Supplementary Materials for 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

This is the supplementary material for the main body of "Deep Imputation on Large-Scale Drug Discovery Data", the purpose is to provide addition information and examples around the main text.

Project Activities Test Results

Figure S1 shows a detailed example of one of the Project Activity endpoints included in Figures 1 and 2 of the main text, comparing the RMSEs of the Alchemite Imputation and Virtual models with that of the RF model for this endpoint. The mean value of Alchemite error bars for each model is also plotted. Figure S1 again shows that there is an improvement in accuracy associated with focusing on the most confident predictions for a single endpoint, according to the Alchemite models. In contrast, the RF uncertainty estimates offer less meaningful improvement, especially when focusing on the RF predictions purported to be most confident, indicating the RF ensemble uncertainty estimation is not reliable. In the case of Figure S1, the Alchemite uncertainty estimates tend to be pessimistic, with uncertainty estimates (dotted lines) often slightly larger than the true RMSE. In comparison, the RF uncertainty estimates do not represent the actual RMSE of the predictions.

Figure S1 can be interpreted in combination with Figure S2, which shows the effect of considering only the most confident 75% of the predictions from the Alchemite models. In both cases, this results in the models successfully disregarding the largest outliers. The practical result is very similar for both the Imputation (top of figure) and Virtual (bottom of figure) models. Outliers have correctly

been assigned larger uncertainty estimates, and the most confident 75% of the data are closely clustered to the identity line, explaining the high R^2 values for this endpoint.

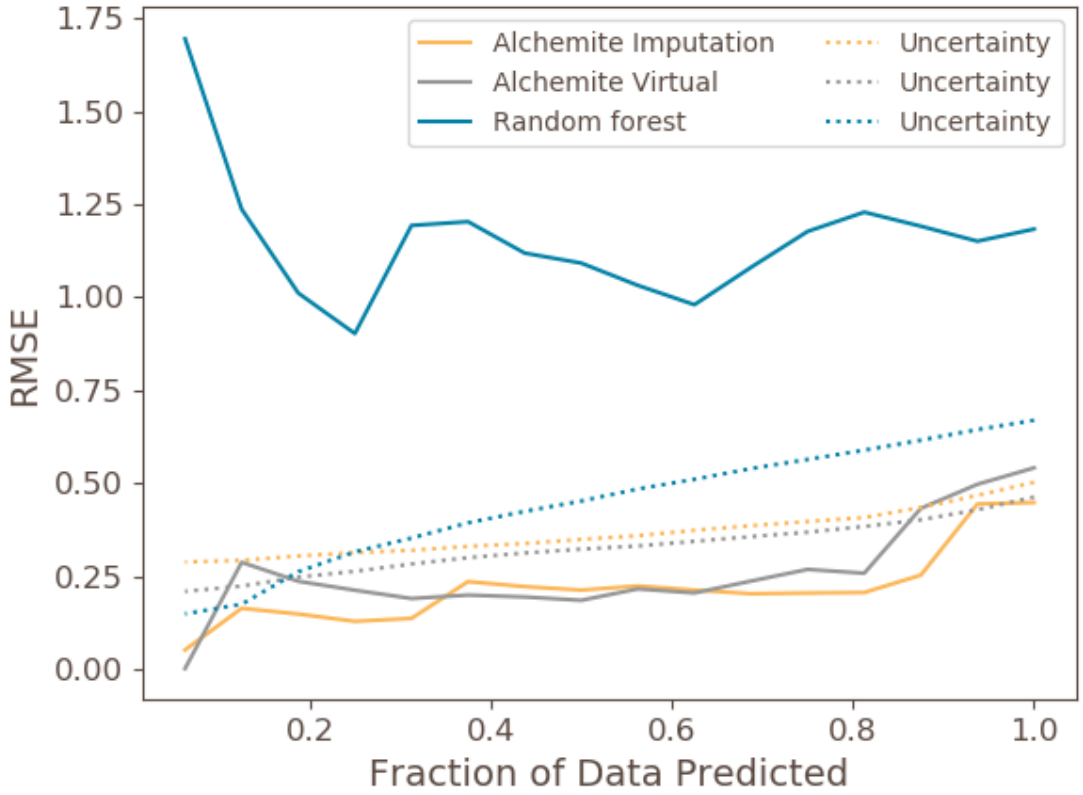

Figure S1: Graph illustrating the relationship between confidence and accuracy of prediction for one of the Project Activity endpoints. The x-axis shows the most confidently-predicted fraction of the test set data, i.e. in moving from right to left only the most confidently predicted values are included. The y-axis shows the root-mean-square error (RMSE) of the fraction of predictions, i.e. a lower value indicates more accurate predictions. The results for the Alchemite Imputation and Virtual models are shown in orange, and grey respectively and those for random forest in blue. The dotted lines shows the mean value of the uncertainty estimates for each model.

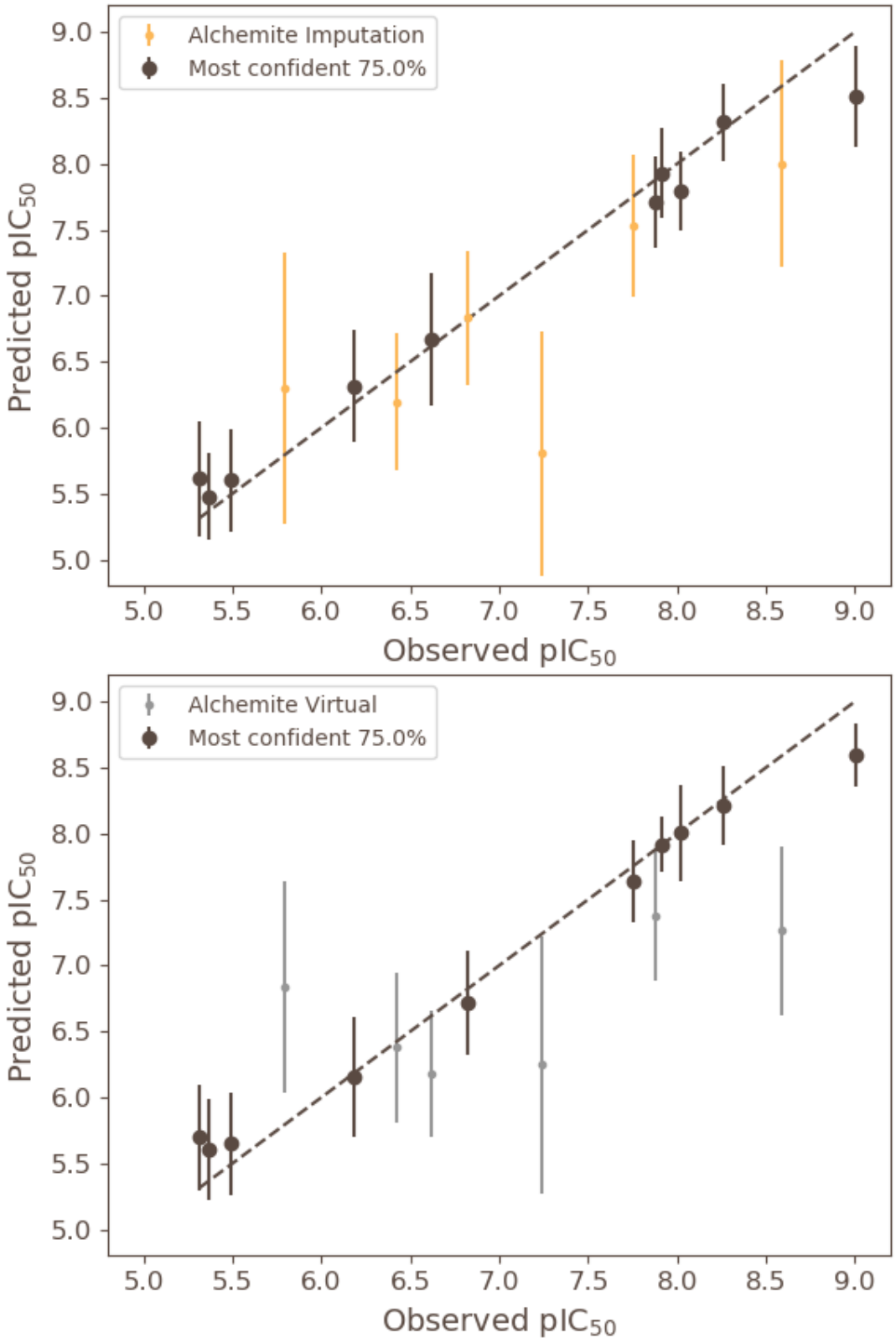

Figure S2: Scatter plots showing predicted versus observed values for the independent test set for the same Project Activity endpoint as in Figure S1. The Alchemite Imputation model result is shown in orange (top) and the Virtual model in grey (bottom). Error bars are shown, corresponding to the Alchemite uncertainty estimate for each point (1 standard deviation). The most confident 75% 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).

ADMET Test Results

Figure S3 shows a similar trend to Figure S1 for an example ADMET endpoint, the logarithm of the basolateral to apical permeability in a cell line (P_{app} B to A). The Alchemite uncertainty estimates correlate strongly with the observed errors in the predictions and are accurate in absolute terms. Conversely, the RF model error bars massively underestimates the RMSE, and are very similar in magnitude to the Alchemite predicted error bars, even though the observed RMSE for the RF model is much higher than either Alchemite model. Furthermore, the rank ordering of the RF error bars is close to random for the RF model because the RMSE is roughly constant (~ 0.5) for all data fractions.

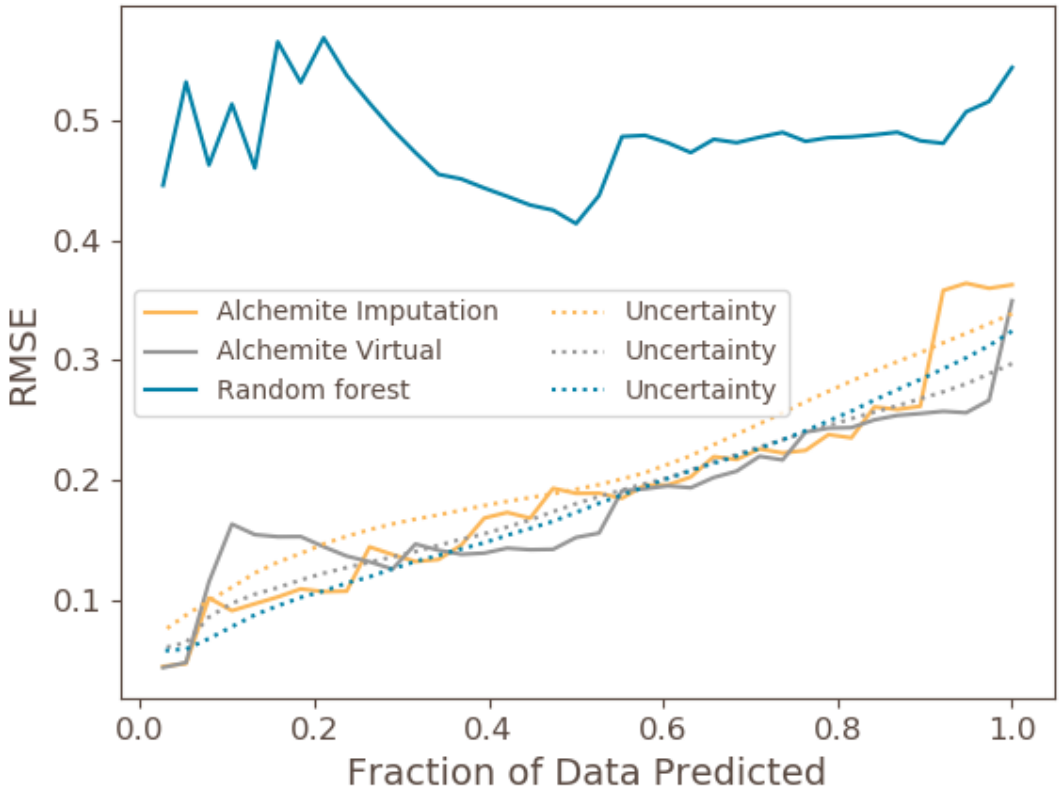

Figure S3: Graph illustrating the relationship between confidence and accuracy of prediction for a single log permeability (P_{app} B to A) ADMET endpoint. The x-axis shows the most confidently-predicted fraction of the test set data, i.e. in moving from right to left only the most confidently predicted values are included. The y-axis shows the root-mean-square error (RMSE) of the fraction of predictions, i.e. a lower value indicates more accurate predictions. The results for the Alchemite Imputation and Virtual models are shown in orange, and grey respectively and those for random forest in blue. The dotted lines shows the mean value of the uncertainty estimates for each model.

Figure S4 again illustrates the same agreement between estimated uncertainty and accuracy for an unrelated pEC_{50} endpoint. This pattern is observed across all types of ADMET data.

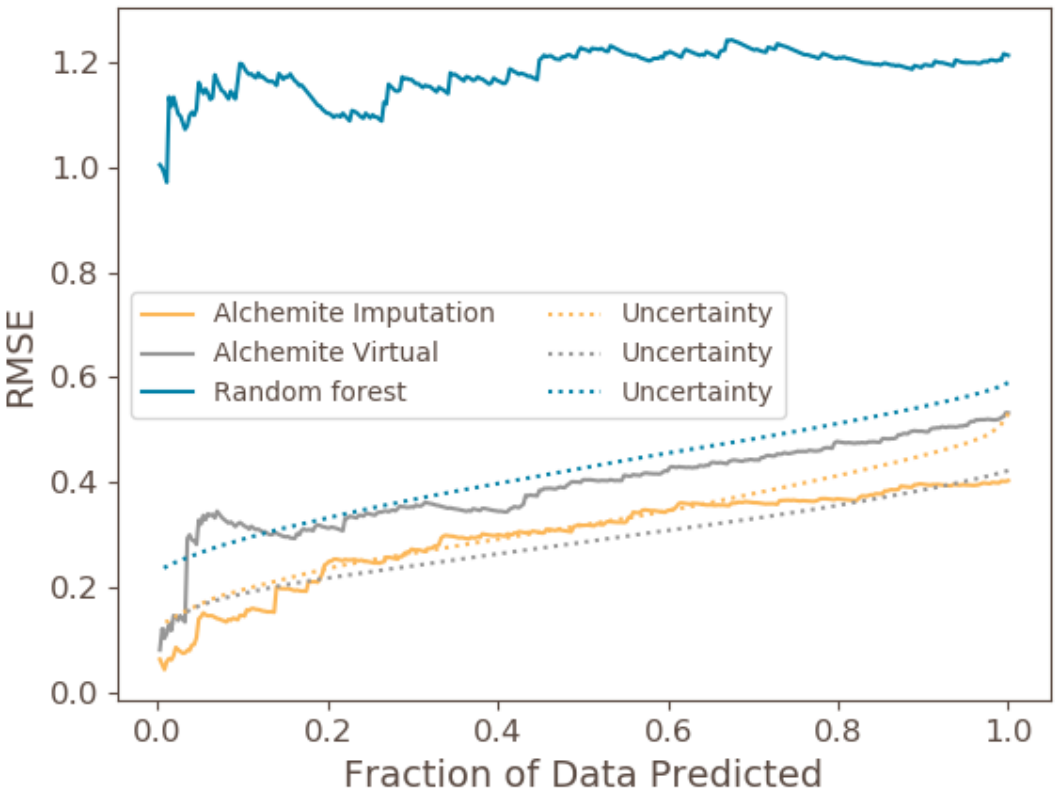

Figure S4: Graph illustrating the relationship between confidence and accuracy of prediction for a single pEC_{50} ADMET endpoint. The x-axis shows the most confidently-predicted fraction of the test set data, i.e. in moving from right to left only the most confidently predicted values are included. The y-axis shows the root-mean-square error (RMSE) of the fraction of predictions, i.e. a lower value indicates more accurate predictions. The results for the Alchemite Imputation and Virtual models are shown in orange, and grey respectively and those for random forest in blue. The dotted lines shows the mean value of the uncertainty estimates for each model.