Deep Imputation on Large-Scale Drug Discovery Data

Benedict W. J. Irwin[†§], Thomas M. Whitehead[‡], Scott Rowland[‖], Samar Y. Mahmoud[†], Gareth J. Conduit[‡§], Matthew D. Segall*[†]

†Optibrium Limited, Cambridge Innovation Park, Denny End Rd, Cambridge, CB25 9PB, UK

‡Intellegens Limited, Eagle Labs, 28 Chesterton Road, Cambridge, CB4 3AZ, UK

‖ Takeda Oncology, 40 Landsdowne St, Cambridge, MA 02139, USA

§University of Cambridge, Cavendish Laboratory, 19 JJ Thomson Ave, Cambridge, CB3 0HE, UK

*ben.irwin@optibrium.com

*matt.segall@optibrium.com

Abstract

More accurate predictions of the biological properties of chemical compounds would guide the selection and design of new compounds in drug discovery and help to address the enormous cost and low success-rate of pharmaceutical R&D. However this domain presents a significant challenge for AI methods due to the sparsity of compound data and the noise inherent in results from biological experiments. In this paper, we demonstrate how data imputation using deep learning provides substantial improvements over quantitative structure-activity relationship (QSAR) machine learning models that are widely applied in drug discovery. We present the largest-to-date successful application of deep-learning imputation to datasets which are comparable in size to the corporate data repository of a pharmaceutical company (678,994 compounds by 1166 endpoints). We demonstrate this improvement for three areas of practical application linked to distinct use cases; i) target activity data compiled from a range of drug discovery projects, ii) a high value and heterogeneous dataset covering complex absorption, distribution, metabolism and elimination properties and, iii) high throughput screening data, testing the algorithm's limits on early-stage noisy and very sparse data. Achieving median coefficients of determination, R^2, of 0.69, 0.36 and 0.43 respectively across these applications, the deep learning imputation method offers an unambiguous improvement over random forest QSAR methods, which achieve median R^2 values of 0.28, 0.19 and 0.23 respectively. We also demonstrate that robust estimates of the uncertainties in the predicted values correlate strongly with the accuracies in prediction, enabling greater confidence in decision-making based on the imputed values.

Introduction

The combination of deep learning and statistical imputation methods is seeing rapidly-growing success in a wide range of scientific domains including high-value materials discovery [1], [2], the development of new chemicals for industrial applications [3], [4], battery development [5], and most importantly for the context of this work *small molecules drug discovery* [6]–[11]. This success can be attributed to the predictive power of the deep learning methodology combined with the flexibility and practical advantages of the imputation framework, which can handle sparse datasets, and use existing, partial assay data to enhance the quality of predictions for missing values in the dataset [8]. Sparse datasets are common in experimental scientific domains, where it is extremely rare that all possible experiments are run on all possible subjects, often due to the cost and time associated with collecting experimental data [6], [8].

In this paper, we will focus on applications of deep learning imputation to the discovery of new drugs. This is a particularly attractive field for the application of artificial intelligence methods of many kinds [12], due to the high costs, long timescales and valuable output of pharmaceutical research and development. The average cost of a novel drug that succeeds in clinical trials and reaches the market is $2.6B [13]. This cost is driven by the high failure rate in the R&D process; only 4% of drug discovery projects result in a marketed drug, and only 12% of candidate drugs that enter expensive and time-consuming clinical trials reach the market [14]. However, the value of an efficacious drug to a patient whose disease is cured or ameliorated may be incalculable, and the associated financial benefit to a pharmaceutical company can be commensurately large; a 'blockbuster' drug will achieve sales measured in billions of dollars per year.

The low success rate of pharmaceutical R&D is, in large part, due to the complexity of the process and the ultimate goal. Drug discovery begins with a biological target implicated in a disease process. This is typically a protein involved in a biological pathway which, if inhibited or stimulated, will treat the disease; for example, inhibiting an essential protein in bacteria, thereby killing the organism, can result in an antibiotic, while stimulating the dopamine receptor in the brain can treat the symptoms of Parkinson's disease. Once a suitable target has been identified, the objective of a drug discovery project is to identify a therapeutic that will achieve the desired effect on the target when dosed to a patient, while avoiding serious side effects. This process often begins by finding initial 'hits' that show activity against the target in an *in vitro* assay, which is followed by an iterative optimization process in which new chemical compounds are synthesized and tested to identify a candidate drug suitable for clinical trials. The design of a high-quality clinical candidate is a complex process, requiring multi-parameter optimization (MPO) of target activity and many other characteristics required in an efficacious and safe drug, often summarised as absorption, distribution, metabolism, elimination and toxicity (ADMET) properties. As compounds progress through the drug discovery process, more complex and expensive experiments are used to assess their likelihoods of success as a drug before a candidate is chosen for clinical trials in humans, subject to approval by the regulatory authorities. The drug discovery process, from selection of a target to nomination of a clinical candidate, takes an average 5.5 years, and the complete R&D process through to launch of a new drug takes 13.5 years on average [14].

Clearly, the abilities to make accurate predictions of the best compounds to synthesize and which to progress to more expensive studies, based on initial experimental results, has the potential to dramatically improve the cost, time and success rate of drug discovery. These objectives are challenging due to the nature of the data available with which to build predictive models. Datasets are typically much smaller and sparser than those seen in traditional machine learning applications, such as image recognition and language processing. It is rare for a compound to have been

measured in all relevant experiments, and no experiments are run on all potential compounds of interest. In a typical pharmaceutical company's database, less than 1% of the possible experimental data points across all compounds of interest will have been measured. In addition to sparsity issues, drug discovery data contain a high degree of experimental noise due to the variability inherent in biological assays [7]. Even when sources of experimental variability are minimized, the results for a given compound in an experimental assay will often vary by 0.5 log units [15].

A wide range of machine learning methods has been applied to predict compound activities and ADMET properties [16]. These quantitative structure-activity relationship (QSAR) models relate features calculated from chemical structures (often referred to as 'descriptors') to one or more target activities or ADMET properties. A comparison of deep learning imputation with the broader field of machine learning in the context of drug discovery is given by Irwin et al. [7]. A successful implementation is the Alchemite method [1]–[5], which has outperformed other data imputation techniques both in terms of accuracy and modelling performance [8] as well as flexibility of implementation and robustness to the challenges associated with practical drug discovery applications [7].

Alchemite demonstrated qualitative benefits over a variety of other machine learning methods, including random forests [17], deep neural networks [18], matrix factorization [19], [20] and purpose-made drug discovery imputation routines [21], [22] on two benchmark drug discovery datasets [6]. These *homogeneous* datasets – purely comprised of target activities – were also designed to mimic the challenging extrapolation expected in drug discovery applications, where the training set is 'known chemistry', and the test set requires extrapolation into 'new chemistry' that has not yet been seen by the models [22]. This application also saw the ability for uncertainty estimates given by Alchemite to allow substantial enhancements to the predictive quality of models: By exploiting the bespoke uncertainty estimate for each prediction and focusing in on the most confident predictions, the effective accuracy of models exceeded a coefficient of determination (R^2) of 0.9, compared to the headline figure of 0.44 on the entire dataset [6], [8]. This focusing effect is impossible for methods which do not provide a robust error bar for each individual prediction [8], and would not yield a benefit where the error bars are of low quality.

The Alchemite method proved more robust than a suite of standard machine learning methods when applied to a real and active drug discovery project [7]. This application demonstrated that – in addition to coping with noise and sparsity – Alchemite could address temporal evolution within datasets and included a mixture of *heterogeneous* endpoints in a single model. These endpoints can either be unrelated, in which case they are treated separately, or related through complex functions of multiple experimental measurements. These benefits of deep imputation were retained on the small-scale datasets typical of a drug discovery project, in contrast with many deep learning methods that rely on large-scale data sets to gain value over simpler machine learning methods in this field.

The method also successfully assisted in finding a novel, active anti-malarial compound when combined with generative methods [8]–[10], [23]. This application relied on so-called 'virtual' models [8], which depend only on calculated molecular descriptors, allowing virtual screening of the generated compounds, which had not yet been synthesized or tested experimentally. The use of Alchemite's robust uncertainty estimates in combination with probabilistic MPO techniques [24] enabled the confident selection of a compound for synthesis and experimental validation.

The above-reported successes have all, to date, been achieved on small- to moderately-sized data sets. For typical drug discovery projects, this would mean hundreds to a few thousand compounds

(rows), and tens of experimental endpoints (columns) in the data matrix [7]. While the Alchemite method has fulfilled the criteria for a robust and practically useful methodology to tackle challenging applications in the field of drug discovery [7], a key requirement is to prove the scalability of the method to large datasets [8], comparable in scale to data available to a moderate-to-large pharmaceutical company. Such a 'large' dataset would contain of order one million compounds and thousands of experimental endpoints.

Scalability has been the focus of other imputation methods, such as the MACAU matrix factorization method [19], which in order to achieve scalability to millions of compounds as desired, results in a linear model which will only capture a shallow degree of the correlation between endpoints [8]. In contrast to this, the Alchemite method presents a non-linear deep learning methodology, which has provably exploited multiple experimental correlations to predict complex and multi-factorial cell-based properties.

Application to large-scale databases of compound data will bring further benefits. Learning from inter-assay correlations across much larger numbers of compounds than would be explored in a single project will enable this information to be leveraged across multiple drug discovery projects and biological targets. This will improve the accuracy of predictions and may reveal unexpected historical correlations between experimental endpoints. Virtual models derived from such a 'global' model will improve the virtual screening of new compounds. Imputation of new activities for existing compounds may reveal opportunities for repurposing compounds for different therapeutic objectives. These applications will unlock enormous value from the wealth of data stored in pharmaceutical companies' data repositories, but whose full potential is, as yet, unrealized. Furthermore, achieving these objectives with a single multitarget model across all compounds and endpoints will reduce computational complexity and cost, versus building and updating models for individual projects on an ad-hoc basis.

In this work we show the first successful application of the Alchemite deep imputation methodology to a pharma-scale data set within a reasonable computational cost. This important step demonstrates that the method meets the requirements for implementation as an overarching modelling method to realize the benefits outlined above. To evaluate the algorithm's ability to successfully handle different kinds of data on the large-scale dataset used in this work, we considered three relevant drug discovery applications:

- **Project Activities:** In a drug discovery project, compounds are assessed using assays to test for activity against one or more biological targets. The Project Activities endpoints were used to evaluate the ability to prospectively predict the activities (measured as the concentration at which half of the maximum inhibition is observed – IC_{50}) of compounds against targets. The data are aggregated across many drug discovery projects, and therefore there are 178 different target columns.
- **ADMET**: This dataset was used to evaluate the ability to predict a broad range of ADMET endpoints, including compound solubility and cell permeability, the extent that compounds inhibit common drug-metabolizing enzymes, metabolic stability and toxicity endpoints.
- **High-throughput Screening (HTS)**: At the inception of a drug discovery project, initial chemical starting points may be identified by a broad and coarse sweep of a wide diversity of compounds to identify those that show indications of activity against the target in question. These HTS campaigns may test hundreds of thousands, or even millions, of compounds. To achieve this at an acceptable cost and time, high-throughput assays are employed that are often noisier than the later project activity assays. The majority of data points will also show little or no activity, which

creates a significant bias in the resulting data set. The objective of this application was to assess the ability to predict HTS activities, despite the limitations in these data. In particular, we wished to test the ability of Alchemite models to predict the activities of a full HTS screen (Assay X) from the results of a much smaller pilot screen.

Methods

The deep imputation method used in this work, Alchemite, is based on the iterative application of a deep learning algorithm to the sparse experimental data to identify and leverage non-linear correlations between endpoints. It has been previously described in detail in Verpoort *et al.* [25] and its application to drug discovery data was described in references [6], [7]. Here we will provide a high-level summary of the method.

Two classes of model are used in this work:

- **Imputation**: These models generate predictions for the test data points using sparse assay data as input, in addition to molecular descriptors, and test an Alchemite model's ability to 'fill in the gaps' in the experimental data for compounds that have been synthesized and tested in some assays.
- **Virtual**: These models are built to expect only molecular descriptors as input. They test an Alchemite model's ability to make predictions based only on compound structure, i.e. for a compound that has not yet been synthesized or tested.

To train an Imputation model, missing values in the sparse experimental data are first given provisional estimates of numbers drawn from a distribution approximating that of the existent experimental data for each endpoint. For each of the N endpoints, the other $N-1$ endpoints and the structural descriptors are used to build models of the experimental data in the endpoint, and this model is used to impute updated values for the initially missing data for each endpoint in parallel to obtain improved estimates for each missing value. This procedure is then iterated, using the estimates from iteration $I-1$ to generate the I^{th} set of estimates. Once the estimates are sufficiently converged, or the desired number of iterations has been carried out (typically two or three iterations) the algorithm returns the latest set of estimates as the predictions for all missing values in the dataset.

The Virtual model is trained similarly, except that the model is constrained not to use experimental endpoints as inputs, mimicking the later application to virtual compounds. This approach still leverages non-linear correlations between endpoints through the iterative procedure, enabling improvements in performance over methods that simplify focus on predicting one endpoint at a time. Predictions can then be made taking as input the chemical descriptors of a compound and iteratively generating estimates for every endpoint, returning the latest set of consistent estimates.

For both Imputation and Virtual models, the underlying modelling of each endpoint is performed using a proprietary 'gradient' kernel. In contrast to standard neural network sigmoid or rectifier activation functions, which can be envisaged as beginning with a large-length-scale approximation of a function and gradually adding more fine detail, the gradient kernel begins with detailed local models and gradually stitches them together into a cohesive whole. This enables more accurate capture of effects like activity cliffs – where a response rapidly varies as a function of the inputs – and is generally on the order of a thousand times quicker to train due to the inherent parallelizability.

One of the most important elements of the deep imputation model is the ability to quantify the uncertainty in predictions. This enables one to separate the most confident predictions from uncertain predictions, targeting future resources only on those compounds with the highest probability of success. An ensemble of sub-models is used to quantify uncertainty for each endpoint at each iteration. Each sub-model is trained on a bootstrap sample of the available data to provide accurate treatment of the variation within the data.

One additional complexity of the drug discovery data used here is that multiple endpoints are frequently measured in the same experimental assay. One endpoint from a given assay should therefore not be used as input to predict another endpoint from the same assay, as at test-time either both endpoints will have been measured for a given compound or neither will be available. To capture this, Alchemite includes generalized, asymmetric constraints on column dependencies. These can also be used to ensure assays that are typically run late in a program are not used as input to predict assays run earlier for a given compound (whilst still allowing the early-stage assays to be used as input to predict the late-stage assays).

Comparison QSAR models

A random forest (RF) model was also constructed for each individual endpoint as representative examples of QSAR methods. These RF models were generated using the scikit-learn implementation of regression RF [26] and take compound descriptors as input only.

A wider comparison of QSAR methods was previously undertaken for project data by Irwin *et al.* [7]. This included partial least squares, RF, Gaussian process, and radial basis function models and found RF models to be broadly representative of the accuracy of QSAR methods. Advanced (multitarget) methods such as deep neural networks and matrix factorization were compared to Alchemite by Whitehead et al. [6]. In all cases, Alchemite's deep imputation method was found to outperform the other approaches in each case significantly.

Metrics

All models were evaluated on an independent test set using two statistics: The coefficient of determination (R^2), defined as

$$R^2 = 1 - \frac{\sum_i (y_i^{pred} - y_i^{obs})^2}{\sum_i (y_i^{obs} - \overline{y^{obs}})^2},$$

which takes values in the range $(-\infty, 1]$, where 1 indicates a perfect model, 0 indicates a model no better than random, and negative values indicate predictions that are worse than random; and the root mean square error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^{pred} - y_i^{obs})^2},$$

where N is a number of compounds in the set, y_i^{pred} is the predicted value and y_i^{obs} is the experimentally observed value for data point i. The RMSE is expressed in the same units as the observed property values. R^2 values were calculated only for endpoints with greater than five data points to give sufficient statistical relevance.

Dataset

All modelling data are proprietary and were provided by Takeda Pharmaceutical Company Ltd. Prior to modelling all qualified and out-of-range data were removed. The remaining data were transformed into units more amenable for machine learning (e.g. log transformations were applied

to columns which varied many orders of magnitude). To maintain full modelling rigour, the dataset was split into training, validation and independent test sets. The full training data set contained 678,994 compounds and 1166 experimental endpoints; the breakdown across the three applications described above is shown in **Table 1**.

Table 1: The breakdown of the training data in terms of the three applications (Project Activities, HTS, ADMET) and the number of endpoints, compounds and data points, along with a measure of sparsity.

Application	Number of Endpoints	Number of Compounds	Total Number of Data Points	Sparsity (% filled)
Project Activities	178	4,501	36,274	4.5
HTS	748	662,635	17,951,700	3.6
ADMET	240	30,495	117,097	1.6

The blind test set contained a total of 17,660 data points across endpoints for each application, as described in **Table 2**.

Table 2: A summary of the number of test points for each dataset (Project Activities, HTS, ADMET) and the selection strategy.

Application	Number of Test Points	Selection Strategy
Project Activities	1167	Temporal, i.e. most recently measured data points
HTS	10396	7858 from a single HTS screen (**Assay X**), else random
ADMET	6097	Random

The independent, blind test sets were prepared by Takeda and withheld during the model building and internal validation process. Predictions for the blind test sets were provided to Takeda before the experimentally observed values were revealed. The test sets were generated by Takeda in the following ways:

- **Project Activities:** The test data points were selected temporally, i.e. the most recently measured data points were withheld, to test the models' abilities to predict the activities of the most recently synthesized compounds and assay results.
- **ADMET**: The test data points were selected randomly to test the models' abilities to predict ADMET properties for a wide diversity of compounds. This selection method gives an even coverage of each type of endpoint and value according to their prevalence in the overall dataset.
- **HTS**: A small proportion of the test data points were selected randomly. However, the majority were derived from a single assay (**Assay X**), for which the results of a pilot screen were provided in the training set, but the results from the remaining compounds in the full screening collection were withheld.

Compound Descriptors

Molecular descriptors were calculated with the StarDrop™ Auto-Modeller™ module. These descriptors can be computed from the atom and bond graph structure of any compound, including virtual compounds, and therefore all descriptors are present for each compound in the dataset. The set of descriptors comprises common whole-molecule properties, including calculated lipophilicity, molecular weight, topological polar surface area, and McGowan's volume, as well as counts of ~300 chemical substructures defined as SMARTS patterns [27], which are essentially regular expressions for chemical subgraph pattern queries.

Results

Here we show the results of the Alchemite models compared with RF models for each endpoint in the independent test set. Because the data for each application represent a large number of endpoints, as shown in Table 1, we show a profile of endpoint results, ordering the R^2 values ordered from highest to lowest for each method (See Figure 1 for example). Alchemite also provides uncertainty estimates for each prediction and we also present a comparison of the uncertainty estimates with the observed errors for the independent test.

Project Activities Test Results

Figure 1 shows the profile over Project Activity endpoints for the independent test data. This plot includes curves showing the performance of Alchemite Imputation and Virtual models relative to the RF models. The median R^2 for the Alchemite Imputation model is 0.69, compared to 0.28 for RF models. The median R^2 for the Alchemite Virtual model is 0.55 which is also substantially higher than the RF models, showing that the multitarget deep learning with sparse experimental data trains a very high-quality virtual model.

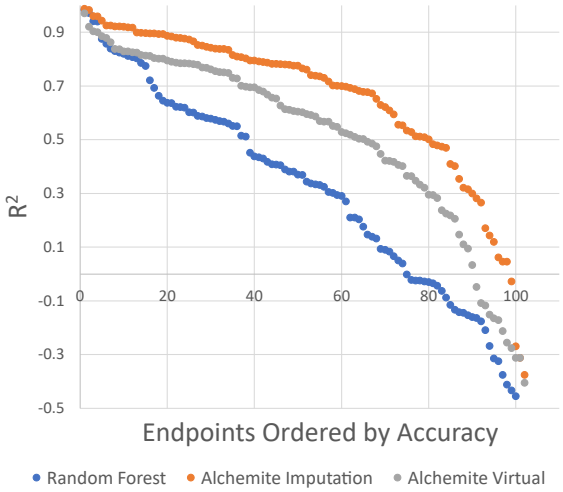

Figure 1: Profile of the coefficient of determinations (R^2) achieved on the independent test set for Project Activity endpoints. The endpoints are ordered from highest R^2 (left) to lowest (right). Alchemite Imputation model (orange), Alchemite Virtual model (grey) and random forest models (blue) are plotted for comparison.

Each of the points in the profile shown in Figure 1 is an R^2 value for predictions of a different endpoint. In the case of the Project Activity endpoints, these are all measurements of activity against a target. Scatter plots and uncertainty analysis are given for a focused example in the

supplementary information better to show how these predictions can be used in practice, while inspecting the quality of the models and the uncertainty predictions.

ADMET Test Results

The results for the independent test set for the ADMET endpoints are shown in Figure 2, and a we can see a similar trend to the above example. The best model is the Alchemite Imputation model with a median R^2 of 0.36, close behind is the Alchemite Virtual model with a median R^2 of 0.32 and finally, RF models achieve a median R^2 of 0.19.

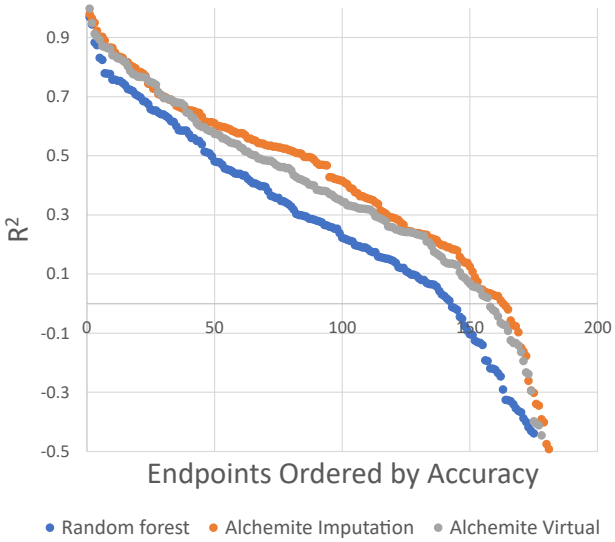

Figure 2: Profile of the coefficient of determinations (R^2) achieved on the independent test set for ADMET endpoints. The endpoints are ordered from highest R^2 (left) to lowest (right). Alchemite Imputation (orange), Alchemite Virtual (grey) and random forest (blue) models are plotted for comparison.

The ADMET endpoints represent a wide variety of different data types, and it is interesting to compare the profile of results for different classes of ADMET endpoints. In particular, Figure 3 shows the accuracy profiles for pIC_{50} and pEC_{50} (the negative base-10 logarithm of the concentration in Molar units which exhibits 50% of the maximum effect) endpoints respectively. From these plots, one can see that the Alchemite models have a much larger advantage over RF for the pEC_{50} endpoints than for pIC_{50} endpoints. The Alchemite Imputation model outperforms the Alchemite Virtual model for the pEC_{50} endpoints, whereas the two models are roughly equivalent for pIC_{50} endpoints. This result is consistent with those seen in smaller project datasets, where Alchemite tends to show the greatest benefit for complex, multi-mechanistic endpoints [7]. While a pIC_{50} measurement relates to the inhibition of a single target protein, pEC_{50} measurements result from more complex assays that may be influenced by multiple factors; for example, the activity of a compound in a cell will relate not only to its activity against a target protein, but also permeability through the cell membrane, solubility in the buffer solution and binding to other proteins in the cellular matrix. This demonstrates one of Alchemite's advantages, namely the ability to learn directly from relationships between experimental endpoints, which may capture these other factors, to make better predictions.

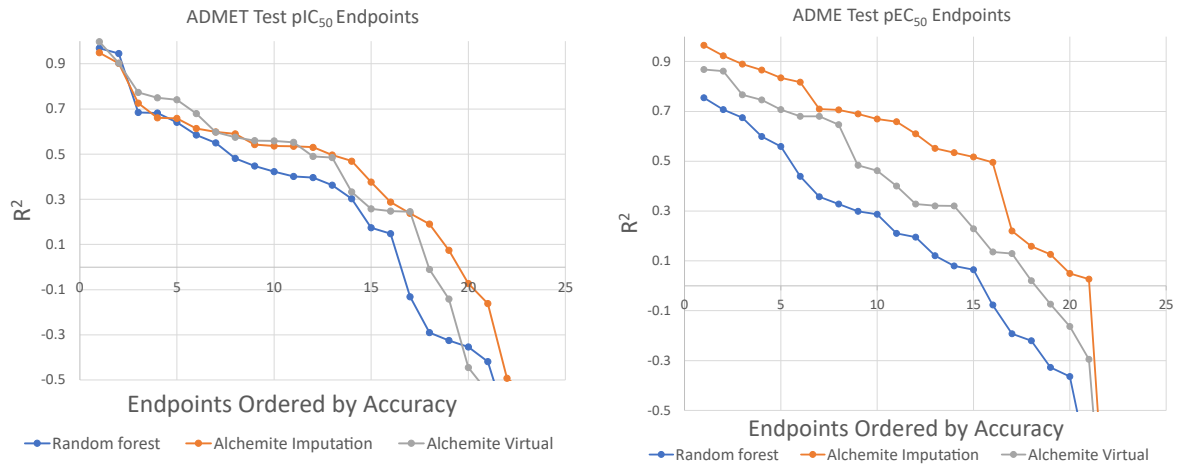

Figure 3: Profile of the coefficient of determinations (R^2) achieved on the independent test set for ADMET pIC_{50} endpoints (left) and ADMET pEC_{50} endpoints (right). The endpoints are ordered from highest R^2 to lowest. Alchemite Imputation model (orange), Alchemite Virtual model (grey) and random forest models (blue) are plotted for comparison.

HTS Test Results

Figure 4 shows the profile of R^2 results over the HTS endpoints in the independent test set. In this case, the Alchemite Virtual model is essentially equivalent to the RF models, whereas the Alchemite Imputation results are an improvement. The median R^2 for RF is 0.23, similar to the Alchemite Virtual model with a median R^2 of 0.27, but lower than the Alchemite Imputation model, which achieved an R^2 of 0.43. This shows that there is additional information in the correlations between endpoints and in the structure of the training data that can be exploited by using Imputation on HTS data.

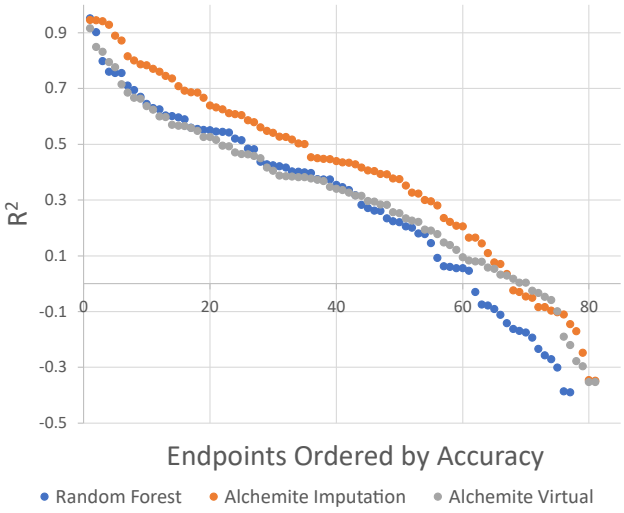

Figure 4: Profile of the coefficient of determinations (R^2) achieved on the independent test set for HTS endpoints . The endpoints are ordered from highest R^2 (left) to lowest (right). Alchemite Imputation model (orange), Alchemite Virtual (grey) and random forest (blue) models are plotted for comparison.

One objective of the HTS test was to assess the ability of the Alchemite models to predict activities for the full screening deck for Assay X, based on the pilot screen data for this assay. However, none of the models considered in this study showed sufficient predictive power for this endpoint, and the R^2 values were close to zero (Alchemite Imputation 0.07, Alchemite Virtual 0.12, RF -0.17). There are several possible explanations for this:

- The overlap in compounds with data measured in Assay X and other endpoints in the training set is lower than is typical for other endpoints in the training set. The maximum overlap corresponds to only 10% of the compounds for Assay X.
- The distribution of percentage inhibition data is challenging to model. For Assay X, the large majority of the measured values are distributed around 0%, plus or minus 10% and only a very small proportion of compounds are measured to have significant activities, as we would expect from HTS. Furthermore, the noise in the measured values for inactive compounds may be affecting the ability of the accuracy metric to distinguish good from poor models and guide the model optimization. It may be possible to transform the percentage inhibition data to reduce the impact of this noise.

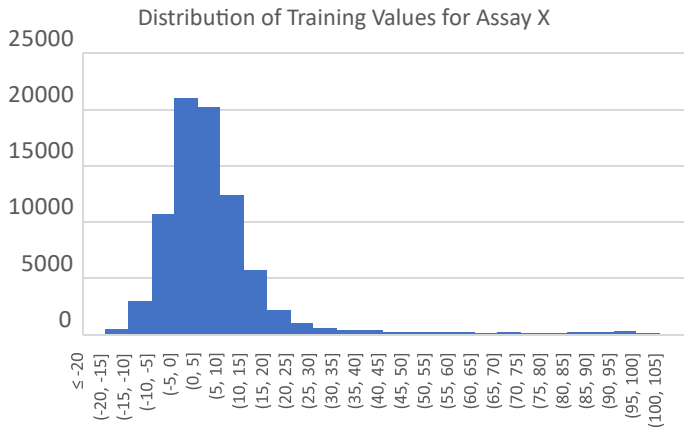

Figure 5: The distribution of training values for Assay X. Most of the values are centred around 0% inhibition, i.e. inactive compounds, and the width of that peak is likely to be noise.

In order to test the impact of the bias in the distribution of observed values, a new version of the model was built in which the active data were oversampled by duplicating the active compounds 15 times relative to the inactives. This did not improve the accuracy of the resulting model. This indicates that the problem is less likely to be due to sampling bias and suggests that the algorithms may be attempting to model the noise in the data rather than the signal.

Taking Uncertainties in Predictions into Account

To test the accuracy of the uncertainty estimates produced by Alchemite, we can plot the root-mean-square error (RMSE) in prediction versus the most confidently predicted fraction of the test set, i.e. smaller fractions correspond to the predictions with the smallest error bars (according to the algorithm). If the uncertainties are conveying useful information, we would expect the most confidently predicted fractions of the test set to show better accuracy, i.e. a lower RMSE.

The quality of uncertainty predictions, averaged across all 178 Project Activity endpoints, as a function of the most confidently predicted fraction of the dataset, is shown in Figure 6. All of the Project Activity endpoints are in the same units (pIC_{50}, the negative base-10 logarithm of the concentration in Molar units which exhibits 50% of the maximum target inhibition), and this average is well-defined. For comparison, the uncertainties in the RF predictions were calculated as the standard deviation of predictions from the ensemble of decision trees. Figure 6 shows that the error bars produced by all three methods, on average, provide some useful information in identifying more accurate predictions. However, the benefit from RF error bars is much smaller than that of the Alchemite uncertainty estimates. The decrease in RMSE correlates much more strongly with the error bars for both the Virtual and Imputation Alchemite models. In addition, the absolute RMSE is

much higher, on average, for RF predictions. The Alchemite Virtual model has a lower RMSE, and the Imputation model has the best RMSE. An example of a similar analysis for an individual Project Activity endpoint is provided in the supplementary information.

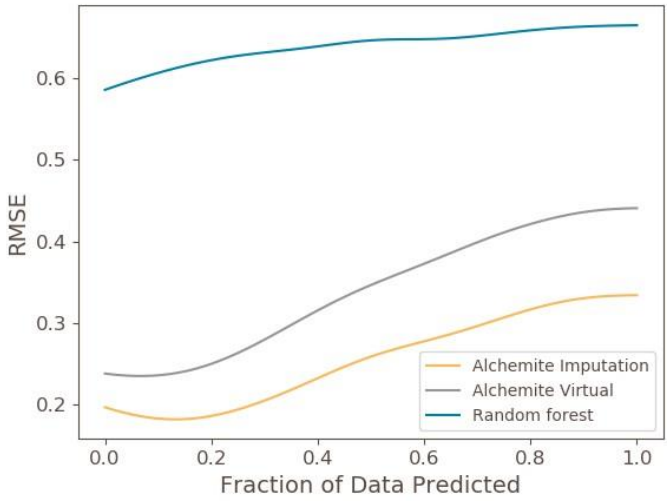

Figure 6: Graph illustrating the relationship between confidence and accuracy of prediction for the Project Activities test set. The x-axis shows the most confidently-predicted fraction of the test data, i.e. in moving from right to left only the most confidently predicted values are included. The y-axis shows the root-mean-square error (RMSE) of the fraction of predictions, aggregated over 178 Project Activity endpoints. A lower RMSE value indicates more accurate predictions. Results for random forest models are shown in blue, the Alchemite Imputation model in orange and the Alchemite Virtual model in grey. For ease of visualization, Gaussian smoothing has been applied to the accuracy calculated at each sampled fraction.

We can also explore the ability of Alchemite to focus on the most confident predicted values in the more heterogeneous ADMET endpoints. Unlike the Project Activity endpoints, the mixed units across the ADMET endpoints mean that the error analysis cannot be summarised in a single graph for the full data set in analogy to Figure 6. We consider some illustrative examples of individual endpoints in Figures 7 and 8. Figure 7 reflects the performance of Alchemite permeability models – the logarithm of the basolateral to apical permeability in a cell line (P_{app} B to A) – by comparing the most confident 50% of predictions with all predictions for both the Imputation (left) and Virtual (right) models. The most confident predictions are more closely clustered to the identity line, and the clear outliers have been dropped.

Figure 8, plots the predictions for an unrelated pEC_{50} endpoint and shows that, while the predictions for this endpoint follow the observed values quite well, there is more scatter in the predictions, and some points have large uncertainty estimates. That is to say, predictions for this endpoint are accurate, but not precise. If more precision is required, we can again focus in on the most confident predictions. In this instance, we show the most confident 25% of predictions according to the Alchemite error bars, and these predictions are clustered around the identity line in a tighter grouping than the baseline model.

We can see that the correlation between the most confident and accurate results is also strong for these models, even when the baseline R^2 is not high. This importantly allows the models to be useful even in situations where they would be otherwise discarded due to a poor correlation for the full test set. The correlations of RMSE with the confidence in predictions data for these endpoints are shown Figures S3 and S4 in the supplementary information and show a similar result to that of Figure 6. These confirm the benefits of quantifying uncertainty for diverse endpoints.

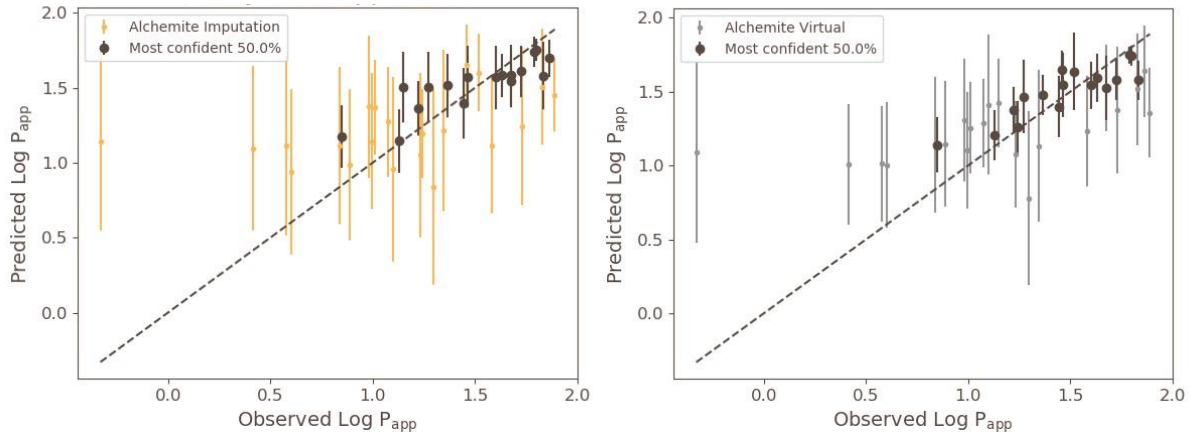

Figure 7: Scatter plots showing predicted versus observed values for all points in the independent test set for a for a single log permeability (P_{app} B to A) ADMET endpoint. The Alchemite Imputation model result is shown in orange (left) and the Virtual model in grey (right). Error bars are shown, corresponding to the Alchemite uncertainty estimate for each point (1 standard deviation). The most confident 50% of the predictions for each model, according to the associated Alchemite uncertainty estimates, are highlighted as bold points. The identity line is shown for comparison (dashed).

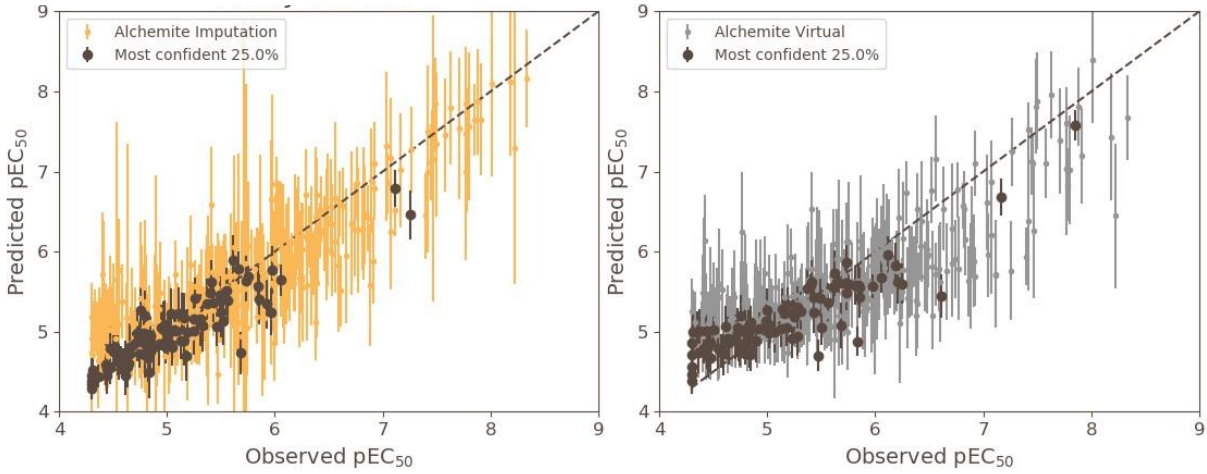

Figure 8: Scatter plots showing the predicted versus observed values for all points in the independent test set for a single pEC_{50} ADMET endpoint. The Alchemite Imputation model result is shown in orange (left) and the Virtual model in grey (right). Error bars are shown, corresponding to the Alchemite uncertainty estimate for each point (1 standard deviation). The most confident 25% of the predictions for each model, according to the associated Alchemite uncertainty estimates, are highlighted as bold points. The identity line is shown for comparison (dashed).

Computational Resources

Depending on hyperparameters, it takes 8-16 hours to train this model on an AWS EC2 m5.4xlarge instance, (64 GiB of Memory, 16 vCPUs). A larger cost is incurred for hyperparameter optimisation, for which thetime can be estimated as $T_{hyp} \approx 1.3 \times N_{fold} N_{samples} T_{base}$, where T_{base} is the base training time for a dataset, $N_{samples}$ is the number of hyperparameter optimisation samples required for convergence (usually 20-50), and N_{fold} is the number of cross-validation folds. However, following an initial hyperparameter optimisation the model can be updated with new data in the time taken for training, unless there is a significant change in the overall structure of the data set. Furthermore, the hyperparameter optimisation process can be further parallelised over the cross-validation folds to reduce the overall time by a factor of N_{fold}.

Conclusions

Some general conclusions can be drawn across all applications. Alchemite Imputation models consistently outperform RF models, and generally outperform Alchemite Virtual models. This

highlights a benefit of deep learning imputation, which can learn directly from the relationships between experimental endpoints and gain valuable information, even from very limited experimental data, to more accurately fill in missing experimental values. In all applications, the Alchemite Virtual model performed better than or equivalently to RF. The Alchemite algorithm is competitively fast when compared to other deep learning methods and was applied to a pharma-scale data set within a reasonable computational cost.

We've demonstrated that the Alchemite uncertainty estimates correlate strongly with the accuracy of the corresponding predictions, unlike those derived from random forest ensemble-based uncertainties. This result is particularly exciting because generating robust and objectively useful uncertainty estimates from neural networks remains a major challenge [28]. Valid uncertainty estimates are essential to the effective use of models; understanding where a result is likely to be sufficiently accurate enables high-quality compounds to be identified with confidence while avoiding missed opportunities by incorrectly discarding a potentially good compound due to an uncertain prediction [15].

There were endpoints which could not be modelled by any method for all applications (i.e. the rightmost points in Figures 1-4). Without heavy preprocessing, all large data sets will have such endpoints, especially on the repository-wide scale. We should not expect to be able to model all endpoints, particularly when the data are noisy or where few data points are available. However, it is notable that the inclusion of noisy and uncorrelated endpoints in the data set did not have a detrimental effect on the performance of the Alchemite models for the majority of endpoints. This contrasts with other multitarget modelling approaches that benefit where there are strong correlations between endpoints, but suffer a detrimental effect from the introduction of uncorrelated endpoints into the data set [29].

There are also some more specific conclusions we can draw for each of the three individual applications:

Project Activities Conclusions: For the Project Activity endpoints, all of the Alchemite models significantly outperform RF, showing the method is very effective on activity type endpoints. The results from the independent test were consistent with those from the internal validation; this is remarkable because the test set selected by Takeda Pharmaceuticals was *temporally* based, representing the most relevant and recent compounds in the corresponding project endpoints. Therefore, the results indicate the consistency and utility one could expect when deploying Alchemite models in real projects.

As expected, the Alchemite Imputation model slightly outperforms the Virtual model because the former model has access to more information, in the form of sparse experimental data. This shows that the cross-correlations between experimental endpoints offer significant practical utility, and it is sensible to exploit this where possible.

HTS Conclusions: The Alchemite Imputation model outperforms the Alchemite Virtual and RF models on HTS data, which represent some of the most challenging and noisy data. However, the prediction of the full screening collection for ‘Assay X’ based on an initial pilot screen was not possible with any of the models. One approach to addressing this may be to apply a classification method, but this is beyond the scope of this study.

ADMET Conclusions: For ADMET endpoints, both Alchemite Imputation and Virtual models outperform RF on the full data set, and the Imputation model achieves higher accuracies than the

Virtual model. Individual ADMET endpoints are less likely to correlate, as the different endpoints are often quite distinct, and therefore we see a smaller improvement of the Alchemite Imputation model over the Alchemite Virtual model and RF. However, more complex experimental endpoints often depend on multiple factors that may be captured by endpoints derived from simpler assays, and, in these cases, we see a more substantial improvement of the Alchemite models over conventional QSAR methods such as RF. This represents a sizeable potential value because the simpler assays are typically lower cost and higher throughput than the more complex, cell-based assays and hence are often used earlier in a drug discovery project. Imputation can better leverage the results from these earlier assays to more accurately select the best compounds for more expensive, downstream studies.

We note that endpoints derived from the same experimental assay are not used as input to predict one another in this study. However, in this study, we have taken a very conservative approach to constrain the dependencies between assay endpoints; for example, endpoints that are not necessarily measured together, such as solubility measurements at different pH values, have also been constrained not to impute one another. This does not need to be the case; solubilities measured at different pH values could be defined as independent inputs, allowing solubilities at one pH value to be used to predict solubilities measured at different pH values. We would expect the use of such related, but independent endpoints to further improve the accuracy of the Alchemite models. Such a use could, for example, enable accurate prediction of solubilities at multiple pH values to be made based on a single measurement, further reducing the experimental resources required.

Supporting Information:

There is supporting information detailing further specific examples of the benefits of uncertainty estimation.

Corresponding Author Information:

Benedict W. J. Irwin, ben@optibrium.com

Matthew D. Segall, matt@optibrium.com

Notes:

BWJI, SYM and MDS are employees of Optibrium Ltd.. TMW and GJC are employees of Intellegens Ltd.. SR was an employee of Takeda until December 2020.

References

[1] B. D. Conduit, N. G. Jones, H. J. Stone, and G. J. Conduit, “Design of a nickel-base superalloy using a neural network,” *Mater. Des.*, vol. 131, pp. 358–365, Oct. 2017.

[2] B. D. Conduit, N. G. Jones, H. J. Stone, and G. J. Conduit, “Probabilistic design of a molybdenum-base alloy using a neural network,” *Scr. Mater.*, vol. 146, pp. 82–86, Mar. 2018.

[3] P. Santak and G. Conduit, “Predicting physical properties of alkanes with neural networks,” *Fluid Phase Equilib.*, p. 112259, 2019.

[4] P. Santak and G. Conduit, “Enhancing NEMD with automatic shear rate sampling to model viscosity and correction of systematic errors in modeling density: Application to linear and light branched alkanes,” *J. Chem. Phys.*, vol. 153, no. 1, p. 014102, Jul. 2020.

[5] M.-F. Ng, J. Zhao, Q. Yan, G. J. Conduit, and Z. W. Seh, “Predicting the state of charge and health of batteries using data-driven machine learning,” *Nat. Mach. Intell.*, vol. 2, no. 3, pp. 161–170, Mar. 2020.

[6] T. M. Whitehead, B. W. J. Irwin, P. Hunt, M. D. Segall, and G. J. Conduit, “Imputation of Assay Bioactivity Data Using Deep Learning,” *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1197–1204, Mar. 2019.

[7] B. W. J. Irwin, J. Levell, T. Whitehead, M. Segall, and G. Conduit, “Practical Applications of Deep Learning to Impute Drug Discovery Data,” *J. Chem. Inf. Model.*, 2020.

[8] B. W. J. Irwin, S. Mahmoud, T. M. Whitehead, G. J. Conduit, and M. D. Segall, “Imputation versus prediction: applications in machine learning for drug discovery,” *Futur. Drug Discov.*, vol. 2, no. 2, p. FDD38, Apr. 2020.

[9] B. Irwin, A. Wade, and M. Segall, “Guiding Drug Optimisation Using Deep Learning Imputation and Compound Generation,” *Int. Pharm. Ind.*, 2020.

[10] E. G. Tse *et al.*, “Predicting Bioactivity when there is No Target: Performance of Methods in an Open, Crowdsourced Competition (Submitted),” 2021.

[11] S. Mahmoud *et al.*, “Imputation of Sensory Properties Using Deep Learning (submitted),” 2021.

[12] K.-K. Mak and M. R. Pichika, “Artificial intelligence in drug development: present status and future prospects,” *Drug Discov. Today*, vol. 24, no. 3, pp. 773–780, Mar. 2019.

[13] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, “Innovation in the pharmaceutical industry: New estimates of R&D costs,” *J. Health Econ.*, vol. 47, pp. 20–33, May 2016.

[14] S. M. Paul *et al.*, “How to improve R&D productivity: the pharmaceutical industry’s grand challenge,” *Nat. Rev. Drug Discov.*, vol. 9, no. 3, pp. 203–214, Mar. 2010.

[15] M. D. Segall and E. J. Champness, “The challenges of making decisions using uncertain data,” *J. Comput. Aided. Mol. Des.*, vol. 29, no. 9, pp. 809–816, Sep. 2015.

[16] Y. Lo, S. E. Rensi, W. Torng, and R. B. Altman, “Machine learning in chemoinformatics and drug discovery,” *Drug Discov. Today*, vol. 23, no. 8, pp. 1538–1546, 2018.

[17] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[18] G. Hessler and K.-H. Baringhaus, “Artificial Intelligence in Drug Design,” *Molecules*, vol. 23, no. 10, p. 2520, Oct. 2018.

[19] J. Simm *et al.*, “Macau: Scalable Bayesian factorization with high-dimensional side information using MCMC,” *IEEE Int. Work. Mach. Learn. Signal Process. MLSP*, vol. 2017-Septe, pp. 1–6, 2017.

[20] A. P. Singh and G. J. Gordon, “Relational Learning via Collective Matrix Factorization Categories and Subject Descriptors,” 2008.

[21] E. J. Martin, V. R. Polyakov, X.-W. Zhu, P. Mukherjee, L. Tian, and X. Liu, “All-Assay-Max2 pQSAR: Activity predictions as accurate as 4-concentration IC50s for 8,558 Novartis assays,” *bioRxiv*, no. 4218, p. 620864, 2019.

[22] E. J. Martin, V. R. Polyakov, L. Tian, and R. C. Perez, “Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds,” *J. Chem. Inf. Model.*, vol. 57, no. 8, pp. 2077–2088, 2017.

[23] C. Swain, M. Todd, S. Kanza, and J. G. Frey, “AI3SD, OSM & RSC-CICAG Predicting the activity of Drug Candidates when there is no target Workshop Report,” 2020.

[24] M. D. Segall, “Multi-Parameter Optimization: Identifying High Quality Compounds with a Balance of Properties,” *Curr. Pharm. Des.*, vol. 18, no. 9, pp. 1292–1310, Mar. 2012.

[25] P. C. Verpoort, P. MacDonald, and G. J. Conduit, “Materials data validation and imputation with an artificial neural network,” *Comput. Mater. Sci.*, vol. 147, pp. 176–185, May 2018.

[26] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, Jun. 2011.

[27] “Daylight SMARTS.” [Online]. Available: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. [Accessed: 16-Dec-2019].

[28] L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay, and C. W. Coley, “Uncertainty Quantification Using Neural Networks for Molecular Property Prediction,” *J. Chem. Inf. Model.*, vol. 60, no. 8, pp. 3770–3780, Aug. 2020.

[29] Y. Xu, J. Ma, A. Liaw, R. P. Sheridan, and V. Svetnik, “Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships,” *J. Chem. Inf. Model.*, vol. 57, no. 10, pp. 2490–2504, 2017.