Statistical and Machine Learning Approaches to Predicting Protein-Ligand Interactions

Lucy J. Colwell
Department of Chemistry, Cambridge University, Cambridge, UK

Abstract

Data driven computational approaches to predicting protein-ligand binding are currently achieving unprecedented levels of accuracy on held-out test data sets. Up until now, however, this has not led to corresponding breakthroughs in our ability to design novel ligands for protein targets of interest. This review summarizes the current state of the art in this field, emphasizing the recent development of deep neural networks for predicting protein-ligand binding. We explain the major technical challenges that have caused difficulty with predicting novel ligands, including the problems of sampling noise and the challenge of using benchmark datasets that are sufficiently unbiased that they allow the model to extrapolate to new regimes.

Introduction

Molecular recognition is a fundamental requirement of biological systems. The interactions between proteins and small molecules are central to biology, allowing cells to sense their surroundings and respond appropriately. Estimates place the number of small molecules that can be synthesized at \( \approx 10^{60} \), yet just a small fraction of potential protein-ligand interactions have been explored. Finding novel interactions is of great importance to drug discovery and basic biology. Given the enormity of the search space, computational approaches can narrow down the possibilities. However, despite three decades of computational effort, biochemical experiments are still essential to determine the efficacy of
ligand binding to a protein target [1, 2]. The results of computational analysis have been decidedly mixed: it is challenging to use even experimentally well-characterized ligand protein interactions to computationally design novel interactions [1, 2], much less explore the vast space of possibilities.

There are three increasingly demanding tasks in protein-ligand binding prediction: virtual screening predicts whether a ligand binds to a given target; affinity prediction predicts the binding affinity; and pose prediction identifies the molecular interactions causing binding to occur. In this review we focus on the first; the others have been reviewed elsewhere [3, 4]. Approaches to virtual screening can be categorised as physical or statistical. The idea of using first principles physical models to describe protein-ligand interactions is attractive, however timescale and computational resource constraints mean that simplified descriptions of features such as protein flexibility, and solvent are necessary. Even the most sophisticated docking algorithms cannot accurately reproduce large numbers of known interactions, much less predict new ones. Scoring functions can be empirical [5, 6, 7, 8, 9] or knowledge-based [10, 11, 12, 13, 14]. Significant expertise is required to encode physico-chemical interactions through the use of hand-tuned features and parameters, and can be highly specific to the system that they are designed for [15].

Recently, the use of high throughput methods to screen large libraries of proteins and small molecules and quantify their interactions has made it possible to correlate activity with representations of proteins and small molecules, to infer predictive models. Techniques from machine learning and artificial intelligence have been introduced, allowing both the parameters and the model to be learned from the wealth of experimental data available in databases such as Chembl [16, 17, 18, 19]. Increasingly publications are demonstrating that data driven approaches have the potential to make significant contributions to these problems [20**, 21*, 22*, 23*, 24, 25, 26, 27].
Machine learning - potential and limitations

The aim of any machine learning or statistical approach is to identify patterns among training examples that can be used to make predictions outside the training set. An algorithm achieves this by mapping the training set representation to a space in which active and inactive ligands segregate - this mapping can be guided by physical models [13][23] but is more often learned directly from the data without addition of extensive physico-chemical knowledge [21*, 22*][26]. Once this mapping has been learned, it is hoped that the location of new examples in this space - clustered with training set actives or inactives - will accurately predict their activity. To compare the performance of different algorithms, a test set of ligands with known activity is held-out during the training process. An algorithm that can make accurate predictions for this unseen data is presumed to extrapolate well.

In particular, approaches that use deep neural networks (DNNs) have been shown to make predictions on held-out test sets with eye-opening levels of accuracy, exceeding 0.95 AUC (Area under the Receiver Operator Characteristic) on benchmark datasets [23][22*]. However, the extent to which such results extrapolate is not yet clear - are they overfit to the training data [29, 1][30**]? A number of studies posit that the test/train protocol is not as exacting as it might appear due to the non-uniformity of the distribution of ligands in chemical space. If training set actives are closer to test set actives than to training set inactives, by some metric that is not fully predictive of protein-ligand binding activity such as molecular weight, or number of ring systems, then the algorithm can appear to make accurate predictions for test set molecules without being able to extrapolate this predictive ability [31, 19, 32]. Machine learning performs best when abundant data drawn uniformly from the space of interest is available, but in this setting human chemists choose which molecules to work with, often based on clear similarities to known success stories [33, 34].

Definitively showing that these approaches generalise is perhaps the outstanding challenge facing this field today. In the search for novel pharmacen-
Figure 1: **Schematic of a deep neural network for predicting properties of small molecule ligands.** In this schematic, the input molecule is first encoded in a binary molecular fingerprint of fixed length. Convolutional architectures are a popular choice among practitioners and typically require that a fixed length representation be either used or learned from the molecular graph. The neural network simply consists of a very large number of simple functions (or neurons) of the form indicated in the schematic. Deep neural networks have large numbers of layers of neurons, resulting in a huge number of parameters that are learnt during training from the data, which is typically split into distinct training, cross-validation and test datasets.

The goal of this article is to review statistical approaches to molecular recognition in the context of protein-ligand binding, focusing on recent results that exploit DNNs (see figure 1). Here we briefly outline the basic steps of machine learning algorithm that predicts protein-ligand binding.
Molecular Representation

There are almost as many choices for representation of the input data as there
are for the machine learning algorithm employed \cite{35, 36}. The simplest involve
counting the numbers of different heavy atoms present in a ligand, together with
other features such as hydrogen bond donors/acceptors, chiral centers and ring
systems \cite{19, 30**}. Some information about the chemical structure is retained
by descriptors such as atom pairs or donor-acceptor pairs \cite{37, 38} where each
element has the form \((\text{atom type } i) - (\text{distance in bonds}) - (\text{atom type } j)\).
More information is encoded by chemical fingerprints, for example MACCS
keys \cite{39} and ECFP fingerprints \cite{40}; fixed length binary descriptors which can
be generated by the package RDKit \cite{41}. Here, each non-hydrogen atom is used
as a centre from which fragments are generated by extending radially from the
centre along bonds to neighbouring atoms; the maximum radius considered \(N\)
is encoded in the name as ECFP2\(N\). A unique identifier is assigned to each
fragment, and the set of identifiers for a molecules is mapped to a fixed length
bit vector to yield the molecular fingerprint.

This abundance raises the question of which representation is most useful
for different prediction tasks. Recently the suggestion has been made that it
may be more effective to also learn the molecular representation itself, alongside
the metric and corresponding embedding space used to distinguish active from
inactive ligands \cite{21*, 22*, 27, 42}. However, counter-intuitively it has also been
reported that the use of more complex molecular descriptors can result in little
gain of predictive ability \cite{43, 30**}

Representation and Sampling Noise

One rational for the finding that more complex representations can result
in little improvement is noise due to finite sampling. The basic premise of any
predictive algorithm is that similarities among known interaction partners can
reveal the requirements of the binding site, and thus predict novel interactions.
A straightforward approach is to compile the set of ligands known to bind to a
protein receptor of interest, and identify those features that show statistically significant enrichment among this set [44]. However, because there are only finitely many samples (i.e. known ligand binders), some features will be enriched purely by chance. This chance similarity increases with the number of variables, so representations that have more variables will lead to greater random similarity between features. For DNNs in particular, representations with thousands or even millions of features have recently been employed [24][21*, 22*]; although these algorithms have the ability to share information between targets it is still important that the level of chance similarity between small molecules is quantified and accounted for.

This phenomena has been carefully studied in the field of random matrix theory, which provides a null distribution that describes the similarity between samples (ligands) that can be expected by chance due to finite sampling as a function of the number of samples available, and the number of variables present in the ligand descriptor [45][46*]. A simpler method for generating this null distribution simply involves computing the covariance matrices of multiple sets of \( n \) random ligands, where \( n \) is the sample size, using the same ligand descriptor for each set, and so obtaining the distribution of the largest entries that occur due to finite sampling noise. A similar approach can be taken for any measure of molecular similarity, defining a statistical null distribution with which to compare putative active molecules.

**Benchmark Datasets**

The application of machine learning algorithms to problems such as protein-ligand binding is facilitated by the availability of large experimental datasets. For both algorithms that require 3D complexes and those that work with 2D ligand structures, a number of benchmark datasets have emerged that allow researchers to compare the performance of different algorithms. The PDB-bind benchmark [47] comprises 1300 diverse protein-ligand complexes with high quality structural and binding data. Another popular choice is the Directory
of Useful Decoys (DUD) [48] or Directory of Useful Decoys Enhanced (DUDE) benchmark [18], which have been carefully constructed to attempt to control for dataset biases that enable a machine learning algorithm to learn features that distinguish actives from inactives without providing information about the relevant molecular interaction. The construction of benchmark datasets that are unbiased, and so provide an accurate indication of how well an algorithm can extrapolate is an important research direction that requires attention [49*][50][30*].

**Choice of Algorithm**

Early linear regression methods were soon super-seeded by nonlinear models constructed using early versions of neural networks, among other techniques. Algorithms such as support vector machines [51][52], Gaussian Procceses [53][54] and Random Forests (RF) [55][18][56][43] have all received significant attention within the community. One advantage of these approaches is that many allow some level of interpretability, enabling the medicinal chemist to rationalise and evaluate the resulting model. This is more challenging for DNNs (see figure 1), which have recently been found to be highly successful on a variety of different tasks including computer vision [57] and speech recognition [58]. In the last few years there has been a burst of activity involving the construction of DNNs to predict protein-ligand binding [20**, 21*, 22*] [24, 25, 26, 28, 59, 60, 61, 62].

A key breakthrough came in 2012 with the Merck sponsored Kaggle contest, in which the winning entry submitted by G. Dahl improved the mean squared Pearson correlation coefficient ($R^2$) between predicted and observed binding activities over 15 proprietary Merck datasets from 0.42 (attained using RFs) to 0.49 [20**]. Although a combination of DNNs [63*], gradient boosting machine [64] and Gaussian process regression [53] were used by the winning entry, Dahl suggests that the DNN was the main cause of the performance increase. A careful comparison of DNNs and RFs at protein-binding prediction over 15 datasets found that the best $R^2$ values over all parameter sets are obtained
using DNNs for five datasets, and using RFs for one, with mixed results for the remaining datasets. For refined parameter ranges, this increased to nine datasets for DNNs compared to one for RFs, leading the authors to conclude that DNNs have the potential to outperform RFs generally.

A crucial issue with DNNs is the set of algorithmic parameters affecting predictive performance that must be decided by the researcher, such as network size, choice of activation functions and use of dropout, which are distinct from parameters that are learned from the data. In general it is not computationally feasible to search the space of parameter values, in particular because different parameters are not independent of each other. Some analysis has been presented by varying one parameter at a time, and recording the $R^2$ values of the resulting DNNs.

**Conclusion**

Advances in technology have led to enormous opportunity for using machine learning algorithms to learn models of protein ligand interactions. Machine learning algorithms are inherently flexible and with a sufficient amount of carefully curated data, the algorithms have the potential to find patterns that humans cannot find on their own. However, besides the issues already discussed with this review, other challenges must be overcome: the parameters learned from the data can lead to overfitting - the unintended specialisation of the trained model to the dataset used for training. For example, if the training and test datasets were identical, the algorithm would simply memorise all details of the training data, and achieve perfect performance on the test dataset. However, it would likely be unable to make accurate predictions for additional molecules that it had not previously seen before.

For chemical data, a key problem is to define what constitutes a distinct partitioning between test and training data. Simply requiring that the molecules are different from one another is not enough. Machine learning algorithms are extremely effective at learning any feature that distinguishes
samples from different classes (such as binding and non-binding). This means that it is necessary to go further, and require that active molecules do not share any nuisance similarity or property that distinguishes them from inactive molecules, while not being the crucial molecular feature that enables binding. The situation is complicated by the fact that in most cases the crucial feature that enables binding is not known - indeed identifying these feature(s) is the ultimate goal of the exercise. However relatively simple devices such as synthetic datasets, in which binding is said to occur if particular logical combinations of features are present, or bias measures that quantify simple molecular similarities can be surprisingly effective.

Despite these challenges, this is an exciting time for building models of protein–ligand interactions. One could hope that a review written five years hence would be filled more with victories than warnings.

References


learning molecular fingerprints, in: Advances in neural information processing systems, 2015, pp. 2224–2232. * Introduces an approach in which molecular descriptors are learned directly from the molecular graph. The authors report that these data-driven features are more interpretable, and display better performance on a number of prediction tasks.


[30] I. Wallach, A. Heifets, Most ligand-based benchmarks measure overfitting rather than accuracy, arXiv preprint arXiv:1706.06619 (2017). ** Using a novel metric the authors demonstrate that many benchmark datasets are biased. This implies that machine learning algorithms trained on these datasets may not generalize beyond the domain of the training data as well as might be expected, given the reported performance on held out test data.


[43] P. J. Ballester, A. Schreyer, T. L. Blundell, Does a more precise chemical description of protein–ligand complexes lead to more accurate prediction


[46] A. A. Lee, M. P. Brenner, L. J. Colwell, Predicting protein–ligand affinity with a random matrix framework, Proceedings of the National Academy of Sciences 113 (48) (2016) 13564–13569. * This paper applies random matrix theory to develop a novel approach that identifies the spurious enrichment of features among active or inactive ligands that can be caused by finite sampling effects.


