Abstract

Motivation: Recent large-scale omics initiatives have catalogued the somatic alterations of cancer cell line panels along with their pharmacological response to hundreds of compounds. In this study, we have explored these data to advance computational approaches that enable more effective and targeted use of current and future anticancer therapeutics.

Results: We modelled the 50% growth inhibition bioassay end-point (GI50) of 17,142 compounds screened against 59 cancer cell lines from the NCI60 panel (941,831 data-points, matrix 93.08% complete) by integrating the chemical and biological (cell line) information. We determine that the protein, gene transcript and miRNA abundance provide the highest predictive signal when modelling the GI50 endpoint, which significantly outperformed the DNA copy-number variation or exome sequencing data (Tukey’s Honestly Significant Difference, \( P < 0.05 \)). We demonstrate that, within the limits of the data, our approach exhibits the ability to both interpolate and extrapolate compound bioactivities to new cell lines and tissues and, although to a lesser extent, to dissimilar compounds. Moreover, our approach outperforms previous models generated on the GDSC dataset. Finally, we determine that in the cases investigated in more detail, the predicted drug-pathway associations and growth inhibition patterns are mostly consistent with the experimental data, which also suggests the possibility of identifying genomic markers of drug sensitivity for novel compounds on novel cell lines.

Contact: terez@pasteur.fr; ab454@ac.cam.uk

Supplementary information: Supplementary data are available at Bioinformatics online.
genetic makeup of a cell on a large scale. The US National Cancer Institute (NCI) pioneered these efforts by assembling the NCI60 tumour cell line panel, which, to date, has been assayed for its sensitivity to over 130,000 compounds and has been extensively profiled at the biological level (Shoemaker, 2006). Although these cell line collections have proved valuable to identify genomic markers of drug sensitivity (Barretina et al., 2012; Garnett et al., 2012) and even led to the development of new drugs (Adams and Kauffman, 2004), the question now arises of how these pharmacogenomic dataset can be meaningfully mined, both to discover cancer subtype-specific drugs, and ultimately to aid in the design of personalized cancer treatments.

Previous studies of the NCI60 panel include the identification of drug mechanism of action (MoA) (Weinistein et al., 1992), visualization tools for drug sensitivity data (Paull et al., 1989; Weinstein et al., 1997) and drug sensitivity predictions based on the cell line profiling data (Kutalik et al., 2008; Riddick et al., 2011; Staunton et al., 2001; Szakacs et al., 2004). Beyond algorithmic differences, the conceptual limitation shared by these models was the unfeasibility of extrapolating the data to novel compounds and cell lines simultaneously, as the cell line profiling data and chemical information were separately used as predictive features and were not integrated into a single machine learning model. To overcome this limitation, two recent studies have pioneered the combination of drug and cell line information [gene expression, gene copy number variations (CNVs) and mutation profiles] on the data from the Genomics of Drug Sensitivity in Cancer project (GDSC) for drug sensitivity prediction (Garnett et al., 2012). The first study (Menden et al., 2013) modelled the sensitivity of 608 cell lines to 131 drugs with Neural Networks and Random Forests, thereby obtaining a coefficient of determination ($R^2$) of 0.64 on an external set. The second study (Ammad-ud-din et al., 2014) applied kernelized Bayesian matrix factorization to model the sensitivity of 650 cell lines to 116 drugs, and obtained an $R^2$ value of 0.32 when predicting the activity for new drugs on a set of cell lines in the training set. The authors showed that the combination of chemical and cell line information indeed improved model performance, which permitted the extrapolation of drug activities to complete the missing entries in the bioactivity matrix of 482 cell lines by 116 drugs.

In this study, we now propose the simultaneous modelling of chemical and cell line information in a single machine learning model to predict the 50% growth inhibition bioassay end-point ($GI_{50}$) of 17,142 compounds screened against 59 cancer cell lines from the NCI60 panel. Conformal prediction was implemented to predict the confidence intervals (CI) for individual predictions. The integration of these different, yet complementary, streams of information is often termed proteochemometrics or pharmacogenomic (PGM) modelling (Cortes-Ciriano et al., 2015b; Wheeler et al., 2013). In typical PGM models, each compound–cell line interaction is numerically encoded by the concatenation of the compound, $\Phi_c(c, d)$, and cell line, $\Phi_d(c, l)$, descriptors into a single vector, namely $\Phi_{cld}(c, d, l)$. Although this encoding approach is widely used, other approaches exist (Su et al., 2010; Yamanishi et al., 2010). In fact, in the case of kernel methods, each pair $\Phi_{cld}(c, d, l)$ is encoded by the tensor product of $\Phi_c(c, d)$ and $\Phi_d(c, d)$ (Jacob and Vert, 2008). Subsequently, these descriptors are related in a single machine-learning model to the specific biological readout of interest (Cortes-Ciriano et al., 2015b; van Westen et al., 2011). In practical terms, PGM thereby helps us to understand complex relationships between the compound structures and cell line features and enables the estimation of the bioactivity for (novel) compounds on (novel) cell lines (Fig. 1a).

We curated cell line profiling data consisting of 59 cell lines that were characterized using various genetic and proteomic profiling methods. From this, we assembled 14 profiling datasets, denoted here as dataset views. We benchmarked their predictive signal, and demonstrated that the simultaneous modelling of the compound and cell line information improves the prediction of compound potency across the NCI60 panel compared with models trained on only compound information. Our modelling approach permits us to interpolate and extrapolate compound bioactivities to novel cell lines and tissues in the NCI60 panel, and, to a lesser extent, to chemically dissimilar compounds. Finally, we demonstrate that the predicted bioactivities can be used to estimate growth inhibition patterns across the NCI60 panel and that the predicted drug-pathway associations are consistent with the experimental data published in the literature.

2 Methods

2.1 Datasets
The raw $pGI_{50}$ values (−log$_10$ $GI_{50}$, $M$) for 17,142 compounds were downloaded from CellMiner (database version 1.4) (Reinhold et al., 2012). The mean $pGI_{50}$ values were calculated when several measurements were available for the same compound–cell line combination and the standard deviation of the replicates was considered as the experimental uncertainty of the measurements. The representation of the chemical structures (with respect to the aromaticity or stereochemistry) was normalized with the StandardiseMolecules function of the R package camb using the default values (Murrell et al., 2015). The final dataset consisted of 17,142 compounds versus 59 cell lines (the NCI60 panel, except for ME.MDA_N) and a total of 941,831 data points, which corresponds to a matrix completeness of 93.08%. Table 1 summarizes the details of the profiling datasets used to describe the cell lines.

2.2 Compound descriptors
The compounds were described with circular Morgan fingerprints in count format, which encode the compound structures via radial atom neighbourhoods (Bender et al., 2004). The fingerprints were calculated with the R package camb (Murrell et al., 2015). The size of the fingerprints was set to 256 bits, and the maximum radius of the substructures considered was set to 2. We used the circular fingerprints because they have provided high retrieval rates in comparable studies (Bender et al., 2009; Koutsoukas et al., 2013).

2.3 Compound clustering
The compounds were clustered using periodic two-dimensional Self-Organizing Maps (SOMs) (Bouvier et al., 2014). Two dimensions, here 50 × 50, determine the map size in a periodic manner, whereas the third dimension contained the compound fingerprints. Each vector along the third dimension is called a neuron, $\nu$. The same fingerprints used to train the PGM models served as input vectors to the SOMs. The SOM values were initialized from a uniform distribution spanning the values present in the input vectors. At each training step, the most similar neuron to the input vector considered was updated. The conventional Unified distance matrix (U-matrix) was calculated to delineate the clusters. The U-matrix value associated with a given neuron, $U_{\text{height}(\nu)}$, is defined as the average Euclidean distance between that neuron and its eight closest neighbours: $U_{\text{height}(\nu)} = \frac{1}{8} \sum_{\phi(v) \in N(\nu)} E_{\phi}(\nu, \mu)$, where $N(\nu)$ is the set of neighbours and $E_{\phi}$ is the Euclidean distance between neurons. A distance threshold value was then applied to the U-matrix to define the contours of the compound clusters (Cortes-Ciriano et al., 2015c).
2.4 Model training

Random Forest (RF) models (Breiman, 2001) were trained with the ensemble.RandomForestRegressor module of the python library scikit-learn (Pedregosa et al., 2011) using the following parameters: (i) number of trees in the forest: 100 (Sheridan, 2013); (ii) criterion to assess the quality of a split: mean squared error; (iii) minimum number of data points to split a node: 1; (iv) minimum number of data points in a leaf to keep a given node split: 1; (v) maximum number of randomly selected descriptors considered when splitting a node (mtry): dimensionality of the input space. Radial-kernel Support Vector Machine (SVM) models were trained using the svm module of the python library scikit-learn (Pedregosa et al., 2011).
The parameter values were optimized with a 10-fold cross-validation (CV) and a grid search. The parameter grid was composed of the following values: (i) $C \in \{2^{-8}, 2^{-6}, \ldots, 2^2\}, 10^3, 10^2, 10^1\}$ and (ii) $\sigma \in \{2^{-8}, 2^{-6}, \ldots, 2^2\}$. All calculations were performed on 16 Intel Xeon E5-2670 processors and a total memory of 256 GB. The training times on the complete dataset ranged between 6 and 8 h.

2.5 Learning strategies

2.5.1 PGM models

In the following, we define ‘PGM models’ as those models simultaneously trained on (i) compound descriptors and on (ii) one cell line profiling data set view.

- Assessment of the interpolation power with a 10-fold CV (10-fold CV PGM model): a given dataset was randomly divided into (i) a training set comprising 90% of the data points and (ii) a test set comprising the remaining 10% of the data (Fig. 1b). This process was repeated 10 times, each time leaving out a random subset of the data, which enabled the generation of predicted values for all data points. Thus, although the test set is comprised of compounds and cell lines that are also present in the training data, the training and test sets are comprised of different compound–cell line pairs. The interpolation power of these 10 models was evaluated on a per cell line and on a per compound cluster basis. To this end, the RMSE and $R^2$ values were calculated on subsets of the test set grouped by cell line (cell line–averaged performance) or by the compound cluster (compound cluster–averaged performance).

- Leave-One-Compound-Cluster-Out (LOCCO) models (Fig. 1c): all data points annotated on a given chemical cluster were left out as a test set; a PGM model was trained on the training set, and the values for the test set were predicted. These steps were repeated for each compound cluster. This scenario reflects the practical situation where a model predicts pIC$_{50}$ values for chemically novel compounds, and thus permits the assessment of the extrapolation capabilities of a PGM model in chemical space.

- Leave-One-Cell-Line-Out (LOCO) models (Fig. 1d): this validation scenario is similar to LOCCO, except that the data left out for the test set are now all of the data obtained for a particular cell line. This scenario intends to mimic the situation where the PGM model is employed to extrapolate the data to novel cell lines.

- Leave-One-Tissue-Out (LOTO) models: The performance of the PGM models was further evaluated by extrapolating to cell lines whose tissue of origin was not present in the training set. This scheme is similar to LOCO, except for the fact that all of the cancer cell lines originating from the same tissue were left out from the training set each time. It was employed because cell lines from the same tissue, despite their genetic differences, often still share commonalities with regard to compound activity.

To quantify the improvement of the PGM models over models exclusively trained on compound fingerprints (compound-only models), the following one-space learning strategies were explored (Brown et al., 2014).

2.5.2 Compound-only (QSAR) models

- Family Quantitative Structure-Activity Relationship (QSAR F): the models were exclusively trained using the compound fingerprints as the input features (Fig. 1c). A QSAR F model serves to assess whether the explicit inclusion of the cell line information improves the prediction of compound activity; if the PGM model displays a performance very similar to the QSAR F model, the inclusion of cell line information would add little benefit to the modelling of compound activity.

- Individual QSAR models per cell line: one QSAR model per cell line was exclusively trained on compound descriptors, which is, therefore, not able to include information from the biological side or measurements against other cell lines. In this case, the comparison was made with the compound cluster-averaged interpolation power of the PGM models to evaluate whether the integration of the compound and cell line information leads to higher predictive ability with respect to the per cell line QSAR models.

2.5.3 Cell line kernels

To assess whether the explicit inclusion of the cell line profiling data as input variables provides a higher predictive signal than the cell line kernels (Jacob et al., 2008), which encode cell lines with a vector quantifying the similarity among cell lines in a given space (e.g. gene expression), we trained models with two sources of information, namely: (i) compound descriptors and (ii) one of the following cell line kernels:

- Dirac kernel (Jacob et al., 2008) or cell line identity fingerprints (CLIFP) (Brown et al., 2014). CLIFP are binary descriptors of length equal to the number of different cell lines considered, where each bit position corresponds to one cell line. Formally, CLIFP are defined as $\text{CLIFP}(i, j) = \delta(i - j)$ ($i, j \in 1, \ldots, N_{\text{cells}}$), where $\delta$ is the Kronecker delta function and $N_{\text{cells}}$ is the number of distinct cell lines. These descriptors are simply cell line indicator variables, and, therefore, no information on compound activity across different cell lines is shared during model training (Jacob et al., 2008).

- Multitask (MLT) kernel (Jacob et al., 2008). This kernel encodes the cell lines in the following form: $\text{MLT}(i, j) = (1 + \delta(i - j)) (i, j \in 1, \ldots, N_{\text{cells}}))/2$. When coupled to SVM, this kernel decomposes the model as a sum of two linear functions. The first function, common to all cell lines, learns the shared patterns of compound activity against all cell lines (e.g. substructures enriched for compound activity). The second function describes the aspects of compounds specific to each cell line. The cell lines are located at an equal distance, and information across cell lines can be shared during the learning phase (though their biological similarities and differences are not explicitly taken into account).

- Cor. Transcriptome kernel. This kernel encodes cell lines by defining a $N_{\text{cells}} \times N_{\text{cells}}$ matrix where each $i, j$ entry corresponds to the Spearman’s Rank Correlation coefficient ($r_s$) of the transcript levels (log$_2$) of 19 965 genes (Reinhold et al., 2010) between cell lines $i$ and $j$.

- Cor. Proteome kernel. This case is similar to the previous kernel, with the difference that it considers the correlation of the protein levels of 8113 distinct proteins (Gholami et al., 2014) for all cell line pairs.

2.6 Model validation

The predictive power of the models was assessed on the test set according to the RMSE$_{\text{test}}$ and $R^2_{\text{test}}$ values (Golbraikh and Tropsha, 2002):

$$R^2_{\text{test}} = 1 - \frac{\sum_{i=1}^{N}(y_i - y'_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2},$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - y'_i)^2}{N}}.$$
where $N$ represents the size of the test set; $y_i$ the observed, $\hat{y}_i$ the predicted, and $\bar{y}$ the average pGI50 values of those data points included in the test set; and $\hat{y}_i^0 = s \bar{y}$, with $s = \sum y_i \bar{y}_i / \sum y_i^2$.

2.7 Assessment of the maximum and minimum achievable model performance

To assess the maximum and the minimum achievable RMSE$_\text{test}$ and $R^2_\text{test}$ values based on the experimental uncertainty of the pGI50 values, we generated the simulated data as follows:

- **Maximum performance.** A sample, $A$, of the same size as the test set was randomly extracted from the vector containing the whole set of pGI50 values. Subsequently, the experimental uncertainty (multiplied by $-1$ in half of the cases) was added to each data point in $A$, thus defining the sample $B$. We define experimental uncertainty as the standard deviation of replicate pGI50 measurements. In cases where no experimental uncertainty was available for a data point, the mean of the available replicate-averaged experimental uncertainties, namely 0.27 pGI50 units, was used. Next, the RMSE$_\text{test}$ and $R^2_\text{test}$ values were calculated for $A$ with respect to $B$. These steps were repeated 1000 times, which led to the distributions of the maximum achievable RMSE$_\text{test}$ and $R^2_\text{test}$ values.

- **Minimum performance.** The procedure was the same as in the previous case, except that sample $A$ was randomized before calculating the RMSE$_\text{test}$ and $R^2_\text{test}$ values.

2.8 Conformal prediction

Conformal prediction builds upon past experience, i.e. the data points in the training set, to define the confidence intervals (CI) for individual predictions, $y_i$, with an error of $1 - \epsilon$, where $\epsilon$ is the tolerated error. The non-conformity score, $S_{\text{new}}$, for a new data point, $x_{\text{new}}$, quantifies how distant it is with respect to the data points from the training set, $X = \{x_i\}_{i=1}^N$. Here, the non-conformity scores were calculated as:

$$S_{\text{new}} = \frac{x_{\text{new}} - \hat{y}_i}{\tilde{\sigma}_i},$$

where $\tilde{\sigma}_i$ is the predicted error for $x_{\text{new}}$ with an error model. The main advantage of conformal prediction is that the CI are always valid, i.e. a confidence of 0.8 means that the predicted confidence regions will contain the observed value in at least 80% of the cases (Norinder et al., 2014).

To calculate and validate the predicted CI, the following protocol (Norinder et al., 2014) was implemented. First, the entire dataset was randomly divided into an external set (20% of the data) and a training set (80%). The latter was subsequently split into a calibration set (30%) and a proper training set (70%). Thus, these three sets include data from the entire sets of cell lines and compounds. Two models were trained on the proper training set, the first of which predicted the pGI50 values (point prediction model), whereas the second predicted the errors in the prediction (error model), $\hat{\sigma}$. Both models were trained with the compound fingerprints and the ‘G.1T 1000 genes’ dataset view as input features using the complete dataset. The point prediction model was generated by training an RF model on the proper training set with a 10-fold CV and with pGI50 values as the dependent variable. The generated CV predictions served to calculate the residuals (i.e. prediction errors) for the data points in the proper training set. Subsequently, the error model was generated by training an RF model on the proper training set using these residuals as the dependent variable. The pGI50 values and prediction errors were subsequently predicted for the calibration set and employed to calculate the vector of non-conformity scores for this set, which, after being sorted in increasing order, led to:

$$z_{\text{calib}} = \{z_{\text{calib}}\}_{i=1}^{N_{\text{calib}}},$$

where $N_{\text{calib}}$ corresponds to the number of data points in the calibration set. The $z$ value for a given confidence level, namely $z_{1-\epsilon}$, was calculated as follows:

$$z_{1-\epsilon} = z_{\text{calib}} \text{ if } \epsilon \equiv N_{\text{calib}} \times (1 - \epsilon),$$

where $\epsilon$ indicates identity. This corresponds to traversing the set $z_{\text{calib}}$ at its $z_{1-\epsilon}$ element, namely $1-\epsilon$. Finally, the prediction errors ($\hat{\sigma}_i$) and the pGI50 values ($\bar{y}$) were predicted for the test set. The CI was calculated as:

$$\bar{y} \pm z_{1-\epsilon} \times \hat{\sigma}_i.$$

2.9 Pathway–drug associations

The average log$_2$ gene transcript level for the genes in each pathway of the MSigSB C2 Canonical Pathways gene set (Liberzon et al., 2011) was defined as the expression level of each pathway. To assess the association between the drug response (pGI50) and the expression of a given pathway, we fitted a linear model considering the tissue type, $T_i$, as a blocking factor (Haibe-Kains et al., 2013), which was defined as:

$$\text{pGI50} = \beta_0 + \beta_1 T_i + \epsilon,$$

where $T_i$ corresponds to the average expression of pathway $i$ in a given cell line, $T_i$, to the tissue of origin of that cell line, and $\epsilon$ to the error term. The $\beta_0$ values determined the strength of the associations, whereas their significance was estimated by the statistical significance of $\beta_0$ (two-sided t-test, $z = 0.05$).

3 Results

3.1 Summary of the cell line profiling dataset views

We collected seven profiling datasets for 59 cell lines from the NCI60 panel, excluding ME-MDA-N due to the lack of gene transcript microarrays (Table 1). These molecular/phenotype datasets were combined in a variety of ways, which we termed dataset views in analogy to database views (Costello et al., 2014). The dataset views may consist of (i) a profiling dataset, (ii) a subset thereof, e.g. gene transcript levels of gene sets or (iii) a modification of the dataset to which prior knowledge is added, e.g. the calculation of pathway expression levels based on knowledge of the cell signalling networks. A total of 14 dataset views were defined, which are summarized in Table 1 and provided in the Supplementary Material.

In addition to the complete dataset, comprising all available data and including 17 142 distinct compounds (Supplementary Table S8) and 941 831 data points, we assembled three additional datasets: (i) the uncorrelated bioactivities 0.5 dataset, comprising 3641 distinct compounds (199 940 data points) whose bioactivity distributions on the 59 cell lines exhibit standard deviations higher than 0.5 pGI50 units (Supplementary Table S9); (ii) the uncorrelated bioactivities 1 dataset, comprising 165 distinct compounds (9 376 data points) whose bioactivity distributions on the 59 cell lines exhibit standard deviations higher than 1 pGI50 unit (Supplementary Table S10); and (iii) the high confidence dataset, exclusively comprising data points averaged over at least two experiments (304 212 data points and 5302 distinct compounds). The uncorrelated bioactivities 0.5 and 1 datasets served to assess the models’ performance on compounds displaying a dynamic range of bioactivities across the cell line panel, while the latter dataset served to evaluate whether the models’ performance improves when using replicate-averaged cell line sensitivity data (Supplementary Table S11 for details).

3.2 Chemical space characterization

To assess the chemical diversity in the data, we clustered the 17 142 compounds with SOMs (Supplementary Figure S1 and Section 2.3), which resulted in the definition of 31 distinct chemical clusters. Several clusters, e.g. 4 and 18, are highly homogeneous, as highlighted by the high inter-neuron similarity (blue areas in Supplementary Fig. S1). In contrast, other clusters are comprised of
more diverse compounds (shown in red in Supplementary Fig. S1). For instance, cluster 2 (Supplementary Table S8) is composed of polycyclic aromatic compounds with diverse halogen substituents and topologies. The definition of structural clusters will be employed later to examine the extrapolation power of the PGM models to chemically dissimilar structures.

3.3 Comparison between RF and SVM

Although the statistical robustness of RF to model the chemical and biological data has been empirically observed (Menden et al., 2013), the mathematical notions by which information is shared across compounds and cell lines are poorly understood. Given that a comparison between the multi-task learning capabilities of RF and SVM in the field of pharmacogenomics is lacking, we first compared their performance in 10-fold CV when trained on the uncorrelated bioactivities 1 dataset using: (i) compound fingerprints and (ii) either gene expression data or a cell line kernels as the input features (Fig. 2a). We did not find statistically significant differences in the performance between the RF and SVM 10-fold CV PGM models trained on compound fingerprints and the dataset view ‘G.t.l. 1000 genes’ (P < 0.05). The performance of these models was comparable to that of the 10-fold CV PGM models encoding cell lines with the cor. GE and MLT kernels trained with RF and SVM, respectively (P < 0.05, labelled with ‘a’ in Fig. 2). Hence, in this dataset, encoding cell lines with kernels that are naïve with respect to the nature of the genomic makeup of the cell lines, in this case the MLT kernel, provides virtually the same predictive signal as the gene expression data, i.e. the ‘G.t.l. 1000 genes’ dataset view. Interestingly, the performance of SVM and RF across the cell line kernels explored was not constant, as the RF models outperformed SVM when using the cor. Proteome and Dirac kernels to encode cell lines, whereas the SVM models outperformed RF in the case of the MLT kernel, thus highlighting that information across compounds and cell lines is not shared in a similar way across RF and SVM. Based on the comparable predictive power of RF and SVM, we used RF to train all of the models presented in the following sections because RF (i) requires smaller training times than the kernel methods and (ii) are robust with respect to the value of their parameters (Supplementary Fig. S2 and Sheridan, 2013).

3.4 PGM model validation

The PGM models trained on the complete dataset with 10-fold CV using compound fingerprints and the dataset view ‘G.t.l. 1000 genes’ as input features exhibited mean RMSE_{test} and \( R^2_{0 \ test} \) test values of 0.40 ± 0.00 pgI(0) units and 0.83 ± 0.00 \((n = 10)\), respectively. These values were consistent with the theoretical maximum and minimum achievable performance, which were 1.42 and 0.33 pgI(0) units, respectively, for the RMSE_{test} and 0.96 and –0.96, respectively, for the \( R^2_{0 \ test} \) (Supplementary Fig. S3). Moreover, the model performance did not stem from chance correlations, as \( R^2_{0 \ test} \) test values became negative when 75% of the pgI(0) values were randomized (Clark and Fox, 2004) (Supplementary Fig. S4). Modelling the high confidence dataset led to a similar performance, with an RMSE_{test} of 0.45 pgI(0) units and an \( R^2_{0 \ test} \) value of 0.84. This indicates that the predictive power of RF does not decrease when we included the data points measured in only one experiment, which is in agreement with a recent benchmarking study on the noise sensitivity of machine learning algorithms in bioactivity modelling (Cortes-Ciriano et al., 2015a). The 10-fold CV PGM models were further evaluated on the uncorrelated bioactivities 0.5 dataset, where RMSE_{test} and \( R^2_{0 \ test} \) test values of 0.58 pgI(0) units and 0.79, respectively, were observed. These RMSE_{test} and \( R^2_{0 \ test} \) test values were also in agreement with the maximum achievable performance estimated for the uncorrelated bioactivities 0.5 dataset (Supplementary Fig. S5).

3.5 Comparisons between the PGM and QSAR models

To test whether the PGM models display higher predictive power than per cell line QSAR models, we trained (i) the individual QSAR 10-fold CV RF models with increasingly larger training sets, whose performance was assessed on the held-out data (orange in Fig. 2), on the uncorrelated bioactivities 1 dataset and (ii) the 10-fold CV PGM
RF models on the compound and cell line descriptors (‘G.t.l. 1000 genes’ dataset view). In the latter case, the training set comprised an increasingly larger percentage of the data annotated on a given cell line (x-axis in Fig. 2) and an increasingly larger percentage of the data annotated on the remaining cell lines (indicated in the legend of Fig. 2). We consistently obtained lower RMSE\text{test} values with the PGM models than with the individual QSAR models, indicating that PGM modelling enables the sharing of information in the biological and chemical space. Likewise, the RMSE\text{test} values systematically decreased as more data were added to the training set in the PGM models. The biggest RMSE\text{test} difference between the individual QSAR and the PGM models were obtained when the least amount of data (5 or 10%) was assigned to the training set, whereas the lowest difference was found when almost all of the data available on a given cell line was used for training. This indicates that integrating data across different cell lines seems more beneficial when the data are relatively scarce.

On the complete dataset, the 10-fold CV PGM models slightly outperformed models trained exclusively on compound descriptors, namely QSAR F, with RMSE\text{test} and $R^2$\text{test} values of 0.45 pGI\text{50} units and 0.78, respectively. These results indicate (compared to an RMSE\text{test} value of 0.40 pGI\text{50} units and $R^2$\text{test} = 0.83 for the 10-fold CV PGM model) that the pGI\text{50} values of the compounds are largely correlated across the cell line panel. However, the 10-fold CV PGM models increased the relative performance over the QSAR F models on the uncorrelated activities 0.5 dataset, with respective RMSE\text{test} values of 0.58 versus 0.69 pGI\text{50} units, respectively, highlighting the finding that the PGM models appear to be more suitable for modelling compounds exhibiting an uncorrelated pGI\text{50} value on a cell line panel. Moreover, the 10-fold CV PGM models significantly outperformed individual cell line models, i.e., QSAR (two-sided t-test, $z = 0.05, P < 0.05$), trained on the data points corresponding to a given cell line that exclusively used the compound descriptors as the input features. These individual models displayed an average RMSE\text{test} value of 0.73 ± 0.05 pGI\text{50} units (Supplementary Table S12). Therefore, integrating the biological information from different cell lines in a single PGM model improves the compound sensitivity prediction on the NCI60 panel over the QSAR models, and even more so in cases where activities across cell lines show less of a correlation.

3.6 Benchmarking the cell line profiling datasets
To benchmark the predictive signal across the 14 dataset views, we used the uncorrelated bioactivities 0.5 dataset, as it contains the compounds displaying the least correlated pGI\text{50} values on the cell line panel and thus is more challenging to model. For each dataset view, we trained a 10-fold CV PGM model using that dataset view and the compound fingerprints as the input features, which resulted in a total of 140 models (14 data set views × 10 CV folds). An analysis of variance (ANOVA) on the RMSE\text{test} values (Fig. 3a) yielded significant differences between the models ($P < 1 \times 10^{-17}$). Similar to previous reports (Costello et al., 2014; Jang et al., 2014), we have also observed that the gene transcript levels led to the highest predictive power, with median RMSE\text{test} values in the 0.56–0.58 pGI\text{50} units range (Fig. 3a). However, the combination of transcript levels from different gene sets, e.g. kinases and ABC transporters (Fig. 3a), did not translate into increased performance, suggesting that these dataset views might contain redundant predictive signal. In addition, there were no significant differences in performance between the models trained on the gene transcript levels and those trained on the miRNA abundance or RPLA data [Tukey’s Honestly Significance Difference (HSD)], $P < 0.05$). Interestingly, the performance of the models trained on the CNV or exome sequencing data was significantly worse (Fig. 3a), with RMSE\text{test} values in the 0.63–0.68 pGI\text{50} units range. This poorer performance could possibly be expected given the sparseness of these data, as the percentage of non-zero entries in the CNV and exome-seq matrices is 2.66 and 0.03%, respectively.

3.7 Bioactivity interpolation to the cell lines and compound clusters present in the training set
The interpolation power of the 10-fold CV PGM models was evaluated by calculating the RMSE\text{test} values on subsets of the test set

![Fig. 3. Benchmarking the cell line profiling dataset views for the cell line sensitivity prediction. (a) The predictive power of the 14 dataset views (Table 1) and two cell line kernels, namely cor. Proteome and cor. Transcriptome, was quantified by the RMSE values on the test set. For each dataset view, we trained the 10-fold CV PGM models on the uncorrelated bioactivities 0.5 dataset. We found significant differences among the dataset views (ANOVA, $P < 0.01$). Post-hoc analyses (Tukey’s HSD, $z = 0.05$) were used to cluster the dataset views according to their predictive power. Dataset views sharing a letter label performed at the same level of statistical significance. We consistently found that the gene transcript levels and the abundance of proteins and miRNA led to the most predictive models (labelled with ‘a’). (b) The evaluation of both interpolation and extrapolation power was evaluated on the complete dataset. After finding significant differences across groups (ANOVA, $P < 0.01$), we found that the PGM models interpolate and extrapolate to new cell lines and tissues at the same level of statistical significance (Tukey’s HSD, $z = 0.05$). In contrast, we found statistically significant differences in the performance between extrapolation and interpolation to new chemical clusters. The blue points indicate the median and the interquartile range (25th–75th percentile), whereas the red points indicate the mean RMSE value.](image-url)
grouped by the cell line, cell-line-averaged (Supplementary Fig. S6a), or compound cluster, compound cluster-averaged performance (Supplementary Fig. 6b). Thus, one average RMSE\textsubscript{test} value across the 10-fold models was calculated for each cell line and compound cluster. If not otherwise indicated, the results presented in the following sections correspond to models trained on the uncorrelated bioactivities 0.5 dataset with (i) compound fingerprints and (ii) the ‘G.T.I. 1000 genes’ dataset view as the input features. The cell-line-averaged RMSE\textsubscript{test} values ranged from 0.41 ± 0.01 (U251) to 0.86 ± 0.01 pGI\textsubscript{50} units (HOP-92). We found significant differences for the tissue-averaged performance (Tukey’s HSD, $P < 1 \times 10^{-6}$), with RMSE\textsubscript{test} values ranging from 0.48 ± 0.01 (prostate, displayed in cyan) to 0.70 ± 0.01 (leukaemia, displayed in green) pGI\textsubscript{50} units (Supplementary Fig. 6a). In addition, the learning curves showed that the RMSE\textsubscript{test} values of approximately twice the replicate-averaged experimental uncertainty (0.27 pGI\textsubscript{50} units) and the maximum achievable performance (0.35 pGI\textsubscript{50} units) can be obtained when <10% of the data is used as a training set (Supplementary Fig. S7). In contrast, an ANOVA analysis did not yield significant ($P > 0.05$) differences among the 31 chemical clusters (Supplementary Fig. S6b), with observed median RMSE\textsubscript{test} values in the 0.48–0.65 pGI\textsubscript{50} units range. Hence, overall, it has been found that the interpolation power is constant across the 31 chemical clusters, but not across tissues.

3.8 Extrapolation in the cell line and tissue space

We further evaluated the extrapolation power to novel cell lines and tissues with the LOCO and LOTO models using the complete dataset. The LOCO models exhibited mean RMSE\textsubscript{test} values of 0.43 ± 0.08 pGI\textsubscript{50} units (Fig. 3b and Supplementary Table S12), with the lowest and highest RMSE\textsubscript{test} values of 0.31 and 0.61 observed for cell lines U251 and OVCAR-5, respectively. Notably, we found that the LOCO and cell-line-averaged predictions are highly correlated ($r_s = 0.92$), indicating that interpolation and extrapolation to novel cell lines are correlated. The RMSE\textsubscript{test} values for the LOTO models ranged between 0.35 (prostate) and 0.63 pGI\textsubscript{50} units (leukaemia). Remarkably, the prediction errors were similar across the entire bioactivity range (Supplementary Fig. S8). Overall, we did not observe significant differences in performance among the LOCO, LOTO (Fig. 3b), cell-line-averaged (Supplementary Fig. S6a), and compound cluster-averaged (Supplementary Fig. S6b) results (Tukey’s HSD, $P < 0.05$). We found that the RMSE values between the observed and predicted pGI\textsubscript{50} values calculated using the LOCO and LOTO models for 47 out of 81 drugs, such as Imiquimod (NSC 369100) and Bendamustine (NSC 138783), were below 0.5 pGI\textsubscript{50} units (Supplementary Fig. S9a and b). High RMSE values, between 1 and 1.5 pGI\textsubscript{50} units, were observed for 11 drugs, such as the folate anti-metabolite pemetrexed (NSC 698037) and Irinotecan (NSC 728073). Together, these data indicate that the PGM models extrapolate compound bioactivities to novel cell lines and tissues at the same level of statistical significance as for interpolation within a given cell line or tissue.

3.9 Extrapolation in the chemical space

A markedly different trend was observed for the ability of the PGM models to generalize across the chemical space, as assessed by the LOCCO models using the complete dataset. The LOCCO models exhibited mean RMSE\textsubscript{test} values of 0.83 ± 0.17 pGI\textsubscript{50} units (Fig. 3b and Supplementary Table S12), which differed significantly from the LOCO and LOTO results (Fig. 3b) (Tukey’s HSD, $P < 0.01$), and from the compound cluster-averaged interpolation performance (Supplementary Fig. S6b). Notably, the chemical diversity within the compound clusters was not correlated with model performance, as low RMSE\textsubscript{test} values were consistently obtained for the heterogeneous and homogeneous clusters (Fig. S1 and Supplementary Table S13). The lowest RMSE\textsubscript{test} value of 0.33 pGI\textsubscript{50} units was obtained for cluster 24, which contains 485 compounds with polycyclic ring systems, generally with no more than three fused rings, as well as ring assemblies linked by sulphide, sulfinyl, secondary amine, carboxyl and alkyl groups (Supplementary Table S13). Cluster 16, which was modelled with the highest RMSE\textsubscript{test} value of 1.23 pGI\textsubscript{50} units, contains tri- and tetracyclic systems with hydroxybenzene, methoxybenzene and quinone functionalities. We obtained errors in prediction values below 0.5 pGI\textsubscript{50} units for 13 out of 81 drugs in the complete dataset, and below 1 pGI\textsubscript{50} units for 43 compounds. The worst predicted drugs were depsipeptide (NSC 630176) and the halichondrin B analogue NSC 707389, with respective errors in prediction of 4.29 and 4.35 pGI\textsubscript{50} units (Supplementary Fig. S9c). Taken together, these results indicate that the range of errors is considerably large (>4 pGI\textsubscript{50} units), and, hence, extrapolations in the chemical space remain a challenging task, and certainly on this particular dataset.

3.10 Conformal prediction provides informative CI

Conformal prediction was included in the modelling framework to provide CI for the individual predictions. The percentage of data points for which the predicted value lay within the calculated CI, which were calculated with an increasing confidence level, is highly correlated with the confidence level (Spearman’s $r_s > 0.99$, Supplementary Fig. S10). Therefore, the combination of the Random Forest and conformal prediction estimates the compound activity as a pGI\textsubscript{50} region associated to a user-defined confidence level, which represents an easily interpretable estimate of the reliability of individual predictions.

3.11 Consistency of pathway–drug associations with predicted bioactivities

To investigate whether the bioactivities predicted by the PGM models make it possible to identify genomic markers of drug sensitivity, we evaluated the consistency between the pathway–drug associations inferred from the experimental and predicted bioactivities for the 37 FDA-approved drugs and the 19 compounds in the clinical trials (Supplementary Table S14) present in the uncorrelated bioactivities 0.5 dataset. For each pathway, we fitted a linear model controlled by the tissue source, where the average pathway expression was considered as the predictor of drug sensitivity. Overall, no significant differences (Tukey’s HSD, $P < 0.05$) were observed between the pathway–drug associations calculated with the most predictive 10-fold CV PGM (Supplementary Fig. S11a) and LOCO models, as the median Spearman’s $r_s$ values between the $\beta_p$ coefficients estimated with (i) the experimental cell-line sensitivity data and (ii) the CV and LOCO predictions were in the 0.75–0.91 range (Supplementary Fig. S11b). These results indicate that the 10-fold CV and LOCO predictions reflect the association between gene expression summarized at the pathway level and the cell line sensitivity. However, significant differences were found among these groups and the LOCCO and LOTO models, for which the median Spearman’s $r_s$ values were 0.63 and 0.03, respectively. We obtained similar results when considering only the pathways significantly associated to drug response [false discovery rate (FDR) < 20\%] (Supplementary Fig. S11c and d).
Next, we analysed whether the pathway–drug associations are consistently predicted for drugs exhibiting diverse MoAs. Out of the 56 drugs considered, 26 exhibited median Spearman’s $r_s$ values between the $\beta_p$ coefficients estimated with the experimental cell line sensitivity data and the 10-fold CV predictions in the 0.5–0.75 range, and 18 were above 0.75 (Supplementary Fig. S10b). High Spearman’s $r_s$ values were obtained across the 22 distinct drug MoAs, indicating that no specific MoA is favoured. Notably, most Spearman’s $r_s$ values increased (Supplementary Fig. S11d) when the calculation of the pathway–drug associations was restricted to the pathways significantly associated to drug response (FDR < 20%). Together, we can conclude that the identification of genomic markers of drug sensitivity is significantly dependent on the presence of cell lines originating from the same tissue and structurally similar compounds in the training set.

### 3.12 Prediction of the growth inhibition patterns using the NCI60 panel

The experimental and predicted growth inhibition patterns obtained by the most predictive models were highly correlated (Supplementary Fig. S11e), with median Spearman’s $r_s$ values in the 0.53–0.58 range, and 32 out of the 56 drugs from the uncorrelated bioactivities 0.5 dataset had values higher than 0.5 for (Supplementary Fig. S11f). Nevertheless, the LOCO, LOTO and LOCCO models (Tukey’s HSD, $P$ value < 0.001) displayed a marked decrease in the $r_s$ values (Supplementary Fig. S11e) with respect to the most predictive 10-fold CV PGM models.

The relative growth inhibition values using the NCI60 panel can be depicted in a bar plot with the pGI50 values transformed into z-scores. Figure 4 depicts the observed and predicted growth inhibition patterns for methotrexate (MTX) via the 10-fold CV PGM model. MTX was chosen for illustration, given its complex growth inhibition pattern. (Bar plots for the 56 drugs calculated with the predictions obtained with the 10-fold CV PGM, LOCO, LOTO and LOCCO models are provided in the Supplementary material.) The predictions accounted for 55 out of 59 cases for the relative sensitivity of the cell line. For instance, the six leukaemia cell lines (green turquoise) were predicted to be sensitive to MTX. Moreover, complex inhibition patterns for renal-derived cell lines (light magenta) were accounted for by the predictions, as the TK-10, RXF-393 and A498 renal cell lines were predicted to be highly resistant to MTX, whereas the effect of MTX on sensitive cell lines, namely UO-31, SN12C, CAKI-1 and ACHN, was also correctly predicted (Fig. 4b). Taken together, these data indicate that the drug sensitivity predictions were able to account for the complex patterns of cell line growth inhibition for this particular drug.

### 3.13 Comparisons to previous methods

We compared our results with previous studies (Ammad-ud-din et al., 2014; Menden et al., 2013) by applying our modelling procedure to the GDSC and CCLE datasets using the Morgan fingerprints as the compound descriptors, and the gene transcript levels for the 1000 genes displaying the highest variance across the cell line panel to describe the cell lines (Supplementary Text and Table S15). For the GDSC dataset, we obtained lower mean RMSE$_{test}$ and higher $R^2_{test}$ values, namely 0.75 ± 0.01 and 0.74 ± 0.01 pGI50 units, respectively, with the 10-fold CV models compared with Menden et al. (2013), who obtained an RMSE$_{test}$ of 0.83 and an $R^2_{test}$ of 0.72 pGI50 units, and Ammad-ud-din et al. (2014), who obtained an RMSE$_{test}$ of 0.83 ± 1.00 and an $R^2_{test}$ of 0.32 ± 0.37 pGI50 units. The same trend was observed for the LOTO models, with mean RMSE$_{test}$ and $R^2_{test}$ values of 0.81 ± 0.16 and 0.72 ± 0.08 pGI50 units, respectively, compared with Menden et al. (2013), with RMSE$_{test}$ and $R^2_{test}$ values of 0.99 and 0.61 pGI50 units, respectively. The results obtained here, with RMSE$_{test}$ values of 1.40 ± 0.80 pGI50 units, are comparable to those of Ammad-ud-din et al. (2014), with RMSE$_{test}$ values of 0.83 ± 0.41 pGI50 units, when extrapolating to new compounds. However, it is important to mention that Ammad-ud-din et al. (2014) grouped compounds into sets at random and not based on chemical similarity, which likely made their extrapolation easier.

### 4 Discussion

The primary goal of this study was to capitalize on the increasing amount of in vitro cell line sensitivity and molecular profiling data of cancer cells to predict the growth inhibition patterns of compounds on the NCI60 panel. Although the principles of pharmacogenomic...
modelling are not new, this study represents novelty in the field, as, to our knowledge, it is the first effort to combine large-scale NCI anticancer screening data using the NCI60 panel with the available cell line profiling information, including error bars. Unlike previous modelling studies on the NCI60 panel (Abaan et al., 2013; Paull et al., 1989; Staunton et al., 2001; Szakacs et al., 2004; Wan and Pal, 2014), we simultaneously integrated the chemical information and cell line profiling data, which enables us to predict the growth inhibition patterns and to inter- and extrapolate on the chemical and cell line domains. In addition, coupling conformal prediction to the Random Forest enabled the definition of CI for the individual predictions.

We consistently found the highest predictive signals in the gene expression, miRNA and protein abundance data. The incorporation of prior biological knowledge by summarizing the gene expression data at the pathway level did not provide an additional predictive signal. Interestingly, high predictive power was attained with models trained on gene expression data from the genes displaying the most variable transcript levels across the cell line panel. Encoding cell lines with cell line kernels led, in some cases, to models with comparable predictive power to those trained explicitly on the cell line profiling data. Although encoding cell lines with cell line kernels may be sufficient to model the compound pGI50 values from the NCI60 panel, in our view this would not be the case on datasets comprising highly dissimilar cell lines. In addition, we note, in particular, that the Dirac and MLT kernels do not permit us to extrapolate in the cell line space, likely rendering them less useful in those cases. The sparseness of the CNV and exome sequencing data may be the reason for the poor model performance, and, therefore, it remains to be seen whether the modelling of cell line panels with more comprehensive mutational data leads to better predictions (Costello et al., 2014).

A major challenge to the cell line sensitivity prediction is the extrapolation of compound bioactivities to novel cell lines and to structurally distinct compounds. We did not find significant differences in performance between the interpolation and the extrapolation to new cell lines (LOCO) and tissues (LOTO), with mean RMSE_{test} values in all cases smaller than twice the mean uncertainty value of the bioactivity measurements. This observation enables the prediction of compound activities on cancer cell lines for which little bioactivity data are currently available; however, extrapolation is still improved by the presence of cell lines from the same tissue/ontogeny in the training set. Given that similar compounds exhibit similar growth inhibition profiles (Shivakumar and Krauthammer, 2009), it was expected that the model performance would considerably decrease when predicting the activity of structurally dissimilar compounds. This was indeed the case because mean RMSE_{test} values of approximately three times the average experimental uncertainty were obtained when extrapolating in the chemical space (LOCCO). Although the error in prediction should ideally be close to the experimental uncertainty, this performance may still be useful for compound prioritization. Previous studies have shown that adding physicochemical descriptors or increasing the bit-string length of the Morgan fingerprints (set to 256 here) leads to higher predictive power when modelling a highly diverse set of molecules (De Bruyn et al., 2013). Here, we did not obtain higher predictive power when increasing the bit-string length, when adding physicochemical descriptors to the compound fingerprints, or when using Morgan fingerprints in binary format (Cortes-Ciriano et al., 2014). Our approach displays a higher predictive ability (10% decrease of CV RMSE) than the methods previously applied to the GDSC dataset (Supplementary Text and Supplementary Table S15).

Moreover, the application of our approach to the Cancer Cell line Encyclopedia (CCLE) (Barretina et al., 2012) dataset led to statistically validated models, displaying $R^2_{test}$ values of ~0.74, which were comparable to those obtained on the GDSC and NCI60 datasets. Given that gene expression profiles for the 44 cell lines comprised in the CCLE and the NCI60 panel are highly correlated (Spearman’s $r_s=0.88$) (Supplementary Fig. S12), we suggest that PGM models trained on the NCI60 panel could be applied to the CCLE cell line panel, which could identify new purposes for the 17,142 compounds considered here (Weinstein, 2012).

It is also important to consider how the cell line sensitivity is quantified for the application of the type of model presented here. Fallahi-Sichani et al. (2013) applied a multi-parametric analysis to a dataset comprising the activity of 64 anticancer drugs on 53 breast cancer cell lines. The results of this study indicate that the parameters of the dose–response curve vary systematically and depend on both the cell line as well as the drug class. For instance, the MoA of a drug has a strong influence on drug efficacy ($E_{max}$), potency ($IC_{50}$) and on the steepness of the drug response curve. Overall, this study indicates that parameters other than the potency of the drug response curve should be considered in studies comparing drug activity, as they are likely to provide crucial insights into the biology of the cell line’s response to drug treatment and into the drugs’ MoA.

An additional key aspect is the consistency of the in vitro cell line sensitivity data. A previous study (Haibe-Kains et al., 2013) reported a rather low correlation of the cell line sensitivity data from the CCLE and GDSC datasets. We found low concordance between the NCI60 and the CCLE sensitivity data (Supplementary Fig. S13), and we did not obtain a high correlation between the predictions calculated with a model trained on the NCI60 data for eight drugs on 44 cell lines shared by the CCLE and the subset of the NCI60 cell lines considered here. Actually, the RMSE values for these predictions against the experimental data from the CCLE and, the pIC$_{50}$ values from the CCLE and the pGI$_{50}$ from the NCI60 dataset for these eight drugs and 44 cell-lines was comparable, namely 0.87 log$_{10}$ units, indicating that high predictive power cannot be attained given the low concordance of the sensitivity data from these two datasets. These results come as no surprise, given the different surrogates of cytotoxicity exploited by the assays used to screen the CCLE and NCI60 panels, namely metabolic activity and protein abundance, respectively. Therefore, we conclude, in accord with Haibe-Kains et al. (2013), that models trained on the NCI60 data set are likely to fail on sensitivity data measured with a different experimental procedure, e.g. the CCLE or GDSC datasets.

Finally, although cultured cell lines and primary tumours differ genetically (Borrèll, 2010) and 2D cell line cultures do not recapitulate the complex tumour microenvironment, investigating the extent to which gene expression (and other cell line profiling) data can be used to model in vitro cell line sensitivity can help researchers to develop approaches for the prediction of clinical drug responses using the genomic data of tumour samples (Geelker et al., 2014).

Acknowledgements

I.C.C. is a fellow of the Pasteur-Paris International PhD Programme. The authors thank Dr. Sudhir Varma for his valuable help in gathering and preparing the data. The authors thank the computational resources provided by the Institut Pasteur. Three anonymous referees are thanked for their comments to improve the manuscript. Prof. Véronique Stoven is thanked for fruitful discussions on kernels and multitask learning.

Funding

G.J.P.v.W. and J.P.O. thank the EMBL member states, The Wellcome Trust (WT086151/Z/08/Z) and Marie Curie (COFUND) for funding. I.C.C., T.M.
and M.N. thank the CNRS and Institut Pasteur for support. A.B. thanks Unilever and the European Research Commission (Starting Grant ERC-2013-StG 336159 MIXTURE) for funding. M.N. thanks the Investissement d’Avenir Bioinformatics Program (Grant Bip:Bip) and the European Research Commission (Advanced Grant ERC-2011-StG 294809 BayCellS) for funding.

Conflict of Interest: none declared.

References


