PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Accurate cancer phenotype prediction with AKLIMATE, a stacked kernel learner integrating multimodal genomic data and pathway knowledge

**Vladislav Uzunangelov**[¤], **Christopher K. Wong**, **Joshua M. Stuart***

Department of Biomolecular Engineering, University of California, Santa Cruz, California, United States of America

¤ Current address: Bristol-Myers Squibb, Redwood City, California, United States of America
* jstuart@ucsc.edu



OPEN ACCESS

**Citation:** Uzunangelov V, Wong CK, Stuart JM (2021) Accurate cancer phenotype prediction with AKLIMATE, a stacked kernel learner integrating multimodal genomic data and pathway knowledge. PLoS Comput Biol 17(4): e1008878. https://doi.org/10.1371/journal.pcbi.1008878

**Editor:** Anna R. Panchenko, Queen's University, CANADA

**Received:** September 23, 2020

**Accepted:** March 15, 2021

**Published:** April 16, 2021



**Data Availability Statement:** All relevant data are within the manuscript and its Supporting information files. Additional data availability can be found in S1 Text.

**Funding:** The study was supported by grants to J. M.S. from NCI (U24-CA143858, 1R01CA180778, NHGRI (5U54HG006097), and NIGMS (5R01GM109031). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Abstract

Advancements in sequencing have led to the proliferation of multi-omic profiles of human cells under different conditions and perturbations. In addition, many databases have amassed information about pathways and gene "signatures"—patterns of gene expression associated with specific cellular and phenotypic contexts. An important current challenge in systems biology is to leverage such knowledge about gene coordination to maximize the predictive power and generalization of models applied to high-throughput datasets. However, few such integrative approaches exist that also provide interpretable results quantifying the importance of individual genes and pathways to model accuracy. We introduce AKLIMATE, a first kernel-based stacked learner that seamlessly incorporates multi-omics feature data with prior information in the form of pathways for either regression or classification tasks. AKLIMATE uses a novel multiple-kernel learning framework where individual kernels capture the prediction propensities recorded in random forests, each built from a specific pathway gene set that integrates all omics data for its member genes. AKLIMATE has comparable or improved performance relative to state-of-the-art methods on diverse phenotype learning tasks, including predicting microsatellite instability in endometrial and colorectal cancer, survival in breast cancer, and cell line response to gene knockdowns. We show how AKLIMATE is able to connect feature data across data platforms through their common pathways to identify examples of several known and novel contributors of cancer and synthetic lethality.

## Author summary

We describe a new method that incorporates multimodal molecular measurements with gene pathway information for classification and regression prediction tasks. The method combines powerful machine learning methodologies such as ensemble, kernel and stacked learning. A key new contribution is the creation of empirical kernels from pairwise

**Competing interests:** The authors have declared that no competing interests exist.

similarities of sample predictions and sample paths along the trees of a random forest specific to a particular gene module. We demonstrate the method performs as well as, or better than, top approaches on three very different cancer genomics applications.

This is a *PLOS Computational Biology* Methods paper.

## Introduction

The drop in sequencing cost has made it common for biological experiments to generate multi-omic profiles under a variety of conditions and perturbations. For example, the Cancer Genome Atlas (TCGA) contains thousands of patient samples with simultaneous copy number, mutation, methylation, mRNA, miRNA and protein measurements [1]. The analysis of multi-omic experiments produces feature sets that capture the genomic and transcriptomic changes relevant to a specific condition, sample subtype, or biological pathway—this "prior knowledge" eventually accumulates in a growing number of databases [2–6]. However, the interpretable integration of such prior knowledge with multi-omic data from new experiments remains an important challenge (see Materials and methods section for a discussion on challenges and previous approaches).

In this study, we introduce the Algorithm for Kernel Learning with Integrative Modules of Approximating Tree Ensembles (AKLIMATE)—a novel approach that combines heterogeneous data with prior knowledge in the form of gene sets. AKLIMATE can evaluate the predictive power of individual features (e.g. genes) as well as feature sets (e.g. pathways). It harnesses the advantages of Random Forests (RFs; native handling of continuous, categorical and count data, invariance to monotonic feature transformations, ease of feature importance computation), Multiple Kernel Learning (MKL; intuitive integration of overlapping feature sets), and stacked learning (improved accuracy) while avoiding many of their shortcomings. AKLIMATE relies on three major computational insights. First, it leverages biological knowledge and multiple data types by building a separate RF model for each distinct gene set (e.g. biological process, transcriptional signature, genomic location), with the ability to incorporate heavily overlapped feature groups without the undesirable side effects of other approaches (e.g. Group Lasso; see Materials and methods). Second, it uses co-classification patterns of sample pairs across decision trees within an RF model to compute a kernel similarity matrix (RF kernel) that is data driven yet capable of capturing complex non-linear feature relationships. Third, the RF kernel naturally re-weights the input features so that more informative ones make bigger contributions to kernel construction. This final property is key when dealing with feature sets in which some, but not all, of the features are informative for classification. We demonstrate that these insights lead to a general improvement in performance when AKLIMATE is compared against state-of-the-art algorithms on various classification and regression tasks.

## Results

We evaluated AKLIMATE on multiple prediction tasks—microsatellite instability in endometrial and colon cancer, survival in breast cancer, and small hairpin RNA (shRNA) knockdown viability in cancer cell lines. We benchmarked AKLIMATE against comparable methods that have performed well in recent Dialogue on Reverse Engineering and Assessment Methods (DREAM) challenges. We chose both classification and regression tasks as well as various levels of data availability—a single data type, multiple data types (including inferred data), or multiple data types with clinical information.

## Microsatellite instability

We first tested AKLIMATE on predicting microsatellite instability in the colon and rectum adenocarcinoma (COADREAD) and uterine corpus endometrial carcinoma (UCEC) TCGA cohorts. Microsatellite instability (MSI) arises as a result of defects in the mismatch repair machinery of the cell. Tumors with MSI (often accompanied by higher mutation rates) represent a clinically relevant disease subtype that is associated with better prognosis. MSI is also an immunotherapy indicator as such tumors produce more neoantigens. MSI can be predicted with high accuracy from expression alone [7], providing a straightforward benchmark for AKLIMATE performance with a single feature type in a binary classification setting.

We used expression data (upper-quantile normalized RSEM counts of RNA-Seq) and MSI annotations as described in [8] for the COADREAD and UCEC TCGA cohorts. The UCEC cohort consisted of 326 patients, of which 105 exhibit high microsatellite instability (MSI-H) and the remaining 221 are classified as either low (MSI-L) or stable (MSS). The COADREAD cohort included 261 samples, with 37 MSI-H and the remaining 224 classified as either MSI-L or MSS. In both tumor types, we trained models to distinguish MSI-H patients from MSI-L +MSS patients on 50 phenotype-stratified partitions of 75% training and 25% test folds. We then computed area under the ROC curve (AUROC) for each set of test fold predictions.

We compared AKLIMATE to Bayesian Multiple Kernel Learning (BMKL) because it performed well in several DREAM challenges, in particular winning the NCI-DREAM Drug Sensitivity Prediction Challenge [9]. Furthermore, its pathway-informed extension [7] shares several similarities with AKLIMATE—it is a multiple kernel learning method that operates on pathway-derived kernels. In particular, BMKL uses expression-based Gaussian kernels computed on features from the Pathway Interaction Database (PID) pathway collection [10]. We tested four versions of BMKL—sparse single-task BMKL(SBMKL), dense single-task BMKL (DBMKL), sparse multi-task BMKL (SBMTMKL), and dense multi-task BMKL (DBMTMKL). Sparse BMKL models are the focus of [7]—they use sparsity-inducing priors to train models with few non-zero kernel weights (we used the hyperparameters specified in [7]). We added DBMKL and DBMTMKL to the comparison because dense MKL models (almost all kernels receive non-zero weights) tend to produce higher predictive accuracy in many experimental settings (e.g. [11]). Their parameters were identical to the ones for the sparse models except for $(\zeta_\kappa, \eta_\kappa)$, which were set to (999, 1) in the dense models ($(\zeta_\kappa, \eta_\kappa) = (1, 999)$ in the sparse ones). Finally, the single-task models were trained separately on the UCEC and COADREAD cohorts, while the multitask versions learned MSI status on the two TCGA cohorts jointly, with each cohort representing a separate task. All train/test splits were matched across methods. All BMKL models used 196 PID gene sets; model parameters, kernel computations and data filtering steps matched the setup in [7].

Since AKLIMATE uses a much larger gene set compendium ($S = 17,273$ feature sets, see S1 Text), we controlled for this prior information imbalance as a possible source of performance bias. For that purpose, we created a reduced version (AKLIMATE-reduced) that is restricted to the same set of 196 input PID sets as used by the original publication of BMKL. We refer to the full unrestricted model of AKLIMATE as simply AKLIMATE in this comparison.

AKLIMATE predicted MSI status in the UCEC cohort significantly better than AKLIMATE-reduced or any of the BMKL models (Fig 1A, mean AUROCs 0.885±0.051, 0.92±0.029, 0.903±0.039, 0.92±0.03 for SBMKL, DBMKL, SBMTMKL, and DBMTMKL respectively). In particular, AKLIMATE achieved mean AUROC of 0.962±0.019 compared to 0.938±0.031 for AKLIMATE-reduced ($P < 4.1e-08$; paired Wilcoxon signed rank test), suggesting a predictive benefit to AKLIMATE's larger collection of gene sets. A larger gene set collection is both more likely to contain sets derived specifically to describe the MSI process and more flexible in

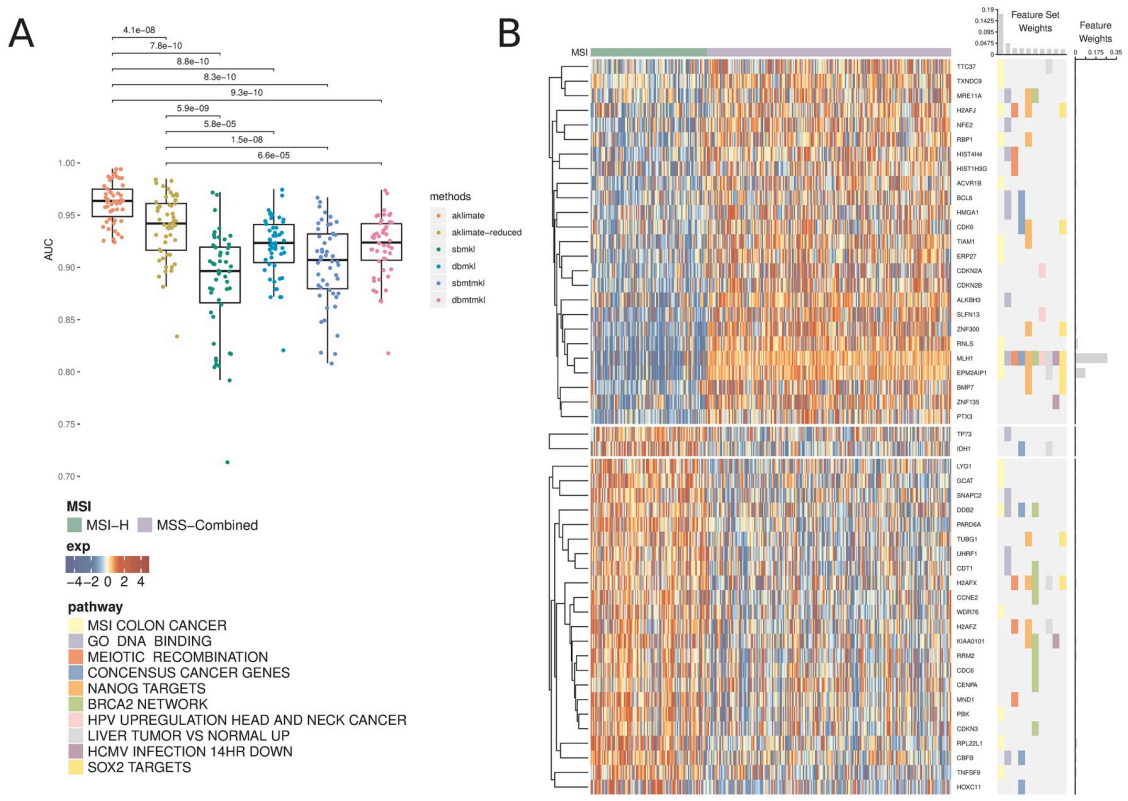


**Fig 1. AKLIMATE performance on predicting MSI in UCEC TCGA.** (A) Performance of AKLIMATE and BMKL on classifying MSI-H vs MSI-L+MSS. AUROC computed for 50 75%/25% stratified train/test splits. P-values for paired Wilcoxon signed rank test. Methods compared: aklimate—AKLIMATE on full collection of feature sets; aklimate-reduced—AKLIMATE with 196 PID pathways; sbmkl—sparse single-task BMKL; dbmkl—dense single-task BMKL; sbmtmkl—sparse multi-task BMKL; dbmtmkl—dense multi-task BMKL. Multi-task BMKL models are trained to simultaneously predict MSI status on UCEC and COADREAD cohorts. (B) Top 10 predictive AKLIMATE feature sets and top 50 predictive features. Expression of top 50 features (left heatmap); Membership of most predictive features in most predictive feature sets (right heatmap). Features are organized by KNN clustering into 3 groups, followed by hierarchical clustering within each cluster. Feature set model weights scaled to sum up to 1 (barplot, top of right heatmap). Feature model weights scaled to sum up to 1 (barplot, right of right heatmap). Feature and feature set weights averaged across 50 train/test splits.

https://doi.org/10.1371/journal.pcbi.1008878.g001

terms of the possible combinations of component gene sets. Indeed, the most informative feature set according to AKLIMATE was "MSI Colon Cancer" (16.9% relative contribution to model explanatory power)—a gene expression signature for MSI-H vs MSI-L+MSS in COADREAD cohorts [12] (Fig 1B). Futhermore, the next two most informative sets were "GO DNA Binding" (4.55% relative contribution) and "REACTOME Meiotic Recombination" (2.4% relative contribution), both of which are strongly relevant to DNA mismatch repair (MMR). Of note, the MutL Homolog 1 (MLH1) gene was the top-ranked single feature in AKLIMATE (26.7% relative contribution) and was present in the ten top ranked gene set kernels (Fig 1B). MLH1 is a key MMR gene involved in meiotic cross-over [13]—loss of MLH1 expression, usually through DNA methylation, is known to cause microsatellite instability. This demonstrates that AKLIMATE was able to pinpoint individual causal genes as it sifts through thousands of gene sets. In contrast, the meta-pathway constructed by AKLIMATE-reduced represents a poorer approximation to the underlying biological process, as evidenced by its lower AUROC. This is likely due to the fact that, out of the 196 PID pathways used, only "PID P53 Downstream" (35.2% relative contribution) contained MLH1 S2 Fig), limiting the influence of this key gene on the prediction task.

Interestingly, even though AKLIMATE-reduced used the same feature sets as BMKL, it achieved a statistically significant improvement over all BMKL varieties (Fig 1A). This included both the non-sparse and multi-task BMKL varieties, despite the fact that the latter drew benefit from an entire additional training set of COADREAD data. In this case, the difference in kernel representations may have contributed to the improved performance of AKLIMATE-reduced (see Discussion). For the COADREAD MSI classification task, we found that all methods performed equally well, with AKLIMATE-reduced having marginally lower accuracy compared to DBMKL & DBMTMKL and the full AKLIMATE model showing marginally higher mean AUROC compared to all BMKL varieties (S3 Fig). We suspect the strong MSI signal in the feature data makes this prediction task somewhat more facile, limiting our ability to discriminate among methods based on their predictive performance.

## Breast cancer survival

For our second benchmark, we considered the task of predicting survival in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort [14]. This problem is much more challenging than predicting MSI status, as demonstrated by the DREAM Breast Cancer Challenge [15] and elsewhere [16]. Another difference between this task and MSI inference is that the METABRIC cohort is annotated with curated clinical data. In fact, the clinical features are quite informative for survival prediction—pre-competition benchmarking by the DREAM Challenge organizers found that models that used exclusively clinical features significantly outperformed ones that used only genomic features, and performed only marginally worse than models in which clinical features were augmented by a subset of molecular features selected through prior domain-specific knowledge [17]. In addition, the best pre-competition clinical and molecular feature model had better accuracy than all but the top 5 models in the actual challenge [15].

A breast cancer model that foregoes the use of clinical data would clearly suffer from inferior performance as well as reduced relevance in real-world medical settings. To achieve clinical data integration, AKLIMATE introduces a special category of "global" features, which are added to the "local" features of each feature set prior to the construction of its corresponding Random Forest. Global features can therefore be interpreted as a uniform conditioning step applied to all AKLIMATE component Random Forests. In our METABRIC analysis, all clinical features are treated as global.

We compared AKLIMATE to two state-of-the art METABRIC survival predictors. The first one, which we refer to as the "Breast Cancer Challenge" (BCC) method, was the top-performer in the Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge [15, 18]—an ensemble of Cox regression, gradient boosting regression, and K-nearest neighbors trained on different combinations of clinical variables and molecular-feature derived metagenes. The second one is a multiple kernel learning (MKL) method—Feature Selection MKL (FSMKL) [16]—that implements a pathway-informed extension of SimpleMKL [19] using linear and polynomial kernels created from clinical data and molecular features in pathways of the Kyoto Encyclopedia of Genes and Genomes (KEGG) [20]. In FSMKL, pathway features from different data types lead to the construction of separate kernels—in the case of METABRIC, each pathway produces distinct expression and copy number kernels (in contrast, AKLIMATE learns one kernel matrix from the combined pathway features across all data types). In addition, FSMKL treats each clinical feature as a singleton pathway that produces an individual kernel. To make our results directly comparable to FSMKL and BCC as presented in [16], we used a subset of the patient cohort ($N = 639$) and a reduced set of clinical variables to match the dataset used in that publication (see S1 Text).

We cast the problem as a classification task where molecular and clinical input features are used to predict whether a patient is alive or not at the 2000 day mark. Based on the 2000 day cutoff, there were 387 survivors and 252 non-survivors in the reduced cohort. Similar to the MSI analysis, we performed 50 stratified repeats of 80% train and 20% test partitions; AKLIMATE was trained on each training split and its accuracy computed on the respective test samples. To decrease computation time, AKLIMATE's kernel construction step used 1000 trees instead of the default 2000 (see S1 Text for all other hyperparameter settings).

The full AKLIMATE model had higher mean accuracy than BCC (74.1±3.3% vs 73.3±0.2%) and was on par with FSMKL (74.1±3.3% vs 74.2±1.8%) with a slightly higher standard error across cross-validation folds compared with FSMKL (S4 Fig). The AKLIMATE accuracy was largely unchanged when it was restricted to use the same KEGG pathways as FSMKL (S5 Fig). All three algorithms selected clinical variables as important for prediction.

Of note, two of AKLIMATE's main advantages were not fully utilized under these experimental settings. First, the benefit of incorporating prior knowledge is reduced because of the relatively low information content of genomic features. Second, the clinical variables may lack complex interactions, which may explain why modeling them with simpler linear techniques is just as effective. In fact, the most explanatory feature in the full AKLIMATE model (16% relative contribution) was the Nottingham Prognostic Index (NPI) [21]—a linear combination of tumor size, tumor grade, and number of lymph nodes involved. Even with these disadvantages, however, AKLIMATE achieved performance on par with two state-of-the-art methods that were specifically optimized for the METABRIC data.

In this task, clinical information proved to be the most influential data type for survival prediction. AKLIMATE models with clinical features alone were more accurate than AKLIMATE models with genomic features alone (p-val = 1.9e-08, paired Wilcoxon signed rank test, Fig 2A). This is even more striking considering the models used tens of thousands of genomic features versus only 15 clinical ones. However, the AKLIMATE models using both clinical and genomic features outperformed each single-component model (p-val = 1.4e-09 for full versus genomic, p-val = 6.7e-04 for full versus clinical, paired Wilcoxon signed rank test, Fig 2A), suggesting that the inclusion of genomic features contained complementary signals with respect to the clinical variables. The mean relative contributions of each data type in the full models were 61.6% for clinical, 29.6% for expression, and 8.8% for copy number.

Importantly, while the full AKLIMATE and FSMKL models achieved similar mean accuracy, the influence on prediction of individual features and pathways in each model were quite different. AKLIMATE heavily favored clinical variables, with 54.5% of the model's relative explanatory power carried by just five features—NPI, tumor size, lymph node involvement, age and treatment (Fig 2B). FSMKL ranked tumor size as the most important clinical feature (second overall), with other clinical variables also generally considered relevant—e.g. NPI (9th), age (11th), histological type(14th), tumor group(34th) and PAM50(40th) [16]. Most clinical variables, however, were considered less informative than FSMKL's top ranked KEGG pathway-based kernels. FSMKL's highest weighted kernel was "Intestinal Immune Response for IgA production", followed in importance by "Arachidonic acid metabolism", "Systemic lupus erythematosus", "Glycerophospholipid metabolism", and "Homologous recombination"—all of these KEGG-based kernels scored higher than any clinical variable with the exception of tumor size [16].

AKLIMATE's most informative feature sets were enriched for breast cancer progression and response to treatment, with 3 of the top 10 and 8 of the top 20 (S1 Table) related to these functional groups. For example, the most informative feature set ("BREAST CANCER ER-/PR- DN") represents a signature that is correlated with reduced protein abundance of the estrogen (ER) and progesterone (PR) hormone receptors [22]. Similarly, the 9th most

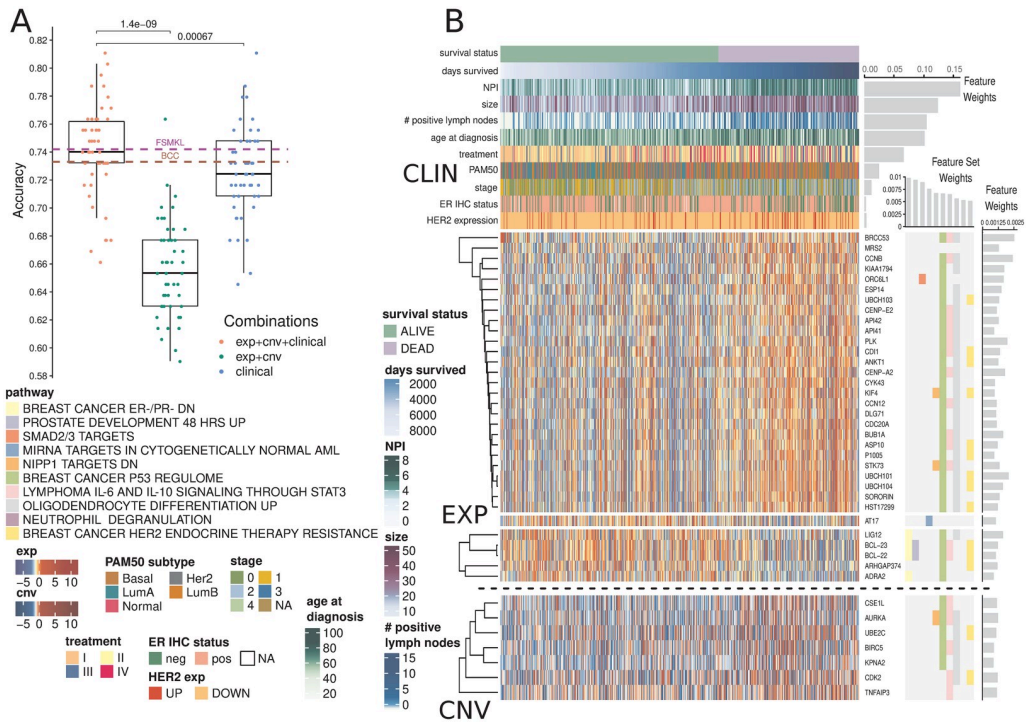


**Fig 2. AKLIMATE performance on predicting survival at 2000 days in the METABRIC cohort.** (A) Performance of AKLIMATE under different data type combinations. EXP+CNV—AKLIMATE with genomic features only; clinical—an RF model run with the clinical variables only; EXP+CNV+CLINICAL—AKLIMATE with genomic features as "local" variables and clinical features as "global" ones. FSMKL and BCC dashed lines show mean performances for the two models under 5-fold cross-validation as shown in [16]. (B) AKLIMATE results highlighting the top 10 predictive feature sets and top 50 predictive features. Figure organized as Fig 1. Clinical variables shown as column annotations; included only if among the 50 most informative features in the model. Clinical variables are ranked from top to bottom by relative predictive contribution. Survival status is a binary variable representing survival at 2000 days (labels) while days survived shows actual duration of survival. Samples sorted by days survived within the two classes. Feature and feature set weights averaged across 50 train/test splits.

https://doi.org/10.1371/journal.pcbi.1008878.g002

informative pathway ("BREAST CANCER HER2 ENDOCRINE THERAPY RESISTANCE") [23] captures transcriptome changes associated with the development of resistance to targeted therapies. Both of these signatures serve as proxies for highly relevant information not available for the METABRIC cohort (protein activity for PR and therapy resistance). Furthermore, the latest American Joint Committee on Cancer breast cancer staging manual introduces tumor grade and ER/PR/HER2 receptor status among the key breast cancer biomarkers, which already include tumor size and lymph node engagement [24]. All of these appeared as highly informative AKLIMATE model features, either directly or via a proxy genomic signature.

Breast cancers are typically subtyped into one of several main pathological groups that heavily influence treatment selection (e.g. luminal / ER-positive, HER2-amplified, basal / triple-negative, normal-like). Appreciable differences in AKLIMATE's METABRIC performance can be seen when the samples are divided up by their subtypes (based on PAM50 signatures [25], S6 Fig). The basals were associated with the least accurate (65.6% ± 8%, n = 131) and the luminal-As with the most accurate (80.8% ± 5.4%, n = 200) results, a trend that is expected given the known higher heterogeneity levels of less differentiated basal relative to luminal-A tumors.

## shRNA knockdown viability

We tested AKLIMATE's ability to integrate multiple data types and solve regression tasks to predict cell viability post shRNA knockdown. In this case, the prediction labels were continuous values representing cell line survival after the shRNA-mediated mRNA degradation of a particular gene. Gene profiles were computed with the Analytic Technique for Assessment of RNAi by Similarity (ATARiS) method [26] that creates consensus scores by combining viability phenotypes from multiple shRNAs targeting the same gene. We selected 37 such profiles for different genes across 216 cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) [27]. We chose these 37 tasks (out of 5711 available consensus profiles) because they had at least 10 cell lines showing strong (>2 standard deviations from mean) knockdown viability response and were in the top quartile by variance of all consensus profiles (see S1 Text). Based on the results of the DREAM9 gene essentiality prediction challenge [28], we expected this task to be more difficult than the other case studies.

We used expression and copy number measurements from CCLE [29] as predictive features. We augmented these two data types by adding discrete gene copy number calls made by GISTIC2 [30] and activities for 447 transcriptional and post-transcriptional regulators inferred by hierarchical VIPER [31, 32]. We focused on the 206 cell lines for which we had knockdown profiles, copy number and expression features.

We compared AKLIMATE's performance to three of the five top performing methods in DREAM9 [28] as well as three baseline algorithms. We briefly describe the DREAM9 subchallenge 1 top performers next. Multiple Pathway Learning (MPL) [31] took 5th place (see https://www.synapse.org/#!Synapse:syn2384331/wiki/64760 for challenge results)—it used elastic-net regularized Multiple Kernel Learning with Gaussian kernels based on feature sets from the Molecular Signature Database (MSIGDB) [3]. MPL and Random Forest Ensemble (MPL-RF) took 2nd place—its prediction was based on averaging a Random Forest classifier with MPL. MPL and MPL-RF were our contributions to the DREAM9 challenge. Both methods were run with the same hyperparameters and pathway collections as in DREAM9 [28, 31]. Kernelized Gaussian Process Regression (GPR) took 3rd place—it used extensive filtering steps to reduce the input feature dimensionality, followed by principal component analysis, and finally Gaussian Process regression with covariance computed from the principal components [28] (see also https://www.synapse.org/#!Synapse:syn2664852/wiki/68499 for implementation and model description). We downloaded the code from the Synapse URL and ran it with the published DREAM9 hyperparameters.

To provide performance baselines, the DREAM9 winners were augmented by standalone Random Forest (RF), Generalized Linear Model (GLM) with lasso penalty (GLM-sparse), and GLM with L2 regularization (GLM-dense). RF was run with the *ranger* R package [33] with the following hyperparameters—sampling without replacement with 70% of the samples used for tree construction, minimum node size of 10, 1500 trees and 10% of the features randomly sampled for each node split. GLM-dense and GLM-sparse were run using the *glmnet* R package [34] with the response family set to "gaussian" and the strength of regularization $\lambda$ optimized through cross-validation. The elastic net tradeoff between the lasso and ridge penalties $\alpha$ was set to $\alpha = 0.8$ (GLM-sparse) and $\alpha = 0.001$ (GLM-dense).

We did not include the nominal first place winner of DREAM9—an ensemble of four kernel ridge regression models with kernels trained through Kernel Canonical Correlation Analysis and Kernel Target Alignment—because we could not re-run the source code supplied with the challenge submission. We feel this omission is not material as the top 3 methods were declared joint co-winners—their results were shown to be statistically indistinguishable from each other but separable from the rest of the entries [28]. Furthermore, experiments in [31]

suggest that this method underperforms MPL, MPL-RF and GPR when only high-quality shRNA knockdown profiles are considered.

To save computational time, we compared methods on a single stratified train/test split for each ATARiS profile (a different split for each profile) where 67% of the cell lines were used for training and 33% were withheld for testing. Each method was run with its recommended parameters and filtering steps—if no filtering steps were specified, the method used all available features. AKLIMATE's prediction binarization quantile was set to its default value of $q = 0.05$ (see Methods, also S1 Text for all other settings).

AKLIMATE achieved average Root Mean Squared Error (RMSE) of 1.031 vs 1.048 for GPR, 1.056 for RF, 1.066 for GLM-dense, 1.07 for MPL, 1.071 for GLM-sparse and 1.081 for MPL-RF. While the improvements were modest, the mean RMSE difference was statistically significant in all but one case (Fig 3A). AKLIMATE was also the top performer when we considered the number of times an algorithm achieved the best RMSE on an individual prediction task (Fig 3B, AKLIMATE retained top spot under non-RMSE metrics as well—S7 Fig). AKLIMATE performed better than average across nearly all tasks (S8 Fig); its advantage was particularly pronounced in predicting the essentiality of key regulators—Beta-Catenin (CTNNB1), Forkead Box A1 (FOXA1), Mouse Double Minute 4 Homolog Regulator of P53 (MDM4), Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (PIK3CA)—or housekeeping genes—Proteasome 26S Subunit ATPase 2 (PSMC2), and the fifth ATPase subunit (PSMC5). As these gene classes are heavily studied and thus over-represented in our pathway compendium, AKLIMATE's enhanced accuracy may be due to the relatively higher abundance of relevant prior knowledge.

For example, AKLIMATE's ability to predict MDM4 knockdowns benefited from MDM4's function as a p53 inhibitor—many of the most informative feature sets in the MDM4 model relate to p53's regulome and its role in controlling apoptosis, hypoxia and DNA damage response (Fig 3C). The top features were also functionally linked to p53—Cyclin Dependent Kinase Inhibitor 1A (CDKN1A, 28.6% relative contribution) is a kinase that controls G1 cell cycle progession and is tightly regulated by p53; Mouse Double Minute 2 homolog P53 binding protein (MDM2, 8% relative contribution) participates in a regulatory feedback loop with p53; Zinc Finger Matrin-Type 3 (ZMAT3, 3.7% relative contribution of expression; 3.5% of copy number) interacts with p53 to control cell growth. In addition, the four p53 features from each data type were all present among the 50 most informative ones. Individually they carried little signal (relative contribution: 0.5% expression, 0.3% copy number, 0.3% inferred protein activity, 0.2% GISTIC score) but taken together they clearly implicated p53 as one of the top 10 most informative genes. The ability to capture such multi-omic interactions is one of AKLIMATE's main strengths, made possible by the use of RF kernels that fully integrate all data types. The synergy between different p53-based features is impossible to observe in methods that assign individual kernels to each data type (e.g. BMKL, FSMKL and MPL). Encouragingly, AKLIMATE dominated MPL/MPL-RF on almost all tasks even though they share the same MKL solver (Figs 3A and S8 Fig).

FOXA1 knockdown offered another example of AKLIMATE showing superior predictive performance on a well-studied gene. FOXA1 dysregulation is an essential event in breast cancer progression and subtype characterization. Due to breast cancer's prevalence and clinical importance, there is an extensive list of relevant signatures in our feature set compendium. Eight of the top 10 (and 12 of 14 overall) feature sets in the FOXA1 AKLIMATE model were directly related to breast cancer experiments under different conditions (S10 Fig). As expected, FOXA1 features were the most informative (relative contribution: 31.5% FOXA1 expression; 3.4% FOXA1 inferred activity), with the androgen receptor (AR) also among the top 5 most informative genes (relative contribution: 2.8% AR expression).

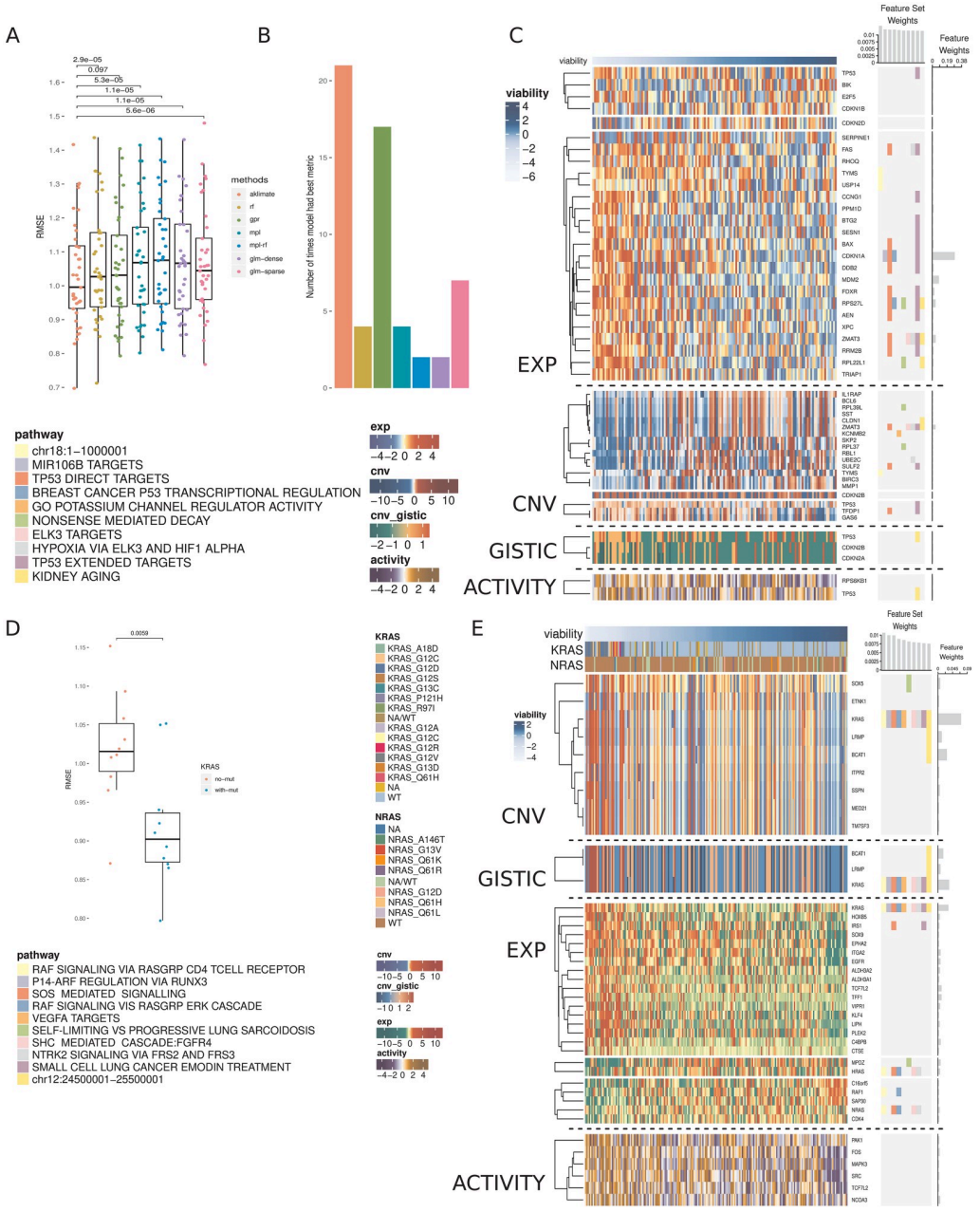


**Fig 3. Prediction of cell line viability after shRNA gene knockdowns.** (A) RMSEs of AKLIMATE and competing methods on 37 consensus viability profiles from the Achilles dataset. P-values for RMSE comparisons from paired Wilcoxon signed rank test. Methods: Random Forest (RF), Gaussian Process Regression (GPR), Multiple Pathway Learning (MPL), ensemble of MPL and Random Forest (MPL-RF), L2 regularized linear regression (GLM-dense), L1 regularized linear regression (GLM-sparse). (B) Number of times an algorithm produced the best RMSE on a prediction task. To prevent small relative RMSE differences from having a biasing effect on the win counts, for each task we consider all algorithms with RMSE within 1% of the min RMSE to be joint winners. For that reason, the total win counts add up to more than the number of regression tasks. (C) AKLIMATE top 10 informative feature sets and top 50 informative features for the task of predicting MDM4 shRNA knockdown viability. Organized as Fig 1. (D) RMSEs of KRAS AKLIMATE models with and without the use of mutation profiles for eight key regulators. Results shown for 10 matched stratified train/test splits where 80% of the cohort is used for training and 20% for testing. (E) AKLIMATE top 10 informative feature sets and top 50 informative features for the task of predicting KRAS shRNA knockdown viability when mutation features are used (feature and feature set weights averaged over 10 train/test splits). Organized as Fig 1.

https://doi.org/10.1371/journal.pcbi.1008878.g003

Out of the 37 shRNA prediction tasks we considered, shRNA knockdown of the Kirsten Rat Sarcoma Viral Oncogene Homolog (KRAS) was the most obvious example of a task involving a well-characterized gene that did not experience discernible accuracy improvement over competing methods (S8 Fig). Our hypothesis is that a true biological "driver" is absent from the set of molecular features presented to AKLIMATE. To test this, we added a mutation data type containing information for 8 key regulators (KRAS, NRAS, PIK3CA, BRAF, PTEN, APC, CTNNB1 and EGFR), 3 of which (KRAS, CTNNB1, PIK3CA) have knockdown profiles among the 37 shRNA prediction tasks. Even with the addition of these limited 8 features we observed dramatic improvement in KRAS shRNA prediction accuracy across all metrics (Fig 3D, mean RMSE 0.918±0.025 vs 1.02±0.024; mean Pearson 0.665±0.023 vs 0.529±0.025; mean Spearman 0.57±0.039 vs 0.453±0.034). Furthermore, the KRAS mutation feature was by far the most informative (23.5% relative contribution, Fig 3E), followed by KRAS copy number (6.9%), KRAS GISTIC (3.3%) and KRAS expression (3.1%). While KRAS expression, copy number and GISTIC features all appeared among the 50 most informative ones in the "no mutation" run, their combined relative contribution was only 3.33% (1.5% copy number, 1% expression, 0.8% GISTIC, S11 Fig). Thus, the addition of the KRAS mutation feature was not only key to improving the predictive performance of the model on its own, but it also helped "lift" the influence of KRAS features form other data types.

Our training features are measured at the gene level, but AKLIMATE has no constraints on the inclusion of more granular data. For example, it can evaluate the importance of mutations at particular amino acid positions or the type of resulting substitutions. In the case of KRAS, glycine replacement by either aspartic acid or valine in the 12th amino acid position appears to have the biggest negative effect on cell viability post-shRNA knockdown (Fig 3E). G12 is a well-known KRAS mutation hotspot [35]—AKLIMATE's ability to prioritize relevant hotspots can be a key advantage in modeling drug response or recommending treatment strategies.

The addition of mutation data did not yield any predictive benefit in modeling PIK3CA or CTNNB1 post-knockdown viability (S12 and S13 Figs). CTNNB1 had only nine mutations in the cohort, all of them in different codons. PIK3CA had 30 mutations, with some hotspots—the lack of improvement in this case may be due to the change in protein sequence not being biologically relevant, the ability of other genomic features to fully capture the mutation signal, or the fact that the "no mutation" models were quite accurate to begin with.

## Discussion

Recent surveys of cancer genome landscapes have shown that alterations of a particular pathway can involve many different genes and many different kinds of disruptions—for example, RB1 mutation, RB1 methylation, or CDKN2A deletion can all lead to aberrant cell proliferation [36]. Consequently, many bioinformatics approaches seek to combine data at the level of a biological process to benefit machine-learning applications in the cancer genomics setting. However, data platform diversity often prohibits such integration—variables can be of different scales (e.g. copy number vs gene expression) or different types (continuous DNA methylation, binary mutation calls, ordinal inferred copy number estimates). AKLIMATE's early integration approach is a potential solution to capturing complementary, and possibly highly non-linear, information spread across data modalities—all data types are considered, and potentially used, when supervised training constructs a Random Forest-based kernel for genes that function together in a common cellular process. In contrast, MPL, FSMKL and other MKL approaches that employ unsupervised kernel construction compute segregated data-type specific kernels for each pathway and let the linear combination "meta-kernel" determine their optimal combination. This may result in sub-optimal solutions—features that belong to the

same pathway but are in different data types can now only interact with each other on the kernel level and not individually. AKLIMATE's approach creates a richer interaction model that is flexible enough to capture same-gene, cross-gene, and cross-data type interactions.

Another limitation of current pathway-informed kernel learning methods is that a single informative feature can go undetected if only present in large pathways—if all member features contribute equally to kernel construction, the importance of the relevant feature is obscured by the non-relevant majority. In contrast, AKLIMATE's RF kernels effectively allow individual features to influence the model. This advantage is illustrated by the improvement in AKLIMATE performance over BMKL on the MSI prediction task. BMKL's Gaussian kernels treat each feature in a feature set as equally important in the computation of the respective kernel matrix. As MLH1 appears only once in the PID pathway compendium, its contribution is masked by less informative features. On the other hand, due to the supervised manner of their construction, AKLIMATE's RF kernels inherit an RF's ability to prioritize features based on their relevance to the classification task—informative features are by definition overrepresented among tree node splitting variables. As all components of the RF kernel are derived from properties of the RF trees, informative features exercise proportionately higher influence over the RF kernel construction. Therefore, if only a small subset of a pathway's features are truly relevant, they can be clearly distinguished from (thousands of) non-relevant ones. For example, both AKLIMATE and AKLIMATE-reduced pick MLH1 as the most informative feature (26.7% and 16.4% relative contribution respectively), with a steep decline in the importance of the next best feature (AKLIMATE—EPM2AIP1, 8% relative contribution; AKLIMATE-reduced—PARD6A, 4.8% relative contribution).

In addition to filtering out non-relevant genes from a gene set definition, an AKLIMATE model can also facilitate the repurposing of seemingly unrelated prior knowledge—e.g. a gene set collected from one study is predictive in another one where the connection between the two studies is not obvious. For example, in the UCEC MSI prediction task, AKLIMATE selected a gene set that is upregulated upon HPV infection in head-and-neck cancer. The connection between HPV infection (which knocks out TP53) and UCEC microsatellite instability involves the DNA damage and repair machinery. In this case, it makes sense that a gene set associated with DNA repair would be selected as a predictive feature of MSI status (even for a different tumor type). In other cases, the relationship between an informative feature and the prediction task itself may not be as obvious. As collections of gene sets continue to grow, especially due to contributions from less curated sources like the results of high-throughput studies (e.g. gene clusters from an RNA-Seq experiment), the number of gene combinations that could serve as predictive biomarkers increases—leveraging this form of prior knowledge will thus continue to grow in importance.

A further key advantage of AKLIMATE is its ability to accommodate variables that do not readily map to genome-based feature sets (e.g. clinical data). In cases such as METABRIC, where clinical features provide much of the predictive power, the most salient question is how to identify orthogonal genomic features conditioned on a set of given clinical data. Posing the problem this way reflects the practical situation in a hospital setting where clinical information is nearly always on-hand to treating physicians. AKLIMATE and FSMKL illustrate two different ways of incorporating such information. FSMKL treats each clinical variable as a feature set of size one and constructs a kernel for each of them individually. This approach is viable, but quite restrictive in how it models clinical variable interactions. While AKLIMATE can accommodate such a setup, it also permits a more complex representation of the way features interact—each clinical feature is of a special "global" type that gets included in every feature set. The "global"-"local" feature hierarchy allows flexibility in modeling interactions among clinical variables and between clinical variables and genomic features. Such a hierarchy is

necessary when features work on different biological scales—for example, tumor grade is an organ-level characteristic that captures a snapshot of the behavior of millions of cells and is therefore likely to have a different information content compared to the copy number status of an individual gene.

An important aspect of AKLIMATE's use of prior knowledge is its ability to identify relevant features even in cases where many confounders exhibit high collinearity. This problem is similar to the one encountered in genome-wide association studies where an allele conferring a phenotype of interest can exist in a large haplotype block containing the alleles of many other irrelevant "hitchhiking" genes. In such situations, prior knowledge often helps researchers select the true causal variant among potentially many false positives. Consider AKLIMATE's top two most informative features for the MSI prediction task—MLH1 and EPM2AIP1. EPM2AIP1 is on the DNA strand opposite to MLH1, shares a bi-directional CpG island promoter with it and can be concurrently transcribed [37]. The transcriptional profiles of the two genes are nearly identical (Fig 1B)—in the absence of other information it would be extremely difficult to prioritize the "driver" (MLH1) over the "passenger" (EPM2AIP1) using expression data alone. AKLIMATE's feature sets provide the necessary prior knowledge—while EPM2AIP1 is indeed deemed the second most informative feature, its relative contribution is over three times smaller than that of MLH1.

AKLIMATE's robustness to false positives is not limited to expression features—it can prioritize relevant genes even if they are subject to large-scale copy number events and thus have almost identical copy number profiles with many other genes. We observed this effect in the MDM4 knockdown prediction task (Fig 3C) as well as KRAS knockdown prediction with mutation data (Fig 3E). In the former, there is a clear large scale copy number event that involves 7 of the 50 most predictive genes, but ZMAT3 is given by far the highest weight because of the biological prior of the feature sets. Similarly, in the latter 9 copy number features have very similar profiles, but the KRAS one is prioritized as the most important. The KRAS GISTIC feature is also favored among a group of genes affected by large-scale events—an indicator that the robustness to collinearity extends beyond continuous to categorical features.

Making use of prior knowledge, such as the results of past experiments, is often a key component in the success of machine-learning applications in genomics analysis [17, 38]. AKLIMATE uses a biologically-motivated prior distribution on the feature space—as demonstrated, this approach often outperforms methods that use a uniform prior over input features. AKLIMATE updates its prior information in a data driven manner—therefore, the feature set compendium does not have to be tailored to the problem at hand, avoiding the need for problem-specific filtering heuristics. When high quality relevant prior experiments are available, AKLIMATE tends to perform better. For example, the most important feature set for MSI prediction in endometrial cancer is a previously published signature characterizing MSI in colon cancer (Fig 1B). From that perspective, AKLIMATE acts as a framework for prioritizing past experiments that are most relevant to the interpretation of a new dataset.

By aggregating feature weights within individual data types, AKLIMATE can also be used to rank data type contributions. For example, while expression is generally most important in predicting shRNA knockdowns (mean importance 71.6% across tasks), there are cases where copy number is more informative, such as PSMC2(61.7%), RPAP1(55%) and CASP8AP2 (47%) (S9 Fig). Furthermore, inferred protein activity varies tremendously in terms of its contribution—from $< 1\%$ in predicting PSMC2 knockdowns to 27.7% for STRN4 (S9 Fig). AKLIMATE's ability to zero in on information-rich data types could help in designing targeted future experiments.

From a stacked learning perspective, AKLIMATE augments standard level-one cross validated predictions with topological aspects of the base RFs—how often data points end up in

the same leaf node and whether they tend to end up in early-split or late-split leaves. Such augmentation improves performance—in all the cases we studied, AKLIMATE outperforms both the best individual base learner and the ensemble that averages base learner predictions (S1 Fig). In addition, AKLIMATE does better than a standard Super Learner (regularized linear regression meta-learner) although without achieving statistical significance on some of the data sets (S1 Fig). This suggests that propagating additional information from a base learner (beyond the predictions it makes) into the level-one data can lead to a more accurate meta-learner. While AKLIMATE provides a blueprint for tree-based algorithms, using other base learners might yield better results.

AKLIMATE requires minimal feature pre-processing, can query tens of thousands of feature sets and its main steps are trivially parallelizable. It performs as well as, or better than, state-of-the-art algorithms in a variety of prediction tasks. AKLIMATE can natively handle continuous, binary, categorical, ordinal and count data, and its feature sets are easily extendable. For example, gene expression data can be augmented by epigenetic measurements (e.g. DNA methylation or ATAC-Seq), mutations in promoters or enhancers, splice variant proportions, and so on. Furthermore, slight modifications to AKLIMATE could incorporate prior importance scores for features within a feature set as well as structured relationships such as known feature-feature interactions (e.g. transcription factor to target gene). To simplify the exposition, we have focused our derivations and applications to regression and binary classification problems. However, AKLIMATE is readily extendable to the multi-class setting and the current public code base provides such an implementation.

While AKLIMATE demonstrates enhanced predictive power, future improvements are certainly possible that could lead to a bigger performance boost. For example, a kernel that captures the topological distance between samples in RF trees as suggested in [39] could provide a more accurate replacement for the $K_2$ kernel we introduced. Sample-weighting schemes during kernel construction could be investigated as well. For example, [40] suggests a weighted RF kernel where sample contributions depend on the predictive accuracy of their assigned leaves or the classification error rate for an individual sample.

Finally, even though we present examples drawn from the field of bioinformatics, AKLIMATE can be applied to any task with multi-modal data and prior knowledge in the form of feature groups, particularly when the feature groups have evidence that span data types.

## Materials and methods

### Background

Before describing AKLIMATE in detail, we briefly summarize existing approaches to which AKLIMATE's design and performance can be compared. There are three main challenges inherent to bioinformatics algorithms for supervised and unsupervised learning—prior knowledge integration, heterogeneous data interrogation, and ease of interpretation. New methods try to address some aspects of these three challenges [41–43]. Several common approaches emerge for supervised learning tasks. A popular way of increasing interpretability is to train models with regularization terms that constrain the number of included features—sparse models are considered easier to interpret than ones containing thousands of variables. Common regularization penalties are the lasso [44] and the elastic net [45]. More sophisticated regularization schemes can control the model behavior at the feature set level (e.g. the group lasso (GL) [46] and the overlap group lasso (OGL) [47]), allowing the incorporation of prior knowledge in the form of feature sets. However, both have drawbacks that make them less suitable for biological analysis where genes often participate in multiple processes. GL requires that if a feature's weight is zero in one group, its coefficients in all

other groups must necessarily be zero. OGL tends to assign positive coefficients to entire feature sets, making it less suitable in situations where only some of a set's features are relevant.

Network-regularized methods [48, 49] represent a different approach where gene-gene interaction networks are used as regularization terms on the $L_2$-norm of feature weights. While they do use pathway-level information and are straightforward to interpret, they generally focus on an individual data type.

Multiple Kernel Learning (MKL) approaches [7, 16, 19, 50–52] can incorporate heterogeneous data by mapping each set of features through a kernel function and learning a linear combination of the kernel representations. Each kernel represents distinct sample-sample similarities providing flexible and powerful transformations to access either explicit or implicit feature combinations. To prevent overfitting, MKL methods generally include a regularization term—e.g., $L_1$ sparsity-inducing norm on the kernel weights [19] or the elastic net [51]. Prior knowledge can also be integrated by constructing individual kernels from a pathway's member features within each data type [7, 16, 28, 31, 52]. Indeed, MKL methods with prior knowledge integration [9, 28] have won several Dialogue on Reverse-Engineering Assessment and Methods (DREAM) [53] challenges, including a predecessor of the approach described here [28, 31]. Nevertheless, MKL suffers drawbacks when the contributions of individual input features need to be evaluated. Except in trivial cases, it is generally impossible to assign importance to the original features once the method is trained in the kernel function feature space, thus limiting interpretability of solutions. In addition, feature heterogeneity necessitates the construction of separate kernels for each data type, limiting the ability of MKL to capture cross-data-type interactions.

All of these methods can be used as components in more complex ensemble learning models. Ensembles combine predictions from multiple algorithms into a more robust, and often more accurate, "wisdom of crowds" final prediction. The simplest and most common ensemble technique is averaging the predictions of component models. Averaging over uncorrelated models or models with complementary information can improve performance—it is one of the main reasons for the emergence of collaborative competitions such as the DREAM challenges [9, 54]. In fact, such ensemble methods often win DREAM challenges [18] or outperform competitors in genomic prediction tasks [55]. While they provide a boost in predictive accuracy, their interpretation is quite challenging—ensembles often combine different model types, making the computation of input feature importance impossible in the large majority of cases. A notable exception is the Random Forest (RF) [56]—an ensemble of decision trees that has been widely applied to bioinformatics problems.

A more general approach to combining the predictions of multiple models is model stacking [57]. In stacking, component (base) learner predictions are used as inputs to an overall (stacked) model which produces the final calls. To reduce overfitting and improve generalization, base learner predictions are generated in a cross-validation manner—a sample's predicted label comes from a model trained on all folds except the one that includes the sample in question. Importantly, if the stacked model is a weighted combination of the base models (a Super Learner), it is asymptotically guaranteed to perform at least as well as the best base learner or any conical combination of the base learners [58, 59]. Stacked models exhibit identical pros and cons as standard ensemble models, with diminished interpretability traded off for increased accuracy. In some cases, however, interpretability can be tractable—e.g. [60] uses RF base learners with a stacked least square regression to compute a final weighted average of the base RF predictions. Although not examined in [60], we demonstrate that a similar setup can lead to an intuitive derivation of feature importance scores.

## Overview

AKLIMATE is a stacked learning algorithm with RF base learners and an MKL meta-learner. Each base learner is trained using only the features from a pre-defined feature set representing a known biological concept or process—e.g. a biological pathway, a chromosomal region, a drug response or shRNA knockdown signature, or disease subtype biomarkers (Fig 4). Each feature in a feature set can be associated with multiple data modalities—i.e. a gene will have individual features corresponding to its copy number, mutation status, mRNA expression level, or protein abundance. Furthermore, although feature sets normally consist of genes, they are easily extendable to a more comprehensive membership. For example, they can be augmented with features for mutation hotspots, different splice forms, or relevant miRNA or methylation measurements. AKLIMATE's MKL meta-learner trains on RF kernel matrices (see the *RF Kernel* section), each of which is derived from a corresponding base learner (Fig 5). An RF kernel captures a proximity measure between training samples based on the similarity of their predicted labels and decision paths in the trees of the RF. The MKL learning step finds the optimal weighted combination of the RF kernels. The MKL optimal solution can be interpreted as the meta-kernel associated with the most predictive meta-feature set derived from all interrogated feature sets.

The key contributions of AKLIMATE are

1. The introduction of an integrative empirical kernel function that combines similarity in predicted labels with proximity in the space of RF trees (RF kernel).
2. The extension of kernel learning to a stacking framework. To our knowledge, AKLIMATE is the first stacked learning formulation to incorporate base kernels.

In the next sections we give a detailed description of each AKLIMATE component.

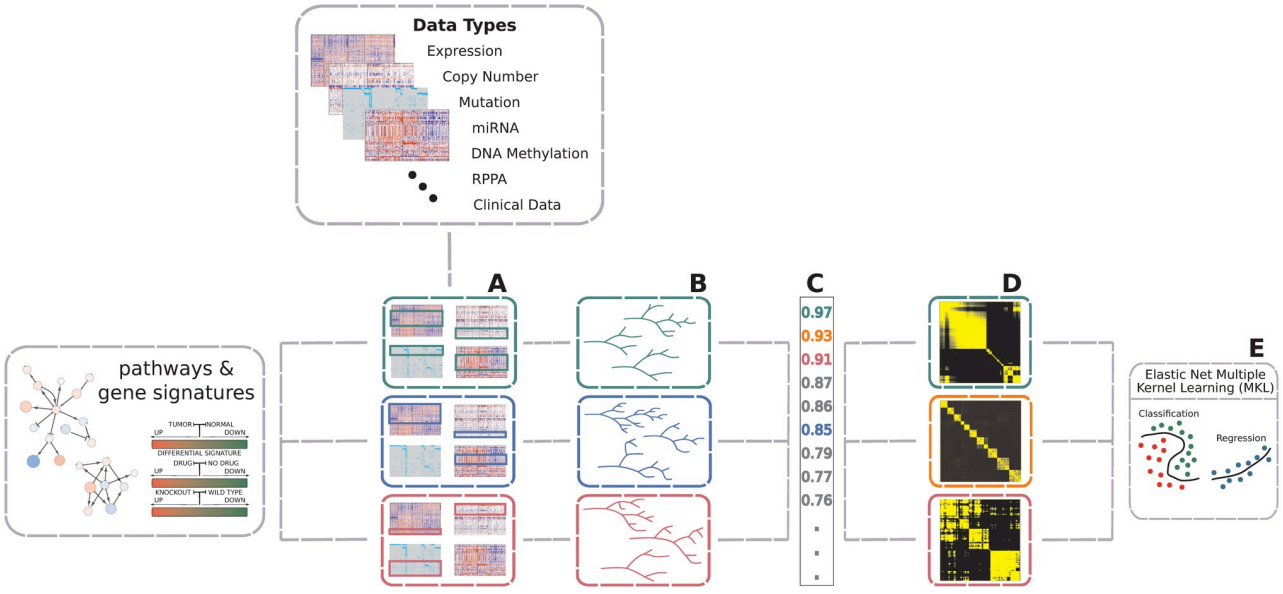


**Fig 4. Overview of AKLIMATE.** AKLIMATE takes as inputs multiple data types and a collection of feature sets. (A) Each feature set is overlapped with the available data types to identify all features that map to it (some feature sets might not extract features from all data types). (B) A feature-set-specific Random Forest model is trained on the multi-modal data identified in A. (C) RF models are ranked based on their predictive performance (metrics for mock RF models in A have matching color). (D) The top ranked RF models are converted to RF kernels (mock kernels shown for RF models with top 3 performance scores). (E) An elastic net MKL meta-learner finds the optimal combination of RF kernels and computes the final predictions. Elastic net hyperparameters are optimized via cross-validation.

https://doi.org/10.1371/journal.pcbi.1008878.g004

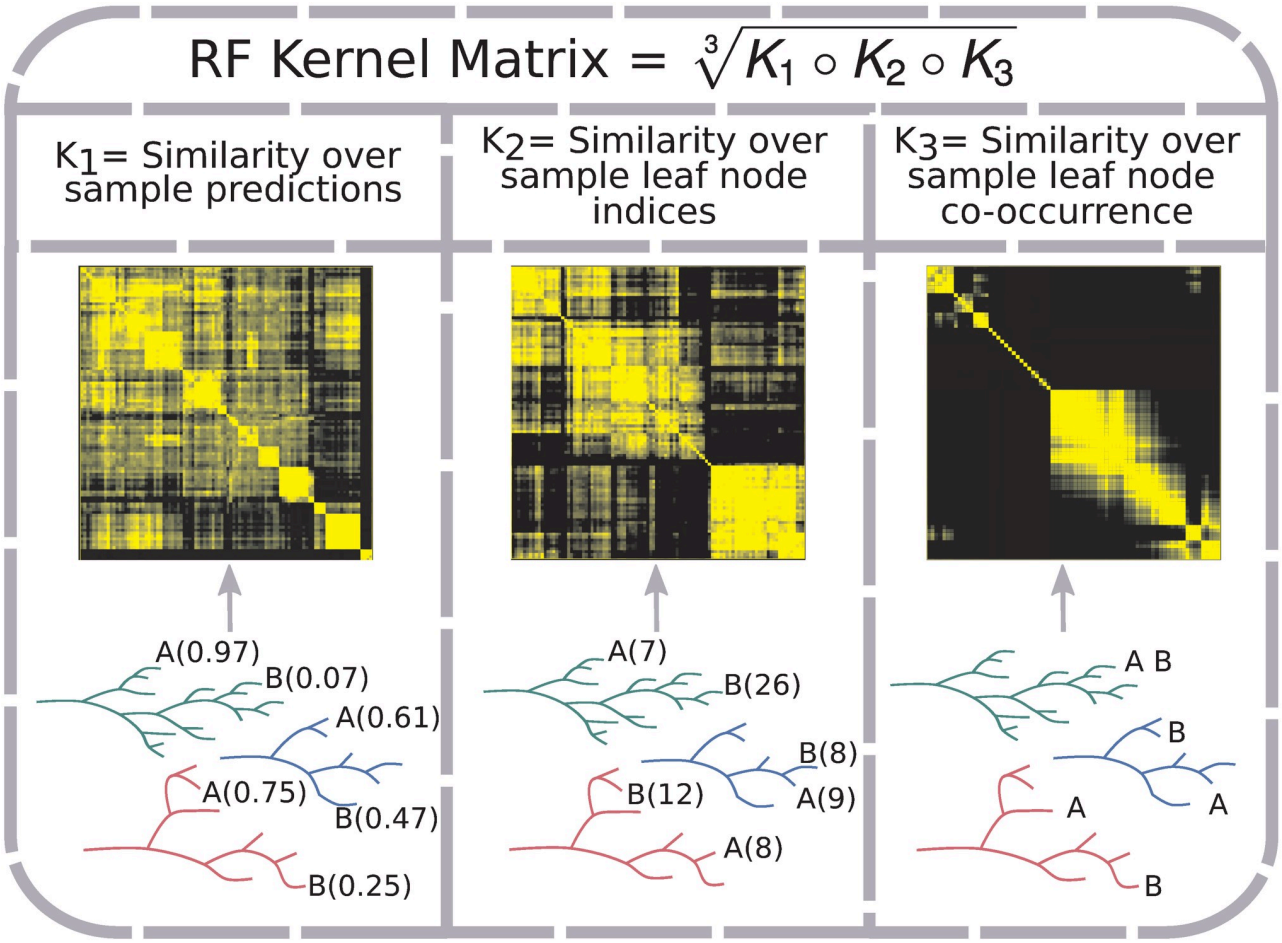


**Fig 5. RF kernel matrix construction.** The RF Kernel matrix is a geometric mean of the Hadamard product of three component similarity matrices. Each component captures a different aspect of a RF model. $K_1$ captures the similarity over RF tree predictions for sample labels (in the case of classification, probability of belonging to a class); two samples (A and B) show different predicted probabilities of belonging to the positive class (probability estimate in parentheses). $K_2$ represents the similarity over RF tree leaf indices to which samples are assigned; predicted leaf indices shown in parentheses for the two samples. $K_3$ reflects the proportion of times samples are assigned to the same RF tree leaf; e.g. the two samples end up together in one out of three example trees.

https://doi.org/10.1371/journal.pcbi.1008878.g005

## Random forest

An RF learner is an ensemble of decision trees each of which operates on a perturbed version of the training set. The perturbation is achieved by two randomization techniques—on one hand, each tree trains on a bootstrapped data set generated by drawing samples with replacement. On the other hand, at each node in a decision tree, only a subset of all features are considered as candidates for the next split. Tree randomization tends to reduce correlation among predictions of individual trees at the expense of increased variance of error estimates [61]. However, due to the averaging effect of ensembles of decorrelated models, it generally reduces the overall error variance of the ensemble RF [61]. That effect is partially negated by an increase in prediction bias, but broadly speaking the more decorrelated the trees the better the RF performance [61].

As each tree is constructed from a subsample of the training data, there is a tree-specific set of excluded, or out-of-bag (OOB) samples. The OOB predictions for the training set of a regression task are computed by walking each sample along the subset of trees in which that

sample is OOB and averaging their predictions:

$$\hat{f}(X_i) = \frac{1}{|OOB_i|}\sum_{t \in OOB_i} f^t(X_i)$$
$$OOB_i = \{t :\in OOB(Tree_t); t = 1 \ldots T\}, \quad (1)$$

where $X_i$ are the input features associated with sample $i$, $OOB_i$ is the set of trees in which sample $i$ is OOB, $T$ is the total number of trees in an RF and $f^t(X_i)$ is the classification of the $i^{th}$ sample in tree $t$. Similarly, for classification tasks the averaging of tree OOB predictions is replaced by majority vote.

It has been shown that the OOB error rate provides a very good approximation to generalization error [56]. For that reason, AKLIMATE uses RF kernels based on OOB predictions as level-one data for the MKL meta-learner training (see *RF Kernel* and *Stacked Learning* below).

## Kernel learning

Kernel learning is an approach that allows linear discriminant methods to be applied to problems with non-linear decision boundaries [62]. It relies on the "kernel trick"—a transformation of the input space to a different (potentially infinite dimensional) feature space where the train set samples are linearly separable. The utility of the "kernel trick" stems from the fact that the feature map between the input space and the feature space does not need to be explicitly stated—if the kernel function is positive definite (i.e. it has an associated Reproducing Kernel Hilbert Space), we only need to know the functional form of the feature space dot product in terms of the input space variables [62, 63]. A kernel function represents such a generalized dot product (**Note**: We define positive definiteness as in [62]—a kernel is positive definite if $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$ for $\forall x_i, x_j \in \mathcal{X}$ and $c_1, \ldots, c_n \in \mathbb{R}$).

Commonly used kernel functions generally have an explicit closed form, such as a polynomial or a radial basis function. Kernel functions that encode more complex object relationships also exist, particularly if the objects can be defined in a recursive manner (ANOVA kernels, string kernels, graph kernels, see [64]). However, knowing the closed form kernel function is not a necessary condition—if the kernel positive definiteness can be verified and the kernel matrix of pairwise dot products (similarities) can be computed, we can still utilize the "kernel trick". In fact, the Representer Theorem [62, 65] guarantees that the optimal solution to a large collection of optimization problems can be computed by kernel matrix evaluations on the finite-dimensional training data, ensuring the applicability of kernel-based alogirthms to real-world problems. More formally, for an arbitrary loss function $L(f(x_1)\ldots f(x_N))$, the minimizer $f^*$ of the regularized risk function

$$R(f) = L(f(x_1)\ldots f(x_N)) + \Omega(\|f\|_H^2) \quad (2)$$

can be expressed as

$$f^* = \sum_{i=1}^{N} \alpha_i k(x_i, ), \alpha_i \geq 0 \quad (3)$$

provided that $k(\cdot, x)$ is a positive definite kernel and the regularization term $\Omega(\|f\|_H^2)$ is a monotonically increasing function.

AKLIMATE uses the dependency structure of a RF trained on a (possibly multi-modal) feature set to define an implicit empirical kernel function. It then computes an RF kernel matrix (see *RF Kernel* below) of pairwise training set similarities to use in a kernel learning algorithm. As different feature sets can contribute complementary information for a given prediction

task, AKLIMATE utilizes a Multiple Kernel Learning approach to integrate the available RF kernels into an optimal predictor.

## Multiple kernel learning

Kernel learning can lead to large improvements in accuracy, provided that an optimal kernel is selected. That choice can be difficult, however—the best kernel function for a multi-modal data set may be non-obvious, may entail tuning several hyperparameters, or may even lack a closed form expression. MKL addresses the problem of kernel selection by constructing a composite kernel that is a data-driven optimal combination of candidate kernels [50, 66–68]. Linear combinations of kernels of the form $K(\alpha) = \sum_{i=1}^{M} \alpha_i K_i$ with either a conical ($\forall i, \alpha_i \geq 0$) or convex ($\forall i, \alpha_i \geq 0$ and $\sum_{i=1}^{M} \alpha_i = 1$) sum constraint on the kernel weights are by far the most commonly used because of their many desirable properties (although algorithms that allow non-linear combinations do exist—see [67]). Some important advantages are:

1. Conical/convex combinations of positive definite kernels are positive definite.
2. The composite kernel is associated with a feature space that is the concatenation of all individual kernel feature spaces.
3. The Representer Theorem is readily extendable to the the conical/convex combination case —the optimal solution $f^*$ takes the form $f^* = \sum_{m=1}^{M} \sum_{i=1}^{N} k_m(x, x_i)\alpha_{m,i}$ [51, 62].

MKL algorithms admit different forms of regularization depending on what norm of the kernel weights is chosen. AKLIMATE uses an elastic-net [45] regularizer on the norm of the individual kernel weights of the form $\lambda_1 \sum_{i=1}^{M} \|\alpha_m\|_{K_m} + \lambda_2 \sum_{i=1}^{M} \|\alpha_m\|_{K_m}^2$, where $\lambda_1, \lambda_2 \geq 0$, $\alpha_m = (\alpha_{m,1} \ldots \alpha_{m,N})^T$, and $\|\alpha_m\|_{K_m} = \sqrt{\alpha_m K_m \alpha_m}$ is the definition of a kernel norm as in [51] and [69]. The elastic net regularization provides the flexibility to find both sparse and dense solutions with appropriately tuned λ's—i.e. $\lambda_1 >> \lambda_2$ gives few non-zero kernel weights (sparse) while $\lambda_1 < \lambda_2$ gives many non-zero kernel weights (dense). Also note that $\lambda_1 = 0$ and $\lambda_2 > 0$ allows for a uniform kernel weight solution.

The explicit form of AKLIMATE's optimization problem is:

$$\min_{\alpha \in \mathcal{R}^{NM}, b \in \mathcal{R}} \sum_{i=1}^{N} \mathcal{L}(y_i, \sum_{m=1}^{M} \sum_{j=1}^{N} k_m(x_i, x_j)\alpha_{m,j} + b) + \lambda_1 \sum_{i=1}^{M} \|\alpha_m\|_{K_m} + \lambda_2 \sum_{i=1}^{M} \|\alpha_m\|_{K_m}^2 \quad (4)$$

and is solved using the SpicyMKL algorithm described in [51]. The optimal kernel weights are recovered by:

$$w_m = \begin{cases} 0 & (\|\alpha_m^*\|_{K_m} = 0), \\ \frac{\|\alpha_m^*\|_{K_m}}{\lambda_1 + \lambda_2 \|\alpha_m^*\|_{K_m}} & (\text{otherwise}) \end{cases} \quad (5)$$

where $\alpha^*$ is the solution to Eq (4). The model weights are re-scaled to satisfy $\sum_{i=1}^{M} w_i = 1$.

## RF kernel

The most common approach to kernel matrix evaluation is to specify an explicit data dependency model for the kernel function—e.g. polynomial, Gaussian, ANOVA or graph kernels [64]. However, choosing the dependency structure a priori can lead to lack of robustness, particularly when it is not obvious what the right dependency structure is. AKLIMATE uses a RF to create an empirical approximation of the true dependency model and to evaluate the

(implicit) kernel function associated with it (the RF kernel). RF kernels are robust to overfitting, generalize well to new data, and facilitate the integration of signals across data types.

Defining a kernel through a RF is not a new idea—in fact the concept was introduced at the same time as RFs [70]. In particular, for a random forest of trees with an equal number of leaves and uniform predictions in each leaf, a positive definite kernel can be defined based on the probability two samples share a leaf [70]:

$$K(x_i, x_j) = \lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T} I(i, j \in R_t(\theta_m)), \tag{6}$$

where $M$ is the number of trees in the forest, $T$ is the number of leaves in a tree, $\theta_m$ is a variable capturing the random selection of training set samples and features in the process of constructing the $m$th tree, $R_t$ is the $t$th leaf of the $m$th tree, and $I(\cdot)$ is the indicator function. The finite approximation of Eq (6) ($M < \infty$; Fig 5, $K_3$ kernel) is positive definite [71]. Kernels generated from RFs and their theoretical properties are also discussed in [72] and [73].

AKLIMATE's RF kernel extends the original definition (Eq (6)) by incorporating two additional RF-derived statistics. The intuition behind the first one is that two samples predicted to have the same label within a tree should be considered alike even if they fall into different leaves (Fig 5, $K_1$ kernel). The reasoning for the second one is that earlier node splits in a tree generally separate more distinct sample groups, while later splits tend to fine tune the decision boundary, highlighting more subtle differences. Thus, irrespective of the predicted labels, two samples that end up in different tree depths—one early- and one late-split leaf—should be considered less similar than samples landing in two late-split leaves (Fig 5, $K_2$ kernel). Incorporating these three patterns, AKLIMATE's RF kernel matrix is computed using the following steps:

1. Calculate similarity over predictions across RF trees:

$$K_1(x_i, x_j) = \exp\left(-\frac{\|p_i - p_j\|^2}{\sigma}\right), \tag{7}$$

   where $p_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,M})$ is a vector of predictions for data point $i$ from the $M$ trees in the RF. $p_i$ always has continuous entries—either the actual predictions in a regression setting, or the probabilities of class membership for classification problems. The $\|p_i - p_j\|^2$ distances are divided by the scaling constant $\sigma = \max_{i,j}\|p_i - p_j\|^2$ so that they map to the [0, 1] range. Exponentiation of the negative distances converts them to similarities.

2. Compute similarity over leaf node indices across RF trees:

$$K_2(x_i, x_j) = \exp\left(-\frac{\|t_i - t_j\|^2}{\sigma}\right), \tag{8}$$

   where $t_i = (t_{i,1}, t_{i,2}, \ldots, t_{i,M})$ is a vector of the leaf indices of sample $i$ across the RF. Tree nodes are indexed starting from the base node and then sequentially across each depth level of the tree. The scaling constant $\sigma$ is set in the same manner as in $K_1$.

3. Calculate similarity over leaf node co-occurrence using the finite approximation of Eq (6) with variable number of leaves per tree:

$$K_3(x_i, x_j) = \exp\left(\left(\frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T_m} I(i, j \in R_t(\theta_m))\right) - 1\right), \tag{9}$$

   where the exponential transformation is added to keep $K_3$ on a similar scale to $K_1$ and $K_2$.

4. Calculate the RF kernel as the geometric mean of the element-wise product of $K_1$,$K_2$ and $K_3$:

$$K(x_i, x_j) = \sqrt[3]{K_1(x_i, x_j) \circ K_2(x_i, x_j) \circ K_3(x_i, x_j)}. \quad (10)$$

Importantly, since the components $K_1$, $K_2$, and $K_3$ are positive definite, so too is $K$. $K_1$ and $K_2$ are Gaussian kernels over $\mathcal{P} \times \mathcal{P}, \mathcal{P} \subseteq \mathcal{R}^m$ and $\mathcal{T} \times \mathcal{T}, \mathcal{T} \subseteq \mathcal{I}^m$ respectively, which ensures their positive definiteness [64]. $K_3$ is the exponent of a positive definite kernel, i.e. a positive definite kernel as well [64]. Finally, $K$ is the Hadamard product of three positive definite kernels raised to a positive power—both of these operations preserve positive definiteness [64].

## Stacked learning

Stacked learning is a generalization of ensemble learning in which the stacked model (meta-learner) uses the prediction output of its components (base learners) as training data for the computation of the final predictions [57, 74]. What makes stacked learning unique is that the base learner predictions (level-one data) are generated in a cross-validated manner that excludes each sample from the training set that produces its predicted label. More specifically, if our training set (level-zero data) is $X = \{X_i : i = 1 \ldots N, X_i \in \mathcal{R}^p\}$ with labels $Y = \{Y_i : i = 1, \ldots, N, Y_i \in \mathcal{R}\}$ and we have a collection of base learners $\boldsymbol{\Delta} = \{\Delta_1, \ldots, \Delta_S\}$ then the stacked generalization proceeds as follows [59, 75]:

1. Randomly split the level-zero data into $V$ folds of roughly equal size—$h_1, \ldots, h_V$ ($V$-fold cross validation). Note that each $h_v$ defines a set of indices that select a subset of the samples in $X$.
2. For each $\Delta_s$ base learner, train $V$ models, collectively denoted as $\hat{\Delta}_s = \{\hat{\Delta}_{s,1} \ldots \hat{\Delta}_{s,V}\}$. Each model $\hat{\Delta}_{s,v}$ uses $X \setminus X_{h_v}$ to train on and generates predictions for $X_{h_v}$.
3. Concatenate the $V$ sets of predictions from $\hat{\Delta}_s$ into a vector $z$ of length $N$. The $N \times S$ matrix $Z$ of such vectors for all $S$ base learners becomes the feature matrix for level-one training.
4. Train a meta-learner $\Psi$ on $Z$ with hyperparameter tuning if necessary, again using $Y$ as the labels.
5. The base learners used in the final stacked model are created using all training samples, so one final (non-cross-validated) training round is needed. To this end, train each $\Delta_s$ base learner on the full level-zero training set $X$.
6. Form the stacked model using the base-learners trained on the full data set and the meta-learner to obtain $(\{\Delta_s: s = 1, \ldots, S\}, \Psi)$. To predict on a new data point $X_{new}$, compute a $1 \times S$ vector $Z_{new} = \{\Delta_s(X_{new}): s = 1, \ldots, S\}$ and use $\Psi(Z_{new})$ as the stacked model's prediction.

The predictive performance of the meta-learner is generally improved when the base learner predictions are maximally uncorrelated. This is achieved either by using different algorithms, or by varying the parameters of a particular modeling approach [58, 74]. AKLIMATE generates diversity through the use of feature subsets built around distinct biological concepts and processes.

**Super learner.** Stacked learning imposes no conditions on the choice of meta-learner $\Psi$. The disadvantage of such flexibility is the lack of theoretical results for the improved empirical performance of stacking. A Super Learner [58] is a type of stacked learner with restrictions on $\Psi$ that give provable desirable properties. The main such constraint is that the optimal meta-

learner $\Psi^*$ is the minimizer of a bounded loss function. A Super Learner for which $\{\Delta_1, \ldots, \Delta_S\}$ and $\Psi$ have uniformly bounded loss functions exhibits the "oracle" property—$\Psi^*$ is asymptotically guaranteed to perform as well as the optimal base learner $\Delta_s^*$ under the true data-generating distribution [76, 77]. Furthermore, if we constrain the choice of $\Psi$ to $\Psi = \sum_{i=1}^{S} \alpha_i \Delta_i, \forall \alpha_i \geq 0$, $\Psi^*$ asymptotically converges to the performance of the optimal conical combination of $\{\Delta_1, \ldots, \Delta_S\}$ [58, 59]. This is the main reason why $\Psi$ often takes the form of regularized linear or logistic regression.

For example, if the aim is to predict a continuous variable, one can set $\Psi = \sum_{i=1}^{S} \alpha_i \hat{\Delta}_i$ and solve the regression problem:

$$\min_{\alpha} \sum_{i=1}^{N} (Y_i - \sum_{j=1}^{S} \alpha_j \hat{\Delta}_j(X_i))^2, \tag{11}$$

which can be regularized or subjected to a convex sum constraint on the $\alpha$ weights (e.g. $\forall \alpha_i \geq 0, \sum_{i=1}^{S} \alpha_i = 1$) [58, 59]. Similarly, the squared error loss of Eq (11) can be replaced with the logistic loss to solve a classification problem:

$$\min_{\alpha} \sum_{i=1}^{N} \log(1 + \exp(-Y_i * (\sum_{j=1}^{S} \alpha_j \hat{\Delta}_j(X_i)))) \tag{12}$$

maintaining all theoretical Super Learner results.

## AKLIMATE

To our knowledge, AKLIMATE is the first instance of a kernel-based stacked learner. The base learners $\{\Delta_1, \ldots, \Delta_S\}$ are RFs, each of which is used to produce an associated kernel. The meta-learner $\Psi$ is an elastic-net regularized MKL that can be interpreted as the kernel learning counterpart of linear regression. Before describing the algorithm, we first discuss how the level-one RF kernels are constructed using OOB samples.

**AKLIMATE level-one (OOB) kernel construction.** AKLIMATE uses RF kernels as the level-one training data. Normally, level-one data is generated with cross-validation to help the meta-learner avoid overfitting. Analogously, AKLIMATE utilizes out-of-bag (OOB) samples to generate the components of the RF kernels. For each pair of samples in the RF, the kernel similarity matrix is calculated using only those trees for which both of the samples have been withheld, with the following procedure:

1. For a RF with $M$ trees, define $OOB(m)$ as the set of samples that are OOB in the $m$th tree. Let $I_{OOB}$ be the tree-level indexing function recording when both samples $i$ and $j$ are simultaneously OOB in a given RF; i.e.:

$$\begin{aligned} I_{OOB}(i,j) &= (\gamma_{1ij}, \ldots, \gamma_{mij}), \text{ where} \\ \gamma_{mij} &= \begin{cases} 1 & i,j \in OOB(m), \\ 0 & \text{otherwise} \end{cases}. \end{aligned} \tag{13}$$

2. Compute the first constituent kernel matrix:

$$K_1(x_i, x_j) = \exp\left(-\frac{\| \langle p_i, I_{OOB}(i,j) \rangle - \langle p_j, I_{OOB}(i,j) \rangle \|^2}{\sigma}\right). \tag{14}$$

3. Compute the second constituent kernel matrix:

$$K_2(x_i, x_j) = \exp\left(-\frac{\left\|\langle t_i, I_{OOB}(i,j)\rangle - \langle t_j, I_{OOB}(i,j)\rangle\right\|^2}{\sigma}\right) . \quad (15)$$

4. Compute the third constituent kernel matrix:

$$K_3(x_i, x_j) = \exp\left(\left(\frac{1}{\sum I_{OOB}(i,j)} \sum_{m=1}^{M} \sum_{t=1}^{T_m} I(i,j \in R_t(\theta_m)) I_{OOB}(i,j)_m\right) - 1\right) . \quad (16)$$

5. Calculate the combined kernel matrix:

$$K(x_i, x_j) = \sqrt[3]{K_1(x_i, x_j) \circ K_2(x_i, x_j) \circ K_3(x_i, x_j)} , \quad (17)$$

with the same notation and $\sigma$ calculations as in Eqs (7)–(10).

AKLIMATE's MKL meta-learner has two elastic-net hyperparameters ($\lambda_1$, $\lambda_2$) that require tuning. This is done by generating a random set of ($\lambda_1$, $\lambda_2$) pairs and ranking them based on $V$-fold (default $V = 5$) cross-validation fit with OOB RF kernels as input. To improve generalization, we use a simplified version of the overfit correction procedure in [78]—instead of selecting the hyperparameters that produce the best CV fit, we choose the ones corresponding to the 90th percentile of the distribution of the CV fit metric.

**AKLIMATE algorithm.** We next describe AKLIMATE's learning algorithm. The method takes training data $(X, Y) = \{(X_i, y_i) : X_i = X_{i1} \cup \cdots \cup X_{id}, i = 1 \ldots N, d = 1 \ldots D; y_i \subset \mathcal{Y}\}$, with $X$ a tensor of $N$ samples and $D$ data types with feature memberships $C = \{C_i\}_{i=1}^{D}$ respectively, and training labels $Y$. In addition, $S$ feature sets (e.g. pathways) are supplied, each containing a list of features $P_s$ such that the complete set is $P = \{P_s\}_{s=1}^{S}$. The algorithm outputs a meta-learner, $\Psi^*$ and a set of selected base learners $\Delta^*$. In addition, a user-supplied parameter $G$ determines the number of top feature sets (pathways) to incorporate for each sample during the base learning step. We use $G = 5$ in practice.

Formally, AKLIMATE operates according to the pseudocode in Algorithm 1. First AKLIMATE trains a separate RF for each feature set using data from all modalities relevant to the features in the set. Model accuracy $\tau$ is stored so that the top $L$ models (by $\tau$) can be selected that correctly predict an individual sample. A final set of relevant RFs $\Delta^*$ is obtained by taking the union over all sample-specific top $L$ models. The helper function RFKERNELOOB creates a kernel from a given RF utilizing the out-of-bag approach described previously. MKL uses these level-one kernels to tune the elastic net hyperparameters $\lambda_1$ and $\lambda_2$ based on cross-validation performance. A final meta-learner $\Psi^*$ is then trained by running MKL with kernels constructed from the full RFs (RFKERNEL helper function) and with the determined elastic net hyperparameters.

**Algorithm 1**: AKLIMATE

```
Input: (X, Y), C, P, G
Output: Ψ*, Δ*
for s ∈ S do
  P̂_s ← ∪_{d=1}^{D}(P_s ∩ C_d);              // list of available features
  RF_s ← RFTRAIN(X_{P̂_s}, Y);                // Train respective base learner
  Ŷ^s ← RF_s(X_{P̂_s});                       // Compute OOB predictions
  τ_s ← FIT(Y, Ŷ^s);          // fit statistic based on OOB predictions
end
for n ∈ N do
```

$\Delta_n^* \leftarrow \{RF_{i_g} : RF_{i_g}(X_n) = Y_n \ \ldots \tau_{i_1} \geq \tau_{i_2} \cdots \geq \tau_{i_g} \geq \max(\{\tau_s\}_{s=1}^S \setminus \{\tau_{i_1}, \ldots, \tau_{i_g}\})\}_{g=1}^G$; // pick top $G$ (ranked by $\tau_s$) RFs predicting sample $n$ correctly

**end**

$\Delta^* \leftarrow \cup_{n=1}^N \Delta_n^*$

$K_{oob} \leftarrow \{\}$

**for** $RF_r \in \Delta^*$ **do**

  $K_{oob} \leftarrow K_{oob} \cup$ RFKERNELOOB($RF_r$, $X$)

**end**

$(\lambda_1^*, \lambda_2^*) \leftarrow \text{argmax}_{\lambda_1,\lambda_2}\{\text{CV}(\text{MKL}(K_{oob}, \lambda_1, \lambda_2, Y))\}$

$K_{full} \leftarrow \{\}$

**for** $RF_r \in \Delta^*$ **do**

  $K_{full} \leftarrow K_{full} \cup$ RFKERNEL($RF_r$, $X$)

**end**

$\Psi^* \leftarrow \text{MKL}(K_{full}, \lambda_1^*, \lambda_2^*, Y)$; // Train MKL meta-learner

**AKLIMATE selection of relevant RFs.** AKLIMATE's selection step for the best RFs $\Delta^*$ in Algorithm 1 can filter the full collection of feature sets down to a subgroup of relevant sets two or more orders of magnitude smaller in size. This makes it possible to incorporate appreciably more feature sets than standard MKL algorithms which require kernel evaluation for all feature sets. When the number of such sets is in the thousands, MKL can be computationally very slow, even for data sets of small sample size. However, usually only a small proportion of the feature set collection is truly explanatory for a given prediction task. Thus, filtering out the non-relevant parts of the compendium does not impact MKL accuracy yet drastically improves computation time.

Algorithm 1 demonstrates the $\Delta^*$ discovery process for a classification task. However, if $Y$ is continuous (i.e. a regression problem), the predictions and labels are not directly comparable for equality. One approach is to take the squared error of the prediction-label differences and use that metric to re-rank RFs for each sample. In our experience, this leads to the selection of suboptimal RFs due to overfitting. Instead, AKLIMATE uses a more robust scheme that shows better results in practice—the vector of predictions for each sample across all RFs $\{\hat{Y}_n^s\}_{s=1}^S$ is binarized into matching and non-matching predictions and then the standard classification case selection rule is applied. The binarization is done as follows:

$$\hat{Y}_{n_{bin}}^s = \begin{cases} 1 & \text{if } |\hat{Y}_n^s - Y_n| \leq quantile(\{|\hat{Y}_n^s - Y_n|\}_{s=1}^S, q), \\ 0 & \text{otherwise} \end{cases}, \tag{18}$$

where $q$ is a user-specified quantile of the empirical distribution of absolute prediction errors $\{|\hat{Y}_n^s - Y_n|\}_{s=1}^S$ (default $q = 0.05$). This setup prioritizes RFs that perform near-optimally on individual data points and optimally when the training set is considered as a whole.

**AKLIMATE importance weighting of individual features and feature sets.** Feature set weights $w^{FS}$, $\sum_i w_i^{FS} = 1$, are recovered directly from the optimal MKL meta-learner $\Psi^*$(Eq (5)). Feature weights can be calculated as $w_i^F = \sum_{k \in \hat{P(\cdot)}_i} w_k^{FS} m(RF_k, i)$ where $\hat{P(\cdot)}_i = \{\hat{P}_s : i \in \hat{P}_s \ \& \ RF_{\hat{P}_s} \in \Delta^*, s = 1, \ldots, S\}$ is the set of all feature sets that have feature $i$ as its member and their associated RF was selected among the set of best RFs $\Delta^*$, while $m(RF_k, i)$ is an $RF$-specific feature importance score computed from the RF associated with the $k$th such feature set.

The simplest way to compute $m(RF_k, i)$ is by averaging the improvement in the splitting criterion over all nodes that used feature $i$ as a splitting variable. For the often used Gini impurity measure, this involves computing the mean difference in impurity before and after each split, with larger mean impurity decreases indicative of more important variables [79]. While fast,

this rule suffers from important shortcomings—for example, it is biased in favor of variables with more potential split points (e.g. continuous or categorical with a large number of categories), particularly when trees train on bootstrapped data [80].

Many alternative $m(RF_k, i)$ rules have been proposed [80–82]. For our work, we choose as default the permutation-based importance calculation described in the original RF paper [56] —the vector of measurements for feature $i$ is randomly permuted and the permuted variable is used in the calculation of OOB predictions; the difference in error rate between the permuted and non-permuted OOB predictions is taken as a measure of feature $i$'s importance. Permutation-based importance is robust and generally performs on par with more complex $m(RF_k, i)$ rules. Its biggest drawback is the higher computational cost. In cases where compute time is the main constraint, we recommend the actual impurity reduction (AIR) metric [83], which is an extension of the pseudodata-augmented approach in [81]. It is similar in speed to the Gini impurity importance, but retains the desirable properties of permutation-based methods. While AIR can lead to a small prediction accuracy penalty, in our experience this effect has been negligible for classification tasks (for regression problems we still recommend permutation-based importance).

## Supporting information

**S1 Text. Technical details of the experiments and the implementation of the AKLIMATE algorithm.**
(PDF)

**S1 Table. Most informative feature sets for breast cancer survival prediction in METABRIC data.** AKLIMATE weights were averaged over 50 train/test splits. The table lists the 20 most relevant feature sets, out of 1836 feature sets with a non-zero weight in at least one train/test split. Weights were normalized to sum to 1. Aliases for the top 10 feature sets are included in brackets—they provide a more descriptive name for the underlying biological function, based on information gathered from the source publication. The aliases are used in Fig 2 and throughout the text.
(PDF)

**S1 Fig. AKLIMATE versus alternative ways of ensembling/stacking component RFs.** To ensure fair comparison, the component RF set for each prediction task was restricted to AKLIMATE's best RFs $\boldsymbol{\Delta}^*$. Ensemble superlearner component RFs—learning a regularized linear regression on predictions from component RFs; ensemble average component RFs—taking the average of the predictions of component RFs; top component RF—only using predictions from the top ranked RF. (A) UCEC TCGA MSI prediction. (B) METABRIC Breast Cancer survival prediction. (C) Achilles shRNA knockdown prediction. P-values for paired Wilcoxon signed rank test.
(EPS)

**S2 Fig. AKLIMATE results for the AKLIMATE-reduced model (using only 196 PID pathways) on UCEC TCGA cohort.** Top 10 predictive feature sets and top 50 predictive features from 50 stratified 75% train/25% test splits are shown. Organized as Fig 1.
(EPS)

**S3 Fig. AKLIMATE performance on predicting MSI-High vs MSI-Low+MSS in TCGA COADREAD cohort.** AUC computed for 50 75%/25% stratified train/test splits. P-values for paired Wilcoxon signed rank test. Methods as in Fig 1. Mean AUROCs of 0.99±0.011, 0.983 ±0.014, 0.976±0.022, 0.988±0.015, 0.981±0.027, 0.989±0.015 for aklimate, aklimate-reduced,

sbmkl, dbmkl, sbmtmkl, and dbmtmkl respectively.
(EPS)

**S4 Fig. Predictive accuracy for selected methods on METABRIC dataset.** Methods include 3 versions of AKLIMATE with different combinations of input data types and some high-performing competitive methods (see main text). Point-and-whiskers plot displays mean and standard error of a method's test accuracy during cross-validation. Mean and standard error for FSMKL and BCC taken from [16].
(EPS)

**S5 Fig. AKLIMATE accuracy on METABRIC survival prediction for models restricted to KEGG pathways as feature sets.**
(EPS)

**S6 Fig. BRCA subtype-specific prediction accuracy on METABRIC survival prediction for AKLIMATE models using expression, copy number, and clinical features.** Samples were assigned to BRCA subtypes using the PAM50 gene expression signature [25].
(EPS)

**S7 Fig. Method performance on predicting cell line viability after shRNA gene knockdowns.** A) Measured by Spearman correlation. B) Measured by the number of times an algorithm produced the best Spearman correlation on a prediction task. C) Measured by Pearson correlation. D) Measured by the number of times an algorithm produced the best Pearson correlation on a prediction task.
(EPS)

**S8 Fig. RMSE scores for 37 shRNA knockdown prediction tasks.** Rows correspond to individual gene knockdowns; columns represent different methods. To highlight differential performance within each task, rows are centered by subtracting the median. Lower centered scores represent lower RMSE (better performance).
(EPS)

**S9 Fig. Data type importance proportions for 37 shRNA knockdown prediction tasks.** Columns correspond to the relative contributions of each input data type across prediction tasks. Relative contributions were computed by summing the model-assigned feature importance scores within individual data types. Columns are normalized to sum to 1.
(EPS)

**S10 Fig. AKLIMATE results highlighting the 10 most informative feature sets and 50 most informative features for the task of predicting FOXA1 shRNA knockdown viability.** Organized as Fig 1.
(EPS)

**S11 Fig. AKLIMATE results highlighting the 10 most informative feature sets and 50 most informative features for the task of predicting KRAS shRNA knockdown viability when no mutation features are used.** Feature and feature set weights averaged over 10 matched stratified 80%/20% train/test splits. Organized as Fig 1.
(EPS)

**S12 Fig. Metrics for PIK3CA AKLIMATE models with and without the use of mutational profiles for 8 key regulators.** Results averaged over 10 matched stratified 80%/20% train/test splits.
(EPS)

**S13 Fig. Metrics for CTNNB1 AKLIMATE models with and without the use of mutational profiles for 8 key regulators.** Results averaged over 10 matched stratified 80%/20% train/test splits.
(EPS)

## Author Contributions

**Conceptualization:** Vladislav Uzunangelov, Joshua M. Stuart.

**Data curation:** Vladislav Uzunangelov.

**Formal analysis:** Vladislav Uzunangelov, Christopher K. Wong, Joshua M. Stuart.

**Funding acquisition:** Joshua M. Stuart.

**Investigation:** Vladislav Uzunangelov, Joshua M. Stuart.

**Methodology:** Vladislav Uzunangelov, Joshua M. Stuart.

**Project administration:** Joshua M. Stuart.

**Resources:** Joshua M. Stuart.

**Software:** Vladislav Uzunangelov, Christopher K. Wong.

**Supervision:** Joshua M. Stuart.

**Validation:** Vladislav Uzunangelov, Christopher K. Wong.

**Visualization:** Vladislav Uzunangelov, Christopher K. Wong.

**Writing – original draft:** Vladislav Uzunangelov, Joshua M. Stuart.

**Writing – review & editing:** Vladislav Uzunangelov, Christopher K. Wong, Joshua M. Stuart.

## References

1. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018; 173(2):291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022 PMID: 29625048
2. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Research. 2011; 39(Database issue):D685–D690. https://doi.org/10.1093/nar/gkq1039 PMID: 21071392
3. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011; 27(12):1739–1740. https://doi.org/10.1093/bioinformatics/btr260 PMID: 21546393
4. Culhane AC, Schröder MS, Sultana R, Picard SC, Martinelli EN, Kelly C, et al. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. Nucleic Acids Research. 2012; 40(Database issue):D1060–D1066. https://doi.org/10.1093/nar/gkr901 PMID: 22110038
5. Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, et al. GeneMANIA Prediction Server 2013 Update. Nucleic Acids Research. 2013; 41(W1):W115–W122. https://doi.org/10.1093/nar/gkt533 PMID: 23794635
6. Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, et al. NDEx, the Network Data Exchange. Cell Systems. 2015; 1(4):302–305. https://doi.org/10.1016/j.cels.2015.10.001 PMID: 26594663
7. Gönen M. Integrating gene set analysis and nonlinear predictive modeling of disease phenotypes using a Bayesian multitask formulation. BMC Bioinformatics. 2016; 17(16):0. https://doi.org/10.1186/s12859-016-1311-3 PMID: 28105911
8. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. Cell. 2014; 158 (4):929–944. https://doi.org/10.1016/j.cell.2014.06.049 PMID: 25109877

9. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature Biotechnology. 2014; 32(12):1202–1212. https://doi.org/10.1038/nbt.2877 PMID: 24880487
10. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the Pathway Interaction Database. Nucleic Acids Research. 2009; 37(Database issue):D674–679. https://doi.org/10.1093/nar/gkn653 PMID: 18832364
11. Tomioka R, Suzuki T. Sparsity-accuracy trade-off in MKL. arXiv:10012615 [stat]. 2010.
12. Jorissen RN, Lipton L, Gibbs P, Chapman M, Desai J, Jones IT, et al. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research. 2008; 14(24):8061–8069. https://doi.org/10.1158/1078-0432.CCR-08-1431 PMID: 19088021
13. Hunter N, Borts RH. Mlh1 is unique among mismatch repair proteins in its ability to promote crossing-over during meiosis. Genes & Development. 1997; 11(12):1573–1582. https://doi.org/10.1101/gad.11.12.1573 PMID: 9203583
14. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486(7403):346–352. https://doi.org/10.1038/nature10983 PMID: 22522925
15. Margolin AA, Bilal E, Huang E, Norman TC, Ottestad L, Mecham BH, et al. Systematic Analysis of Challenge-Driven Improvements in Molecular Prognostic Models for Breast Cancer. Science translational medicine. 2013; 5(181):181re1. https://doi.org/10.1126/scitranslmed.3006112 PMID: 23596205
16. Seoane JA, Day INM, Gaunt TR, Campbell C. A pathway-based data integration framework for prediction of disease progression. Bioinformatics. 2014; 30(6):838–845. https://doi.org/10.1093/bioinformatics/btt610 PMID: 24162466
17. Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, et al. Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling. PLoS Computational Biology. 2013; 9(5):e1003047. https://doi.org/10.1371/journal.pcbi.1003047 PMID: 23671412
18. Cheng WY, Yang THO, Anastassiou D. Development of a Prognostic Model for Breast Cancer Survival in an Open Challenge Environment. Science Translational Medicine. 2013; 5(181):181ra50–181ra50. https://doi.org/10.1126/scitranslmed.3005974 PMID: 23596202
19. Rakotomamonjy A, Bach FR, Canu S, Grandvalet Y. SimpleMKL. Journal of Machine Learning Research. 2008; 9(Nov):2491–2521.
20. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research. 2012; 40(D1):D109–D114. https://doi.org/10.1093/nar/gkr988 PMID: 22080510
21. Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, et al. A prognostic index in primary breast cancer. British Journal of Cancer. 1982; 45(3):361–366. https://doi.org/10.1038/bjc.1982.62 PMID: 7073932
22. Creighton CJ, Osborne CK, van de Vijver MJ, Foekens JA, Klijn J, Horlings HM, et al. Molecular profiles of progesterone receptor loss in human breast tumors. Breast cancer research and treatment. 2009; 114(2):287–299. https://doi.org/10.1007/s10549-008-0017-2 PMID: 18425577
23. Creighton CJ, Massarweh S, Huang S, Tsimelzon A, Shou J, Malorni L, et al. Development of resistance to targeted therapies transforms the clinically-associated molecular profile subtype of breast tumor xenografts. Cancer research. 2008; 68(18):7493–7501. https://doi.org/10.1158/0008-5472.CAN-08-1404 PMID: 18794137
24. Giuliano AE, Connolly JL, Edge SB, Mittendorf EA, Rugo HS, Solin LJ, et al. Breast Cancer—Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA: A Cancer Journal for Clinicians. 2017; 67(4):290–303. PMID: 28294295
25. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of Clinical Oncology. 2009; 27(8):1160–1167. https://doi.org/10.1200/JCO.2008.18.1370 PMID: 19204204
26. Shao DD, Tsherniak A, Gopal S, Weir BA, Tamayo P, Stransky N, et al. ATARiS: Computational quantification of gene suppression phenotypes from multisample RNAi screens. Genome Research. 2013; 23(4):665–678. https://doi.org/10.1101/gr.143586.112 PMID: 23269662
27. Cowley GS, Weir BA, Vazquez F, Tamayo P, Scott JA, Rusin S, et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Scientific Data. 2014; 1:140035. https://doi.org/10.1038/sdata.2014.35 PMID: 25984343
28. Gönen M, Weir BA, Cowley GS, Vazquez F, Guan Y, Jaiswal A, et al. A Community Challenge for Inferring Genetic Predictors of Gene Essentialities through Analysis of a Functional Screen of Cancer Cell

Lines. Cell Systems. 2017; 5(5):485–497.e3. https://doi.org/10.1016/j.cels.2017.09.004 PMID: 28988802

29. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483 (7391):603–607. https://doi.org/10.1038/nature11003 PMID: 22460905

30. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology. 2011; 12(4):R41. https://doi.org/10.1186/gb-2011-12-4-r41 PMID: 21527027

31. Uzunangelov VJ. Prediction of cancer phenotypes through the integration of multi-omic data and prior information. Ph.D. Thesis, UC Santa Cruz. 2019. Available from: https://escholarship.org/uc/item/5cs2x2bz.

32. Robertson AG, Shih J, Yau C, Gibb EA, Oba J, Mungall KL, et al. Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. Cancer Cell. 2017; 32(2):204–220.e15. https://doi.org/10.1016/j.ccell.2017.07.003 PMID: 28810145

33. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software. 2017; 77(1):1–17. https://doi.org/10.18637/jss.v077.i01

34. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010; 33(1):1–22. https://doi.org/10.18637/jss.v033.i01 PMID: 20808728

35. Hobbs GA, Der CJ, Rossman KL. RAS isoforms and mutations in cancer at a glance. Journal of Cell Science. 2016; 129(7):1287–1292. https://doi.org/10.1242/jcs.182873 PMID: 26985062

36. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell. 2018; 173(2):305–320.e10. https://doi.org/10.1016/j.cell.2018.03.033 PMID: 29625049

37. Hitchins M, Rapkins R, Kwok CT, Srivastava S, Wong JL, Khachigian L, et al. Dominantly Inherited Constitutional Epigenetic Silencing of MLH1 in a Cancer-Affected Family Is Linked to a Single Nucleotide Variant within the 5'UTR. Cancer Cell. 2011; 20(2):200–213. https://doi.org/10.1016/j.ccr.2011.07.003 PMID: 21840485

38. The HPN-DREAM Consortium, Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. Nature Methods. 2016; 13(4):310–318. https://doi.org/10.1038/nmeth.3773 PMID: 26901648

39. Englund C, Verikas A. A novel approach to estimate proximity in a random forest: An exploratory study. Expert Systems with Applications. 2012; 39(17):13046–13050. https://doi.org/10.1016/j.eswa.2012.05.094

40. Cao H, Bernard S, Sabourin R, Heutte L. A Novel Random Forest Dissimilarity Measure for Multi-View Learning. arXiv:200702572 [cs, stat]. 2020.

41. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Research. 2018; 46(20):10546–10562. https://doi.org/10.1093/nar/gky889 PMID: 30295871

42. Mallik S, Zhao Z. Graph- and rule-based learning algorithms: a comprehensive review of their applications for cancer type classification and prognosis using genomic data. Briefings in Bioinformatics. 2019. https://doi.org/10.1093/bib/bby085 PMID: 30239597

43. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Frontiers in Genetics. 2017; 8. https://doi.org/10.3389/fgene.2017.00084 PMID: 28670325

44. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996; 58(1):267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

45. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(2):301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

46. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68(1):49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

47. Jacob L, Obozinski G, Vert JP. Group lasso with overlap and graph lasso. In: Proceedings of the 26th Annual International Conference on Machine Learning—ICML'09. Montreal, Quebec, Canada: ACM Press; 2009. p. 1–8. Available from: http://portal.acm.org/citation.cfm?doid=1553374.1553431.

48. Sokolov A, Carlin DE, Paull EO, Baertsch R, Stuart JM. Pathway-Based Genomics Prediction using Generalized Elastic Net. PLOS Computational Biology. 2016; 12(3):e1004790. https://doi.org/10.1371/journal.pcbi.1004790 PMID: 26960204

49. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008; 24(9):1175–1182. https://doi.org/10.1093/bioinformatics/btn081 PMID: 18310618

50. Bach FR, Lanckriet GRG, Jordan MI. Multiple kernel learning, conic duality, and the SMO algorithm. In: Twenty-first international conference on Machine learning—ICML'04. Banff, Alberta, Canada: ACM Press; 2004. p. 6. Available from: http://portal.acm.org/citation.cfm?doid=1015330.1015424.

51. Suzuki T, Tomioka R. SpicyMKL: a fast algorithm for Multiple Kernel Learning with thousands of kernels. Machine Learning. 2011; 85(1-2):77–108. https://doi.org/10.1007/s10994-011-5252-9

52. Manica M, Cadow J, Mathis R, Rodríguez Martínez M. PIMKL: Pathway-Induced Multiple Kernel Learning. npj Systems Biology and Applications. 2019; 5(1):1–8. https://doi.org/10.1038/s41540-019-0086-3 PMID: 30854223

53. Stolovitzky G, Monroe D, Califano A. Dialogue on Reverse-Engineering Assessment and Methods. Annals of the New York Academy of Sciences. 2007; 1115(1):1–22. https://doi.org/10.1196/annals.1407.021 PMID: 17925349

54. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Meth. 2012; 9(8):796–804. https://doi.org/10.1038/nmeth.2016 PMID: 22796662

55. Kim M, Rai N, Zorraquino V, Tagkopoulos I. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nature Communications. 2016; 7. https://doi.org/10.1038/ncomms13090

56. Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

57. Wolpert DH. Stacked Generalization. Neural Networks. 1992; 5:241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

58. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. Statistical Applications in Genetics and Molecular Biology. 2007; 6(1). https://doi.org/10.2202/1544-6115.1309 PMID: 17910531

59. Polley EC. Super Learner In Prediction. UC Berkeley; 2010. 266.

60. Wan Q, Pal R. An Ensemble Based Top Performing Approach for NCI-DREAM Drug Sensitivity Prediction Challenge. PLoS ONE. 2014; 9(6). https://doi.org/10.1371/journal.pone.0101183 PMID: 24978814

61. Louppe G. Understanding Random Forests: From Theory to Practice. arXiv:14077502 [stat]. 2014.

62. Scholkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, MA, USA: MIT Press; 2001.

63. Aronszajn N. Theory of Reproducing Kernels. Transactions of the American Mathematical Society. 1950; 68(3):337. https://doi.org/10.1090/S0002-9947-1950-0051437-7

64. Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge University Press; 2004.

65. Kimeldorf G, Wahba G. Some results on Tchebycheffian spline functions. Journal of Mathematical Analysis and Applications. 1971; 33(1):82–95. https://doi.org/10.1016/0022-247X(71)90184-3

66. Lanckriet GRG, Bie TD, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. Bioinformatics. 2004; 20(16):2626–2635. https://doi.org/10.1093/bioinformatics/bth294 PMID: 15130933

67. Gönen M, Alpaydın E. Multiple Kernel Learning Algorithms. J Mach Learn Res. 2011; 12:2211–2268.

68. Kloft M, Rückert U, Bartlett PL. A Unifying View of Multiple Kernel Learning. In: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al., editors. Machine Learning and Knowledge Discovery in Databases. vol. 6322. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 66–81. Available from: http://link.springer.com/10.1007/978-3-642-15883-4_5.

69. Suzuki T, Sugiyama M. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. The Annals of Statistics. 2013; 41(3):1381–1405. https://doi.org/10.1214/13-AOS1095

70. Breiman L. Some infinity theory for predictor ensembles. UC Berkeley; 2000. 577.

71. Davies A, Ghahramani Z. The Random Forest Kernel and other kernels for big data from random partitions. arXiv:14024293 [cs, stat]. 2014.

72. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006; 63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1

73. Scornet E. Random Forests and Kernel Methods. IEEE Transactions on Information Theory. 2016; 62(3):1485–1500. https://doi.org/10.1109/TIT.2016.2514489

74. Breiman L. Stacked regressions. Machine Learning. 1996; 24(1):49–64.

75. LeDell E. Scalable Ensemble Learning and Computationally Efficient Variance Estimation. Ph.D. Thesis, University of California, Berkeley. 2015. Available from: https://escholarship.org/uc/item/3kb142r2.

76. van der Laan M, Dudoit S. Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples. University of California, Berkeley U.C. Berkeley Division of Biostatistics Working Paper Series; 2003. 130. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.211.5925&rep=rep1&type=pdf.

77. Vaart AWvd, Dudoit S, Laan MJvd. Oracle inequalities for multi-fold cross validation. Statistics & Decisions. 2006; 24(3):351–371.

78. Ng AY. Preventing "Overfitting" of Cross-Validation Data. In: Proceedings of the Fourteenth International Conference on Machine Learning. ICML'97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1997. p. 245–253. Available from: http://dl.acm.org/citation.cfm?id=645526.657119.

79. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Routledge; 1984. Available from: https://www.taylorfrancis.com/books/9781351460491.

80. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007; 8(1):25. https://doi.org/10.1186/1471-2105-8-25 PMID: 17254353

81. Sandri M, Zuccolotto P. A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees. Journal of Computational and Graphical Statistics. 2008; 17(3):611–628. https://doi.org/10.1198/106186008X344522

82. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010; 26(10):1340–1347. https://doi.org/10.1093/bioinformatics/btq134 PMID: 20385727

83. Nembrini S, König IR, Wright MN, Valencia A. The revival of the Gini importance? Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty373 PMID: 29757357