Supplementary Material for:

Practical Applications of Deep Learning to Impute Heterogeneous Drug Discovery Data

Benedict W. J. Irwin*†§, Julian Levell‖, Thomas M. Whitehead‡, Matthew D. Segall*†, Gareth J. Conduit‡§

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

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

‖ Constellation Pharmaceuticals Inc., 215 First St Suite 200, Cambridge, MA 02142, USA

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

Abstract

This additional material supports the paper by describing the method, data set and compounds in more detail.

Methods

Here we provide additional details on the method beyond the description provided in the manuscript, which relies on a reference to a previous paper [1].

With Alchemite, the goal is to solve for the weights and biases of a neural network (depicted in **Figure S1**) that is used within a functional flow chart (depicted in **Figure S2**), where some outputs of the neural network in the first iteration(s) are potentially used as the *inputs* of subsequent iterations. This is solved iteratively in the context of a fixed-point equation $f(\boldsymbol{x}) = \boldsymbol{x}$, for a solution f that is orthogonal to the identity [1,2], for example in terms of matrices and vectors, finding an $Ax = x$. The trivial solution is $A = I$, the identity matrix. To have predictive power we require that $A \neq I$, or more specifically all entries on the leading diagonal of A are zero. For the inputs to the first iteration, missing values are replaced by the mean of the available values of the corresponding endpoint. An iterative expectation maximization algorithm is applied [3], whereby the weights of the neural network are optimized until convergence is reached.

In the applications described herein, the model will have N inputs and outputs, of which $N = N_d + N_e$; where N_d is a number of molecular descriptors and N_e is the number of experimental assay endpoints. The matrix columns corresponding to the descriptor inputs will be complete because these can be computed in advance for any molecular structure. However, the assay endpoint columns may be sparsely occupied; some, or even most, of the potential experimental data may be missing. The output is a complete matrix of assay endpoints in which the missing values have been imputed (the process illustrated in **Figure 1**).

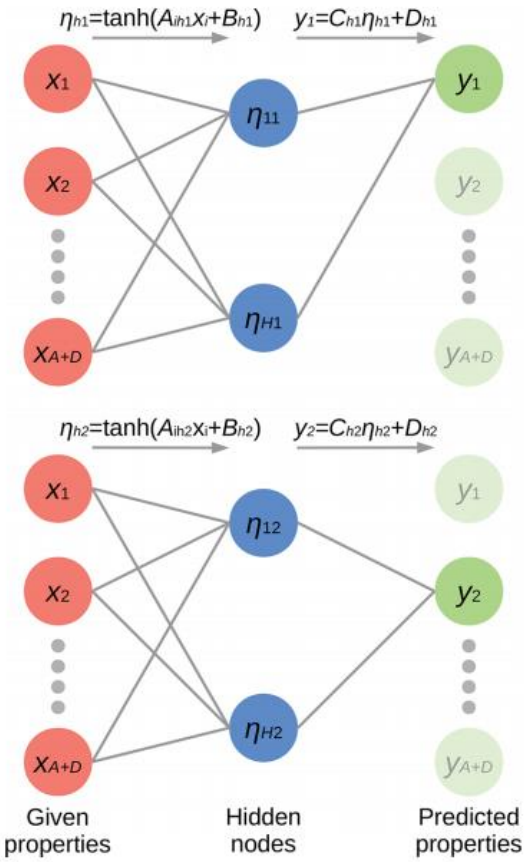

Figure S1. An example network architecture, see Whitehead et al. [1] for more details.

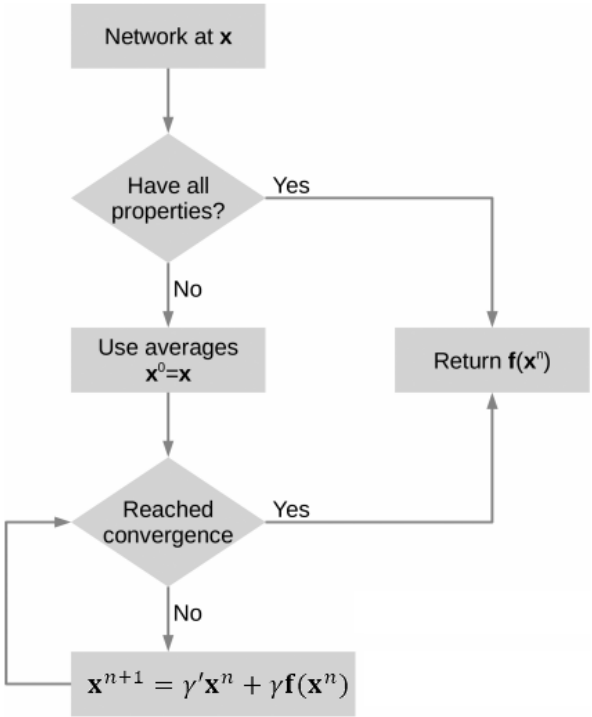

Figure S2. The functional method applied to the network to fill in missing values. Here γ and γ' play the role of a mixing fraction such that $\gamma' = 1 - \gamma$, with γ in the range [0,1].

ADME Assay Definitions and Details

Here we summarise the ADME assay endpoints in detail:

PAMPA : Effective mean PAMPA permeability (Pe) 10^{-6} cm/s
Protocol : PAMPA model using 1.8 % solution (w/v) of lecithin in dodecane (Avdeef, A; Tsinman, O. *European Journal of Pharmaceutical Sciences* **2006**, *28(1-2)*, 43-50)

PPB% : plasma protein binding % bound. Separated by species.
Protocol : Test compound (3µL of 1mM in dmso) incubated at 37°C with plasma (497µL, frozen stock warmed directly to 37°C in water bath) for 6h in CO_2 incubator. Placed 120µL of this sample in dialysis setup against PBS pH 7.4 (120µL) and incubated at 37°C in 5% CO2 atmosphere on rotary shaker at 100rpm. Removed 50µL from plasma and buffer sides of dialysis chamber and diluted these with 50µL of PBS and blank plasma, respectively. Shake at 1000 rpm for 2 minutes and add 400 µL of acetonitrile. Vortex at 1000 rpm for 10 minutes. Centrifuge for 30 minutes at 3,220 g. Transfer 250 µL of the supernatant to new vial and centrifuge again (3,220 g, 30 minutes). Transfer 100 µL of the supernatant to new vial for analysis. Add 100 µL of distilled water to each sample and mix for analysis by LC-MS/MS.

MLM/RLM/DLM/HLM : mouse, rat, dog and human in-vitro clearance assays in microsomes. Intrinsic clearance in uL/min/mg of protein.
Protocol : mouse/rat/dog/human liver microsome stock solution (12.5µL of 20mg/mL solution), NADPH solution (25 µL of 10mM in phosphate buffer), and phosphate buffer (210µL, 100mM, pH7.4) incubated at 37°C for 10mins. Added control (verapamil) or test compound (2.5µL of 100µM solution in dmso) and incubated at 37°C. Removed 25µL aliquots of solution at 0.5,5,10,15,20,30 minutes. Each aliquot was quenched with cold acetonitrile (125µL), centrifuged at 3220 g for 30 minutes. Transfer 100 µL of the supernatant to a new vial, diluted with water (100 µL), mixed well and analyzed using UPLC/MS/MS. Peak areas were determined from extracted ion chromatograms. The slope value, k, was determined by linear regression of the natural logarithm of the remaining percentage of the parent drug vs. incubation time curve. The in vitro half-life (in vitro t½) was determined from the slope (in vitro t½ = -Ln2/k). In vitro intrinsic clearance (in vitro CLint, in µL/min/mg protein) was calculated from the in vitro t½ (in vitro Clint = [(Ln2 x volume of incubation)/(in vitro t½ x amount of protein]).

CYP3A4_%Inhibition / CYP2D6_%Inhibition : single point percentage CYP inhibition at 10 µM of compound.
Protocol : into each well of a 96 deep-well screening plate is placed test compound stock solution (1µL of 10mM solution in dmso) [or vehicle or known standard inhibitor], human liver microsome stock solution (2µL of 20mg/mL solution), testosterone stock solution (1µL, 8mM in acetonitrile) and phosphate buffer (176µL, 100mM, pH7.4). Incubate plate at 37°C for 15 minutes before addition of NADPH solution (20 µL of 10mM in phosphate buffer). Incubate plate at 37°C for 10 minutes. Quench reaction with 3% formic acid in acetonitrile (300µL, cold). Centrifuge the plate at 3220 g for 40 minutes. Transfer 100 µL of the supernatant to a new plate, dilute with water (100 µL), mix well and analyze using UPLC/MS/MS. The inhibition of P450 3A4 enzyme in human liver microsomes is measured as the percentage decrease in the activity of marker metabolite formation compared to non-inhibited controls.

KinSol_pH7.4_ kinetic solubility µM or µg/mL :
Protocol : Compound stock solution (15µL, 10mM in dmso) diluted by addition of pH7.4 PBS (485µL), stirred at 1100rpm at 25°C for 2h, transferred to filter plate, vacuum filtered. Diluted 5µL of filtrate into acetonitrile:water (495µL, 1:1). LCMS comparison of compound concentration vs 3µM standard solution of test compound.

Compound Property and Chemical Space Analysis

In order to illustrate the diversity of compounds explored in the data sets used in this paper, we present chemical space plots to illustrate the diversity of compounds in the data sets relating to each of the projects. These chemical space plots were generated using the t-SNE algorithm [4] applied to the distribution of compounds in the space of compound descriptors used as inputs to the Alchemite and QSAR models.

The chemical space plots below show the distribution of compounds related to Project A only. All of these compounds were included in the initial data set.

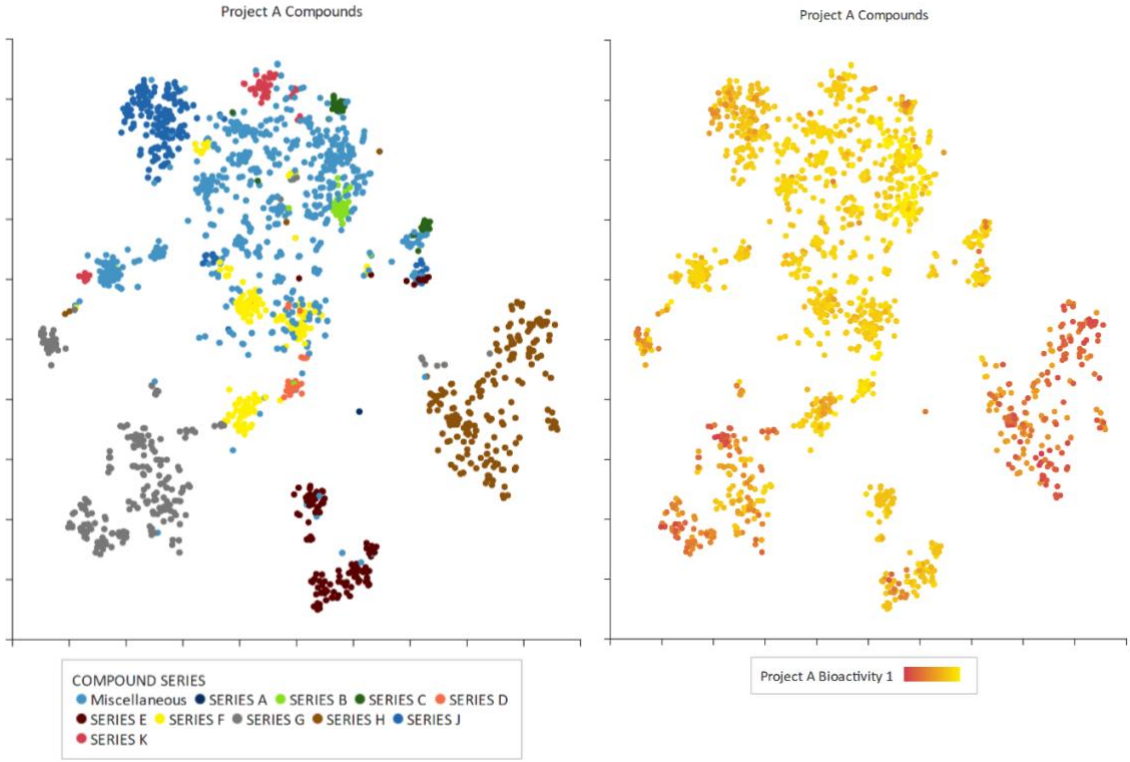

Figure S3. Chemical space plots for the compounds in Project A only. On the left, the points are coloured by the series labels assigned to the compounds by the project chemists. On the right, the points are coloured by the measured Project A Bioactivity 1 value for each compound; red is more active and yellow is less active with the colouring proportional to pIC_{50}.

The chemical space plots below show the distribution of compounds related to Project B only.

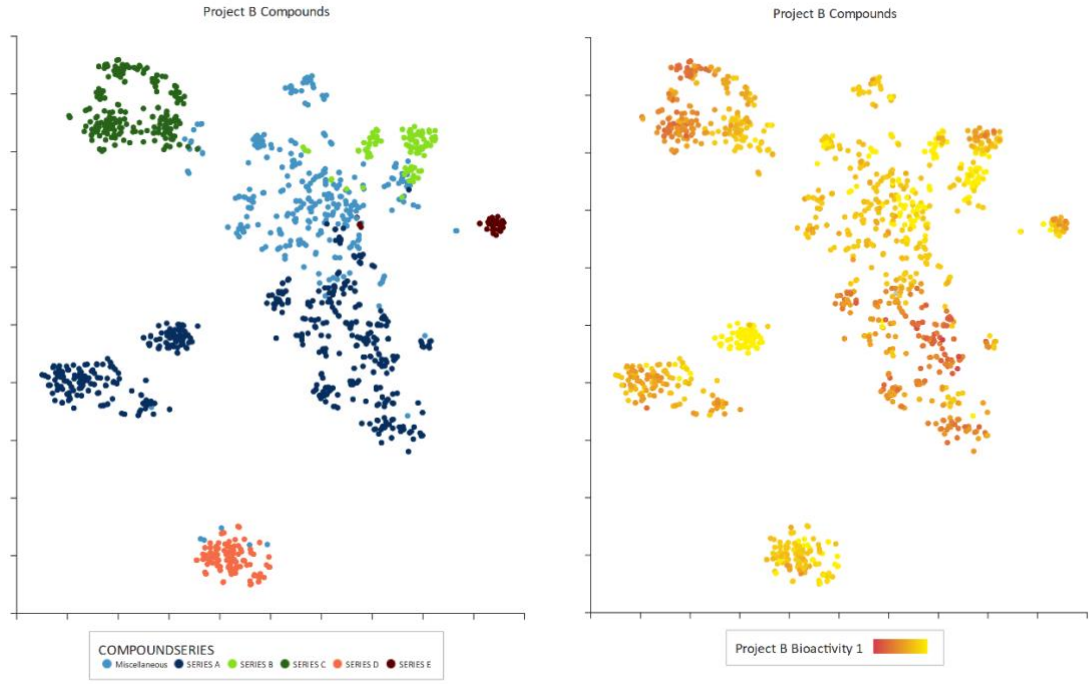

Figure S4. Chemical space plots for the compounds in Project B only. On the left the points are coloured by the series labels assigned to the compounds by the project chemists. On the right, the points are coloured by the measured Project B Bioactivity 1 value for each compound; red is more active and yellow is less active with the colouring proportional to pIC_{50}.

The following table summarises the range of values for the key molecular properties: Molecular weight (MW), calculated logP (clogP) and number of rotatable bonds (RotB), numbers of hydrogen bond acceptors and donors (HBA and HBD) and topological polar surface area (TPSA) calculated using StarDrop [5]A single data set is shown for Project A. For Project B the initial data set, as used in the preliminary comparison, and additional compounds which formed blocks 1, 2 and 3 in the temporal study are shown separately.

Table S1. List of number of compounds in each series for the datasets along with ranges of properties.

Name of Series	Number of Compounds	RotB	LogP	MW
Project A : Misc.	771	0 to 24	-2.8 to 6.6	59 to 549
Project A : Series A	3	35 to 48	-0.9 to -0.1	768 to 995
Project A : Series B	44	1 to 9	-0.9 to 2.7	99 to 225
Project A : Series C	48	2 to 12	0.2 to 6.1	211 to 401
Project A : Series D	36	6 to 14	0.6 to 3.3	177 to 316
Project A : Series E	215	10 to 30	1.4 to 6.3	216 to 627
Project A : Series F	243	6 to 22	0.6 to 5.1	221 to 446
Project A : Series G	486	6 to 20	1.2 to 7.0	293 to 571
Project A : Series H	488	7 to 23	-0.2023 to 5.5	294 to 551
Project A : Series J	258	6 to 19	1.488 to 5.1	324 to 542
Project A : Series K	77	4 to 19	0.6 to 5.9	226 to 484
Project B Original : Misc.	46	0 to 18	-0.7 to 3.9	138 to 419
Project B Original : Series A	252	5 to 26	-0.9 to 4.8	232 to 580

Project B Original : Series B	107	4 to 15	-1.7 to 3.2	164 to 390
Project B Original : Series C	55	9 to 16	1.0 to 4.8	272 to 508
Project B Original : Series D	41	11 to 20	1.5 to 5.6	247 to 454
Project B Original : Series E	37	1 to 15	-0.3 to 2.8	162 to 297
Project B Later : Misc.	284	1 to 23	-0.5 to 5.0	123 to 484
Project B Later : Series A	342	5 to 23	-0.5 to 5.5	302 to 627
Project B Later : Series C	171	9 to 27	1.4 to 4.4	297 to 579
Project B Later : Series D	77	12 to 25	2.1 to 5.4	324 to 556
Name of Series	**Number of Compounds**	**HBD**	**HBA**	**TPSA**
Project A : Misc.	771	0 to 8	0 to 11	15.27 to 197.4
Project A : Series A	3	5 to 7	12 to 15	346.6 to 447.8
Project A : Series B	44	0 to 2	1 to 5	12.03 to 79.29
Project A : Series C	48	0 to 2	0 to 2	15.79 to 82.67
Project A : Series D	36	0 to 1	1 to 3	20.31 to 53.17
Project A : Series E	215	0 to 3	1 to 7	23.55 to 142.4
Project A : Series F	243	0 to 2	0 to 4	16.13 to 92.5
Project A : Series G	486	0 to 3	1 to 7	32.34 to 135.8
Project A : Series H	488	0 to 2	1 to 6	41.13 to 125.2
Project A : Series J	258	0 to 2	2 to 6	22.0 to 109.2
Project A : Series K	77	0 to 3	3 to 10	55.13 to 168.0
Project B Original : Misc.	46	0 to 3	1 to 7	37.3 to 139.8
Project B Original : Series A	252	0 to 4	2 to 9	70.42 to 166.8
Project B Original : Series B	107	0 to 3	1 to 6	43.37 to 104.5
Project B Original : Series C	55	0 to 2	1 to 4	36.36 to 96.17
Project B Original : Series D	41	0 to 1	3 to 5	46.61 to 92.43
Project B Original : Series E	37	0 to 2	2 to 5	55.12 to 87.12
Project B Later : Misc.	284	0 to 3	0 to 8	12.03 to 126.0
Project B Later : Series A	342	1 to 4	2 to 9	84.66 to 180.2
Project B Later : Series C	171	0 to 3	1 to 6	43.78 to 110.2
Project B Later : Series D	77	0 to 1	2 to 6	37.38 to 96.02

From **Table S1** we can see that series A, C and D as well as Miscellaneous compounds were all continued as Project B progressed. The initial data included 252 examples of Project A compounds, but only around 50 compounds for each of C, D and Misc. This indicates that the new compounds in project B expanded significantly on the initial data.

Project A Data Distributions

Figures S5 through S17 show the distributions of the experimentally measured properties for Project A, which had measured values for 3 bioactivities, 2 cell-based activities and 8 ADME properties. The inequality symbols have been removed from these data to highlight the range in the project. This can lead to spikes where many data points have a similar recorded inactive value (e.g. pIC_{50} of 4, which corresponds to 100 μM).

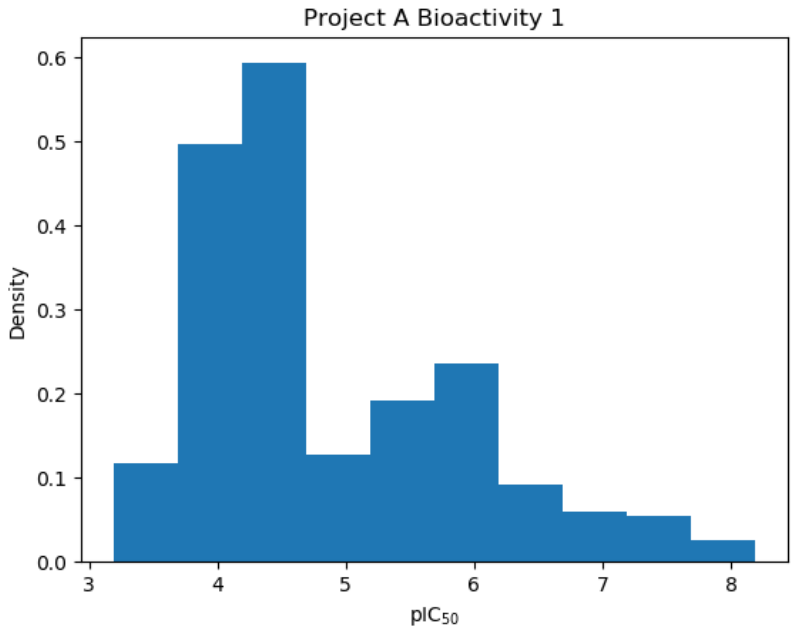

Figure S5. Distribution of experimentally measured Project A Bioactivity 1 values.

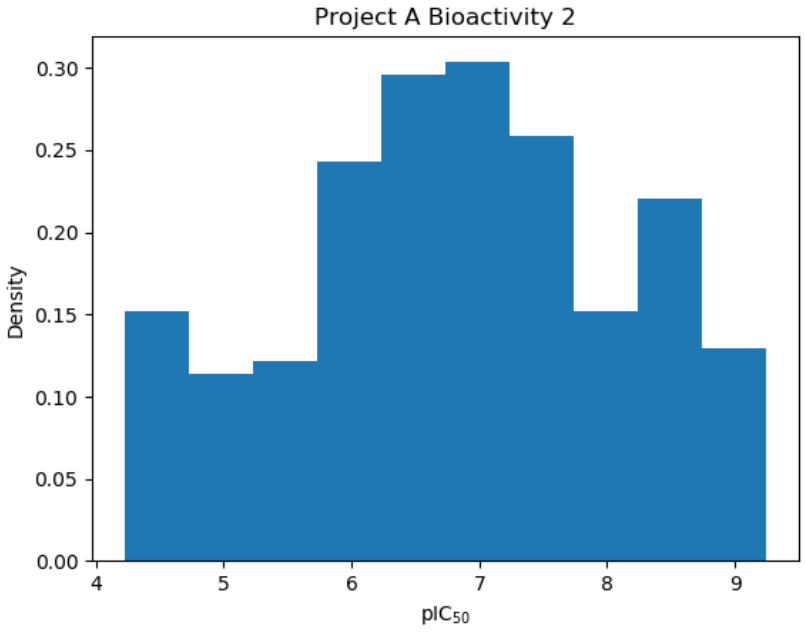

Figure S6. Distribution of experimentally measured values of Project A Bioactivity 2.

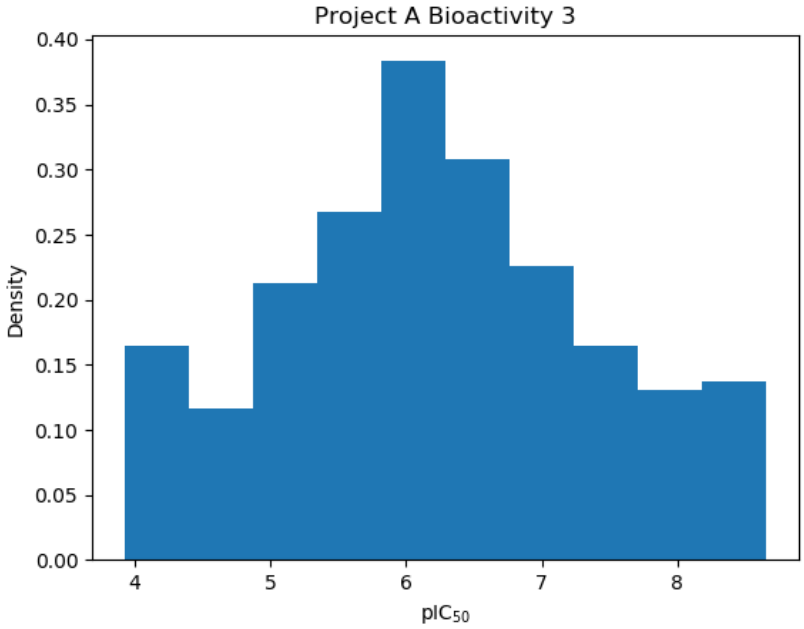

Figure S7. Distribution of experimentally measured values of Project A Bioactivity 3.

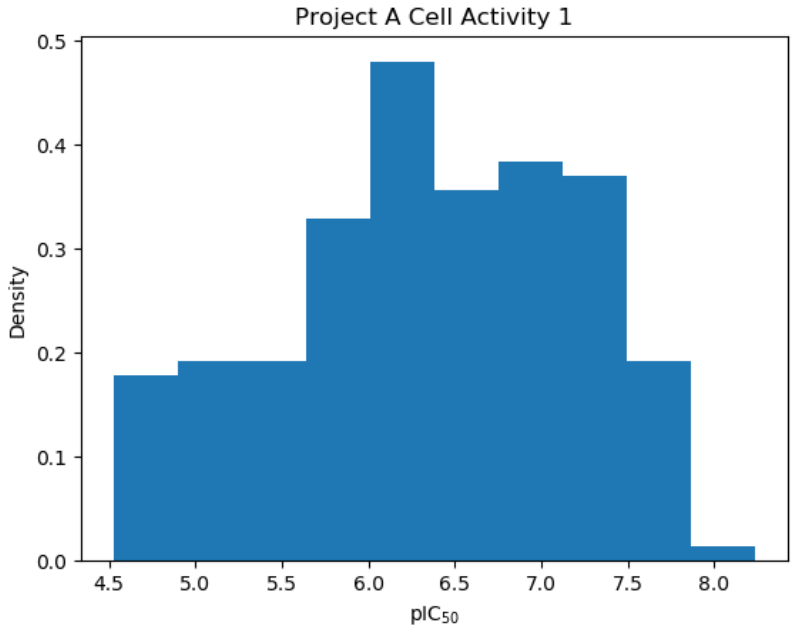

Figure S8. Distribution of experimentally measured values of Project A Cell-based Activity 1.

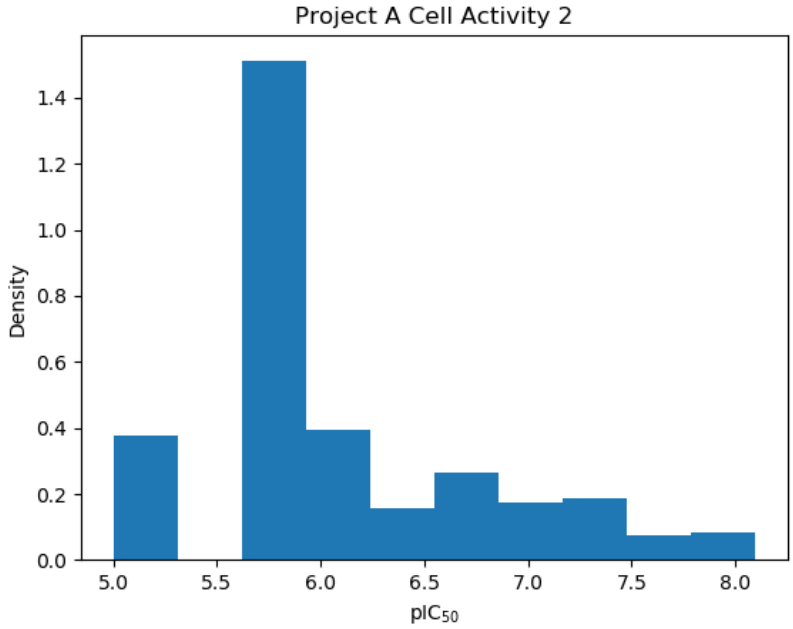

Figure S9. Distribution of experimentally measured values of Project A Cell-based Activity 2.

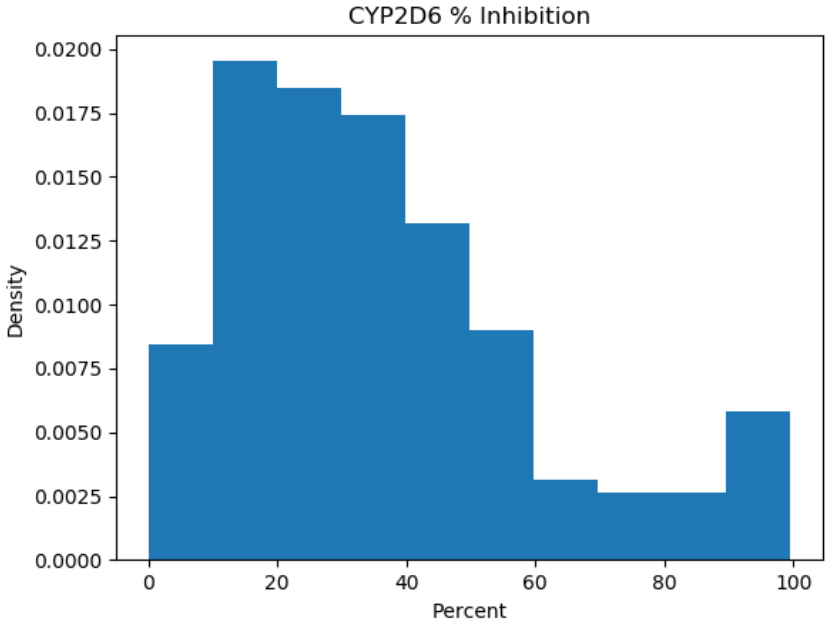

Figure S10. Distribution of experimentally measured values of CYP2D6 Percent Inhibition for Project A compounds.

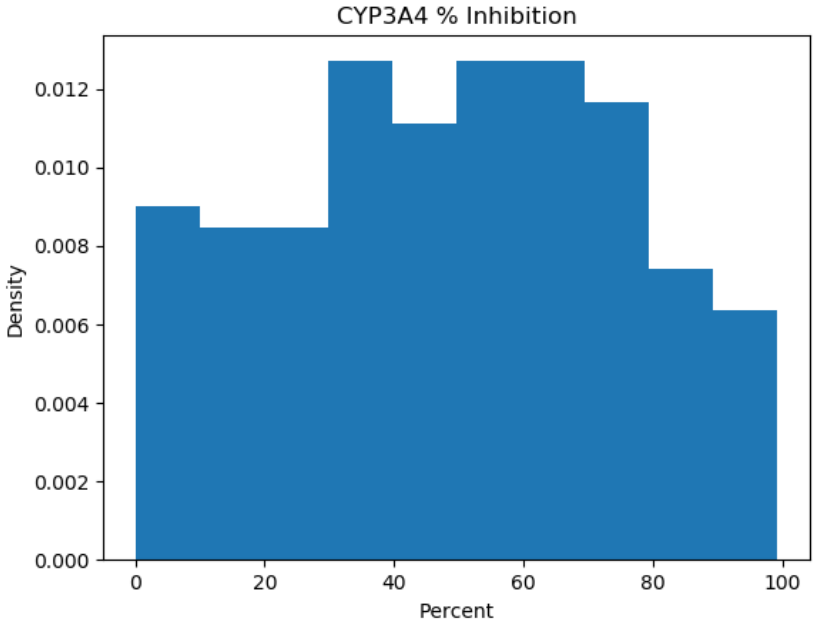

Figure S11. Distribution of experimentally measured values of CYP3A4 Percent Inhibition for Project A compounds.

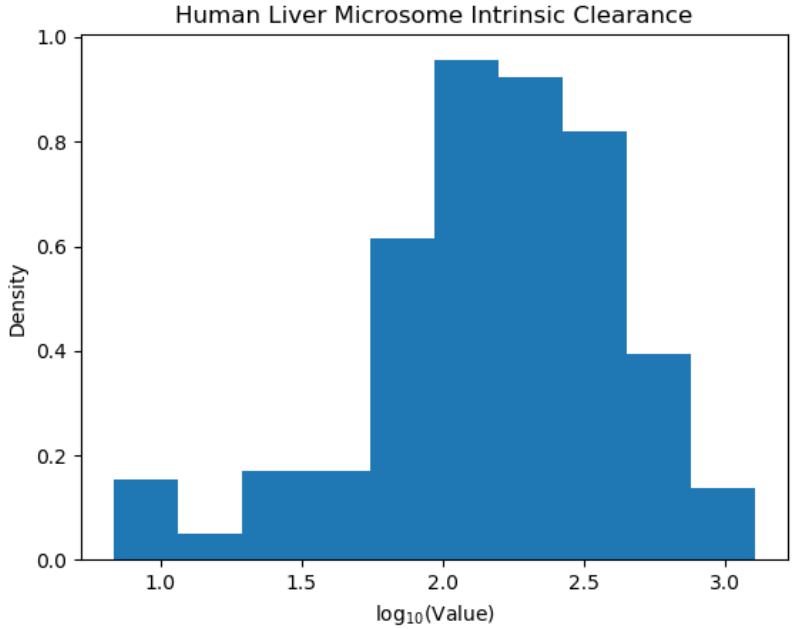

Figure S12. Distribution of experimentally measured values of Human Liver Microsome Intrinsic Clearance for Project A compounds.

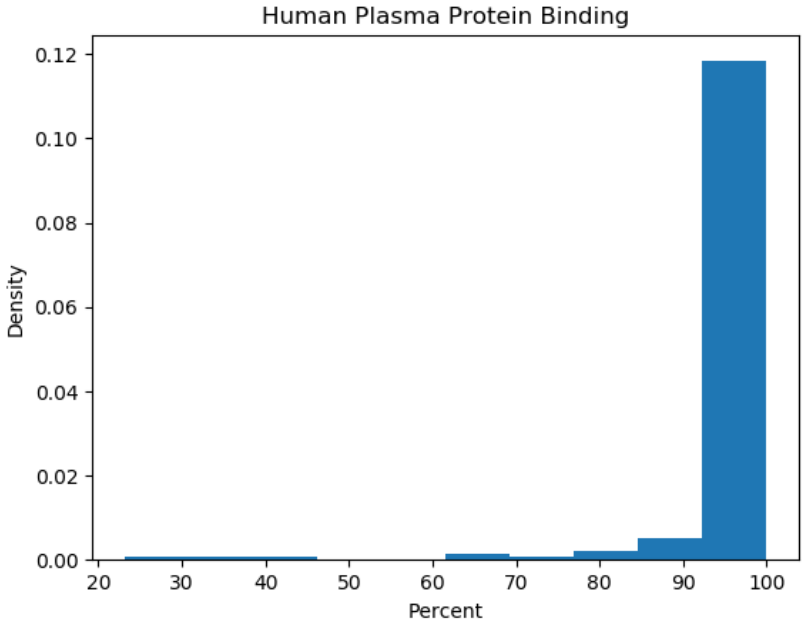

Figure S13. Distribution of experimentally measured values of Human Plasma Protein Binding for Project A compounds.

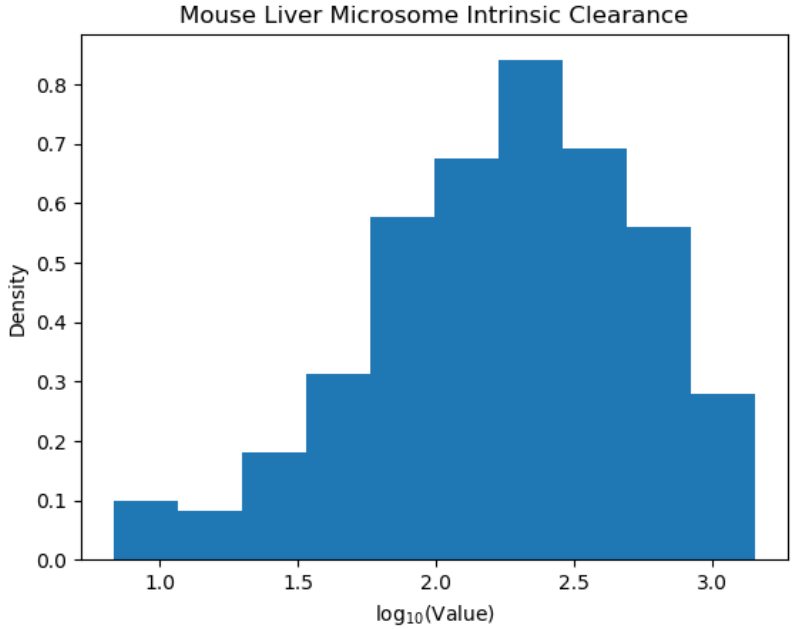

Figure S14. Distribution of experimentally measured values of Mouse Liver Microsome Intrinsic Clearance for Project A compounds.

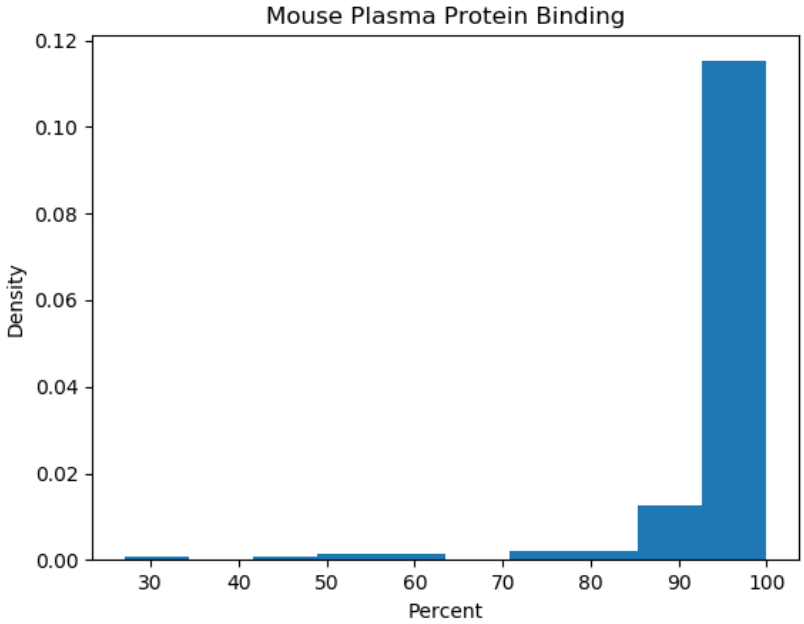

Figure S15. Distribution of experimentally measured values of Mouse Plasma Protein Binding for Project A compounds.

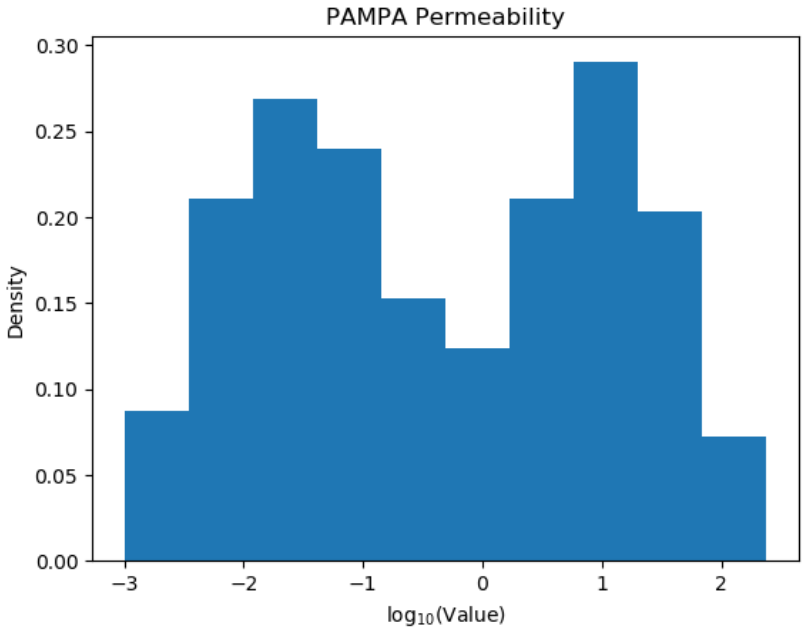

Figure S16. Distribution of experimentally measured values of PAMPA Permeability for Project A compounds.

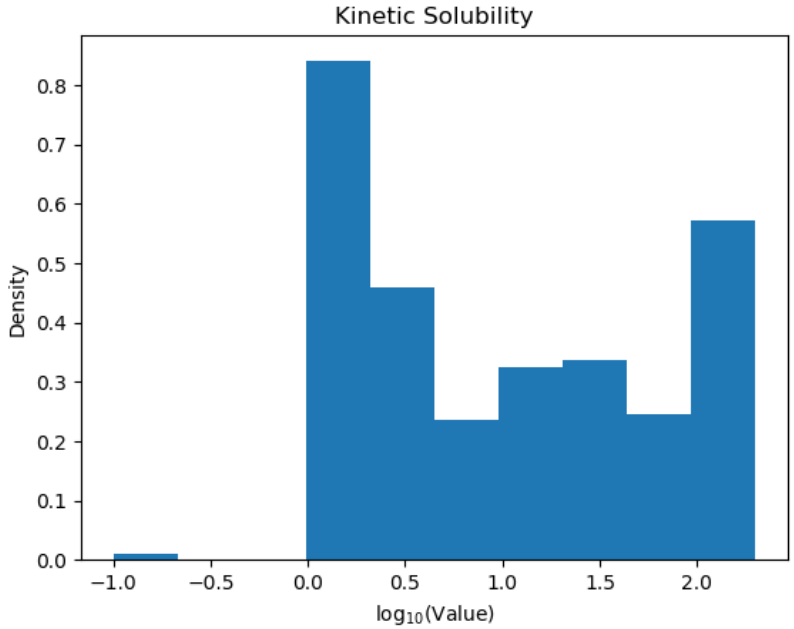

Figure S17. Distribution of experimentally measured values of Kinetic Solubility for Project A compounds.

Project B Data Distributions

Figures S18 through S30 show the distributions of the experimentally measured properties for Project B, which had measured values for 5 bioactivities and 8 ADME properties. The inequality symbols have been removed from these data to highlight the range in the project. This can lead to spikes where many data points have a similar recorded inactive value (e.g. pIC_{50} of 4, which corresponds to 100 μM).

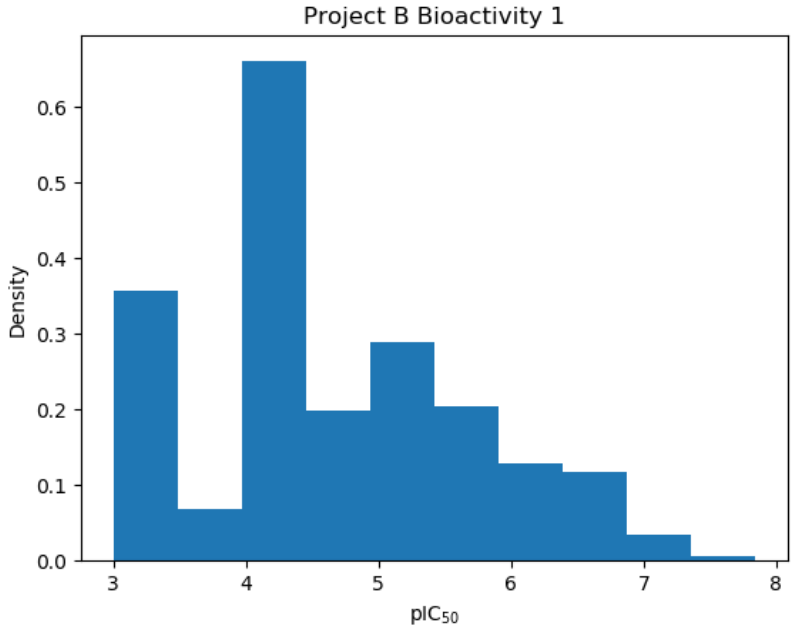

Figure S18. Distribution of experimentally measured Project B Bioactivity 1 values.

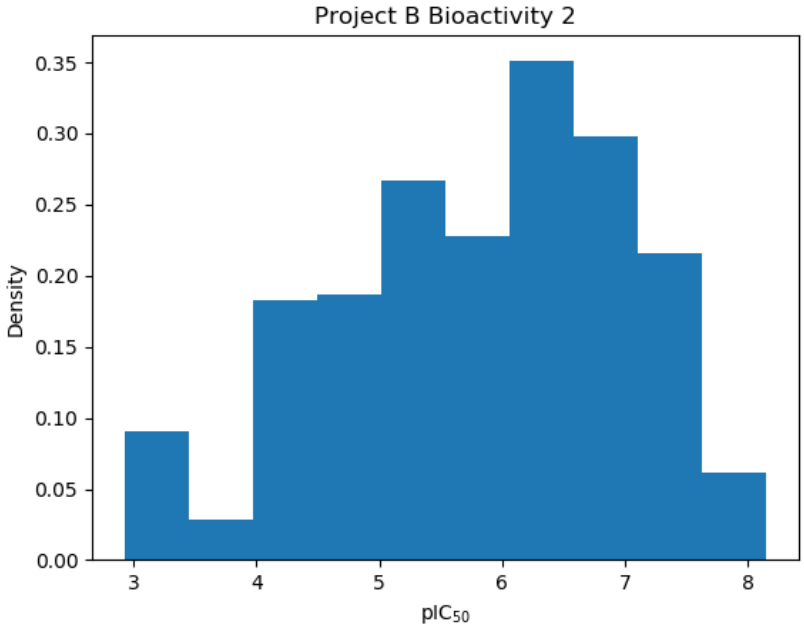

Figure S19. Distribution of experimentally measured Project B Bioactivity 2 values.

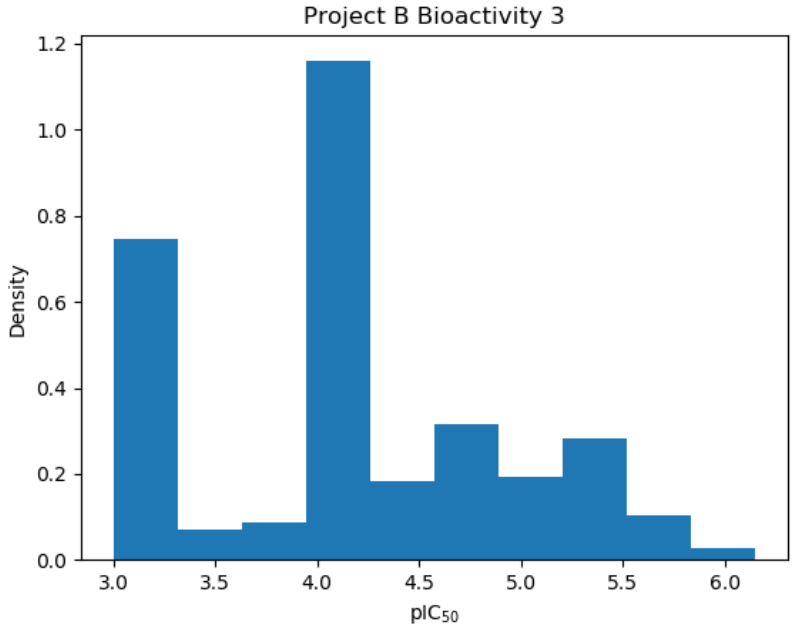

Figure S20. Distribution of experimentally measured Project B Bioactivity 3 values.

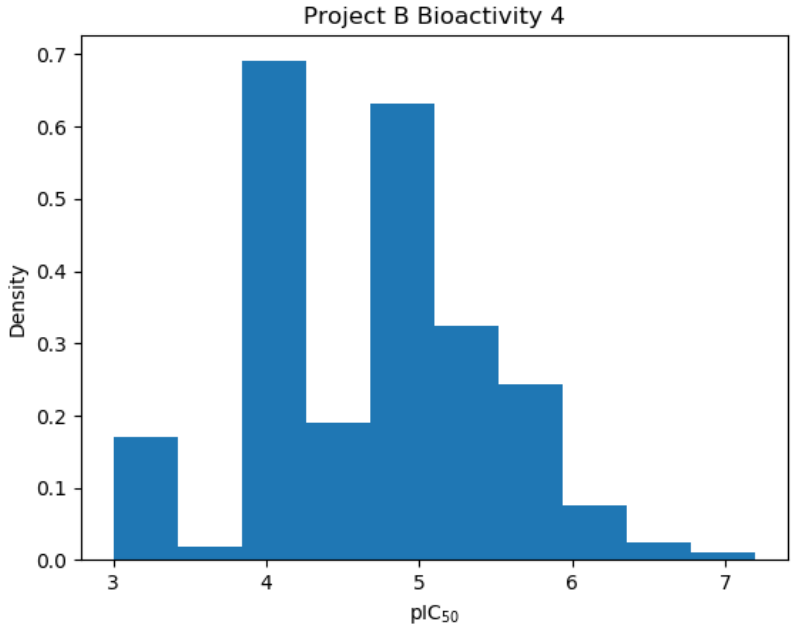

Figure S21. Distribution of experimentally measured Project B Bioactivity 4 values.

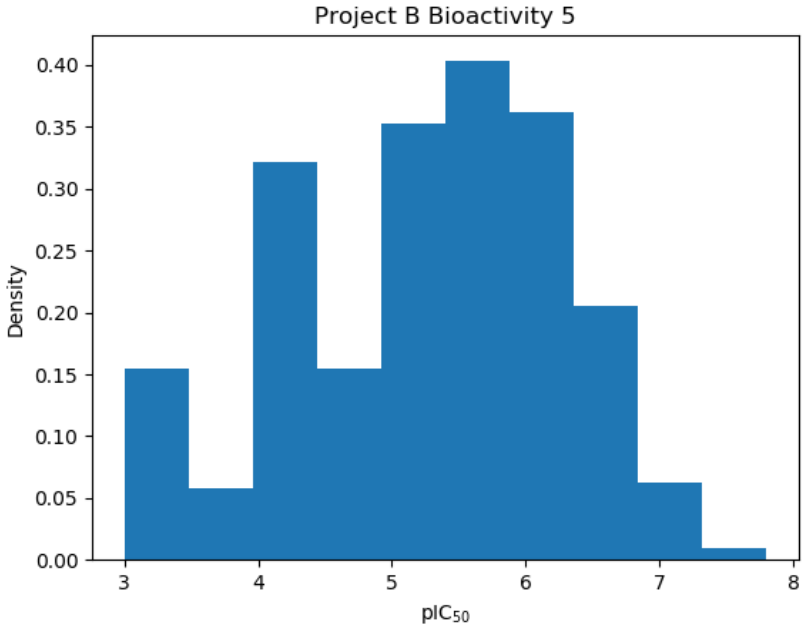

Figure S22. Distribution of experimentally measured Project B Bioactivity 5 values.

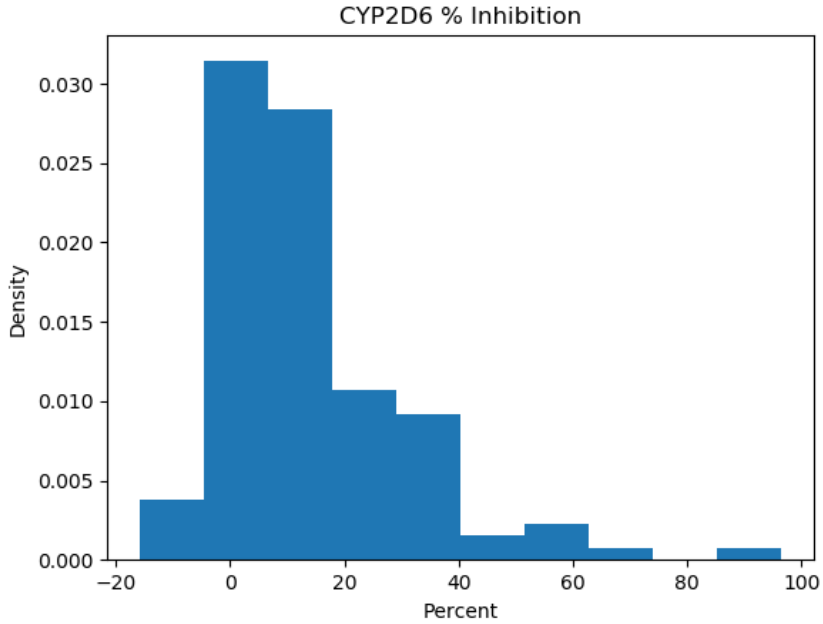

Figure S23. Distribution of experimentally measured values of CYP2D6 Percent Inhibition for Project B compounds.

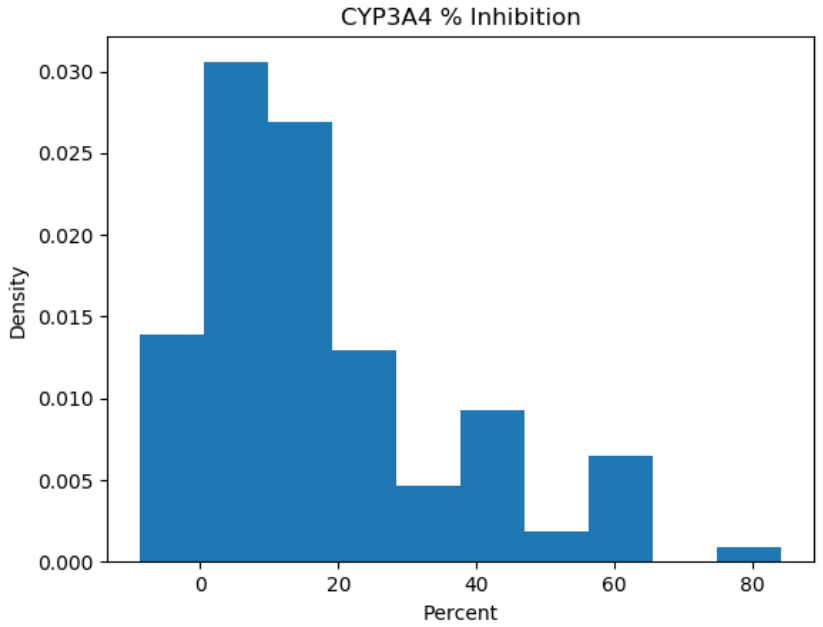

Figure S24. Distribution of experimentally measured values of CYP3A4 Percent Inhibition for Project B compounds.

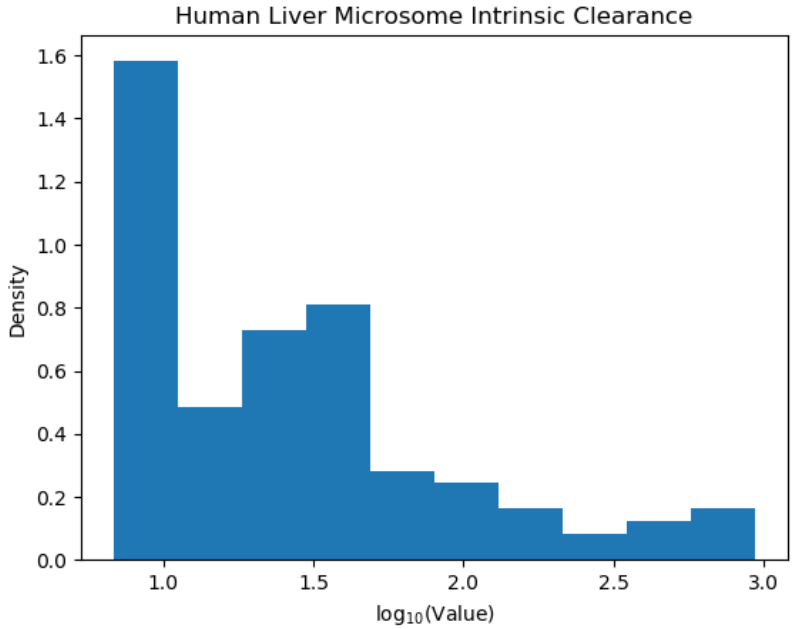

Figure S25. Distribution of experimentally measured values of Human Liver Microsome Intrinsic Clearance for Project B compounds.

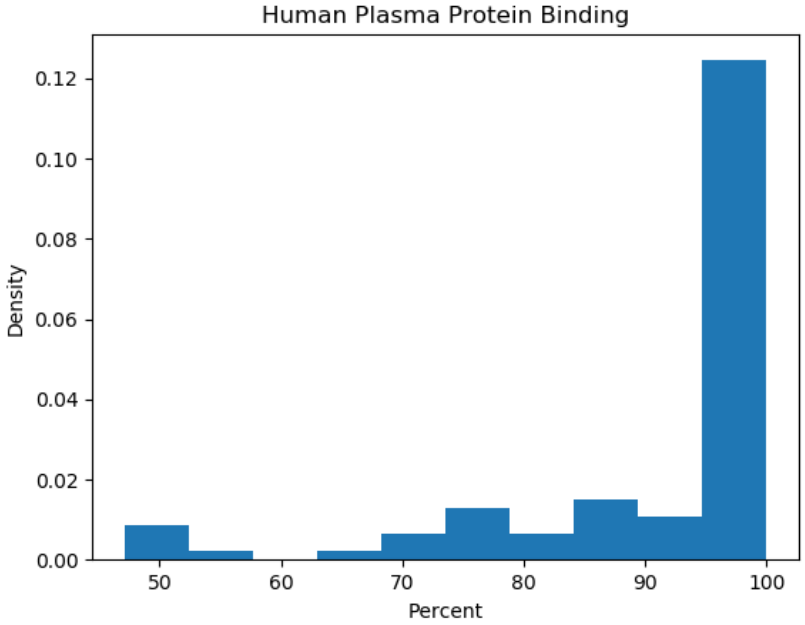

Figure S26. Distribution of experimentally measured values of Human Plasma Protein Binding for Project B compounds.

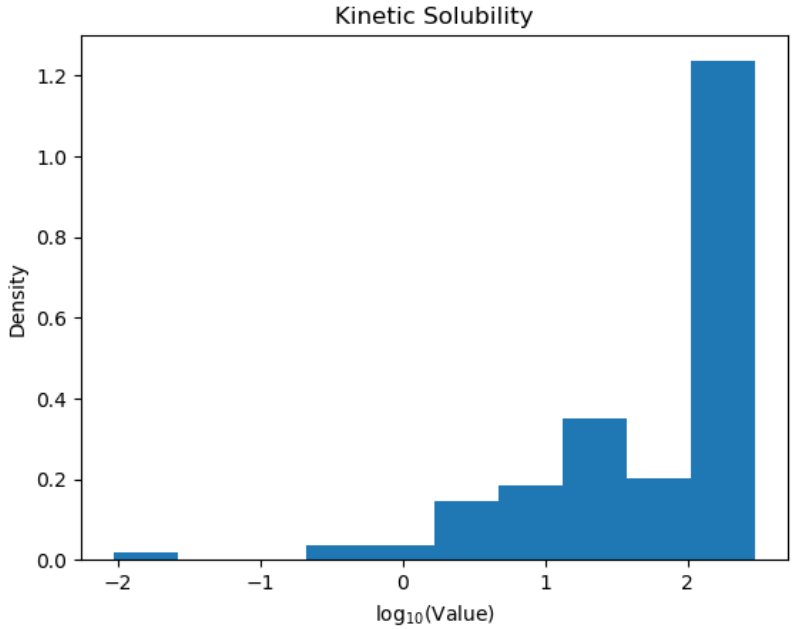

Figure S27. Distribution of experimentally measured values of Kinetic Solubility for Project B compounds.

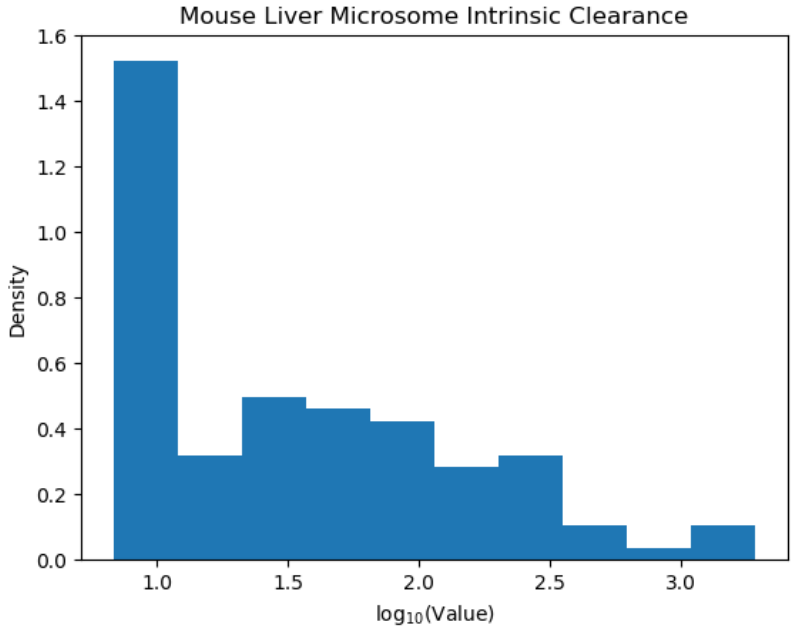

Figure S28. Distribution of experimentally measured values of Mouse Liver Microsome Intrinsic Clearance for Project B compounds.

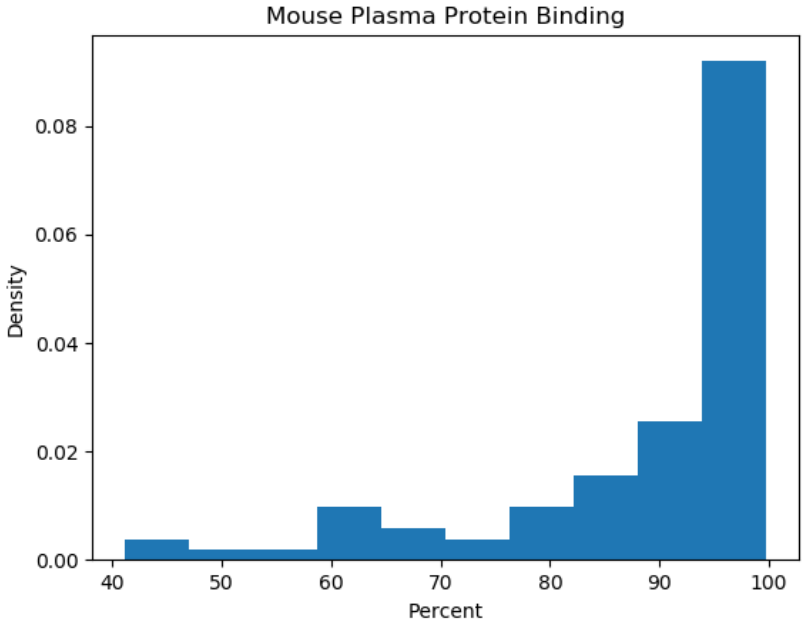

Figure S29. Distribution of experimentally measured values of Mouse Plasma Protein Binding for Project B compounds.

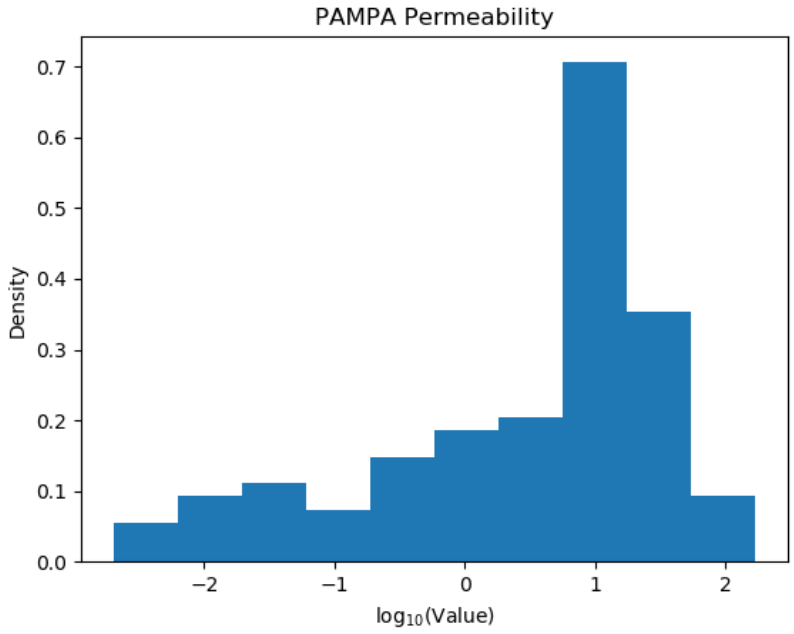

Figure S30. Distribution of experimentally measured values of PAMPA Permeability for Project B compounds.

Temporal Chemical Space Analysis

It is instructive to check the changes in chemical space in the temporal study. The chemical space plot in Figure S31 shows the full data set combining compounds from both Projects A and B, on which the series labels are shown. We note that the series labels generated by the chemists do not correspond perfectly to this particular embedding but captures the majority of compounds. It is also notable that there is an overlap of Miscellaneous compounds between Projects A and B.

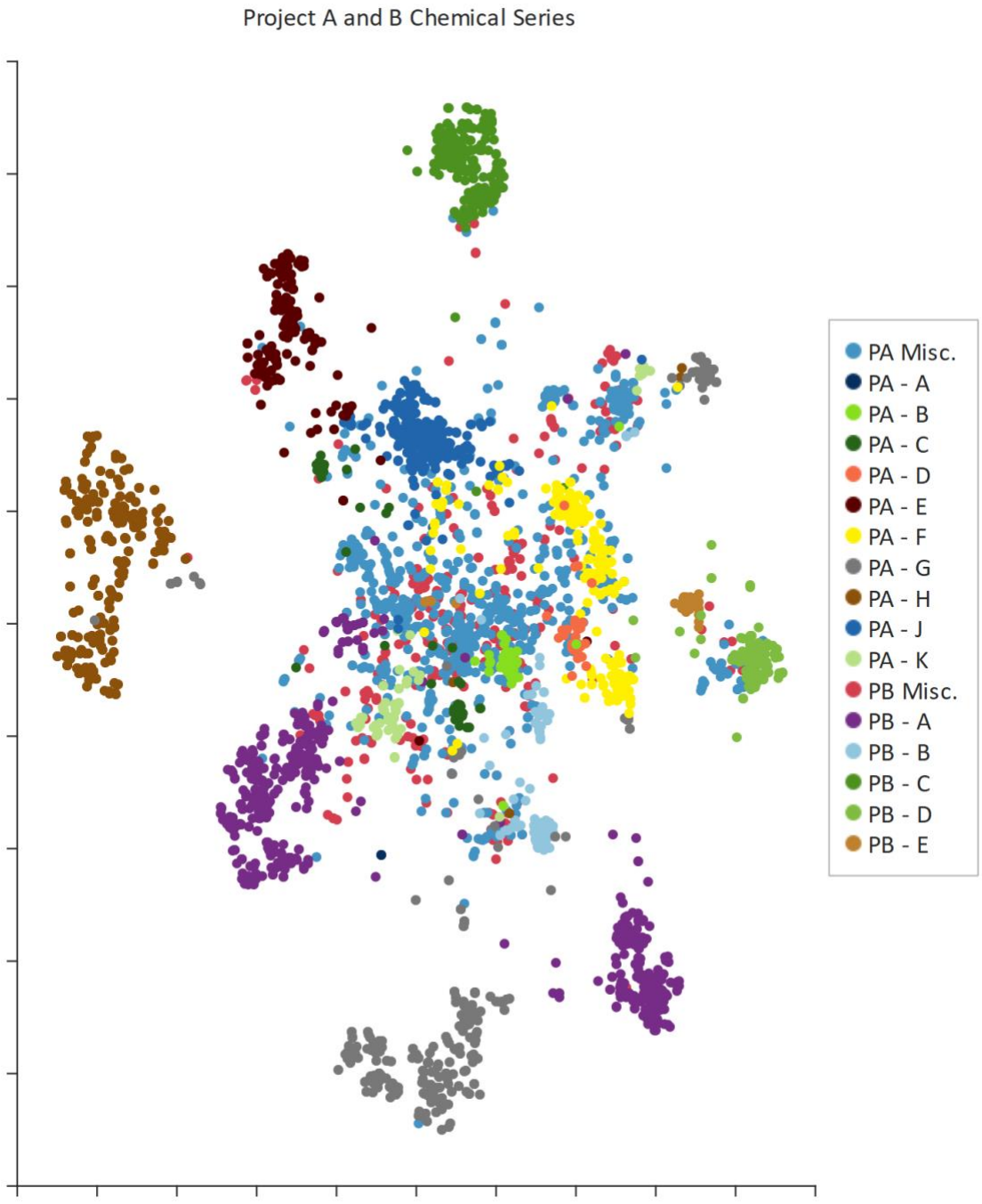

Figure S31. Chemical space of all compounds from all projects. Points are coloured by series with example notation "PA – C" standing for "Project A, Series C".

Figure S32 shows the three blocks of additional compounds, illustrating the changes in chemical space explored by Project B as it progressed. The initial data set used to train the first model for the temporal analysis (labelled Train) is quite diverse. Broadly speaking, the latter three temporally split blocks explored by Project B concentrate in series AB, CB and DB, as well as a number of Miscellaneous(B) compounds.

Figure S32. Chemical space of the three new blocks of compounds in reference to the initial data. From left to right, positions of compounds from block 1,2 and 3 are shown in dark blue against the initial data in light blue.

Block 3, which was used as the independent test set in the temporal analysis, has a reasonable degree of sampling from series PB - C, the left hand cluster of PB - A and an unlabelled sub-cluster of the Miscellaneous structures (top right), but a small number also from series PA- E and PB - D and perhaps a Misc. sub-cluster close to series PB - B.

Chemical Space Uncertainty Analysis

The chemical space above was a good tool visualisation of the uncertainty in predictions and how they are localised with chemical series. There are 18 assays which are fully imputed for Projects A and B using a final model built on all of the data. Each imputed value has an error bar analysis associated with it, and a full cross inspection requires several plots which are shown in Figure S33 to S48. To generate these the final fully imputed matrices of the data set were plotted in the chemical space for both projects (Figure S31) and coloured by uncertainty in prediction.

These indicate that certain assays have specific series for which predictions are more confident. In general, considering predictions with smaller error bars will bring the chemistry back to the descriptor space that is well characterised by the training set. Equivalently extrapolating from the known space will result in an increase in the size of the error bar. This represents the classic trade-off between exploitation of previously well-sampled areas of chemical space (where predictions can be made with higher confidence) and exploration of less-well understood, and hence less confidently predicted, chemical space.

In the following figures, the points are coloured from low uncertainty in red (representing high confidence) and high uncertainty in yellow (representing low confidence). The colour scales in subsequent plots are normalised for each assay.

Project A Activity Uncertainty Plots

Figures S33 to S37 show the distributions of uncertainties in the imputed values for the Project A bioactivites and cell-based activities. Data for these endpoints have only been measured for compounds withing Project A and therefore, as expected, we can see that the greatest confidence in the corresponding imputed values are highest for series related to this project. For example, we can see in Figure S35 that Project A Cell-based Activity 2 is most confidently imputed for series PA - G and PA -H (left and bottom). Cell-based Activity 2 is one of the most locally modelled endpoint, which may partly explain why QSAR models failed to predict it.

All of the Project A activities are more confidently imputed for the clusters corresponding to series PA - G and PA - H. Confidence is exhibited in series PA - E and PA - J for Bioactivities 1,2 and 3 and for Cell-based Activity 1. Bioactivity 2 is confidently imputed for many of the series, including miscellaneous compounds and is approaching a global model for this project.

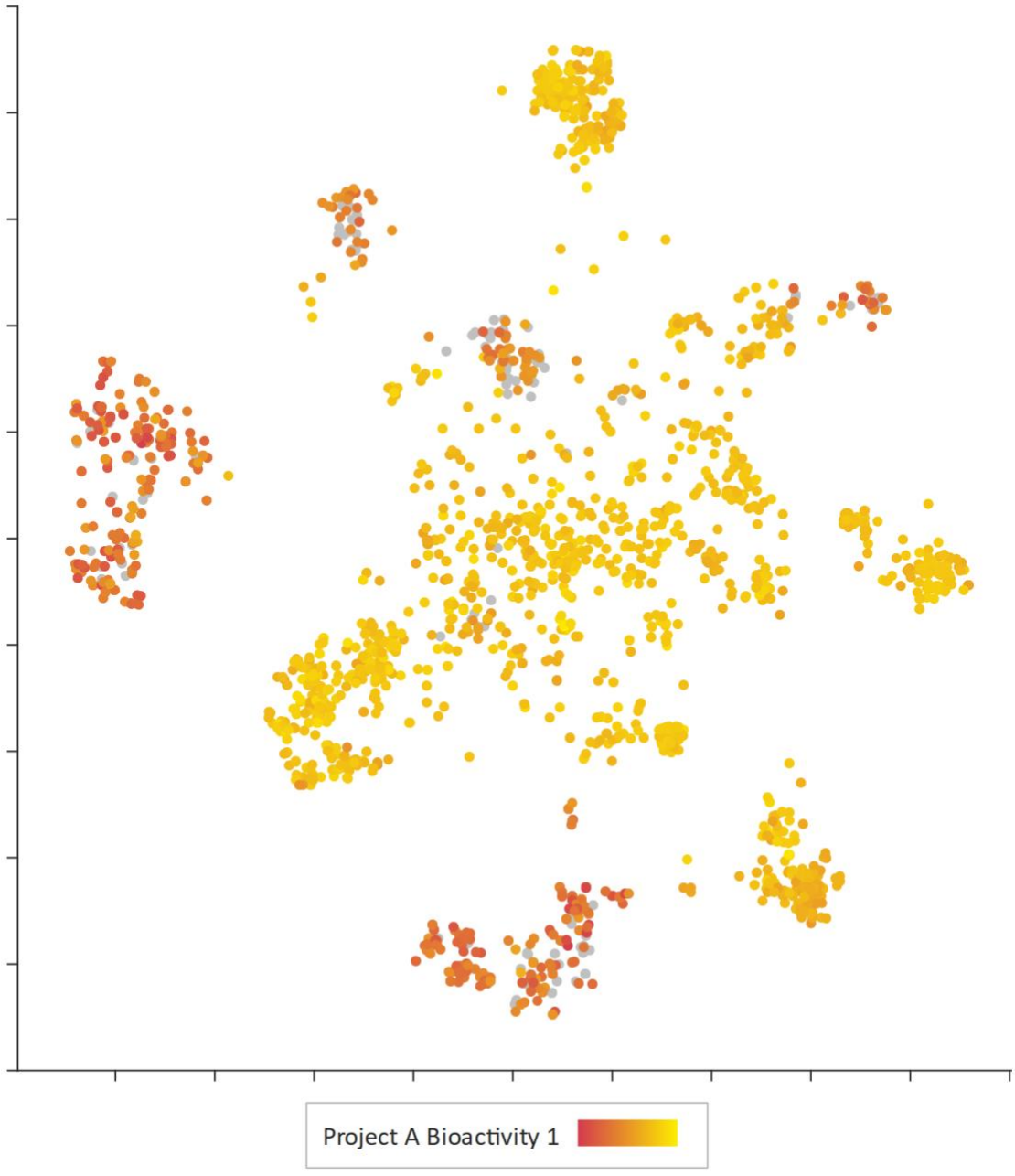

Figure S33. Uncertainty distribution in chemical space for predictions of Project A Bioactivity 1.

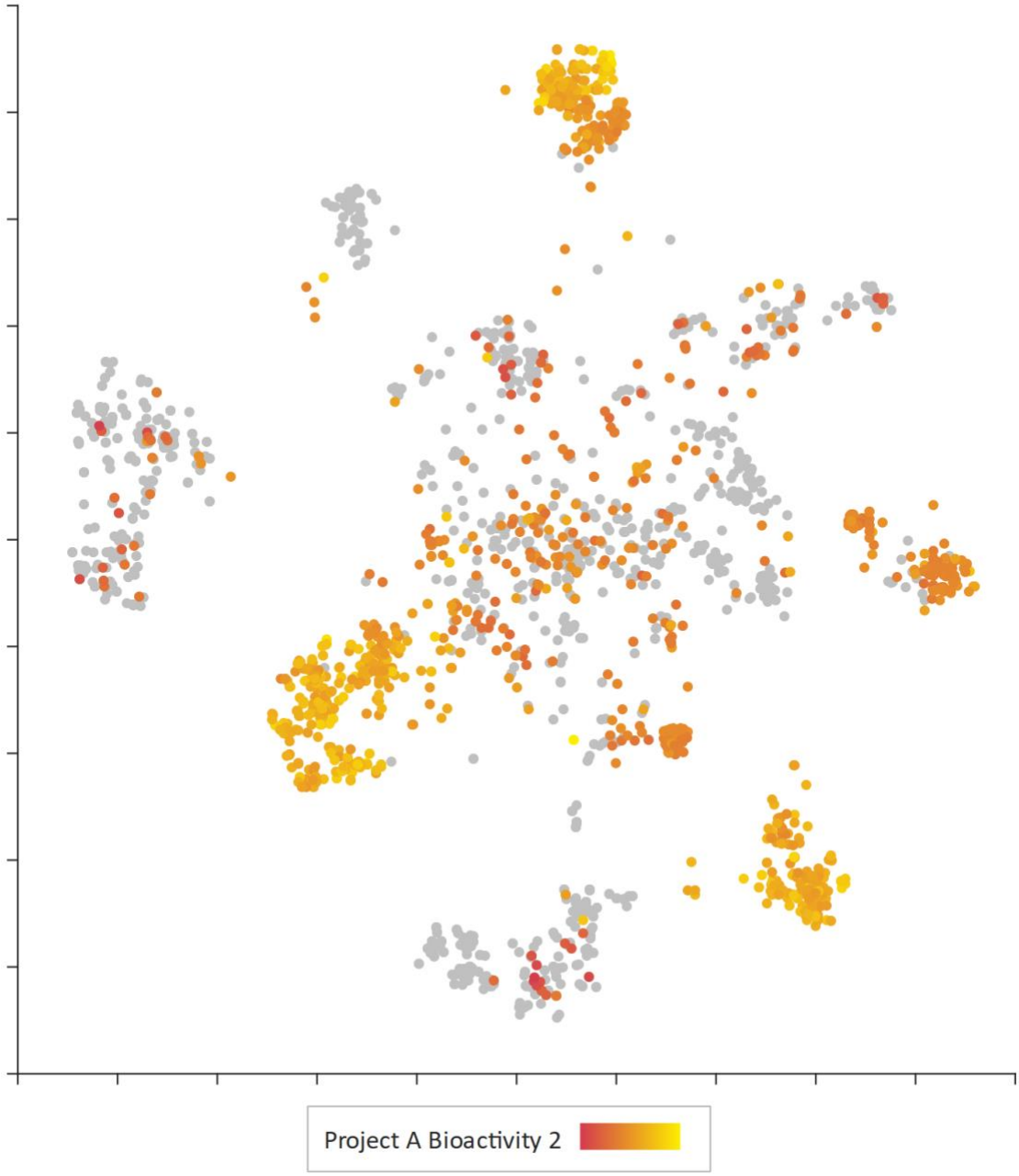

Figure S34. Uncertainty distribution in chemical space for predictions of Project A Bioactivity 2.

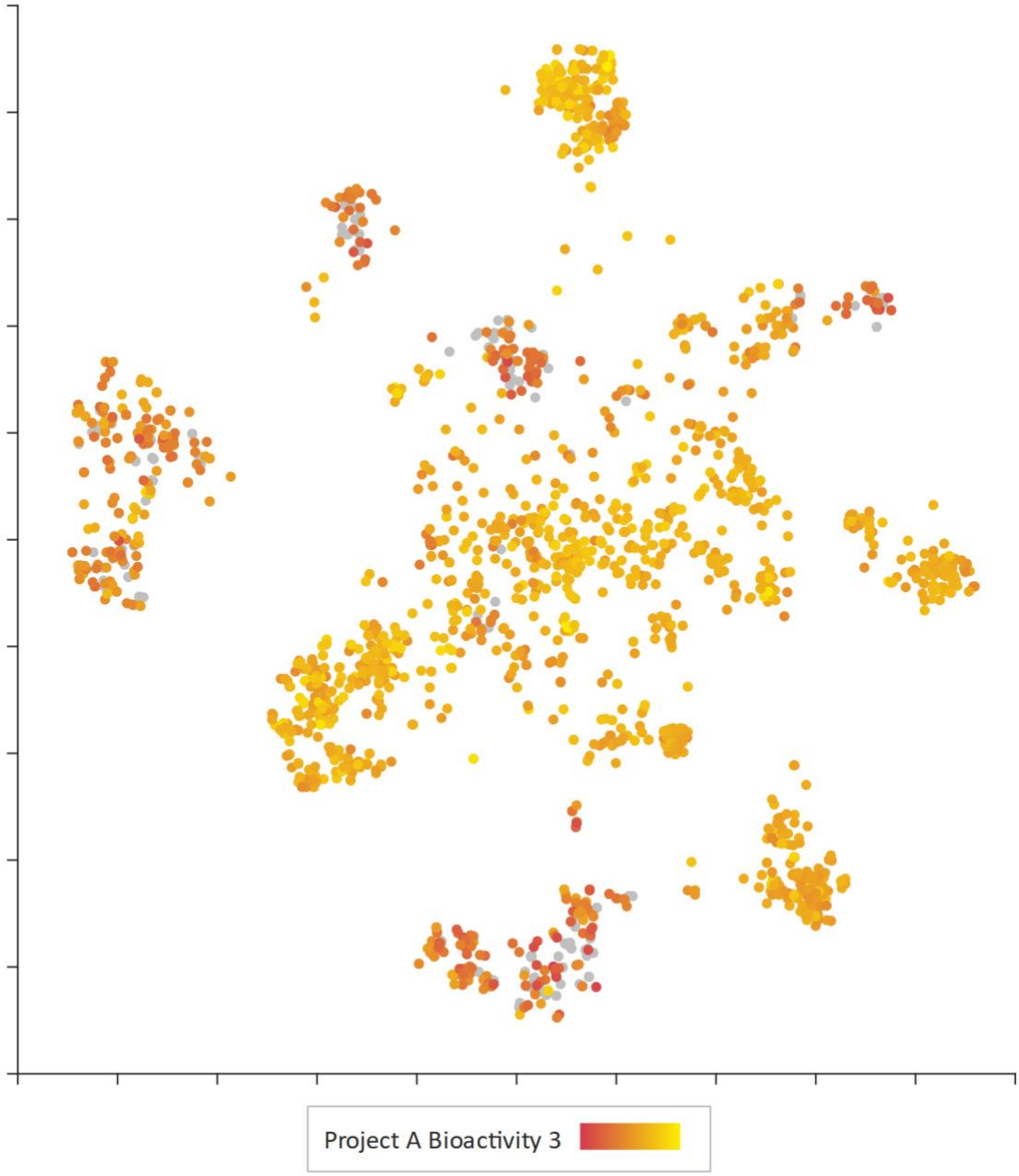

Figure S35. Uncertainty distribution in chemical space for predictions of Project A Bioactivity 3.

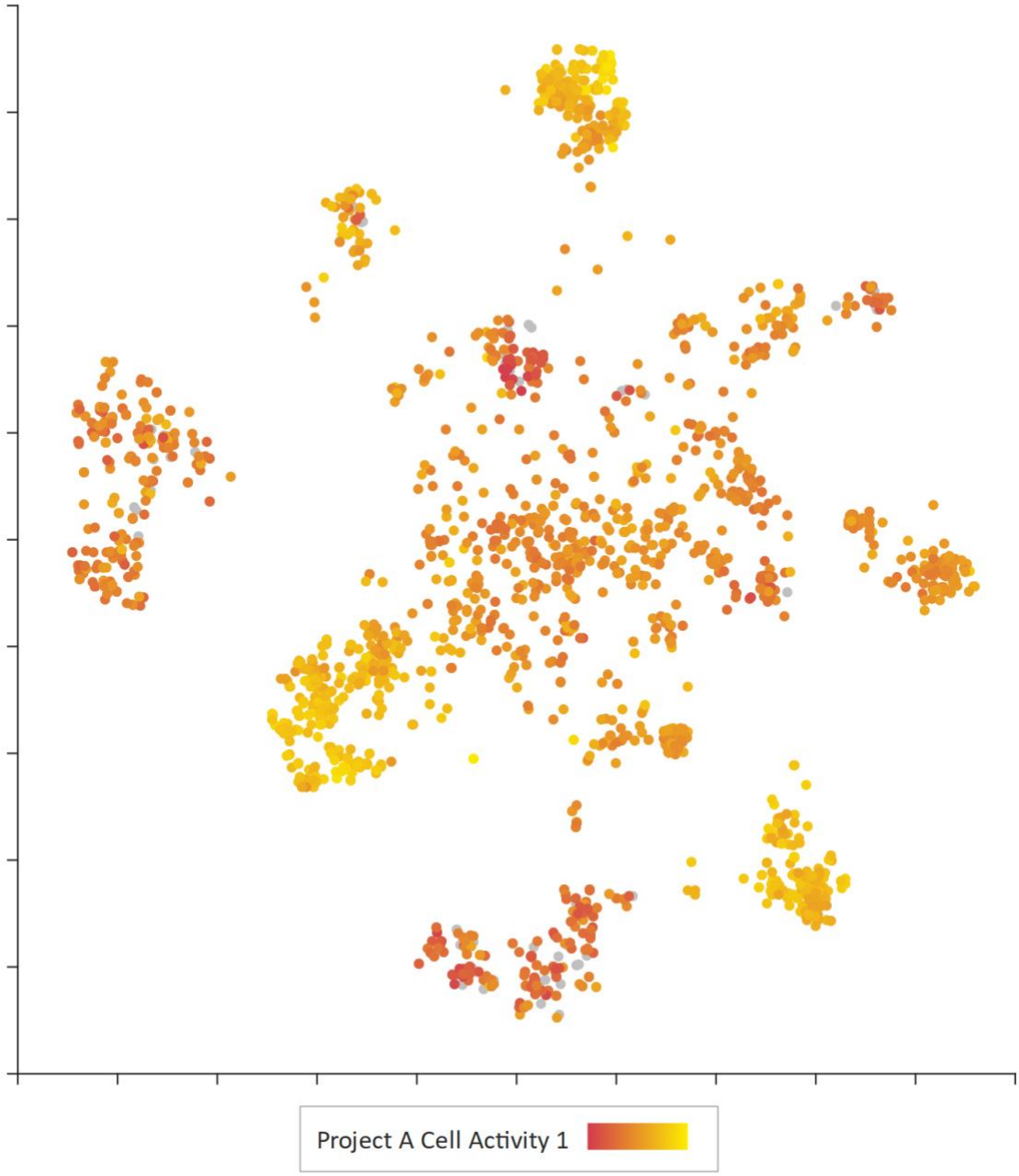

Figure S36. Uncertainty distribution in chemical space for predictions of Project A Cell Activity 1.

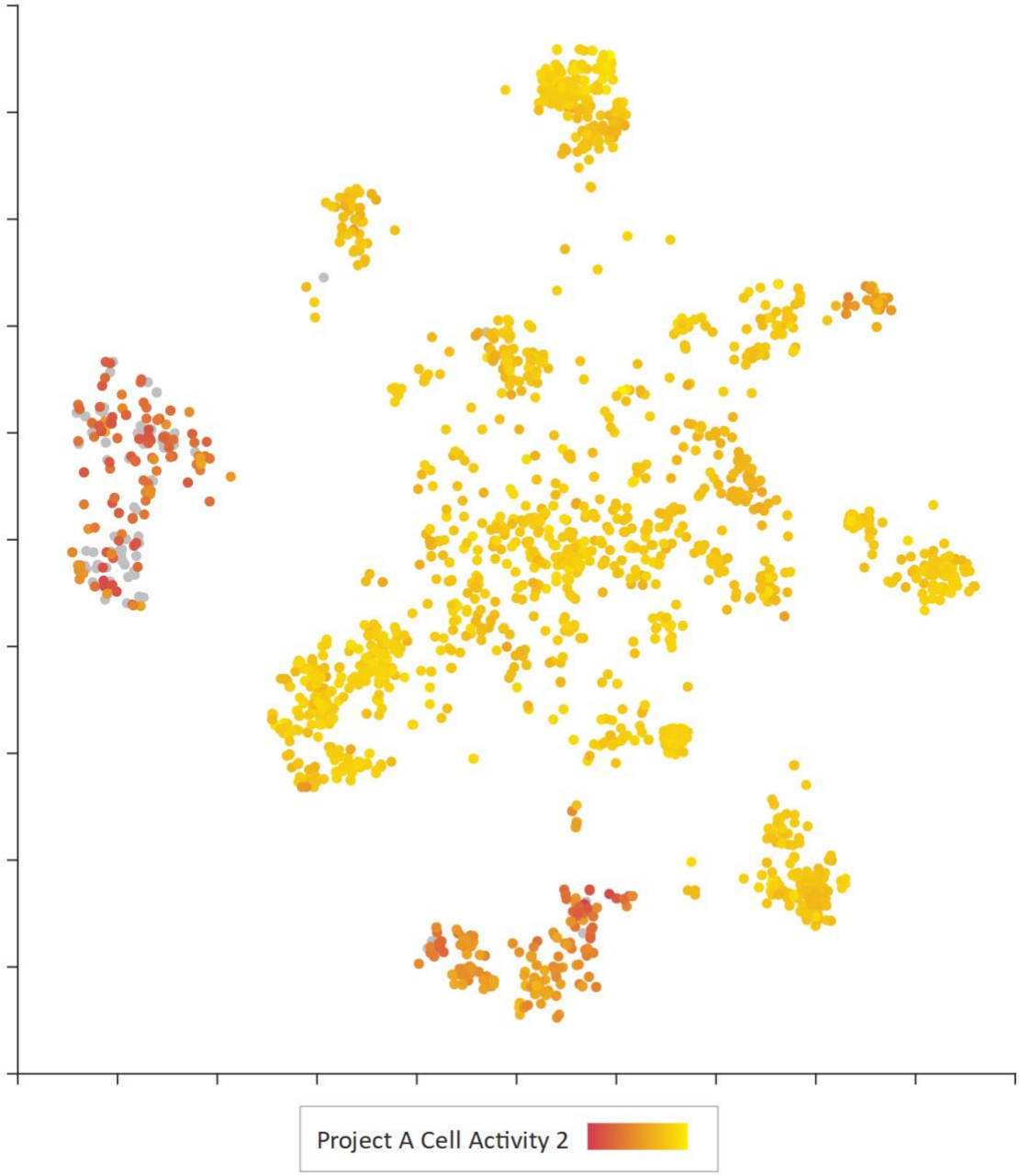

Figure S37. Uncertainty distribution in chemical space for predictions of Project A Cell Activity 2.

Project B Activity Uncertainty Plots

Figures S38 to S42 show the distributions of uncertainties in the imputed values for the Project B activities. Data for these endpoints have only been measured for compounds within Project B and therefore, as expected, we can see that the greatest confidence in the corresponding imputed values are highest for series related to this project. Project B activities are confidently imputed in series AB, CB and sometimes DB. The series labelled AB is split into two clusters in this embedding, and sometimes only half is well predicted. Project B Bioactivity 1 is better predicted across miscellaneous compounds.

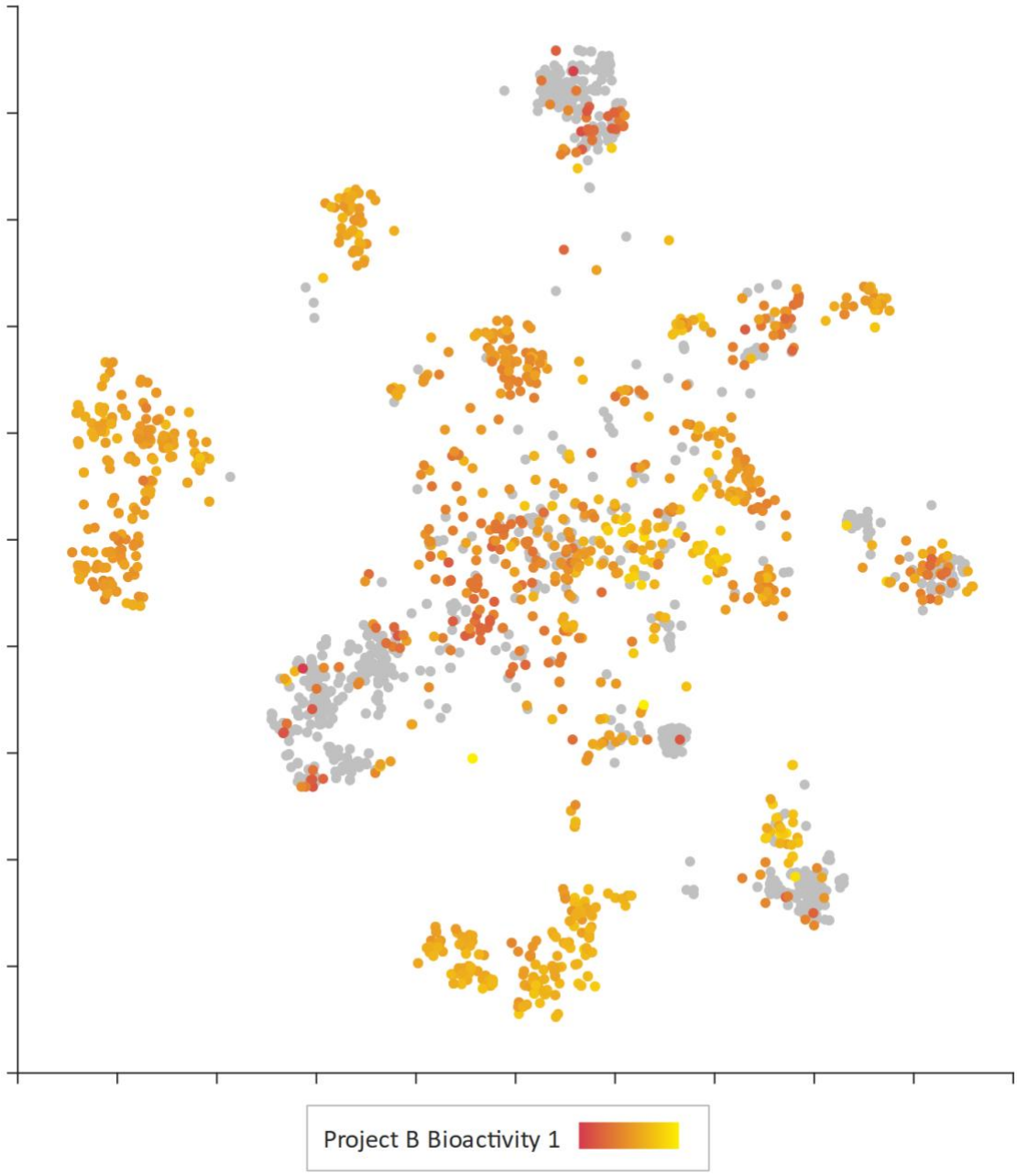

Figure S38. Uncertainty distribution in chemical space for predictions of Project B Bioactivity 1.

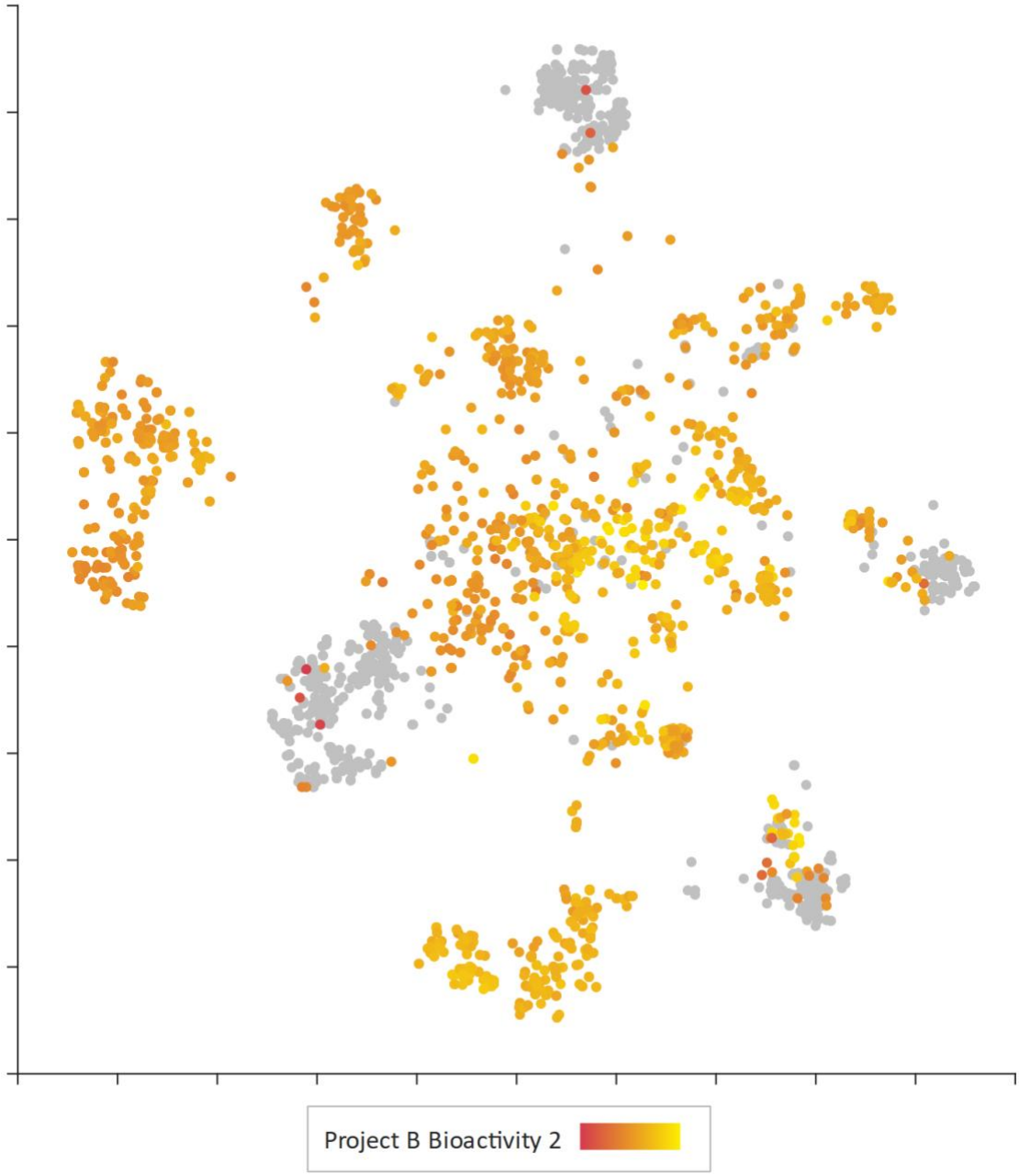

Figure S39. Uncertainty Uncertainty distribution in chemical space for predictions of Project B Bioactivity 2.

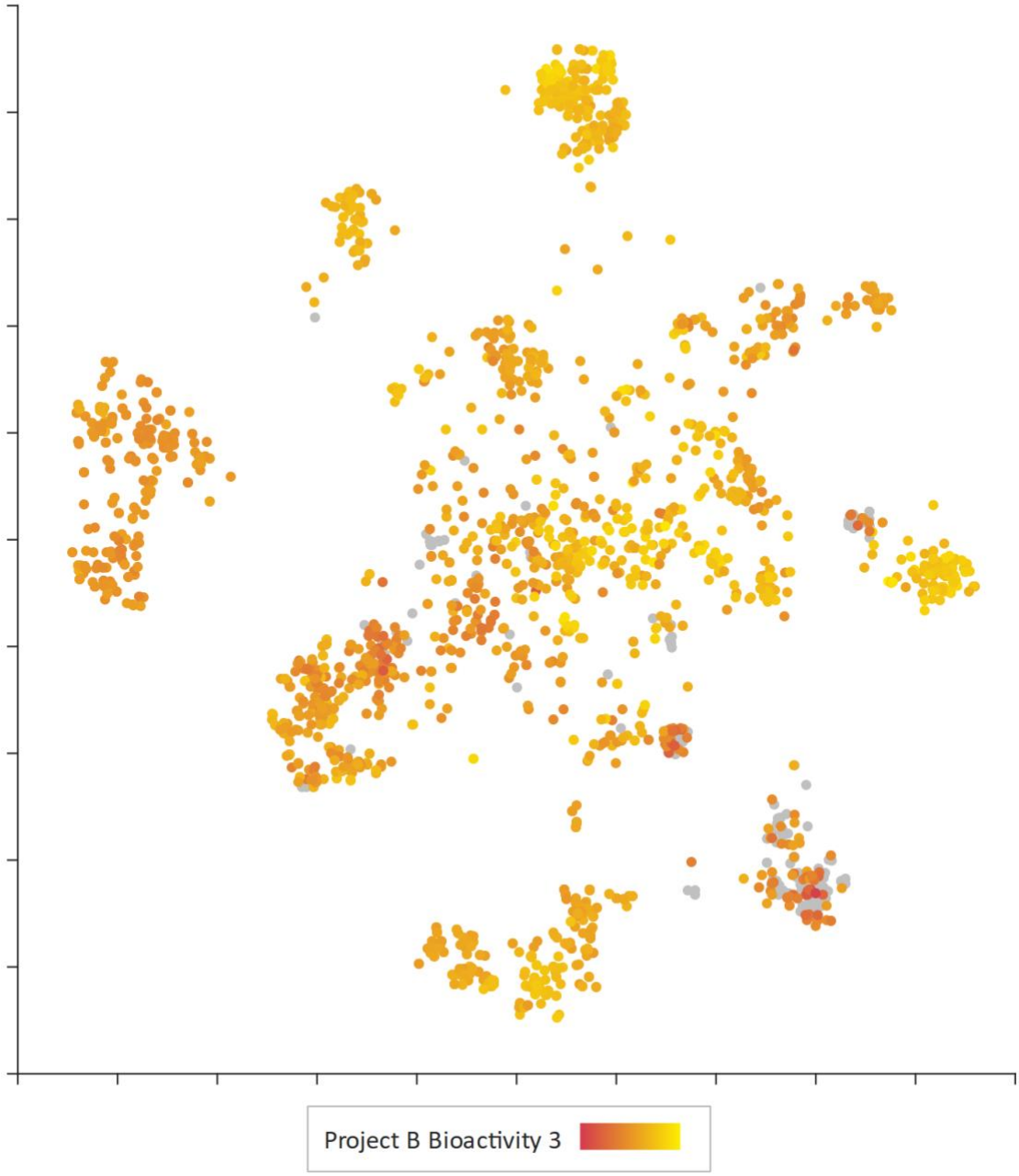

Figure S40. Uncertainty distribution in chemical space for predictions of Project B Bioactivity 3.

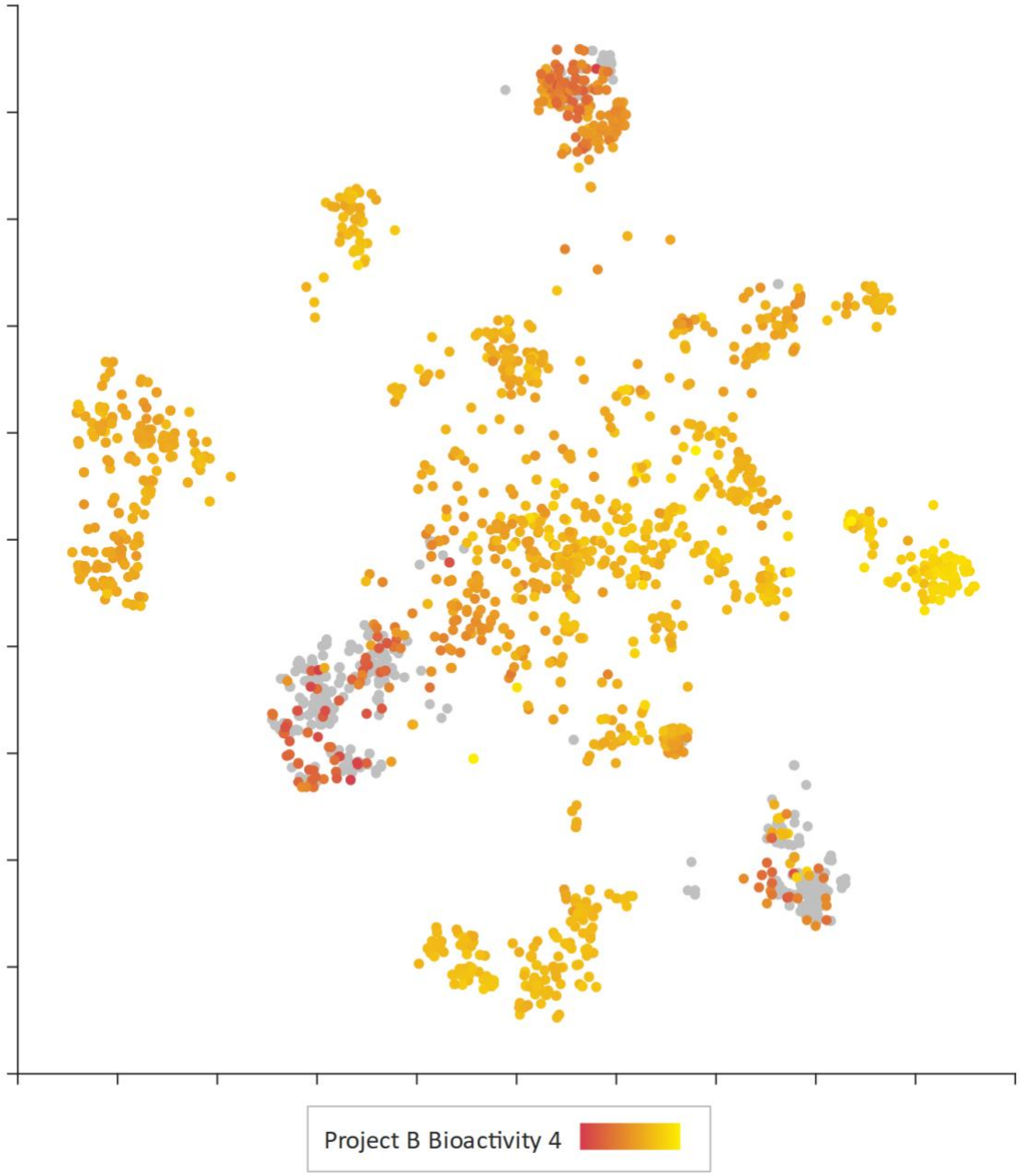

Figure S41. Uncertainty distribution in chemical space for predictions of Project B Bioactivity 4.

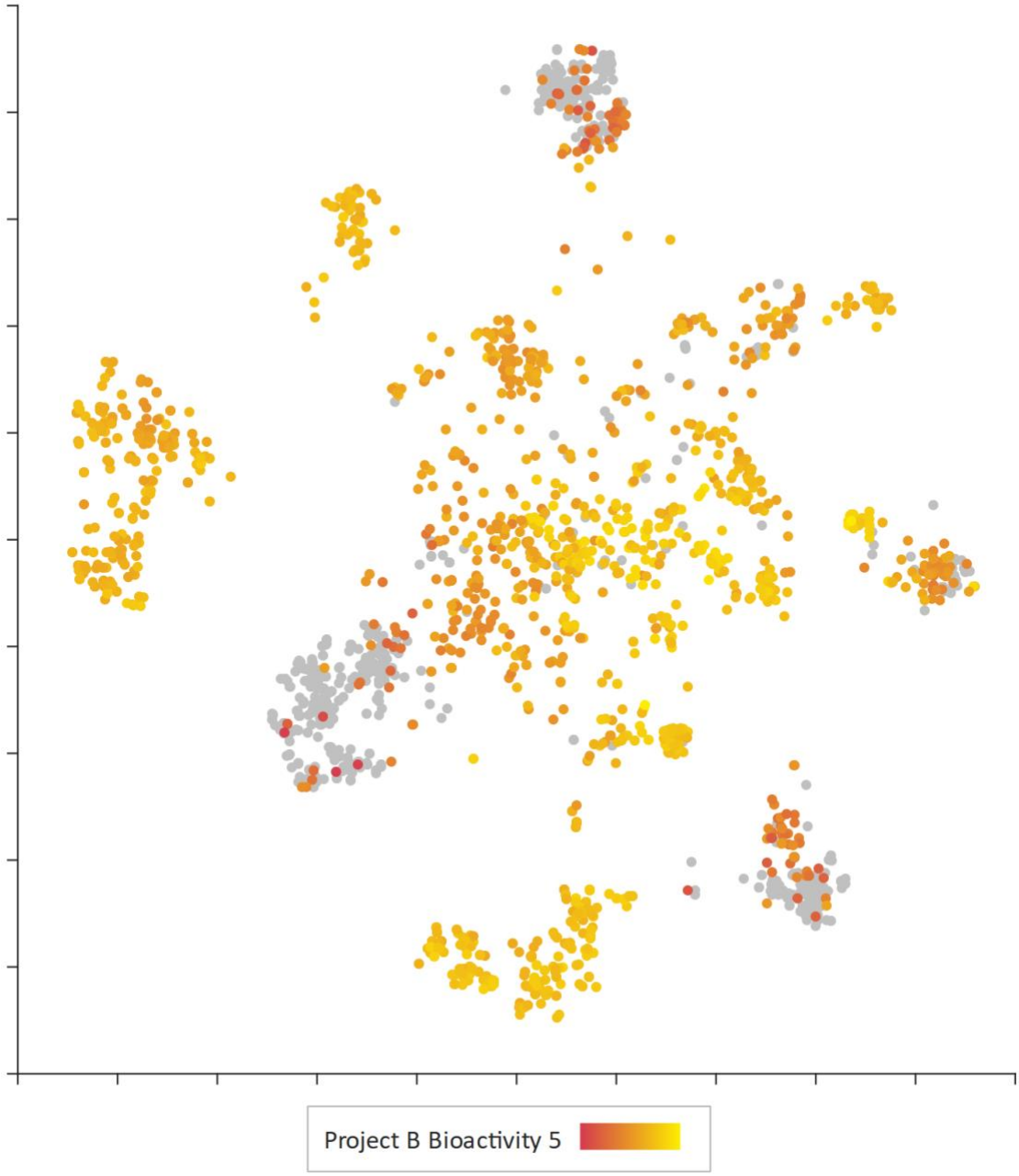

Figure S42. Uncertainty distribution in chemical space for predictions of Project B Bioactivity 5.

Project A and B ADME Endpoints

Figures S43 to S50 show the distributions of uncertainties for the imputed ADME properties. CYP inhibition is less localised than activity but does seem to be more confidently imputed within strongly defined series/clusters. Kinetic solubility does not strongly distinguish between clusters. This suggests the error bars are all of a similar size. Intrinsic clearance models are similar and do not favour any of Project A and B in particular.

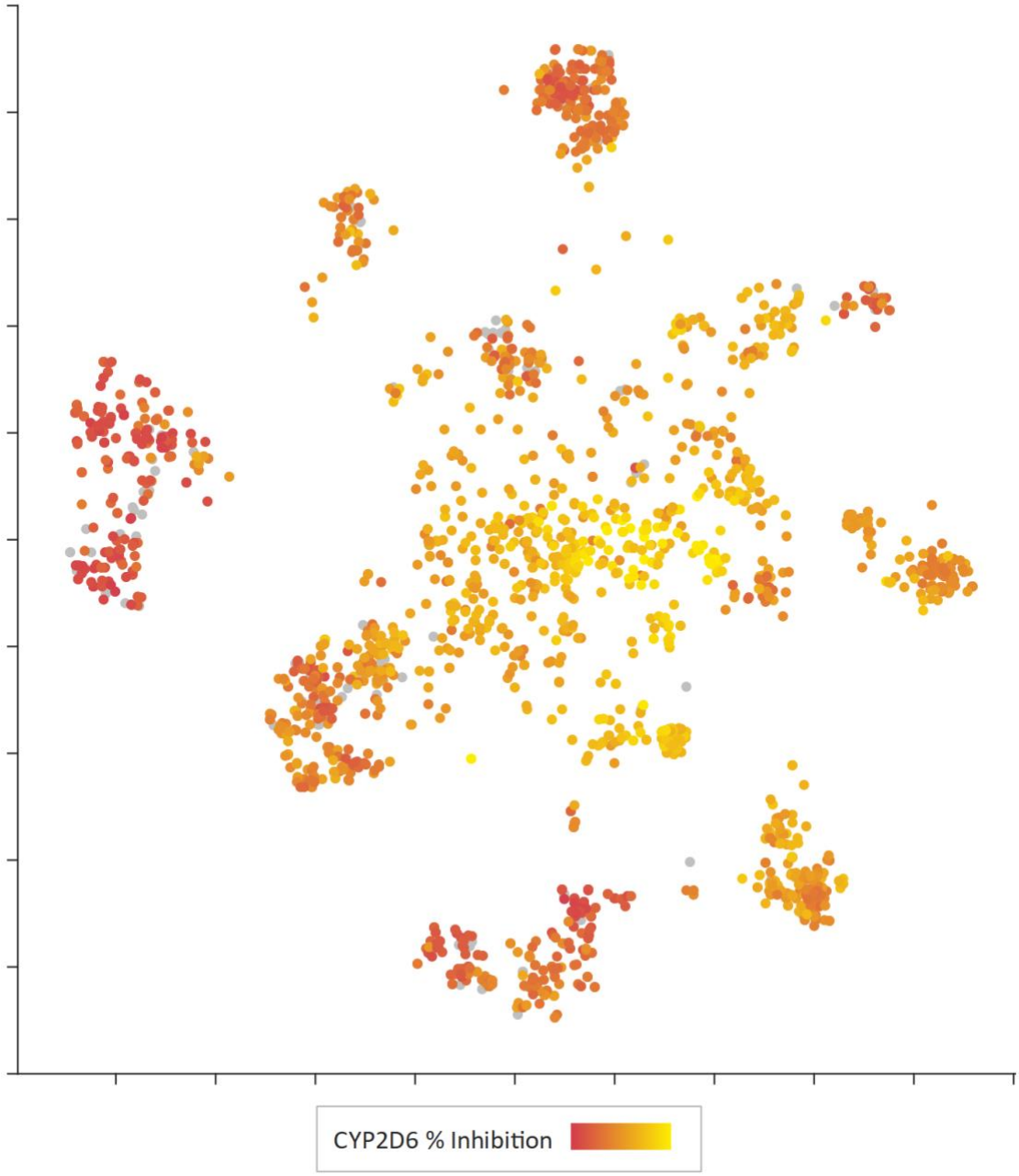

Figure S43. Uncertainty distribution in chemical space for predictions of CYP2D6 Inhibition.

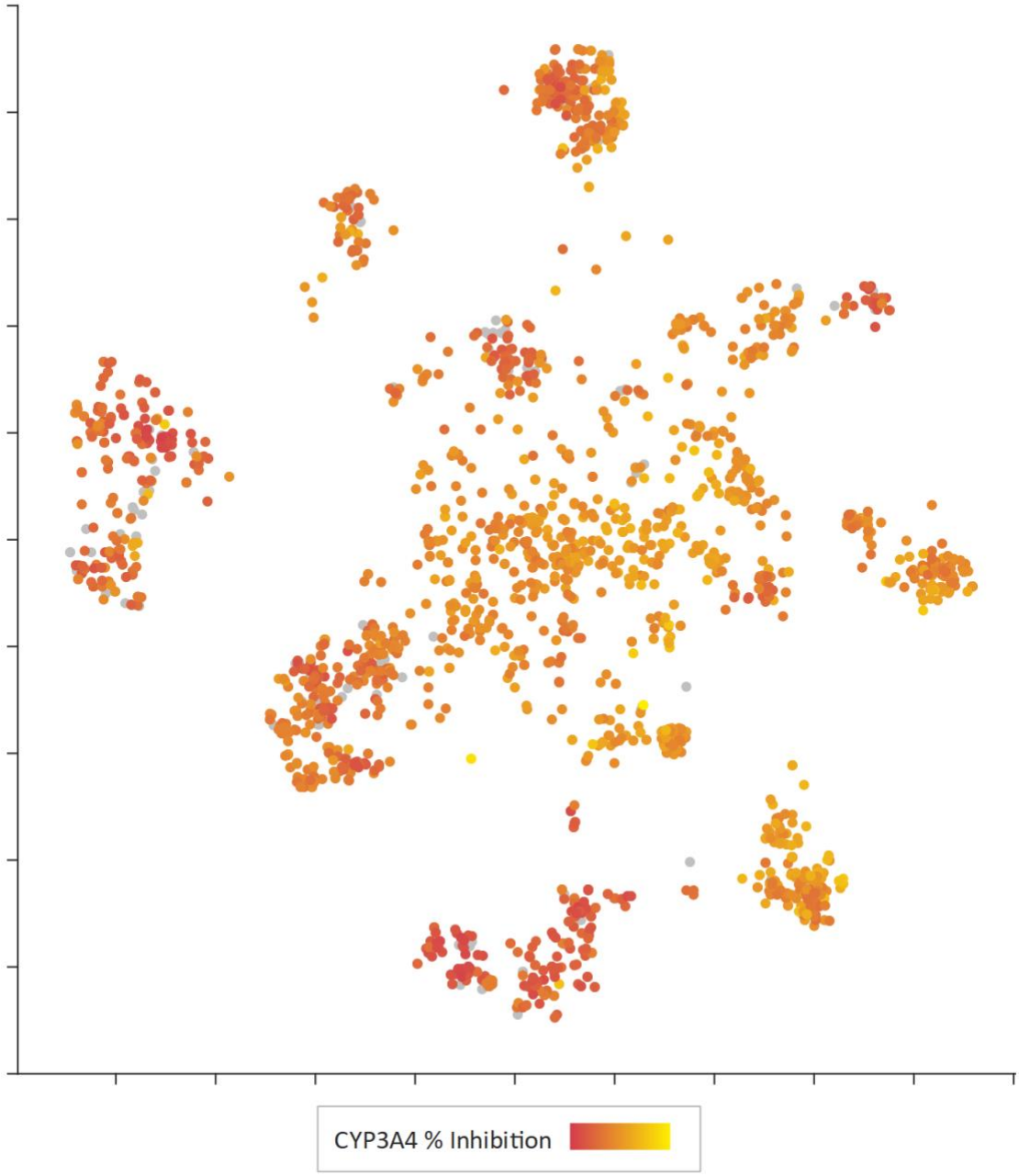

Figure S44. Uncertainty distribution in chemical space for predictions of CYP3A4 Inhibition.

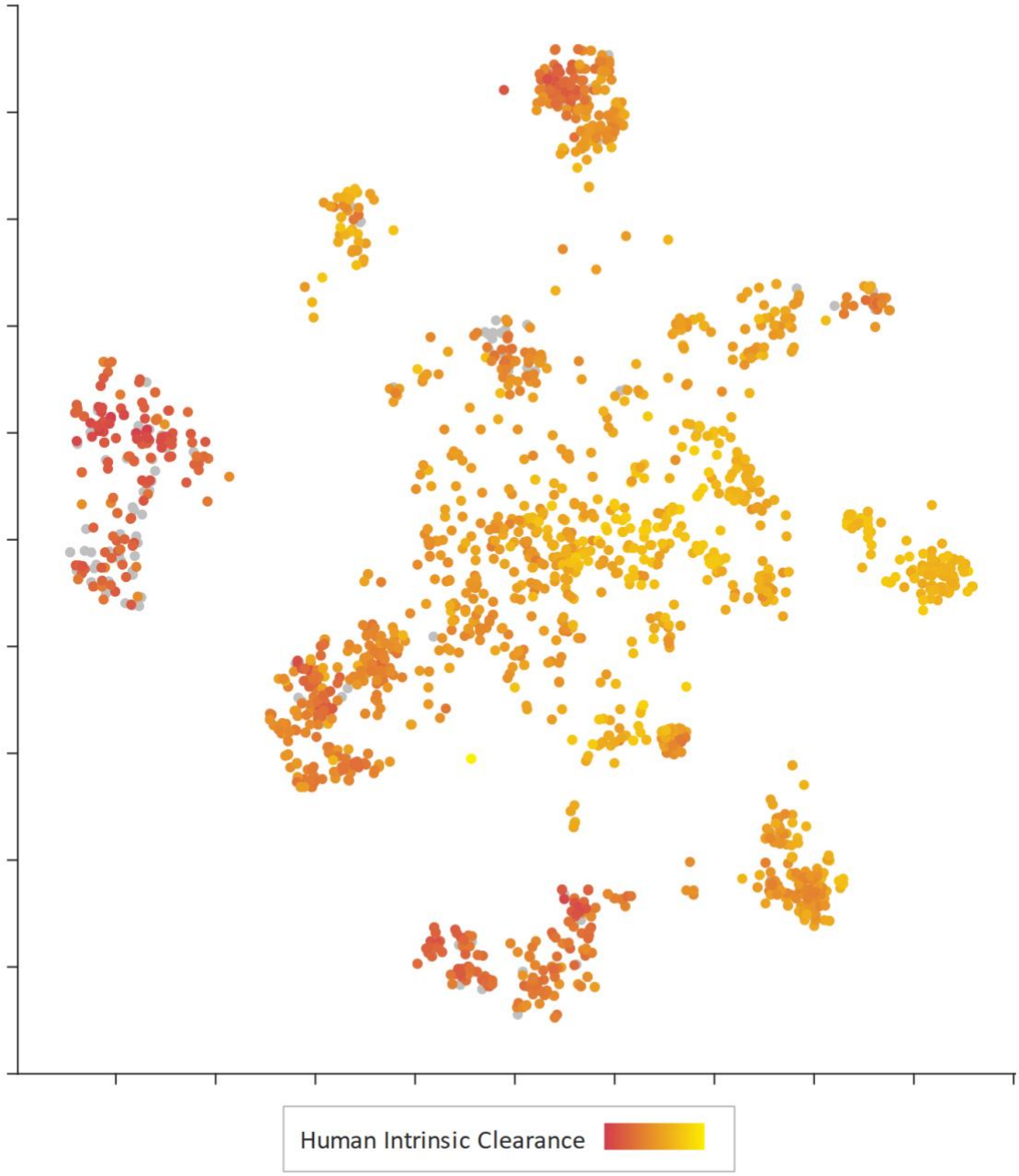

Figure S45. Uncertainty distribution in chemical space for predictions of Human Intrinsic Clearance.

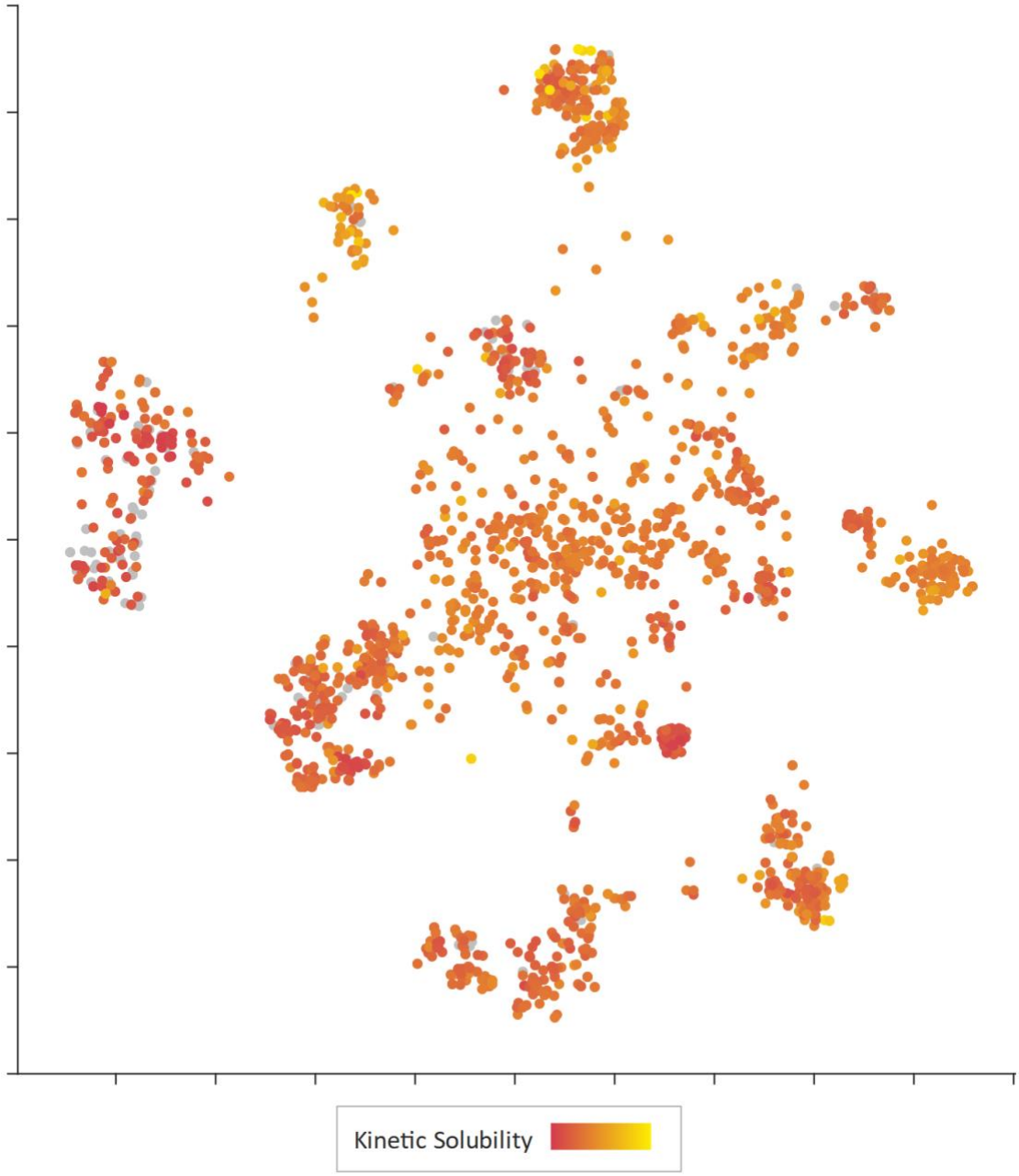

Figure S46. Uncertainty distribution in chemical space for predictions of Kinetic Solubility.

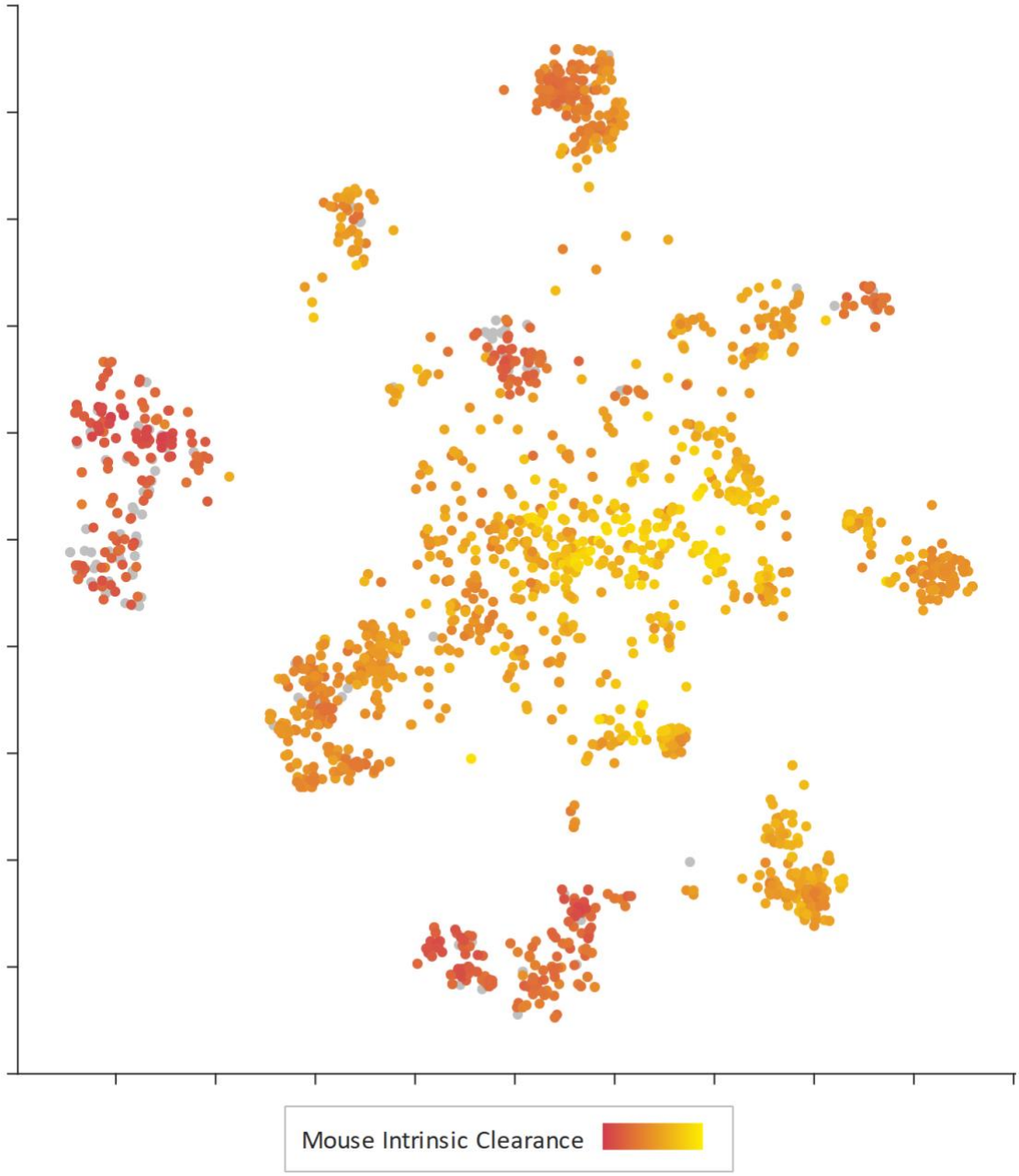

Figure S47. Uncertainty distribution in chemical space for predictions of Mouse Intrinsic Clearance.

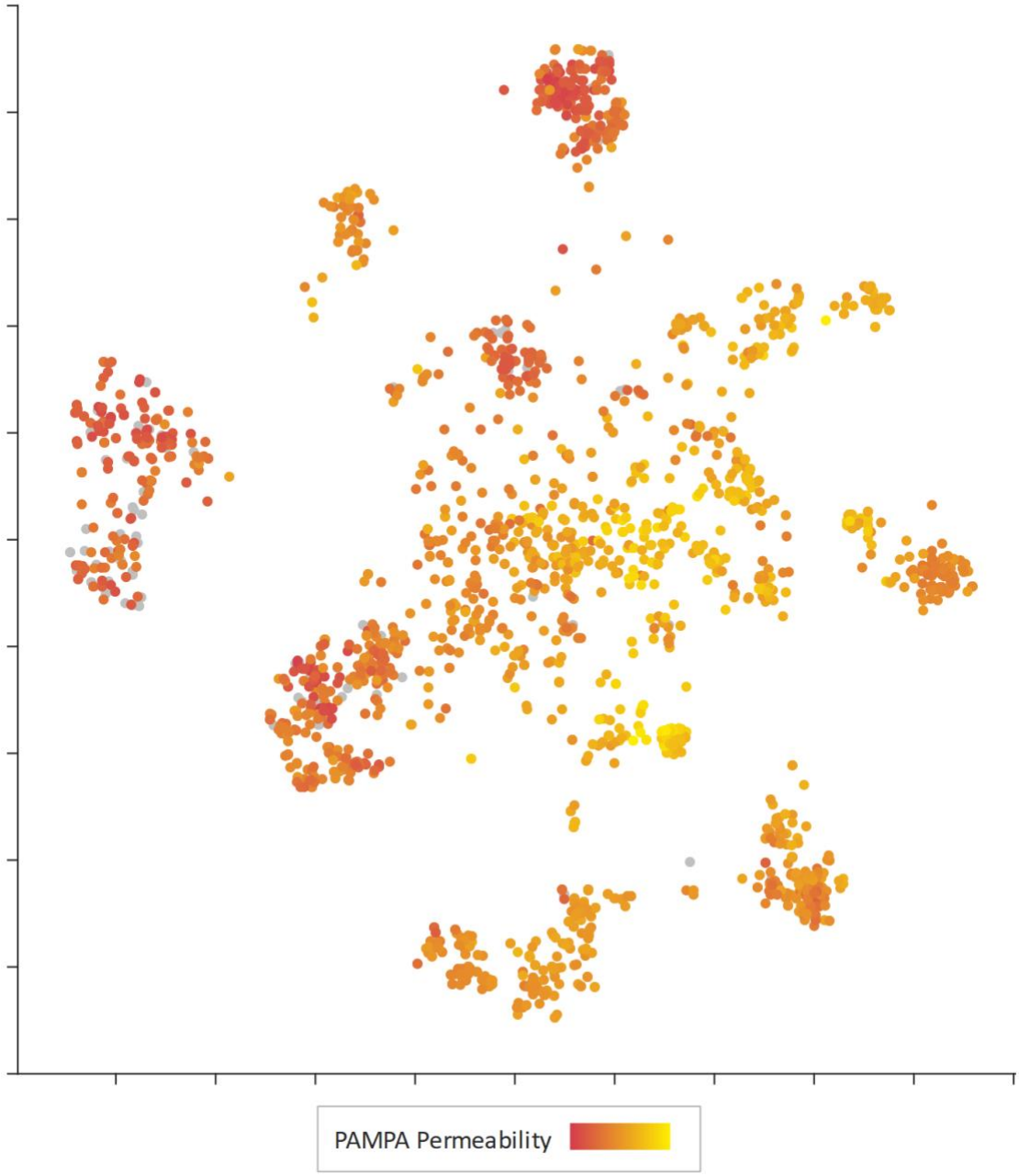

Figure S48. Uncertainty distribution in chemical space for predictions of PAMPA Permeability.

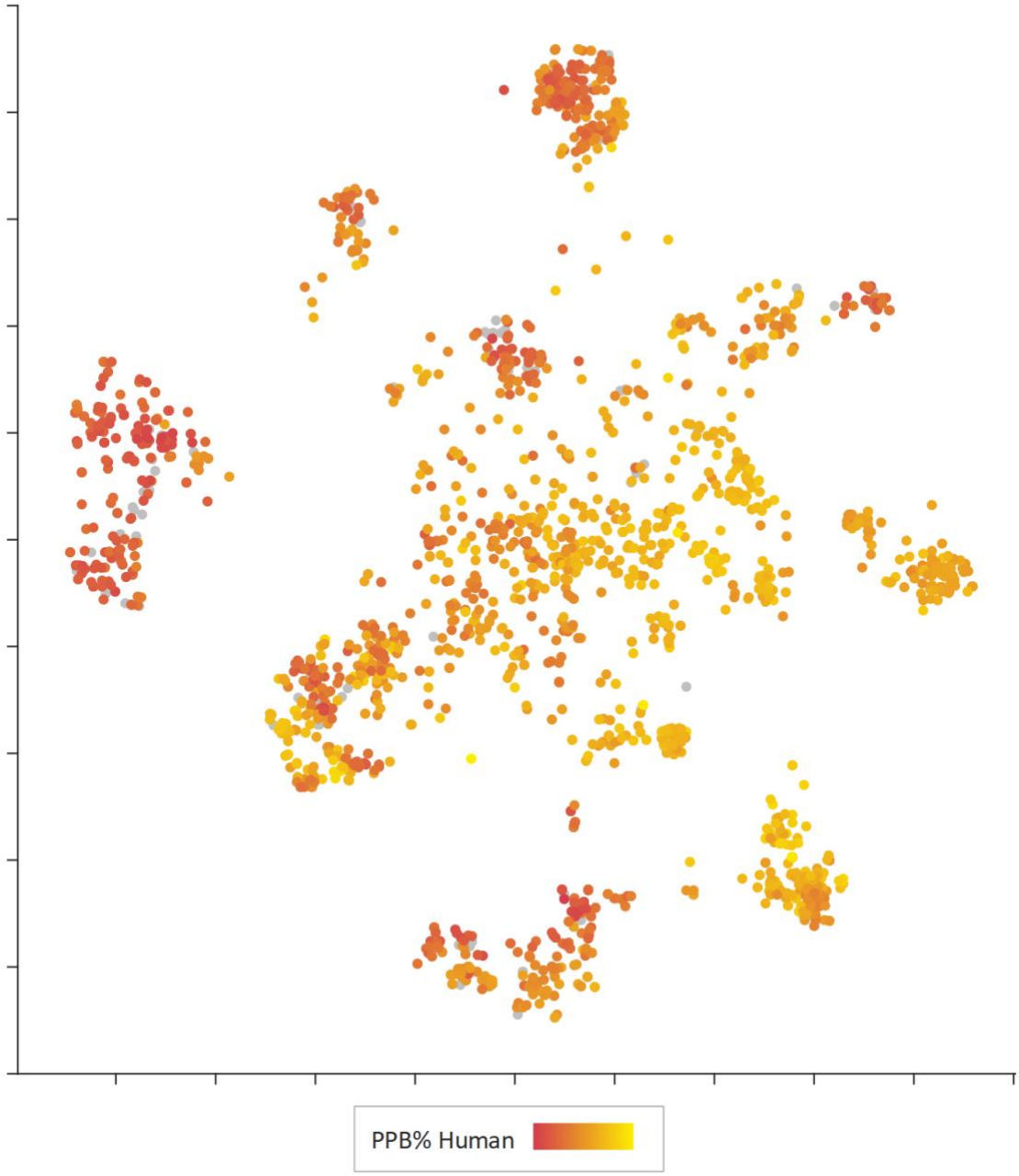

Figure S49. Uncertainty distribution in chemical space for predictions of Human Plasma Protein Binding.

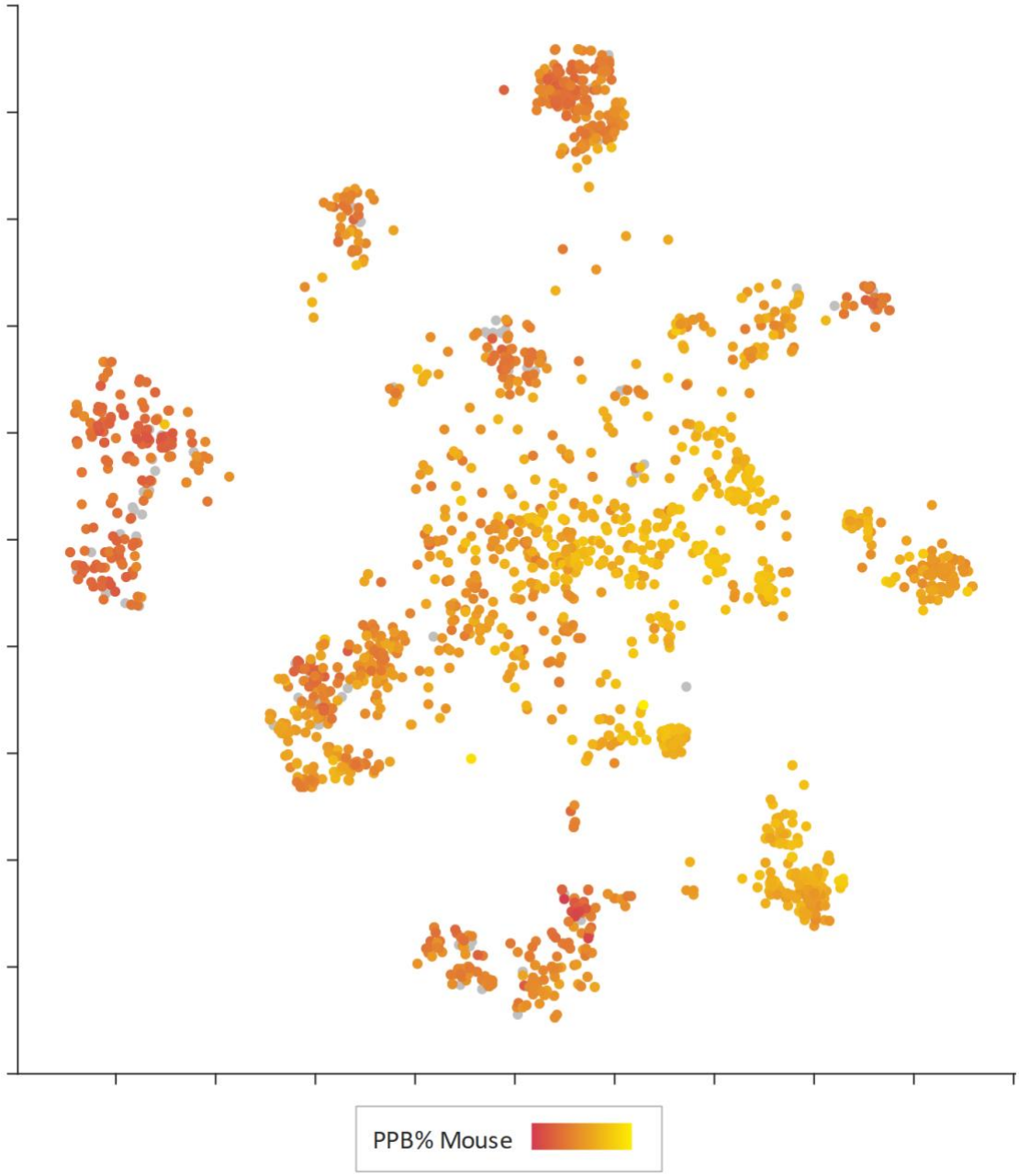

Figure S50. Uncertainty distribution in chemical space for predictions of Mouse Plasma Protein Binding.

Discussion on Random Insertion of Data and Controls

In the process of review there were some very interesting questions: One for example is "What would happen if there was a column of completely random data inserted as a challenge?"

Due to the nature of the SMARTS matching patterns in the StarDrop descriptors any one of the individual descriptors is likely to be essentially random (do not strongly correlate) compared to certain endpoints. Only in combination do they describe complex patterns. There are a few cases to consider for the insertion of random data:

1) If a **full** column of random noise were added, this could be considered as including an additional irrelevant chemical descriptor to the descriptor matrix. We have found the model to be robust to this because, for each endpoint, we select N features which are most relevant to the endpoint for the final model. N is a hyperparameter here. Thus, if a noisy column is added, this will almost certainly be disregarded for meaningful dataset sizes.

2) In terms of model predictions for the noisy column, the R^2 metric compares against the case of random data (with fixed variance). So, we would expect models of $R^2 = 0$ for random columns. If the columns is genuinely filled with noise, it will be impossible for **any algorithm** to predict a held out test set, thus on test data, Alchemite will certainly reproduce a model with $R^2 = \pm\varepsilon$ for some rapidly shrinking ε with large number of test samples.

3) In terms of partially filled random columns (i.e. simulated assays), if such a random sparse column was **co-filled with an experimental column,** such that there were M pairs of data between the two columns, depending on the underlying distribution of the data to achieve a fixed variance one can derive a p-test type quantity to measure confidence in a genuine M-pair correlation over random noise with the same fixed variance; i.e. 3 co-occurring data points are likely to spuriously correlate, but in the limit of large N, this will quickly be ruled out, thus the column will not be selected by the mechanism in case 1).

4) *Pathological columns* of course one could probably engineer some kind of disruptive column. I think it would be an interesting case for adversarial learning to design a random distribution from which column elements are drawn which is maximally misleading for example.

Discussion on Breakdown of Imputation

Another interesting question was "Can one tell when the underlying imputation breaks down?"

Within the bounds of this particular study a useful indicator for telling whether the imputation breaks down is that the error bars will grow rapidly as imputed values are assigned to unknown portions of chemical space. As quoted in the main text:

"Each missing element of the data matrix has an ensemble of predictions filled in from Alchemite and the distribution of this ensemble can take many shapes. The mean and the standard deviation of this distribution gives a unique prediction and error bar for each missing value, where the error bar represents one standard deviation about the mean. In the case where descriptor values far from the training data are provided, the error bar will grow to show the algorithms has no knowledge of that region of chemical space."

To give a preview of a more mathematical formalism, which oversteps this study, one needs to consider that imputation is affected by a few different things:

1) The sparsity of the dataset, (we can try to dictate this with two sparsity parameters $\lambda_1, \lambda_2 \in [0,1]$)

- In the limit that there is no experimental data at all $\lambda_1 = \lambda_2 = 0$.
- If the experimental data only appear as rows (horizontal stripes in the data matrix) $\lambda_1 = 1, \lambda_2 = 0$, then we have full understanding of the cross correlation between assays, but there are now two types of

compounds, virtual compounds with no experimental data (revert to QSAR model), and compounds with full experimental data (no need to impute, however models can be well validated)

- If the experimental data only appear as full columns $\lambda_1 = 0, \lambda_2 = 1$, then these columns are essentially descriptors, and empty columns are modellable.
- Finally, a full block of experimental data $\lambda_1 = \lambda_2 = 1$.
- If $\lambda_1 = \lambda_2 \in [0,1]$, for example a value of 0.5, this is a balanced sparsity, i.e. a uniform random chance that any particular cell is present (or missing), otherwise there is a column/row asymmetry.

So, there will be an underlying percolation or connectivity that is required, based on the granularity or sparsity of present cells, **which is necessary but not sufficient for a good imputation**.

2) The quality of the correlations in the dataset:

- Given the required overlap, as mentioned above, then one needs enough dynamic range in each column to extract a correlation which is useful.
- One also requires enough co-present points that reflect this range, so even if the range of each column is well represented, if the range of the overlap is small, then it will be hard to develop a correlation beyond a simple intersection point.
- If two endpoints only correlate partially (for example activities at different concentrations), then there will be a useable region and an unusable region (consider the case of overlapping functions $\tanh(x)$ and x. This can be captured by a non-local network, but not a linear correlation analysis for example, which would make wrong predictions in the flattened ends of the sigmoidal curve.

3) Is there enough information to recreate the imputed signal accurately

- We might be missing something essential for certain endpoints.
- There will always then be a portion of the signal that cannot be recreated as with other modelling techniques.

Detailed Comparison Between QSAR Models and Alchemite

For the Alchemite model built using the initial data, we used 4 QSAR methods and quoted the best of the four for each endpoint for comparison with Alchemite. As these methods each have their own strengths and weaknesses, it is useful to show a breakdown of each result and this is presented in table S2.

Table S2. Table of coefficient of determination values for each of the four types of QSAR model and Alchemite on each of the activity and ADME endpoints from both projects. Here PLS is partial least squares, RBF is radial basis functions, RF is random forest and GP is Gaussian process. The average and best of the QSAR scores is also reported.

Endpoint	QSAR BEST	QSAR AVERAGE	PLS	RBF	RF	GP	Alchemite
CYP2D6	0.4	0.28	0.08	0.37	0.26	0.4	0.63
CYP3A4	0.26	0.22	0.15	0.24	0.26	0.21	0.3
HLM	0.11	-0.02	-0.08	0.07	0.11	-0.18	0.43
MLM	0.51	0.41	0.31	0.51	0.34	0.49	0.54
KinSol	0.54	0.48	0.4	0.54	0.44	0.54	0.5
PAMPA	0.28	0.22	0.19	0.18	0.24	0.28	0.21

Human PPB	0.6	0.56	0.48	0.56	0.6	0.58	0.72
Mouse PPB	0.56	0.51	0.56	0.49	0.47	0.53	0.63
A Act. 1	0.53	-7.71	0.19	-31.8	0.53	0.23	0.94
A Act. 2	0.67	0.63	0.64	0.56	0.63	0.67	0.79
A Act. 3	0.54	0.44	0.54	0.25	0.5	0.46	0.92
A Cell 1	0.73	0.695	0.73	0.72	0.62	0.71	0.84
A Cell 2	-0.27	-0.565	-0.27	-1.22	-0.29	-0.48	0.57
B Act. 1	0.44	0.3875	0.3	0.43	0.44	0.38	0.65
B Act. 2	0.52	0.415	0.28	0.52	0.46	0.4	0.82
B Act. 3	0.53	0.4475	0.37	0.45	0.53	0.44	0.82
B Act. 4	0.46	0.41	0.3	0.44	0.46	0.44	0.62
B Act. 5	0.57	0.5325	0.47	0.57	0.56	0.53	0.71

Sensitivity to Actives

Some QSAR models are used to predict which analogs to make in the next round of synthesis rather than virtual screening. In this limit resolution between actives is important. We can analyse the model results to this end. An example for Project B Bioactivity 2 is shown in Figure S51. Here for a definition of a cutoff between active and inactive we can see the RMSE of compound predictions for the `actives' (blue) and `inactives' (orange). The RMSE for active predictions is low. The dotted lines around each curve show the standard error from the RMSE estimator. The active curve is also confidently low. Only for very active i.e. pIC_{50} >7 does the RMSE value start to rise as we enter the extrapolative regime.

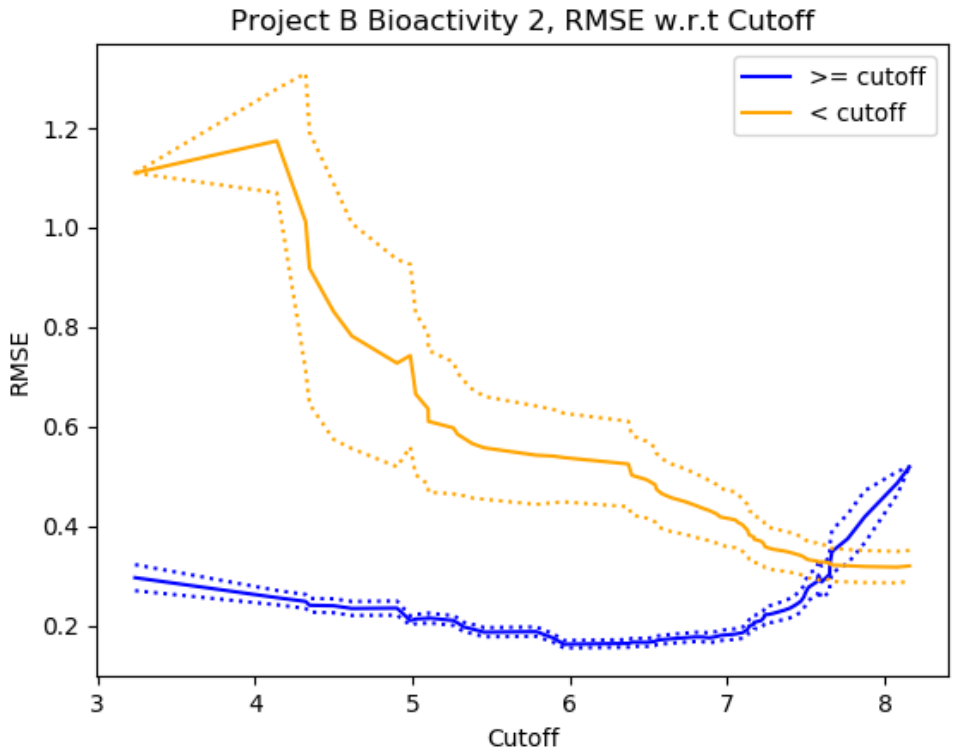

Figure S51. RMSE of all predictions to the left and right of a `cutoff' for activity for Project B Bioactivity 2. The RMSE of actives is low for almost all definitions of cutoff. The dotted lines indicated the standard error in the RMSE.

References

(1) Whitehead, T. M.; Irwin, B. W. J.; Hunt, P.; Segall, M. D.; Conduit, G. J. Imputation of Assay Bioactivity Data Using Deep Learning. *J. Chem. Inf. Model.* **2019**, *59* (3), 1197–1204. https://doi.org/10.1021/acs.jcim.8b00768.

(2) Verpoort, P. C.; MacDonald, P.; Conduit, G. J. Materials Data Validation and Imputation with an Artificial Neural Network. *Comput. Mater. Sci.* **2018**, *147*, 176–185. https://doi.org/10.1016/j.commatsci.2018.02.002.

(3) Mclachlan, G.; Krishnan, T. The EM Algorithm and Extensions , 2nd Edition. **2008**.

(4) van der Maaten, L.; Hinton, G. Visualizing Data Using {t-SNE}. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

(5) StarDrop™. (accessed 16/12/2019)