

Original Article

Improving Compound Selection in Drug Discovery: A Quantitative Approach for Biased Data Modeling

Rohit Singh Raja

¹Associate Director of Quality Engineering, Texas, USA

Received Date: 27 November 2024

Revised Date: 31 December 2024

Accepted Date: 21 January 2025

Abstract: According to the latest findings from the World Health Organization (WHO), cardiovascular disease reigns supreme as the leading global cause of mortality. Detecting heart ailments at an early stage is of paramount importance, as managing the condition often necessitates proactive measures like lifestyle modifications and preventive medications. Failing to address the issue promptly may unleash a cascade of cardiovascular complications, potentially culminating in heart attacks or other life-threatening events that demand immediate medical intervention and exhibit alarmingly high fatality rates. To confront this challenge, an extensive dataset procured from Kaggle, containing a plethora of patient information alongside an identifier indicating the presence or absence of underlying heart disease, will be harnessed. Through the implementation of state-of-the-art optimization techniques, a binary classification machine learning model will be trained to predict the likelihood of new, unseen patients harboring underlying heart disease. Multiple optimization methods will be rigorously compared to unveil the most optimal model, tailored precisely to address this pressing issue.

Keywords: Driven Drug Discovery, Machine Learning Models, Predictive Modeling, Space Optimization.

I. INTRODUCTION

A longstanding goal of artificial intelligence research is to construct algorithms to aid in the design of new drugs [1], [2]. To date, a plethora of machine learning algorithms have been proposed to automatize most of the preclinical drug discovery steps, from the generation of structurally novel molecules showing drug-like properties [3] to the prediction of *in vitro* potency and binding affinities [4]. However, there is still a paucity of integrative and fully data-driven computational pipelines to propose candidate molecules for further experimental testing in an iterative fashion till molecules with a set of desired properties are found.

Here, I present a modeling framework based on a set of multi-parameter optimization algorithms to discover structurally novel and active compounds Using a data set heavily biased towards active compounds as a starting point. Specifically, I discuss in depth two general problems that beset any attempt to use quantitative techniques to direct a drug discovery project, namely: extrapolation using a heavily biased data set as starting point and multi-parameter optimization to identify compounds satisfying a set of quantitative conditions. I use publicly available data on activity against the parasite responsible for malaria with the greatest morbidity and mortality: *Plasmodium falciparum*.

I use the Tres Cantos Antimalarial Set (TCAMS) as a case study of a heavily biased data set. This data set encompasses the 13,533 compounds that inhibited the growth of *Plasmodium falciparum* 3D7 by at least 80% at 2 microM concentration out of roughly 2 million compounds screened (active rate 0.7%) [5]. The structures for the inactive compounds were not reported, and hence, the available structures correspond to only active compounds.

Although I use standard machine learning algorithms in this study, I note that my interpretation of the results is somewhat different from a standard predictive modeling outlook. In this work, I formulate and investigate the tasks of model extrapolation and multi-parameter optimization geometrically, with molecular space M living inside the space $M \rightarrow F = [0, 1]^{128}$ associated with 128-bit Morgan fingerprints, and a similarity metric given by Tanimoto distance.

The first question I examine is model extrapolation: given a set of molecules whose values for some features of interest are known, how well will a model that is trained on such data perform in predicting that same feature for very different molecules? In other words, how well will a model perform when extrapolated away from its training area? This difficulty is particularly acute when trying to build models from public data due to the biases in the way this data is reported [6], [7]. Here, I show how to use the geometry of molecular space embedded in F to correct for this bias. my results indicate a way in which predictive models can be trained on a small and biased dataset but nonetheless produce sensible generalizations when predicting arbitrary inputs.

The second question I address is multi-parameter optimization. A successful drug requires many different properties, and being active against the target is only one of these. It must also be non-toxic and have the correct chemical



properties to allow it to reach the target inside the human body. I argue that optimizing for these multiple objectives can be facilitated by considering virtual screening as a geometry problem. This is again due to the geometry of F . The correlation between any two random vectors is very close to zero due to the size of the dimension of the space. Thus, it is feasible for an optimizer, starting from one point, to move in a ‘good’ direction that is orthogonal to several other directions (which is precisely what a multi-parameter optimizer must do). Let $B(m, \epsilon) \subset M$ be the set of molecules that are Tanimoto distance less than or equal to ϵ from the molecule m . The difficulty comes from the fact that given some molecule $m \in M$, the number of molecules in $B(m, \epsilon)$ may be small for small ϵ , so although it is easy to find a vector direction starting from a particular molecule in which to search for an improvement, there just aren’t any nearby molecules *in that direction*. However, the number of molecules in $B(m, \epsilon)$ grows extremely fast as a function of ϵ . My results indicate that correctly adjusted models will generate informative predictions for values of ϵ sufficiently large that there are, in fact, plenty of molecules to choose from in any desired direction.

II. METHODS

A. Collection and Curation of Antimalarial Screening Data

For this project, I gathered antimalarial screening data from ChEMBL database version 23 using the `chembl_webresource_client` python module [8]. Specifically, I assembled a first data set by downloading potency values against *Plasmodium Falciparum* from the following seven Malaria screening collections available in ChEMBL, namely: GSK TCAMS, St Jude, MMV Malaria Box, Novartis, Harvard, OpenS, and WHO-TDR. IC_{50} values were modeled in a logarithmic scale ($pIC_{50} = -\log_{10}IC_{50}$ [M]). The code required to download directly from ChEMBL all these data sets, as well as the assay IDs for all of them, is available on the accompanying GitHub repository for this article: <https://github.com/owatson/MalariaPaper>.

In addition, I assembled a second data set by downloading both antimalarial screening (i.e., potency) and *in vitro* toxicity data (50% growth inhibition bioassay endpoint values measured by screening the compounds data set against the cell line HepG2) for the compounds in the TCAMS data set. The toxicity values from this data set were used to build the toxicity models described below.

Finally, I used a third data source to explore the commercially available chemical space during the optimization steps. This is the library of (approximately) 7.5M commercially available compounds from Molport. I use these compounds as:

- A proxy for ‘accessible molecular space’.
- A domain from which to select the most promising compounds according to various optimization criteria I investigated (see below).

B. Molecular Representation

I standardize all chemical structures in all datasets described above to a common representation scheme using the Python module standardizer (<https://github.com/flatkinson/standardiser>). Inorganic molecules were removed, and the largest fragment was kept in order to filter out counterions [9]. To represent molecules for subsequent model generation, I computed circular Morgan fingerprints [10] for all compounds using RDKit (release version 2013.03.02) [11]. Specifically, I computed hashed Morgan fingerprints in binary format using the RDKit function `GetMorganFingerprintAsBitVect`, which returns values in F_2^{128} , and in count format, using, in this case, the RDKit function, `GetHashedMorganFingerprint`, which returns values in N^{128} . I decided to use Morgan fingerprints as compound descriptors given the higher retrieval rates obtained with this descriptor type in comparative virtual screening studies [12]. The radius was set to 2 and the fingerprint length to 128. I note that longer fingerprints are associated with higher predictive power [13]. However, I did not observe a large improvement when I used longer fingerprints to model this data set. Hence, I decided to use short fingerprints to decrease the computational footprint of my analyses.

C. Machine Learning

I built Random Forest (RF) and Ridge Regression models using the python library `scikit-learn` [14], as previously described [15]. In order to assess the predictive power of the models, I performed K-fold cross-validation. Specifically, I trained RF and Ridge Regression models on bootstrap sub-samples of the data and used the out-of-sample (OOS) results for each subsample to calculate the RMSE and Pearson correlation coefficients (R^2) values for the predicted against the observed potency or toxicity values. In the vein of generating reproducible research [16], [17], all code and data sets (other than Molport data) used to generate this research, as well as a Jupyter notebook containing all the necessary analyses to reproduce the results and figures reported in this contribution, are available at <https://github.com/owatson/MalariaPaper>.

III. RESULTS AND DISCUSSION

A. Consistency of the Antimalarial Screening Data Publicly Available

I first analyzed the consistency of the antimalarial screening data available in ChEMBL. To this aim, I examined the various ability of the bioactivity measurements for compounds tested using diverse assays. My analysis revealed that it is

important to restrict the data one query from ChEMBL to *specifically* bioactivity data points measured under controlled experimental conditions. Otherwise, one obtains a startling variety of results. For instance, by simply extracting all screening data annotated against *Plasmodium Falciparum*, I found over 1,000 different values for Chloroquine (an antimalarial drug) activity. The standard deviation of the recorded bioactivity values for Chloroquine was greater than that for the data. This is in line with previous large-scale analyses of the concordance of public data[6], [18], [19], and reinforces the need for stringent filtering and curation steps to gather high-quality data prior to modeling.

Given the low concordance of the public data, I decided to only consider the TCAMS data set to represent active compounds (Figure 1). I extended the TCAMS data set with inactive compounds by including data points from ChEMBL corresponding to antimalarial data points against any *Plasmodium Falciparum* strain exhibiting pChEMBL values < 5.5. I note that I *cannot* assume that all other ChEMBL data are inactive against *Plasmodium*, as several well-known drugs (and hence active) are missing from the data. Dealing with this non-trivial bias in the data is a key issue I confront in the model fitting phase.

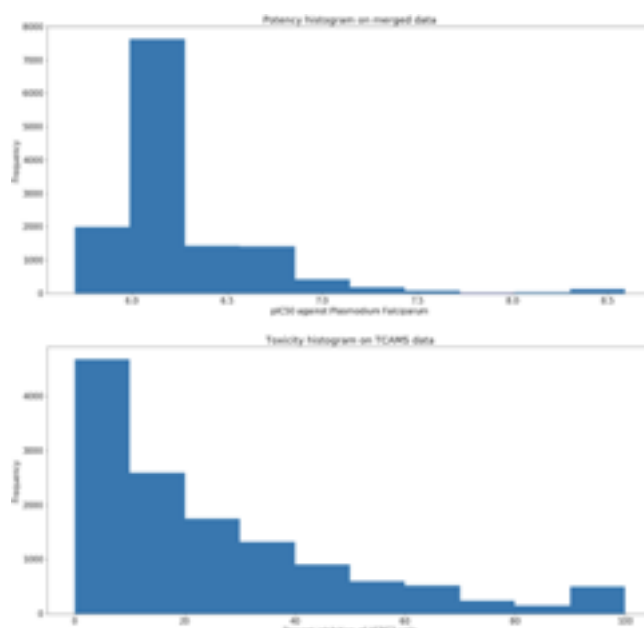


Figure 1: Histograms for Toxicity and Potency Data

B. Modelling compound potency and toxicity against *Plasmodium Falciparum*

Next, I sought to determine the predictive power of RF and Ridge Regression models trained on the full data set. To estimate these quantities, I took 40 random bootstrap samples of my data on which to fit my models (be they ridge regression or random forest) and then used the non-sampled data to compute a value for the quantity of interest (beta and correlation coefficient in this case). Confidence intervals were then taken from the resulting distribution of values.

I show below the mean and the 95% confidence intervals for the R^2 between the predicted and observed potency/toxicity values (I do not show the betas of these models since these were never statistically significantly different from 1.0; i.e., these models neither overfit nor underfit).

Overall, the most predictive models exhibited RMSE and R^2 values of 0.3 and 0.4, respectively. This level of performance is statistically higher than simply predicting the average pIC50 value in the training set for the test set instances ($R^2=0$; RMSE=0.45). Given that a large fraction of the data points in the TCAMS data set have a pIC50 value of around 6, this analysis is important to corroborate that the low RMSE values obtained are not spurious (i.e., a consequence of simply predicting the average value for all test set instances). I note that the reduced range of pIC50 values encompassed in my data set partially underlies the low R^2 values I obtain[20]. These results are in accordance with the expectations of my previous work[15] and models reported in the literature for data sets extracted from ChEMBL: Random Forest models substantially outperform simple linear models when there are no constraints on the in-sample/out-of-sample division.

Together, these results indicate that my choice of algorithms and molecular representation permits to obtain predictive models, although they do not explain a substantial fraction of the variance in the data set.

I also examined the predictive power of models trained using Morgan fingerprints in count format as predictors. These underperformed (non-statistically significantly; Kolmogorov-Smirnov test; $P > 0.05$) relative to binary fingerprints in

predicting potency, and outperformed (again without statistical significance) in predicting toxicity. Given the comparable results of both types of fingerprints, I will only refer to the models based on binary fingerprints for the remainder of the paper.

C. Molecular Similarity and Tanimoto Distance

Result	Mean	Lower 5% Bound	Upper 5% Bound
Potency Ridge Regression R^2	0.111	0.098	0.124
Potency Random Forest R^2	0.304	0.282	0.327
Toxicity Ridge Regression R^2	0.125	0.106	0.139
Toxicity Random Forest R^2	0.299	0.287	0.309

Table I: Modelling Results For Potency And Toxicity Using 128-Bit Morgan Fingerprints In Binary Format

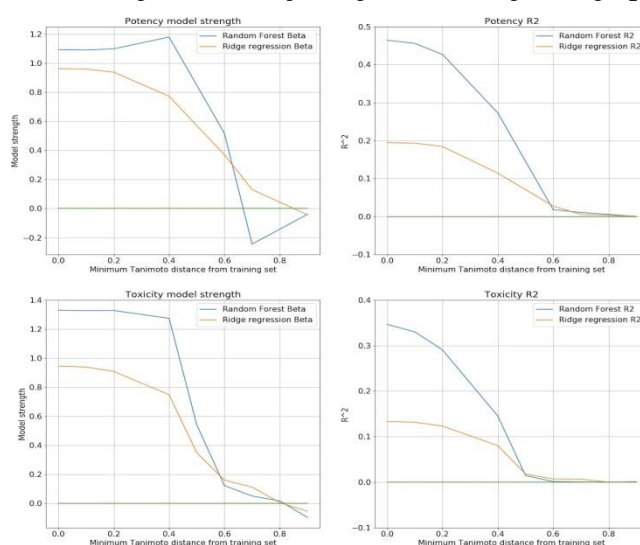


Figure 2: Model Extrapolation as a Function of Tanimoto Distance

Quantifying molecular similarity is key to chemoinformatic applications in general [21] and to my formulation of virtual screening as a geometry problem. The standard distance metric used to measure the similarity and dissimilarity of molecules is the Tanimoto (sometimes called Jaccard) distance [metric\[22\]](#). In this section, I wish to develop some intuition as to why this metric is the right one for my analysis and its basic properties. I map molecules into a binary vector space of some high dimension (128 in this work). In this vector space a ‘1’ implies the existence of some specific molecular substructure in the molecule, and ‘0’ implies its absence (I note that in this work I do not consider the issue that multiple substructures might be mapped to the same fingerprint position).

The Tanimoto similarity of two binary fingerprints A and B is simply the ratio of the size of the intersection of A and B over the size of the union of A and B [22]. In my setting, it is the number of substructures common to the two compounds represented by A and B , divided by the total number of substructures that appear in at least one of the two compounds. The Tanimoto distance is simply $1 - \text{Tanimoto similarity}$. The intuitive plausibility for this being a reasonable metric on molecular space is that two compounds that share no features should presumably be maximally different (unlike in, e.g., the Euclidean space, where two compounds that each only contained one substructure, different in either case, would be very similar).

A good metric in my context will be one for which it is true that molecules close in that metric have similar toxicity and potency values, and I show in the next section that this is indeed the case for these quantities. Suppose that any substructure has a one-in-two chance of being found in a (randomly chosen somehow) molecule. The molecular fingerprints would have each bit randomly being 1 or 0 with equal probability $1/2$. If this were the case, then for two random compounds A and B , there would be approximately $1/4$ of the bits both 1 in their fingerprints and $3/4$ of the bits 1 in at least one of their fingerprints. Hence, the Tanimoto similarity between A and B would be $1/3$, and the Tanimoto distance would be $2/3$. More generally, if one could think of molecules as randomly having any given molecular structure with fixed probability p , then the Tanimoto distance between two random molecules would be given by $(2 - 2p)/(2 - p)$. Empirically, looking at the sets of

molecules I will be dealing with, the average Tanimoto distance between two randomly selected molecules is around 0.7, (see Figure 2), showing that a simple model of 'each substructure is present in any compound with probability (slightly under) 1/2' is not too inaccurate.

Given that I am interested in finding potent molecules structurally dissimilar to those in the training data, I next sought to determine the relationship between Tanimoto distance and compound potency (Figure 2). This allowed us to quantify the rate of change in these quantities as one moves around in molecular space or how correlated the potency and toxicity of two compounds will be, depending on how similar they are. I note, however, that in the case of potency, this correlation needs to be adjusted for bias, as I will describe in a later section. Overall, the fact that the standard deviation increases (and thus correlation decreases) between the potency and toxicity of two compounds as they move further away from each other in Tanimoto distance shows that Tanimoto distance is an appropriate metric to use empirically, as well as theoretically, for this data set. I will pursue this idea in the next subsections.

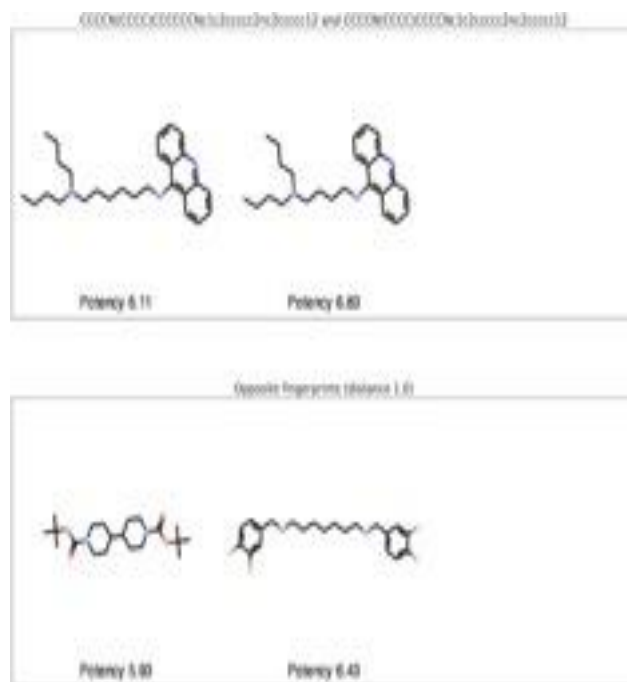


Figure 3: Examples of Compounds Leading to A) Identical Fingerprints but Exhibiting Different Potency Values, and B) Fingerprints at a Tanimoto Distance of 1.0.

D. Co-Variance Of Molecular Properties As A Function Of Tani- Moto Similarity

Next, I investigated the following question. Suppose I know some value of interest (say potency or toxicity) for some compound X. How much uncertainty is there in my knowledge of that value for some other compound Y, depending on how similar X and Y are?

Clearly, suppose Y is very dissimilar to X (essentially a randomly chosen compound). In that case, knowledge of X's properties will tell us nothing about those of Y. Thus, my uncertainty about Y's properties will be the 'base rate' uncertainty for those properties. If on the other hand Y is very similar to X, presumably my uncertainty about Y's properties will diminish (this is in some sense the 'whole point' of doing any kind of QSAR modelling). In an extreme case, I can imagine that Y is *the same* compound as X, in which case the true uncertainty will be 0 (although the data uncertainty, representing noise in assays or data collection, might be substantially higher than 0). In this work, I use 128-bit binary fingerprints as my mathematical representation of compounds. Two compounds can have the same 128-bit representation without being identical, as I show in figure 3 (together with an example of two compounds with 'opposite' fingerprints, or ones that are at maximum Tanimoto distance 1.0 from each other). Understanding the extent of uncertainty reduction as a function of the similarity between two compounds is key to one aspect of quantitative drug discovery. When I select a set of compounds to test, I want to maximize the *information gain* I achieve by testing that set of compounds (subject to minimizing the cost of testing). Clearly, if the compounds I test are too similar, the information I glean is lower (since knowledge of any one of the test outcomes would tell us a lot about the expected outcome of all the other tests). This is a principled way to address the question of 'Exploration vs. Exploitation.' I try to test compounds that I expect to be *good*, but also ones that are not too similar to each other.

E. Model Extrapolation as a Function of Tanimoto Distance

It is intuitively plausible, and well known among practicing chemists that models perform poorly on molecules that are *different* from those that the models were trained upon [23]. In this subsection, I attempt to formalize this notion. I have already seen that Tanimoto distance provides a plausible and empirically successful measure of distance in molecular space. I expect my models to degrade in their accuracy as I use them to predict data points far (in Tanimoto distance) from the training set. I can use the training set to assess the extent of this degradation.

The first step in fitting my models is to transform the response variables in my data into (approximately) normally distributed values with a mean of 0. The mean 0 part is particularly important so that I do not have to worry about whether, when regressing my predictions against my responses, I need to include a constant term. Thus, from my original data set of responses r_i , I obtain normed responses n_i .

The best accuracy I can possibly hope for is given by the ‘Leave one out’ accuracy. This is calculated as follows. For any data point, fit a model on all the other data points, predict for the original data point, and add this prediction to my set. You will then have a set of predictions p_i^N . The ‘Leave-one-out’ accuracy is thus the accuracy of the predictions p_i versus r_i .

I can extend this concept as follows: for any data point, fit a model on all the other data points that are at least distance d from that point. Again, I have a set of predictions $p(d)_{i=1}^N$ and responses r_i . The ‘Distance d ’ accuracy is the accuracy of these predictions. In particular, this method should (with some important caveats that I address in the next section) give a good estimate of the model accuracy when predicting values on a molecule at a distance d away from the training set. If the ‘intuition’ I refer to at the start of this section is correct, then I should see accuracy degrade significantly as d increase.

Let us define $\beta(d)$ by the following equation:

$$r_i = \beta(d)p(d)_i + \epsilon \quad (1)$$

Where r_i is some response value of interest, and $p(d)_i$ is the prediction obtained from a model (of unspecified type) fit on all datapoints $m_k : \text{TD}(m_k, r_i) > d$

Similarly, I can define $R(d)$ as the fraction of the variance of the r_i explained by the equation above (a.k.a. the R -squared of the model) Figure 2 shows plots of $\beta(d)$ and $R(d)$ where the r_i is the log potency in the TCAMS data set, as a function of d . In each plot, I show the functions arising from Random-Forest fits and ridge regression models (both chosen with reasonable parameters, as described in [15]).

These figures, as one of the main results of this paper, deserve some commentary. First, let us note that in line with general wisdom, the fraction of variance explained by the models decreases as you extrapolate further and further away from the dataset. Moreover, the Random Forest model dominates the linear model up to extrapolation distance of around 0.6, where in any case, virtually no variance is explained. The results shown relate to earlier work done by the authors in [23] and [15]. Note that around the extrapolation of Tanimoto distance 0.7, the ridge model still has a small positive beta (i.e., some very small predictive capacity). In contrast, the Random Forest model now has zero or even negative beta. This is in line with [15], which shows that at sufficient levels of extrapolation, the more constrained linear models outperform the more complex machine learning ones. However, at this level of extrapolation, both models are effectively useless since they explain virtually none of the variance in the data. What is perhaps most surprising about these results is the quite large extrapolation distance at which the model quality is *not much diminished*. For example, the Random Forest model at 0.4 extrapolation distance conveys almost half the information as at extrapolation distance 0, where I would have a compound with an identical 128-bit fingerprint in my dataset already.

F. Correcting for Bias in the Training Data

Before I can try to apply my models sensibly in a search for new drug candidates, I need to address the issue of the (extreme) bias in the data I have. Obviously, random compounds are not expected to be active against the malaria parasite. However, it is very clear from looking at a potency histogram of values in my data (Figure 1) that exclusively *active* compounds have been reported. Nor can I assume that any ‘well-known’ (e.g., present in ChEMBL) compound has been screened against malaria and, if not present in my dataset, is inactive against *Plasmodium Falciparum*. In particular, some known drugs (e.g., Doxycycline) are missing from the data, indicating that I am missing some active compounds.

Observing the potency histogram of my malaria data set, I see a peak in pIC50 around 6.5 (Figure 1). A standard ‘inactive’ molecule would have pIC50 value (to within the limits of experimental error) of less than 5 [24]. I will assume all *inactive* compounds have a pIC50 of 3.5 for simplicity.

I need to make sensible and conservative corrections to my model to allow it to make reasonable predictions on any compound input. I do this by breaking the activity prediction into two parts: the activity level (assuming the model is active) and the probability that the compound is active in the first place. I use a Bayesian calculation, where the conditioning

variable is the minimum Tanimoto distance of the compound from any previously seen active ($pIC_{50} > 5$) compound.

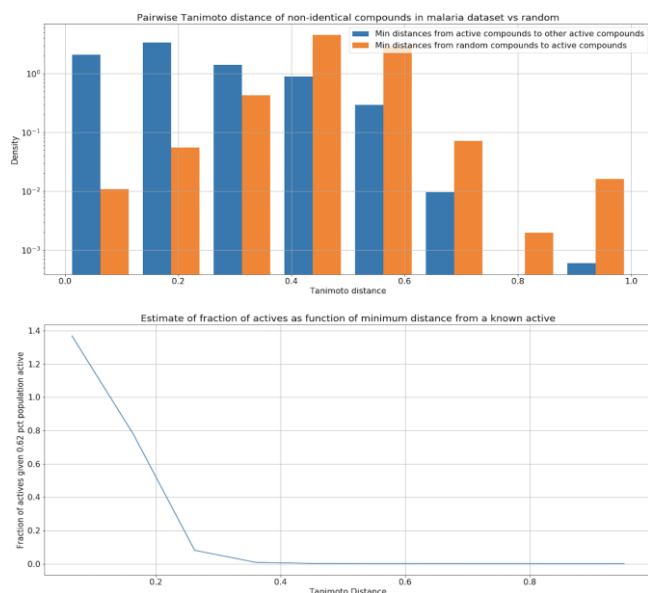


Figure 4: Bias Correction to Estimate Location of Inactive

In Figure 4, I show these two density histograms. The first one corresponds to 100 million pairwise distances between ten thousand randomly chosen compounds from my malaria dataset. Thus, the first histogram represents the distribution of two compounds that are active against *P. falciparum*. The second density histogram was constructed by calculating another 100 million distances between:

- ten thousand randomly selected molecules from the malaria dataset
- ten thousand randomly selected molecules from my Mol- port database.

Thus, these histograms represent the distributions of the distance between a random molecule from my (active) malaria dataset and

- Other random active molecules
- Other random (be it active or inactive) molecules that are reasonable in some way (say readily obtainable).

I note (as might be expected) that, given a molecule active against the malaria parasite, other actives are more likely to be close to it than random molecules are.

I then use these two distributions to estimate the relative density of active vs inactive compounds in the neighborhood of an active compound. I do, however, need a key unobserved value π , namely the overall fraction of active compounds in the full population of 'reasonable' molecules. In my case, I can arrive at an estimate of this unknown value in two ways:

- A 'guesstimate' approach. If I assume that around two million compounds were screened against *Plasmodium Falciparum*[5], which is the sort of size amenable to current high-throughput technologies[25], [5], then I would get a fraction of around 0.01 of all compounds being active in order to create my roughly twenty thousand compound-sized datasets (13,533 out of 2 million gives an active rate of 0.7%).
- Directly from the data, given that I can plausibly assume that compounds very close to an active are themselves active, or at least highly likely to be (e.g., from Figure 4, the covariance in the potency of two compounds with identical fingerprints is low). If I assume this, then I can calculate that the probability of a compound X being (very) close to an active compound Y given that X is active should be $1/\pi$ times the probability of an arbitrary compound Z being close to Y .

These two approaches give roughly the same answer, as shown by the fact that when I use my guesstimate of 0.01 for π in the lower plot of Figure 4, I see that the fraction of estimated actives is close to 1.0 for when the Tanimoto distance from an active tends to 0. I therefore use this value of 0.01 (or one compound in 100 being expected to be active against the malaria parasite) for the rest of this paper.

Thus, at this stage, I can finally write down my model in full general. For an unknown compound C , the predicted potency will be:

$$E[\text{Potency}(C)] = E[\text{Potency}(C \mid C \in A)] \cdot \Pr(C \in A) \quad (2) \\ + I \cdot (1 - \Pr(C \in A)) \quad (3)$$

where A = set of Active molecules and I = activity level of inactive molecules. In this case $E[\text{Potency}(C \mid C \in A)]$ will be a function of:

- My actively fit model prediction on C
- The minimum distance of C from my dataset (as C moves further away from my training set in Tanimoto space, I should, from Figure 2 believe my estimate towards the mean activity level).¹

$\Pr(C \in A)$ is simply a function of the minimum Tanimoto distance of C to the malaria dataset. With the toxicity data, the nature of the bias (if any) is not as clear. Therefore, I will not attempt to correct for bias in the toxicity model and simply take the predictions as they are provided by the model (in particular, I will not believe toxicity prediction towards the mean toxicity as I move further away from the malaria dataset).

G. Multi-Parameter Optimization (MPO)

In this section, I illustrate how the models described above could be used in a prospective drug discovery project. I illustrate how different goals can be traded off against each other using my models. I firstly apply the models fit on the potency and toxicity data to new compounds. To this aim, I use the 7.5M commercially available molecules from Molport (Methods) and look at the top selections from these according to my models along various criteria.

As a starting point, what are the compounds predicted to be most potent in the Molport database (removing stereoisomers or any compounds with identical SMILES)? I show the top five in Figure 5.

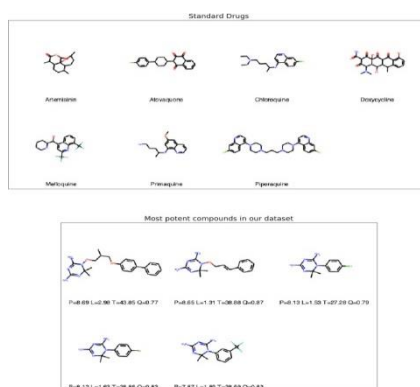


Figure 5: Standard drugs, and my most Potent predicted compounds: P=Potency, L=Lipophilicity, T=Toxicity (%) Q=QED

All these compounds are very slight variations of compounds present in my training dataset, as all show a Tanimoto distance of 0 to compounds in the training set, apart from the fourth compound, which shows a distance of 0.078. Note that the log P values of these compounds are not in the range [2-4] (generally thought to be a good indicator of favorable ADMET properties), and QED values are below 0.5 (another ADMET indicator). The top hit, in fact, is Monensin, an antibiotic that is known to have anti-malarial activity^[26]. If one requires good Lipophilicity and QED values and low predicted toxicity values, one gets as top suggestions the compounds shown in Figure 6.

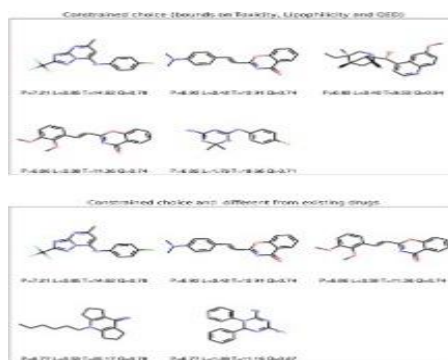


Figure 6: Best Potency, with Lipophilicity and QED thresholds, as well as good Toxicity

A key goal in the fight against malaria is the development of new drugs as the parasites evolve immunity to existing ones. Commercially, one may also aim to develop drugs with improved side-effect profiles or cover novel Intellectual Property space. I still, of course, aim to discover lead compounds that are potent, non-toxic, and have favourable ADMET qualities. For the purposes of a drug discovery program, I want to search as large an area of molecular space as possible and as cheaply as possible. I distil these qualitative goals into the following list of quantitative requirements.

- To represent my goal of finding *novel* candidate drugs, I select seven malaria drugs and require that my selected compounds are at least 0.6 in Tanimoto distance away from any of these.
- As an indication of reasonable ADMET properties, I require logP of my compounds to be within the range [2-4][27].
- As an additional indicator of favorable ADMET properties, I require that the QED of the compounds be > 0.5.
- Predicted toxicity using the toxicity models described above must be < 25%.

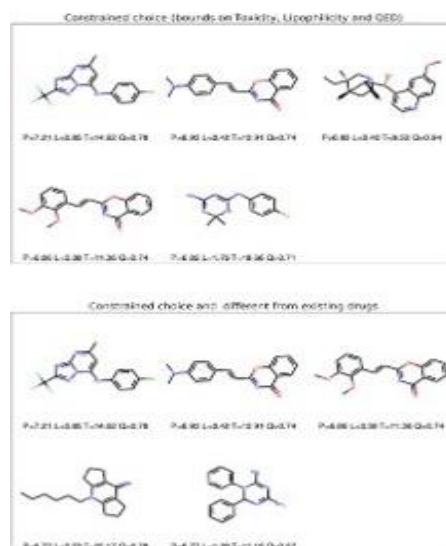


Figure 7: Most potent choices subject to all criteria

This set of requirements picks up a subset of the compounds from my original training set, shown in Figure 7. To the authors' knowledge, these do not show anti-malarial activity, although they satisfy all the criteria that I could think of imposing.

It is of course not surprising that I find compounds in my original training set when I look for 'most potent' compounds, even with all the other criteria I impose.

For the purposes of this study, however, I want to look at the predictions of the model on compounds that are different from any it has seen before. Therefore, I show in Figure 7 my top selections according to the above criteria *that are also at least 0.2 Tanimoto distant from any molecule in the original malaria dataset, thus guaranteeing that these molecules are structurally novel* together with the closest compounds in my original dataset to these choices.

The key points to draw from Figure 7 are that (i) the predicted potency values correspond to moderate activity (obviously at the low end of the range of active compounds since I am deliberately moving away from the training set), and (ii) that given this, I can make sensible probabilistic estimates not only for each compound having potency beyond a certain threshold but also by using the covariance models in Figure ??, for *at least one of my chosen compounds being in a desired range*. Therefore, it is possible to use standard optimization techniques (such as simulated annealing) to choose an optimal set of compounds for testing, and it can even incorporate information as to the cost of purchasing and testing the compounds into this optimization. It seems worth mentioning also that although at 0.2 Tanimoto distance compounds might still share functional groups and chemical moieties, and certainly within the range at which my models are expected to predict very well, even at this short distance, I am finding compounds that are structurally different from their nearest neighbors in the training set. Taken together, these results show that it is possible to satisfy a multi-objective optimization and still find potent candidates, at least when one has a training set as large as the one, I had in this instance.

H. Retrospective Discovery of Highly Active Compounds

Thus far, I have explored the use of my models in making predictions on unseen compounds and choosing sets of compounds according to certain criteria. But how much should I trust these predictions? I have not yet challenged the proposed methodology using shown data on the covariance estimation anywhere. I therefore conclude the body of this paper by showing a simple but representative test to illustrate that I carried out that uses all parts of the methodology.

Specifically, I hold out the eight best compounds (henceforth the Target Compounds) in the training data set and apply the full probabilistic model described in Equation 3 to find these among the N first compounds from the Molport database. The eight 'best' compounds those with pIC_{50} values greater than 7.6, and toxicity less than 5%. Following my previous work [15] that suggested that the best way to separate compounds in terms of Tanimoto similarity is actually to separate on activity levels, I used as my training dataset all the compounds in my malaria set with pIC_{50} values less than 7.5. Indeed, the minimum distance between a target compound and the training set was 0.15, while the maximum was 0.37. This is substantially closer than a random set of 8 compounds, but the model will have to 'extrapolate' somewhat to find them.

I. Objective and Probability of Discovering the Target Compounds by Chance

My objective function was to choose 20 compounds from the $N + 8$ available in such a way as to maximize the likelihood that at least one of those compounds would be a target compound, the value of 20 being chosen arbitrarily. The probability of getting a target compound by choosing a random set of 20 compounds is given by:

$$1 - \frac{(n-8)(n-7)\dots(n-27)}{n(n-1)\dots(n-19)} \quad (4)$$

a) My Analysis Reveals That:

1. If $N = 10000$ and I choose the top 20 compounds by predicted activity, I get three target compounds. The probability of getting at least one by random chance is 1.6%.
2. If $N = 20000$, and I chose the top 20 compounds by predicted activity, I get two target compounds. The probability of getting at least one by random choice is 0.8%.
3. If $N = 70000$, as before, I get one target compound (0.2% probability by chance).
4. If $N = 80000$, I get no target compounds.
5. At first view, these results might seem to indicate that while my models do indeed make my selections better than random, they are not *particularly* impressive.
6. *Using the Covariance Information to Choose the Best Twenty Compounds in the Test Data*
7. The reason for this, however, is that when I choose the top M compounds by predicted potency, I am **not** optimizing for my objective function. I don't want the compounds with the highest expected potency. In the context of the experiment I have set up, I want the compounds with the *highest likelihood of having potency* ≥ 7.6 . To understand how to choose *these* compounds, I need to examine my potency covariance plot in Figure. When I chose the top twenty compounds by predicted potency from a set of 80000, all the compounds that are selected are at Tanimoto distance 0 from my training set (apart from one, which is at a distance 0.05). The top compound by predicted potency has an expected potency of 6.85, but given it has minimum Tanimoto distance 0 from my training dataset, this means (by definition) that *either it was in my training dataset, or it has identical fingerprint to a compound in my dataset*. Looking at the simple fit to potency covariance in Figure ??, this means that the uncertainty in potency is of the order of 0.15 (the value of the y-intercept of the fit).² This, in turn, implies that the probability that this compound has potency ≥ 7.6 is a 5-sigma event, i.e., something with a probability of less than one in one million. Suppose I wish to choose the compounds with a maximum likelihood of having potency greater than some target value. In that case, I need to estimate the uncertainty of my predictions as a function of the (minimum) distance of the compound from my training set. Fortunately, I have all the necessary information to do this. I simply calculate the following: $Var(d) = Var\{p(d)A(d) + (1-p(d))I\}$ (5)
8. where: d is the Tanimoto distance, $p(d)$ is the probability of a compound being active, $A(d)$ is the distribution of actives, and I is the inactive level. When I plot this, I find that the variance peaks at around Tanimoto distance 0.3 from the training dataset (Figure 8).
9. What I should, in theory, do at this point is the following: for any set of twenty compounds, calculate the likelihood that at least one of these compounds has potency greater or equal to my target value of 7.6, and choose that set of twenty. To do this, one needs to use not only the distance of the selected compounds from my training set but also the distance from each other (if I test the same compound twice, we're really only testing 19 compounds rather than twenty, and correspondingly lowering my chances of success).
10. In practice, however, this is too difficult. The objective function has no nice analytic expression and thus needs to be approximated stochastically (e.g., via a Monte Carlo approach). The sample space is a very large discrete space of choices ^{N^{20}} and algorithms such as stochastic annealing work poorly (and slowly) when one can only approximate the objective function. In the particular case I examine here, this does not matter, however. If I simply choose the twenty compounds *that individually are the most likely* to have potency.
11. ≥ 7.6 then I find three target compounds if $N = 300000$ (probability of getting at least one 0.0005). When I set $N = 450000$, I find two target compounds (probability of getting at least one 0.00036). This corresponds to an enrichment factor of 300 and 278, respectively, over a random selection of compounds. In the general case, when using these methods to

choose compounds in a drug discovery project, it would be advisable to check how similar the chosen compounds were to each other. The researcher can then check the value of the objective function on various sets of compounds chosen, using a parameter ϵ to reject any compound if it is within Tanimoto distance ϵ of one already chosen. This should give an easy heuristic to achieve the objective function minimum

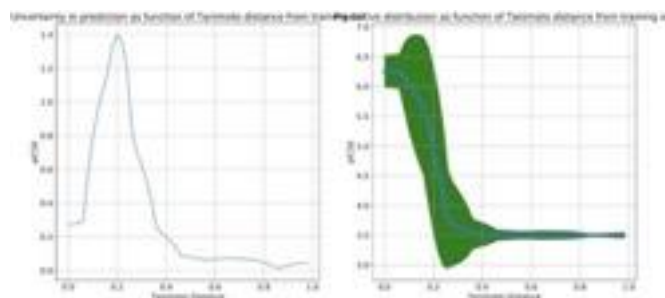


Figure 8: Standard Deviation (Uncertainty in Prediction) as a Function Of Tanimoto Distance.

IV. CONCLUSION

In this contribution, I have introduced an alternative modeling framework to pursue quantitative drug discovery that accounts for the bias in the data. Here, I have used a data set heavily biased towards active compounds. However, the modeling framework I propose here can be applied to data sets showing other biases (e.g., towards inactive molecules). I have shown that models can be fit on molecular data that give reasonable results when applied to any molecular inputs, as shown by the fact that when presented with the full set of commercially available molecules from Molport the models pick up ones they have seen in the training set but also predict plausible values for structurally novel molecules. This is important since unless one has confidence in one's models' predictions for arbitrary inputs, one is forced in optimization to drastically restrict the range in which one searches for a solution. The fact that I have plausible real- value predictions (i.e., predictions that I expect to neither overestimate nor underestimate the value of interest) and can build proper variance models to accompany them means that I can properly specify an objective function to use in searching over molecular candidates and achieve direct feedback from experimental results

V. REFERENCES

- [1] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep Learning in Drug Discovery," *Molecular Informatics*, vol. 35, no. 1, pp. 3–14, jan 2016. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/27491648><http://doi.wiley.com/10.1002/minf.201501008>
- [2] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discovery Today*, vol. 23, no. 6, pp. 1241–1250, jan 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1359644617303598>
- [3] S. Kang and K. Cho, "Conditional molecular design with deep generative models," *Journal of Chemical Information and Modeling*, p. acs.jcim.8b00263, jul 2018. [Online]. Available: <http://pubs.acs.org/doi/10.1021/acs.jcim.8b00263>
- [4] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, sep 2018. [Online]. Available: <https://academic.oup.com/bioinformatics/article/34/17/i821/5093245>
- [5] F. J. Gambo, L. M. Sanz, J. Vidal, C. De Cozar, E. Alvarez, J. L. Lavandera, D. E. Vanderwall, D. V. Green, V. Kumar, S. Hasan, J. R. Brown, C. E. Peishoff, L. R. Cardon, and J. F. Garcia-Bustos, "Thousands of chemical starting points for antimalarial lead identification," *Nature*, vol. 465, no. 7296, pp. 305–310, 2010. [Online] Available: <http://www.nature.com/articles/nature09107>
- [6] T. Kallikokoski, C. Kramer, and A. Vulpetti, "Quality Issues with Public Domain Chemogenomics Data," *Molecular Informatics*, vol. 32, no. 11–12, pp. 898–905, dec 2013. [Online]. Available: <http://dx.doi.org/10.1002/minf.201300051>
- [7] P. Tiikkainen, L. Bellis, Y. Light, and L. Franke, "Estimating Error Rates in Bioactivity Databases," *J. Chem. Inf. Model.*, vol. 53, no. 10, pp. 2499–2505, oct 2013. [Online]. Available: <http://dx.doi.org/10.1021/ci400099q>
- [8] M. Davies, M. Nowotka, G. Papadatos, N. Dedman, A. Gaulton, F. Atkinson, L. Bellis, and J. P. Overington, "ChEMBL web services: streamlining access to drug discovery data and utilities." *Nucleic acids research*, vol. 43, no. W1, pp. W612–20, jul 2015. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/25883136><http://www.medcentral.nih.gov/articlerender.fcgi?artid=PMC4489243>
- [9] D. Fourches, E. Muratov, and A. Tropsha, "Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research." *Journal of chemical information and modeling*, vol. 50, no. 7, pp. 1189–204, 2010 [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/20572635><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2989419>
- [10] D. Rogers and M. Hahn, "Extended-connectivity fingerprints." *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, may 2010. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/20426451>
- [11] G. Landrum, "RDKit: Open-source cheminformatics," <https://www.rdkit.org/> (accessed Jan 12, 2017). [Online]. Available: <http://www.rdkit.org>
- [12] A. Koutsoukas, S. Paricharak, W. R. J. D. Galloway, D. R. Spring, A. P. Ijzerman, R. C. Glen, D. Marcus, and A. Bender, "How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space," *J. Chem. Inf. Model.*, vol. 54, no. 1, pp. 230–242, dec 2013. [Online]. Available: <http://dx.doi.org/10.1021/ci400469u>
- [13] N. M. O'Boyle and R. A. Sayle, "Comparing structural fingerprints using a literature-based similarity benchmark," *Journal of Cheminformatics*, vol. 8, no. 1, p. 36, 2016. [Online]

Available: <http://www.ncbi.nlm.nih.gov/pubmed/27382417><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4932683><http://jcheminf.springeropen.com/articles/10.1186/s13321-016-0148-0>

- [14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vander-plas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duches-nay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [15] A. T. J. W. Oliver Watson, Isidro Cortes, "A decision theoretic approach to model evaluation in computational drug discovery," *Bioinformatics*, vol. In press, 2019
- [16] W. P. Walters, "Modeling, informatics, and the quest for reproducibility," *Journal of Chemical Information and Modeling*, vol. 53, no. 7, pp. 1529–1530, 2013. [Online]. Available: <http://sourceforge.net/>
- [17] G. A. Landrum and N. Stiefl, "Is that a scientific publication or an advertisement? Reproducibility, source code and data in the computational chemistry literature," *Future Medicinal Chemistry*, vol. 4, no. 15, pp. 1885–1887, oct 2012. [Online]. Available: <http://www.future-science.com/doi/10.4155/fmc.12.160>
- [18] T. Kallioikoski, C. Kramer, A. Vulpetti, and P. Gedeck, "Comparability of mixed IC data - a statistical analysis." *PLoS One*, vol. 8, no. 4, p. e61007, jan 2013. [Online]. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3628986&tool=pmcentrez&rendertype=abstract>
- [19] I. Corte's-Ciriano and A. Bender, "How consistent are publicly reported cytotoxicity data? Large-scale statistical analysis of the concordance of public independent cytotoxicity measurements," *ChemMedChem*, vol. 11, no. 1, pp. 57–71, jan 2015. [Online]. Available: <http://doi.wiley.com/10.1002/cmdc.201500424>
- [20] D. L. Alexander, A. Tropsha, and D. A. Winkler, "Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models," *Journal of Chemical Information and Modeling*, vol. 55, no. 7, pp. 1316–1322, 2015. [Online]. Available: <http://pubs.acs.org/doi/10.1021/acs.jcim.5b00206>
- [21] A. Bender and R. C. Glen, "Molecular similarity: a key technique in molecular informatics." *Org. Biomol. Chem.*, vol. 2, no. 22, pp. 3204–3218, nov 2004. [Online]. Available: <http://pubs.rsc.org/en/content/articlehtml/2004/ob/b409813g>
- [22] D. Bajusz, A. Ra'cz, and K. He'berger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?" *Journal of cheminformatics*, vol. 7, p. 20, 2015. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/26052348><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4456712>
- [23] I. Cortes-Ciriano, N. C. Firth, A. Bender, and O. Watson, "Discovering highly potent molecules from an initial set of inactives using iterative screening," *Journal of Chemical Information and Modeling*, vol. 58, no. 9, pp. 2000–2014, 2018. [Online]. Available: <http://pubs.acs.org/doi/10.1021/acs.jcim.8b00376>
- [24] A. Koutsoukas, R. Lowe, Y. KalantarMotamedi, H. Y. Mussa, W. Klaffke, J. B. O. Mitchell, R. C. Glen, and A. Bender, "In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Na'ive Bayes and Parzen-Rosenblatt Window," *J. Chem. Inf. Model.*, vol. 53, no. 8, pp. 1957–1966, 2013. [Online]. Available: <http://dx.doi.org/10.1021/ci300435j>
- [25] Paweł Szyman'ski, Magdalena Markowicz and E. Mikiciuk-Olasik, "Adaptation of High-Throughput Screening in Drug Discovery—Toxicological Screening Tests," *Int J Mol Sci.*, vol. 13, no. 1, pp. 427–452, 2012
- [26] L. S. Ludwig, C. A. Lareau, J. C. Ulirsch, J. D. Buenrostro, A. Regev, and V. G. Sankaran, "Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics," 2019. [Online]. Available: <https://doi.org/10.1016/j.cell.2019.01.022>
- [27] A. Leo, C. Hansch, and D. Elkins, "Partition coefficients and their uses," *Chemical Reviews*, vol. 71, no. 6, pp. 525–616, 1971. [Online]. Available: <https://doi.org/10.1021/cr60274a001>