein-ligand Interactions: I. Derivation of Interaction Potentials. Journal of Computational Chemistry. 2006;27(15):1866–1875.

[116] Zheng Z, Merz Jr KM. Development of the Knowledge-based and Empirical Combined Scoring Algorithm (KECSA) to Score

Protein-ligand Interactions. Journal of Chemical Information and Modeling. 2013;53(5):1073–1083.

[117] Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and Testing of a General AMBER Force Field. Journal of Computational Chemistry. 2004;25(9):1157–1174.

[118] Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, et al. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-atom Additive Biological Force Fields. Journal of Computational Chemistry. 2010;31(4):671–690.

[119] Pérez-Benito L, Keränen H, van Vlijmen H, Tresadern G. Predicting Binding Free Energies of PDE2 Inhibitors. The Difficulties of Protein Conformation. Scientific Reports. 2018;8(1):4883.

[120] Ciordia M, Pérez-Benito L, Delgado F, Trabanco AA, Tresadern G. Application of Free Energy Perturbation for the Design of BACE1 Inhibitors. Journal of Chemical Information and Modeling. 2016;56(9):1856–1871.

[121] Schindler C, Rippmann F, Kuhn D. Relative Binding Affinity Prediction of Farnesoid X Receptor in the D3R Grand Challenge 2 Using FEP+. Journal of Computer-Aided Molecular Design. 2017;32(1):1–8.

[122] Gathiaka S, Liu S, Chiu M, Yang H, Stuckey JA, Kang YN, et al. D3R grand challenge 2015: evaluation of protein–ligand pose and affinity predictions. Journal of computer-aided molecular design. 2016;30(9):651–668.

[123] Gaieb Z, Liu S, Gathiaka S, Chiu M, Yang H, Shao C, et al. D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. Journal of computer-aided molecular design. 2018;32(1):1–20.

[124] Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. Journal of the American Chemical Society. 2015;137(7):2695–2703.

[125] Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: A Web-accessible Database of Experimentally Determined Protein-ligand Binding Affinities. Nucleic Acids Research. 2006;35:D198–D201.

[126] Cournia Z, Allen B, Sherman W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. Journal of Chemical Information and Modeling. 2017;57(12):2911–2937.

[127] Sertkaya A, Wong HH, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. Clinical Trials. 2016;13(2):117–126.

[128] Kola I, Landis J. Can the Pharmaceutical Industry Reduce Attrition Rates? Nature Reviews Drug Discovery. 2004;3(8):711.

[129] Wassermann AM, Lounkine E, Hoepfner D, Le Goff G, King FJ, Studer C, et al. Dark chemical matter as a promising starting point for drug lead discovery. Nature chemical biology. 2015;11(12):958.

[130] Jiménez J, Sabbadin D, Cuzzolin A, Martínez-Rosell G, Gora J, Manchester J, et al. PathwayMap: Molecular pathway association with self-normalizing neural networks. Journal of chemical information and modeling. 2018;.

[131] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27–30.

[132] Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic acids research. 2013;42(D1):D472–D477.

[133] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the universal protein knowledgebase. Nucleic acids research. 2004;32(suppl_1):D115–D119.

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

[135] Schnecke V, Boström J. Computational chemistry-driven decision making in lead generation. Drug discovery today. 2006;11(1-2):43–50.

[136] Schneider G, Fechner U. Computer-based de novo design of drug-like molecules. Nature Reviews Drug Discovery. 2005;4(8):649.

[137] Sanchez-Lengeling B, Outeiral C, Guimaraes GL, Aspuru-Guzik A. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). Harvard University, Chem Rxiv. 2017;.

[138] Elton DC, Boukouvalas Z, Fuge MD, Chung PW. Deep learning for molecular design-a review of the state of the art. Molecular Systems Design & Engineering. 2019;.

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

[140] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in neural information processing systems; 2014. p. 2672–2680.

[141] Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoders. arXiv preprint arXiv:151105644. 2015;.

[142] 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.

[143] 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.

[144] Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: Benchmarking Models for de Novo Molecular Design. Journal of Chemical Information and Modeling. 2019;59(3):1096–1108.

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

[146] Liu Q, Allamanis M, Brockschmidt M, Gaunt A. Constrained graph variational autoencoders for molecule design. In: Advances in Neural Information Processing Systems; 2018. p. 7795–7804.

[147] Schneider G. De novo design–hop (p) ing against hope. Drug Discovery Today: Technologies. 2013;10(4):e453–e460.

[148] Kumar A, Zhang KY. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. Frontiers in chemistry. 2018;6:315.

[149] 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.

[150] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:150203044. 2015;.

[151] Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS computational biology. 2009;5(12):e1000585.

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

[153] Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of cheminformatics. 2018;10(1):39.

[154] Hochuli J, Helbling A, Skaist T, Ragoza M, Koes DR. Visualizing Convolutional Neural Network Protein-ligand Scoring. Journal of Molecular Graphics and Modelling. 2018;84:96–108.

[155] 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.

[156] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one. 2015;10(7):e0130140.