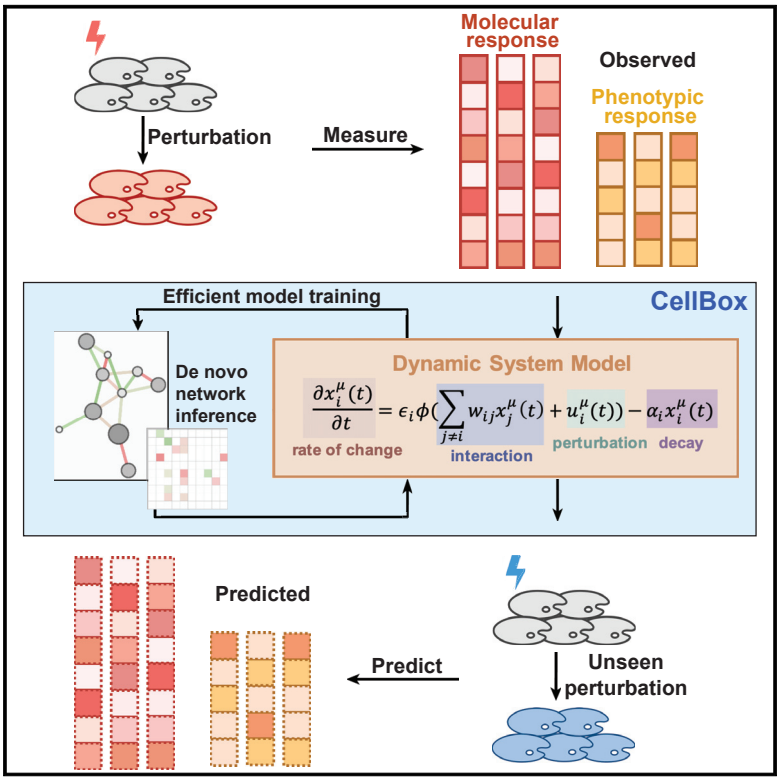
Cell Systems

Article

CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy

Graphical Abstract



Highlights

- CellBox includes explicit models of cell dynamics in a machine-learning framework
- CellBox enables the prediction of system responses to unseen perturbations
- CellBox-derived molecular interactions generally agree with known biological pathways
- CellBox is an example of interpretable scientific machine learning in cell biology

Authors

Bo Yuan, Ciyue Shen, Augustin Luna, Anil Korkut, Debora S. Marks, John Ingraham, Chris Sander

Correspondence

boyuan@g.harvard.edu (B.Y.), c_shen@g.harvard.edu (C. Shen), mathcellbox@gmail.com (C. Sander)

In Brief

The ability to accurately predict cell behavior to previously untested perturbations would benefit the discovery of combination therapies in cancer. To overcome the lack of interpretability of black-box machine-learning models, we developed a hybrid approach called CellBox that combines explicit mathematical models of molecular interactions with efficient parameter inference algorithms adapted from deep learning. The models are data driven and do not require prior knowledge, and their predictive scope scales well with the availability of high-throughput data.

Yuan et al., 2021, Cell Systems *12*, 128–140
February 17, 2021
https://doi.org/10.1016/j.cels.2020.11.013

CellPress

CellPress
OPEN ACCESS

Cell Systems

Article

CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy

Bo Yuan,[1,2,3,7,*] Ciyue Shen,[1,2,3,7,*] Augustin Luna,[1,2,3] Anil Korkut,[4] Debora S. Marks,[3,5] John Ingraham,[6] and Chris Sander[1,2,3,8,*]

[1]Department of Cell Biology, Harvard Medical School, Boston, MA, USA
[2]cBio Center, Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA
[3]Broad Institute, Cambridge, MA, USA
[4]Department of Bioinformatics & Computational Biology, the University of Texas M D Anderson Cancer Center, Houston, TX, USA
[5]Department of Systems Biology, Harvard Medical School, Boston, MA, USA
[6]MIT Computer Science & Artificial Intelligence Laboratory, Boston, MA, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: boyuan@g.harvard.edu (B.Y.), c_shen@g.harvard.edu (C. Shen), mathcellbox@gmail.com (C. Sander)
https://doi.org/10.1016/j.cels.2020.11.013

SUMMARY

Systematic perturbation of cells followed by comprehensive measurements of molecular and phenotypic responses provides informative data resources for constructing computational models of cell biology. Models that generalize well beyond training data can be used to identify combinatorial perturbations of potential therapeutic interest. Major challenges for machine learning on large biological datasets are to find global optima in a complex multidimensional space and mechanistically interpret the solutions. To address these challenges, we introduce a hybrid approach that combines explicit mathematical models of cell dynamics with a machine-learning framework, implemented in TensorFlow. We tested the modeling framework on a perturbation-response dataset of a melanoma cell line after drug treatments. The models can be efficiently trained to describe cellular behavior accurately. Even though completely data driven and independent of prior knowledge, the resulting *de novo* network models recapitulate some known interactions. The approach is readily applicable to various kinetic models of cell biology. A record of this paper's Transparent Peer Review process is included in the Supplemental Information.

INTRODUCTION

The emergence of resistance to single anticancer agents has highlighted the importance of developing combinations of agents as a more robust therapeutic approach to cancer treatment (Fitzgerald et al., 2006; Garraway and Jänne, 2012; Mansoori et al., 2017; Bayat Mokhtari et al., 2017). However, experimental screening of all possible pairwise or higher-order combinations of currently available agents is practically unrealistic. The space of potential new therapeutic targets is even larger and more challenging to explore experimentally. To efficiently narrow down the search space and nominate promising sets of experimentally testable candidates, computational models have been used to predict cellular responses based on sets of perturbation experiments (Azmi et al., 2010; Ryall and Tan, 2015; Cheng et al., 2019), but these have been limited in scope. The ability to model cell biology at a larger scale and to infer causal mechanisms to generalize to unobserved perturbations is critical in facilitating the search for combinatorial, potentially therapeutic candidates.

Perturbation-Response Profiling in Cell Biology

In order to understand cell behavior, a large variety of experimental approaches have been used to profile cellular responses under different perturbations. Biochemical and cell-biological experiments testing relationships of particular protein-protein pairs have, for many years, been successfully used to identify signaling cascades (Cerami et al., 2011; Wrzodek et al., 2013; Croft et al., 2014), but one-by-one perturbation experiments are laborious and the resulting models, although insightfully descriptive, typically do not provide the ability to quantitatively predict detailed molecular or system-level responses. Phenotypic screening collects high-throughput information on whole-cell responses with univariate readouts such as cell viability or growth rate (Lamb et al., 2006; Cheung et al., 2011; Wang et al., 2014; McDonald et al., 2017; Tsherniak et al., 2017). In order to resolve intracellular interactions and provide mechanistic insights, systematic methods have been developed to profile post-perturbational molecular responses with increasing coverage, e.g., changes in transcript (Dixit et al., 2016; Niepel

et al., 2017; Norman et al., 2019) and protein levels (Korkut et al., 2015; Hill et al., 2017). These rich datasets challenge computational methods to comprehensively describe mechanisms and to quantitatively model cell responses to unseen perturbations, e.g., for the design of experiments to test mechanistic hypotheses or for the design of combinatorial therapeutic interventions.

Computational Modeling

Various computational methods have been developed to infer interactions and predict cellular responses (Vanhaelen et al., 2017). Static models use, e.g., co-expression models (Carter et al., 2004; Wang et al., 2009; Babur et al., 2010), maximum entropy networks (Lezon et al., 2006; Locasale and Wolf-Yadlin, 2009), or mutual information related methods (Meyer et al., 2008; Chan et al., 2017), to construct network models of molecular interactions (Şenbabaoğlu et al., 2016; Yi et al., 2017), or use regression models to directly predict cellular responses based on molecular perturbation-response measurements (Dixit et al., 2016; Norman et al., 2019). By contrast, dynamic models, such as Boolean network models (D'haeseleer et al., 2000), fuzzy-logic models (Aldridge et al., 2009), dynamic Bayesian networks (Zou and Conzen, 2005), and ordinary-differential-equation (ODE) network models (Gardner et al., 2003; Nyman et al., 2020), can provide mechanistic insight in terms of the propagation of cellular signals to phenotypic response over time, but typically require prior knowledge of interaction parameters and, thus, currently only work for small systems (Gardner et al., 2003; Klinger et al., 2013). New algorithms to parameterize large-scale mechanistic models, although computationally efficient, require prior knowledge of a set of relevant interactions (Fröhlich et al., 2018). Insufficient prior knowledge is one of the major constraints for modeling large systems, e.g., in that prior information is not available for all components or is aggregated from disparate experimental sources and thus lacks uniform context. A more rigorous approach is to use uniform datasets generated in systematic experiments in one experimental context and then perform *de novo* structure inference of an interaction network specific for that context. Given such data for large systems, the computational challenge is to search for optimal interaction parameter sets in a complex multidimensional solution space. Previous dynamic optimization approaches, such as Monte Carlo (MC) methods and belief-propagation (BP) algorithms, have been used to construct data-driven network models (Bruggeman et al., 2002; Nelander et al., 2008; Hug et al., 2013; Klinger et al., 2013; Korkut et al., 2015). Still, these might not efficiently scale to larger systems (e.g., MC) or might require excessive approximations for the chosen mathematical model to facilitate efficient exploration of solution space (e.g., independent row approximation in BP) (Bruggeman et al., 2002; Korkut et al., 2015). Therefore, to achieve good accuracy of parameter inference for larger systems and to gain the ability to generalize to more sophisticated kinetic models, a more general and potentially more powerful data-driven modeling framework is needed.

Machine Learning and Interpretability

Recently, deep learning has become an effective data-driven framework capable of generating predictions for large and complex systems. Gradient descent implemented with automatic differentiation, which has been broadly used in training graphical models, allows efficient parameter optimization in complex network systems. This framework has been successfully applied to many domains of biomedical research, from pathology image classification (Hou et al., 2016; Esteva et al., 2017) to sequence motif detection (Zhou and Troyanskaya, 2015). Although the predictive power of deep-learning models is often impressive, their interpretation, which is crucial for providing understandable and, therefore, more trustable predictions, remains challenging. The complex multilayer network architecture of most deep-learning models lacks explicit representations and consequent direct interpretation. This difficulty is sometimes called the "black-box" problem (Montavon et al., 2018). To address this problem, we apply a deep-learning optimization approach to learn a data-driven model (called "CellBox") that incorporates an explicitly interpretable network of interactions between cellular components, instead of a black-box neural network, while aiming to maintain a high level of learning performance.

CellBox is designed to be a framework for computational modeling of cellular responses to perturbations that (1) links perturbations to molecular and phenotypic changes in a unified computational model, (2) quantifies time-dependent (dynamic) cellular responses, (3) promises training efficiency and scalability for large-scale systems, and (4) is interpretable in terms of interactions that can be compared with established models of molecular biology, such as signaling pathways. Here, we construct a nonlinear ODE-based model representing a biological network of 99 components connecting perturbations, protein response, and phenotypes to simulate dynamic cellular behavior. The network connections are directly learned from post-perturbational data under 89 experimental conditions with the objective of accurately reproducing the molecular and cellular responses on training data. To reach this objective, we implemented gradient descent with automatic differentiation to infer interaction parameters in the ODE network, which can then be exposed to novel perturbations to predict cell behavior. The key performance criterion for the data-driven model trained with a relatively small set of experiments is whether the model is able to provide reasonably accurate predictions on a large set of unseen perturbation conditions. Anticipating the availability of increasingly informative perturbation-response datasets in diverse areas of cell biology, we present CellBox as a generally applicable framework for modeling a broad range of dynamic cell behavior.

RESULTS

CellBox Model of Perturbation Biology

In order to construct a data-driven model to predict the dynamics of molecular and cellular behavior under combinations of drug treatments, the perturbation data must have (1) paired measurements of changes in protein levels and cellular behavior for a set of perturbations and (2) training and withheld data to test model performance. Here, we used a perturbation dataset for the melanoma cell line SK-Mel-133 (Korkut et al., 2015), which contains molecular and phenotypic response profiles of cells treated with 12 different drugs and their pairwise combinations (Figures 1A and S1). For each of the 89 perturbation conditions, levels of 82 selected proteins and phosphoproteins were measured in cell lysates before and 24 h after perturbation on

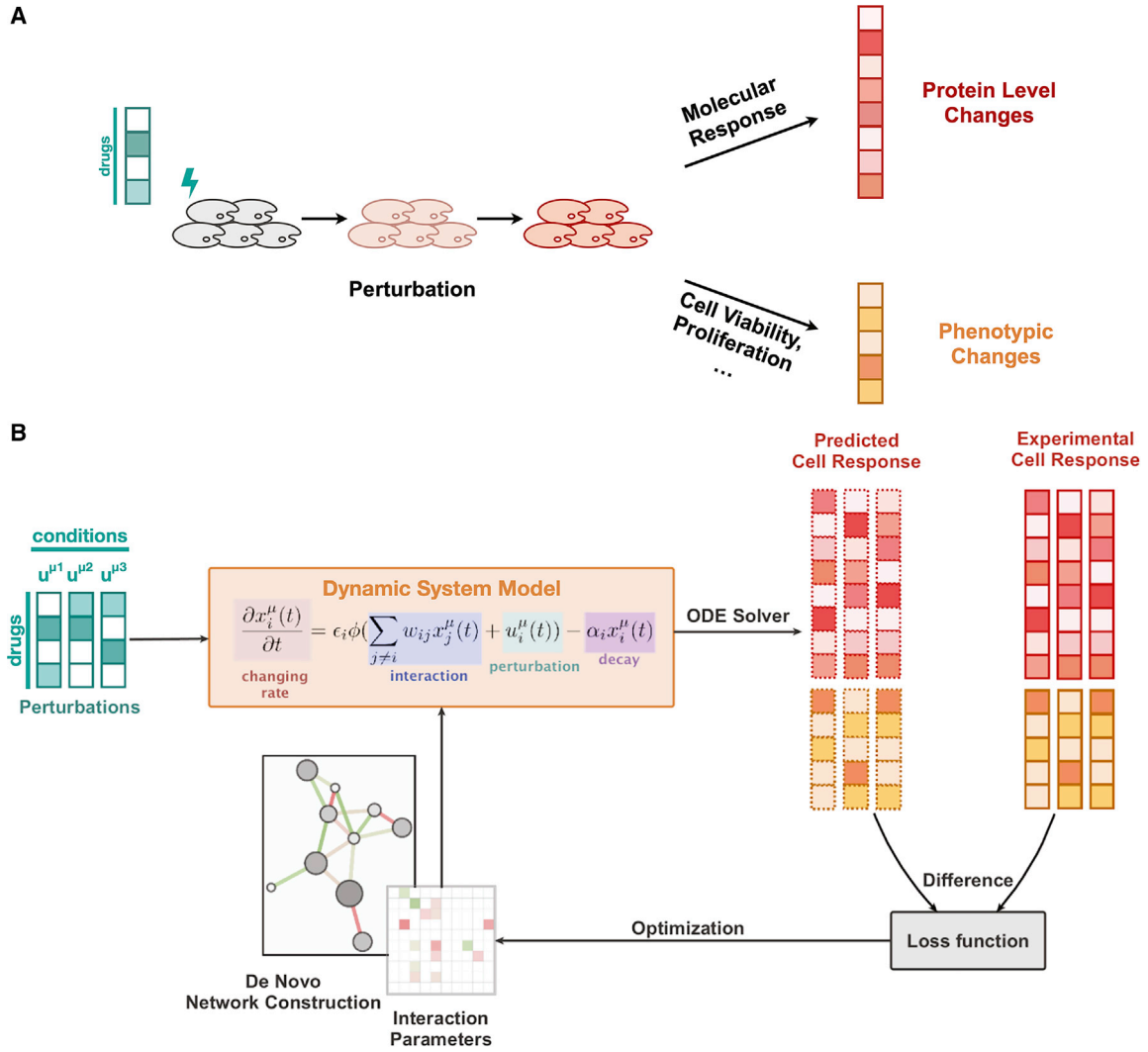


Figure 1. CellBox: Dynamic Modeling of Cellular Systems with Perturbation Data

(A) Perturbations such as drugs were used to disturb the cellular system. The cell responses, including protein- and phosphoprotein-level changes, and phenotypic changes, were measured to provide information for model construction.

(B) Systematic responses of the cells under various drug perturbations were used to construct an interpretable machine-learning model. CellBox models system behavior in terms of interaction parameters connecting molecular (proteins and phosphoproteins) and phenotypic variables using a set of differential equations. CellBox was trained iteratively by optimizing interaction parameters to fit the numerically simulated system response to experimental observations. After training on pairwise data of input perturbation and output system behavior, the CellBox model can be used to predict the cellular response to arbitrary perturbation conditions.

antibody-based reverse-phase protein arrays (RPPA). In parallel, cellular phenotypes were assayed, including cell-cycle progression and cell viability. With parallel measurements of proteomic and phenotypic responses to a systematic set of perturbations, this dataset provides sufficient information to construct network models that quantitatively link molecular changes to cellular responses.

We used a set of ODEs with a nonlinear envelope (Figure 1B) to model the dynamic responses of the system to drug perturbations (STAR Methods). The parameters of the ODEs (w_{ij}, ~10,000 in total) are the interaction strengths between the entities in the network model. The simplicity of the interaction dynamics (Figure 1B), the nonlinear envelope, as well as the restoration term $-\alpha_i x_i(t)$ are computational devices, roughly analogous to mean-field approaches, to account for the fact that the data are limited to a relatively small fraction of all cellular components and to avoid instabilities (Nelander et al., 2008; Molinelli et al., 2013; Korkut et al., 2015). The interaction parameters were randomly initialized and updated throughout the model training process, with the objective of minimizing a loss function. For the loss function, we chose the Euclidean distance between experimental data and the results of the numerical simulation of the ODE model, plus an L1 regularization penalty on network density to avoid overfitting (STAR Methods, Equation 3). We used Heun's ODE solver (Süli and Mayers, 2003) to numerically simulate the ODE system and the Adam optimizer (Kingma and

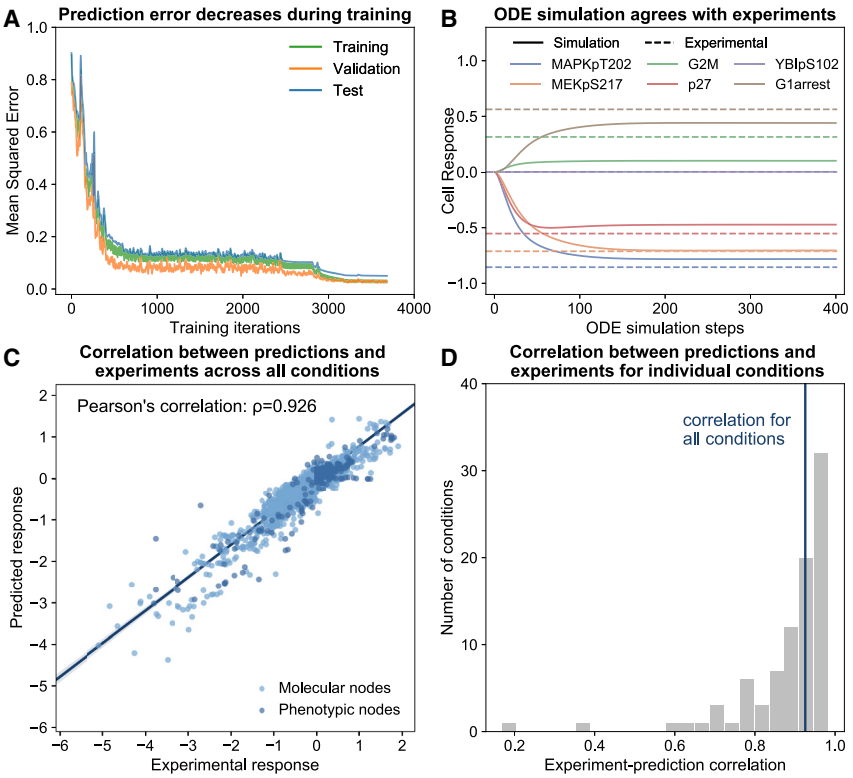


Figure 2. CellBox Convergence and Prediction Accuracy on Randomly Partitioned Training-Test Datasets

(A) Over training iterations, the mean-squared error on the training set (56% of the entire dataset), validation set (14%), and test set (30%) decreased nearly monotonically, and the models converged at the end of the training.

(B) The predicted molecular and phenotypic responses at the steady state of the ODE simulations agree with the experimental data on the test set. A subset of molecular measurements (MAPKpT202, YB1pS102, MEKpS217, and p27) and phenotypic measurements (G2M and G1arrest) are shown. Cell response is defined as the log2 ratio of post- and pre-perturbation measurements. The annotations and identities of the complete set of measurements are in Table S1.

(C) Across 1,000 models trained with different data partitions, the average predicted responses correlate with experimental observations (Pearson's correlation ρ = 0.926, regression line in dark blue with 95% confidence interval). Each point represents one measurement, either molecular or phenotypic, in one perturbation condition.

(D) Nearly all predictions for individual conditions have high correlations with experimental measurements.

Ba, 2014) with automatic differentiation to minimize the loss function. Taken together, we constructed an ODE model of a cell-biological system that was trained by using perturbation data, which we named CellBox.

CellBox Can Be Trained on Perturbation Data to Predict Cell Response Accurately

In order to test the prediction performance of this training scheme, we randomly selected 70% of the perturbation data (n = 62 conditions) for training and withheld the rest 30% (n = 27 conditions) for testing. 20% of the training data were used as a validation set to stop model training when the performance on the validation set did not further improve. We manually fine-tuned the hyperparameters, including learning rate, regularization, and ODE simulation time, to increase the training efficiency (Figure S2). At the end of the training, the numerical solutions of the ODE model converged efficiently to experimental data (Figures 2A and 2B). We repeated the modeling scheme with 1,000 independent random data partitions to construct models for each partition. The average predictions on test sets across all models and all conditions correlate with experimental data with a Pearson's correlation coefficient of 0.93 (Figure 2C). A more refined analysis of individual perturbation conditions showed that the model trains equally well for all conditions (Figure 2D), and that model performance is independent of data scaling that is normally applied to protein-profiling data (Figure S3). The results illustrate that the CellBox models can be efficiently trained with perturbation data to predict cell response to experimentally applied perturbations accurately.

Even though ~70% of the models reached steady solutions of the ODEs (Figure 2B), some models converged to oscillatory solutions (Figure S4A). In order to test whether the oscillation is an artifact of data partitioning during model training, we retrained the models with the same train-test data partitioning but with multiple different random seeds for the computational optimizer (STAR Methods). Each partition of the training data can result in both steady and oscillatory solutions (Figures S4A–S4D). The emergence of oscillatory solutions is independent of ODE solvers (Figures S4E–S4G). On the basis of the assumption that the population average of cell response reaches a stable and non-oscillating steady state 24 h after drug treatment, we excluded the oscillatory models in the following analysis (STAR Methods). Altogether, these results indicate that CellBox, a data-driven ODE-based cellular system model, can be trained to accurately predict dynamics of cell response without any requirement of prior knowledge about the relationship between particular protein levels or phenotypes.

CellBox Model Predicts Cell Response for Single-to-Combo and Leave-One-Drug-Out Cross-Validations

Even though the model makes accurate predictions with different training data, data partitioning, especially random partitioning, raises the concern of information sharing between training and test datasets. Combinatorial conditions in both datasets might share the same drugs such that the test set might not be truly independent of training and, therefore, is suboptimal for rigorous evaluation of the model performance. Moreover, the ability to predict the combinatorial effect of a drug, e.g., dominant, additive, synergistic, when none of its combinations has

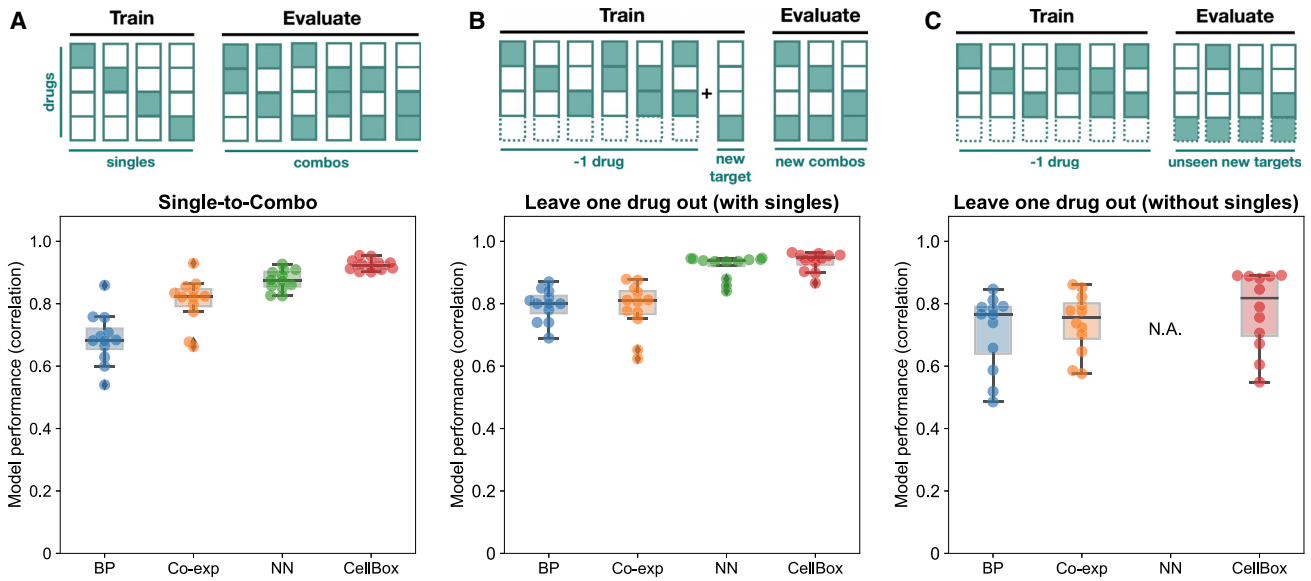


Figure 3. CellBox Can Accurately Predict Cell Response for Single-to-Combo and Leave-One-Drug-out Cross-Validations

(A) When only single conditions were used for training (single-to-combo), the CellBox models predict the effects of combinatorial conditions with high accuracy and outperform the dynamic network model inferred by using BP, the static co-expression network model (Co-exp), and a neural network regression model (NN) trained on the same data.

(B) When combinatorial conditions associated with one drug were withheld from training, the CellBox models retain high accuracy for predicting the effects of unseen drug pairs.

(C) When all conditions associated with one drug were withheld from training, the ODE network models predict the effects of the withheld drug with reduced accuracy, but direct-regression models such as NN cannot generalize to unseen targets at all. For each model type, the performance was evaluated by Pearson's correlation between predicted cell response and experimental cell response. The box charts indicate the group means and standard deviations.

been seen by the model, is a nontrivial challenge in the context of making accurate predictions of experimentally untested drug combinations.

In order to address these points, rather than training the model with random data partitioning, we instead designed more rigorous tasks: single-to-combo (Figures 3A and S7) and leave-one-drug-out cross-validation (Figures 3B and 3C) for each drug. In the single-to-combo analysis, all single-drug-treatment conditions were used for training, and predictions were made on all combinatorial drug conditions. In leave-one-drug-out cross-validation, all the combination conditions containing the treatment of a particular drug with or without the corresponding single-drug conditions were withheld and the rest of the conditions were used for training. In these more stringent tests, we found that the predicted values for withheld data were still highly correlated with the experimental observations (average Pearson's correlation: 0.93 for single-to-combo; 0.94 for leave-one-drug-out with single conditions, similar to that of the training with random partition; 0.79 for complete leave-one-drug-out). Under all three scenarios, on this dataset, CellBox outperforms the BP dynamic model approach previously used in perturbation biology (Korkut et al., 2015) in terms of predictive accuracy. These results indicate that the CellBox model can be trained with a relatively small set of perturbation data and that its predictions can be generalized to unseen combinatorial perturbations. In particular, CellBox models predict more accurately than linear models in the single-to-combo scenario (Figure S7), suggesting that CellBox can capture the nonadditive (synergistic or antagonistic) effects, which is particularly useful in nominating therapeutic drug combinations.

CellBox models are dynamic network models of a cell-biological system. To test whether such interpretable network models of molecular interactions help increase model predictive power, we compared the results to those of a static biological network model and a deep neural network model. The static network model was constructed by learning co-expression correlation for each pair of protein nodes (Co-exp) while the deep neural network model was trained to directly regress phenotypic changes against parameterized perturbations (NN) (STAR Methods). In all three tasks, the static network models had lower accuracy relative to the dynamic CellBox. The NN had comparable performance to CellBox in the cross-validation for individual drugs, but its performance dropped significantly in the single-to-combo analysis (Figure 3A). Furthermore, the NN was unable to generalize to unseen targets whose information is completely excluded from training (Figures 3C and S6). Altogether, because of the lack of mechanistic and dynamic information, static network or direct-regression models appear to be less suitable for facilitating the search for combinatorial targets.

Model Performance Is Robust against Noise and Reduced Training-Set Size

To examine model robustness of the CellBox models against a reduction in training data, we tested the stability of model performance when either the data quality or quantity is compromised. To test the former, we introduced different levels of multiplicative Gaussian noise (STAR Methods) into the input molecular and cellular response data and trained models on the resultant noisy datasets. The assumption behind such multiplicative noise is

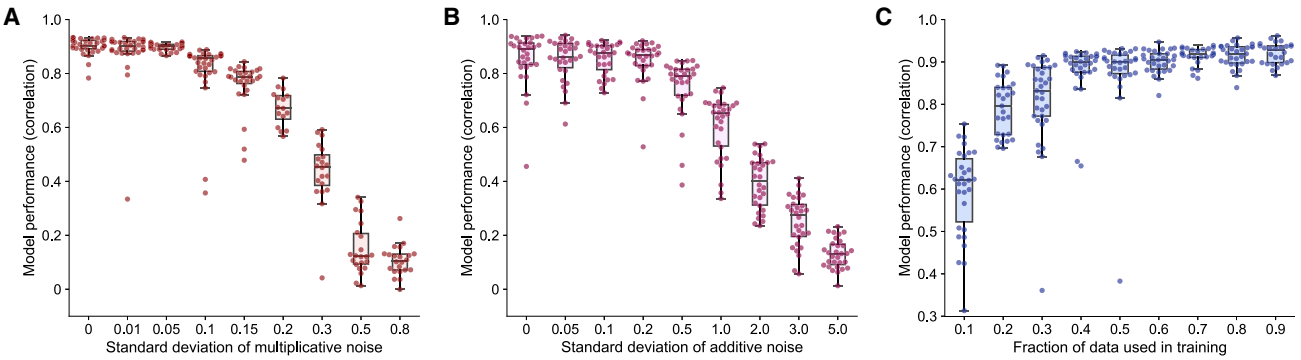


Figure 4. Model Performance Is Stable Against Data Noise and Data Reduction

(A and B) Correlation between predicted responses and experimental responses in the test set decreases as an increased level of multiplicative noise (A) or additive noise (B) is added to the training data (each dot represents one model). The CellBox models can tolerate up to σ_{mul} = 0.05 multiplicative noise or σ_{add} = 0.20 additive noise.

(C) Correlation between predicted responses and experimental responses in the test set increases with an increasing quantity of data used for model training. For the current dataset, the correlation plateaus when 40% of the original dataset is used. The box charts indicate the group means and standard deviations.

that the uncertainty and noise in experimental measurements arise around the true values. When comparing the predicted response in test sets with the experimental data, we found that the predictions from training on the noisy data retain similarly high correlations to experimental data as those trained on the original data, even with the addition of 5% multiplicative Gaussian noise (Figure 4A). As the magnitude of the noise increases, the model performance decreases gradually in terms of both convergence (Figure S8A) and predictive power (Figure 4A). We observed similar behavior when challenging the CellBox models with additive Gaussian noise (Figures 4B and S8B). The predictivity of CellBox models can tolerate an addition of up to σ_{add} = 0.20 additive Gaussian noise, i.e., about half the data standard deviation σ_{data} = 0.46. We conclude that model performance is stable in the presence of moderate experimental errors.

To test the dependency of model performance on data quantity, we trained the model on subsamples of the experimental dataset. We trained models with varying amounts of data (from 10% to 90% in steps of 10%) and found that the models could make accurate predictions of withheld data with as little as 40% of the complete dataset (Figure 4C). We found that increasing the size of the training set further has diminishing returns in terms of model performance. This implies, on the dataset used here with an interaction network of ~100 components, that a comparatively small number of perturbation conditions (40–100, rather than directly testing all ~3,000 possible combinations) is sufficient for constructing reasonably predictive models. This example might be a useful guide for power calculations for systems with hundreds of measured components, which would be of considerable interest.

Comparison of the Network Models with Prior Knowledge about Pathways

We used ODEs as the core framework of the current version of the CellBox mathematical model. Each parameter in the model represents the strength and direction of a biological interaction. In order to investigate whether the inferred interactions are consistent with current knowledge of biological pathways, we used the entire dataset as training data to generate 1,000 full models and examined the resulting *de novo* network edges learned from training. In order to measure both the strength and stability of edge inference, we used t scores (STAR Methods) to assess the statistical significance of each interaction, where a higher absolute t score indicates higher interaction strength and lower variance across the models (Figure 5A). Using the primary targets of the drugs as the ground truth, we first examined the interactions between the drug-activity nodes and their downstream effectors. We found that all 12 drug-activity nodes had significant edge connections to their known primary downstream protein effectors with the interaction directions consistent with their expected effects (Figure 5B), suggesting the models are able to capture the literature-provided interactions between the drug target and their downstream effectors.

To further investigate to what extent the inferred networks represent known pathway interactions, we compared the interactions in the CellBox models to corresponding molecular interactions extracted from the Pathway Commons (PC) resource (Cerami et al., 2011), which is an aggregation of ~20 curated publicly available pathway interaction databases. In this comparison between interactions inferred by CellBox models and prior knowledge, model interactions can, in principle, be found in PC either as one-step (A-B), two-step (A-X-B), or logical (>2 steps) interactions, where the logical interactions can be usefully predictive or perhaps erroneous. Direct interactions (A-B) in PC represent one component (A) affecting the expression, phosphorylation, or state-of-change of the other component (B) (Figures 5C and 5D; Table S2). For example, models and prior knowledge agree for the phosphorylation of the AKT protein kinase (AKTpS473), which negatively affects the insulin receptor substrate 1 (IRS1) (Chandarlapaty et al., 2011), and for the activation of the mitogen-activated protein kinase (MAPKpT202) that affects p27 protein levels (Donovan et al., 2001; Osaki and Gama, 2013). Some model interactions can be found as indirect two-step interactions (A-X-B) in PC, meaning at least one path with one intermediate component (X) can be found between the two components (A and B). For example, retinoblastoma protein (Rb1) affects cyclinD through p21 (Carreira et al., 2005; Lei et al., 2005), and mitogen-activated protein kinase kinase

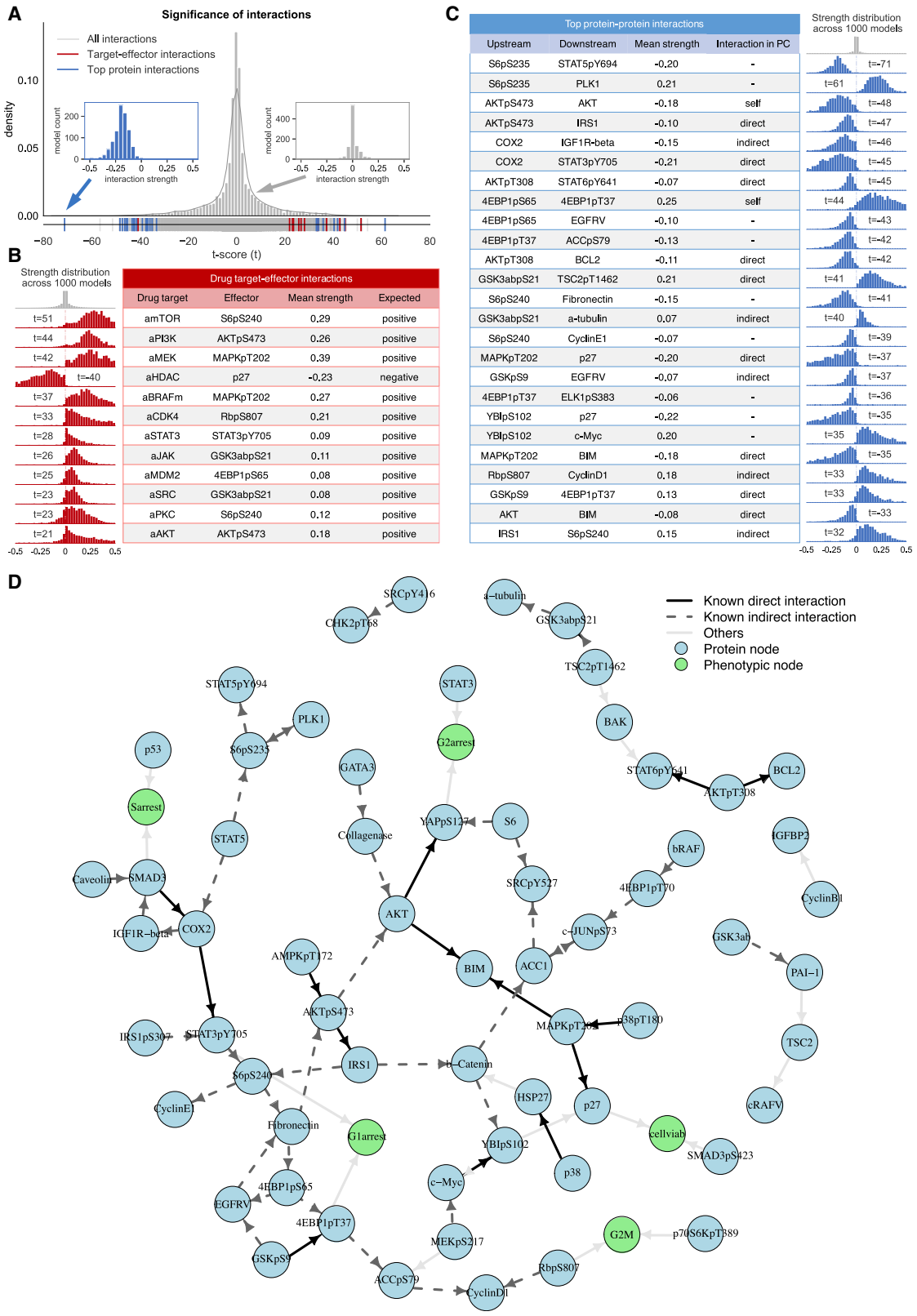


Figure 5. Comparison of the Network Models with Prior Knowledge about Pathways

(A) The t score distribution of all interactions across 1,000 full models suggests that a small fraction of interaction strengths is significantly different from zero. Insets are two examples of interaction strength distributions across models.

(legend continued on next page)

(MEK1) indirectly interacts in two steps with the transcription factor c-Myc by a phosphorylation mechanism via ERK1/2 (Gupta and Davis, 1994; Butch and Guan, 1996; Sears et al., 2000; Aoki et al., 2011). Additional evidence of functional links between proteins identified in the models includes protein-protein edges in the STRING database (von Mering et al., 2005), such as links that represent expression (mRNA) correlation, and PubMed-derived links from text mining (Maglott et al., 2011).

In order to confirm that such agreement between the inferred networks and prior knowledge is not an artifact, we examined solution stability as well as compared the inferred networks to random networks. We conducted stability-selection tests (Meinshausen and Bühlmann, 2010) on the interaction parameters, and the results indicate that the solutions are reproducible and robust (Figure S9A). We compared the inferred networks to networks with randomly drawn interactions (STAR Methods). In the inferred networks, we observed a statistically significant higher number of top interactions consistent with the knowledge in PC, indicating that such agreement is not an artifact of the densely connected signaling networks in the pathway database (Figure S9B). The remaining model interactions are between components that are more than two pathway steps away or cannot be connected by any path in PC. These interactions can be interpreted as either logical interactions important for predictive purposes or potential new physical interactions that are yet to be discovered in molecular experiments. Also, the PC database aggregates interaction information from multiple biological systems, and, therefore, complete consistency is not expected for any particular system, e.g., the melanoma cell line used here. Given that the CellBox network models are constructed in a completely data-driven way without any a priori intention to recapitulate known molecular interactions, the partial agreement between the model-inferred molecular interactions and the experimentally known ones from the literature is evidence of the validity of the modeling approach.

Predictions of Unseen Perturbations Give Candidates for Drug Combinations

Our results so far indicate that the CellBox models can be efficiently trained on a relatively small set of experimental data to parameterize the differential equations that model the behavior of the entire system of nodes and interactions at a reasonable level of predictive accuracy. This model can then predict cell responses to a full range of single and combinatorial unseen perturbations, which would be laborious and costly to test exhaustively by experiments. In order to nominate effective drug combinations for a much reduced number of focused experiments, we used simulations of the 1,000 full models to quantitatively predict the dynamic cell responses to ~110,000 *in silico* perturbations, including different dosages of single perturbations on each protein node as well as all pairwise combinations (STAR Methods). For each perturbation condition, we averaged the predictions across all models and ranked the perturbations by predicted phenotypic changes (Figure 6A).

Previous models on the same dataset, using the same differential equations but parameterized by using BP, had predicted that two drug pairs, MEKi+c-Myc and RAFi+c-Myc, would increase G1 cell-cycle arrest, and this prediction was confirmed by experiments (Korkut et al., 2015). We found that the CellBox model predicts similar effects for these two drug pairs (Figure 6B, panel a, b). In order to identify additional therapeutic candidates, we examined the effects of all possible single and pairwise perturbations on cell-cycle arrest (Figure 6B). The top-ranked candidates included dominant antiproliferative inhibition (uniform colors in rows or columns) of proteins in the Wnt, MAPK, and ERK/MEK pathways, which are known to be cancer related. Besides strong single candidates, synergistic drug pairs are of potential therapeutic interest (Figure 6B; departure from uniform colors). Inhibitory perturbations predicted to have pro-proliferation effects, which are undesirable as such, can also lead to effective antiproliferative candidates via indirectly activating perturbations (top left corner). For example, protein nodes can, in principle, be activated by reducing upstream inhibition or degradation. As the CellBox models are completely data driven, the *de novo* predictions represent system-specific predictions independent of prior knowledge.

DISCUSSION

Quantitative models that are predictive of dynamic cellular responses can be used to design combination therapies in cancer. To provide predictions with sufficient accuracy and potential mechanistic insight, we integrated machine-learning methods with dynamic modeling: we applied an optimization algorithm used in deep learning to a biologically interpretable differential equation (ODE) system. Our model CellBox can be trained efficiently and independently of prior knowledge to predict molecular and phenotypic responses to unseen perturbations with high accuracy. Although trained on a relatively small set of experiments, the model is capable of simulating cell responses to numerous arbitrary combinatorial perturbations applied to nodes repeatedly measured under different perturbation conditions. Ranking of cellular responses to the *in silico* combinatorial perturbations by the desired phenotypic outcome, such as decreased proliferation, potentially leads to specific therapeutic hypotheses.

Interpretability of models that are to be used for practical decisions, such as the design of combination therapy, helps increase confidence and facilitates the design of focused

(B) All 12 interactions between drug target (drug-activity nodes) and their downstream effectors (red bars in A) are significant, and the interaction directions are consistent with the literature. (C) Most of the top significant protein-protein interactions (blue bars in A) can be found as direct or indirect interactions in PC. The distributions of interaction strength across 1,000 models for each interaction in the two tables with corresponding colors are centered away from zero, in contrast to the background distributions of aggregated interactions across models (gray, all interactions with drug-activity nodes in A, all protein-protein interactions in C). All other interactions with t scores and PC information are included in Table S2.
(D) Network visualization of the top interactions (top two interactions acting from each node and onto each node) highlights the level of agreement between model-inferred interactions and those in PC.

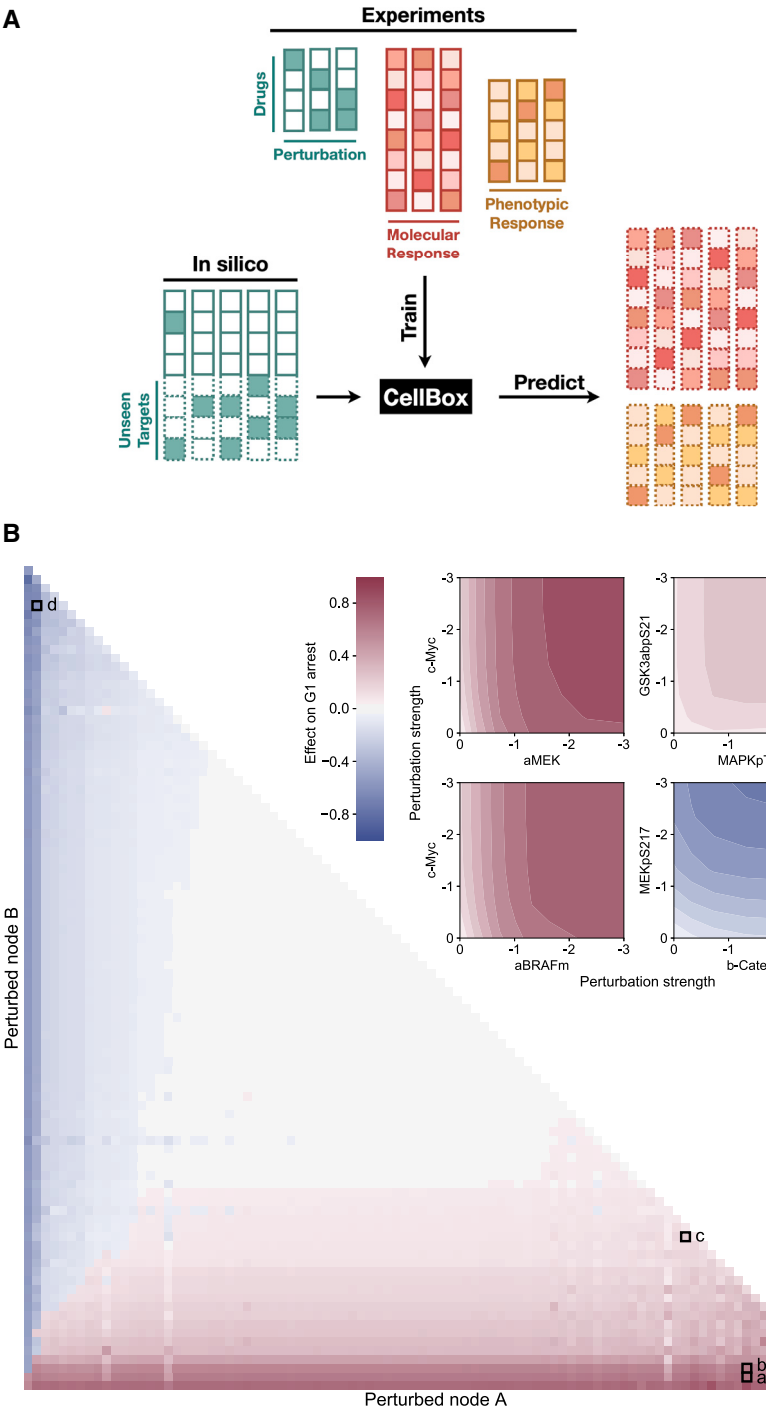


Figure 6. CellBox Provides Testable Predictions of Cell Phenotype under Synthetic Perturbations

(A) For each (phospho)protein node in the network, CellBox was used to simulate all single and paired inhibitions and to predict the phenotypic changes. The phenotypic effects are the average prediction of 1,000 independent models trained on the full datasets.

(B) The antiproliferation effects of two perturbation pairs whose effects on cell -cycle arrest have been experimentally tested were closely examined (left two panels, c-Myc+MEKi [a], and c-Myc+RAFi [b]), as well as two other *in silico* conditions (right two panels, GSK3p+MAPKp [c] and MEKp+β-catenin [d]), by simulating with combinatorial perturbation strengths. The effects on cell-cycle arrest of pairwise combinatorial perturbation of all (phospho) proteins in the network were simulated and can be used to nominate effective pharmaceutical candidates. These *in silico* inhibitory perturbations can result in antiproliferation effects (red, bottom right) or pro-proliferation effects (blue, top left). A complete table of all predicted pairwise combinatorial perturbation effects on other phenotypes is in Table S3.

validation experiments and is, therefore, as important as accuracy (Lipton, 2018). CellBox features interpretability in two aspects: transparency and traceability. *Transparency*: by using a well-defined mathematical model, CellBox is designed to be explicitly interpretable. In the current ODE model, each parameter represents a directed and quantitative interaction between cellular components or phenotypic quantities. *Traceability*: given a perturbation to the cellular system, the ODE simulation indicates how the effects of the perturbation propagate throughout the directed network in a time-dependent manner. The models can, therefore, provide mechanistic hypotheses of how the perturbations cause the observable cellular responses. Our model aims to meet the emerging demands of explainable artificial intelligence (XAI), which is essential for a socially acceptable application of machine-learning models in biomedicine (Thelisson, 2017; Gunning and Aha, 2019).

In principle, CellBox is generalizable to other types of systems and larger systems. Other types of models will presumably benefit from automatic differentiation (AD) combined with stochastic gradient descent that performs optimization directly for any given mathematical ansatz and, therefore, can avoid oversimplified approximations (Baydin et al., 2018). The AD framework allows the flexible use of various mathematical forms of cellular dynamics. Whether alternative dynamics models lead to higher concordance between interactions inferred from the data-driven CellBox models with interactions in prior-knowledge mechanistic models or with biophysical interactions separately verified by direct experiments is to be investigated. We certainly expect increasingly detailed model-inferred interactions and closer agreement with biophysical interactions as larger datasets become available, i.e., with perturbation-response data covering many hundreds if not thousands of changes in molecular levels or covalent modifications.

The ability to model larger systems depends both on the availability of larger datasets and scalable modeling methods. Larger datasets can be obtained by measuring diverse types of molecular data, for example, transcriptomic, epigenomic, and metabolomic changes (Brown et al., 2014; Zaal and Berkers, 2018). A major opportunity for larger datasets might arise from recent cell bar-coding techniques that significantly increase perturbation throughput relative to arrayed experiments (Adamson et al., 2016; Dixit et al., 2016) by measuring levels of transcripts by sequencing, levels of proteins detected by antibodies labeled by oligonucleotides or isotopes at the single-cell level, or levels of both by multiplexing (Frei et al., 2016; Wroblewska et al., 2018; Mimitou et al., 2019; Schraivogel et al., 2020). A key advantage of single-cell perturbation approaches for data-driven inference of dynamic networks is scale, as bar-coded perturbation experiments can be pooled and individual cells identified on the basis of sequence tags.

As CellBox is implemented in the Google TensorFlow framework, it can make use of various advanced machine-learning techniques, such as dropout, mini-batching, and GPU boosting (Bengio, 2012; Liang and Liu, 2015), to improve training efficiency, which partially addresses the issue of scalability (Stapor et al., 2019). As most state-of-the-art single-cell sequencing technologies still suffer from limited sequencing depth and inevitable noise (Pratapa et al., 2020), a new set of computational challenges arises when applying CellBox models to datasets with single-cell readouts. Future efforts are needed to resolve issues of sparsity and stochasticity, e.g., using robust techniques of dimensionality reduction (Lopez et al., 2018; Eraslan et al., 2019; Lotfollahi et al., 2019). Recent work has highlighted mathematical analogies between recurrent neural networks (RNN) and ODEs (Chen et al., 2018), suggesting a potential merge of the two fields for scientific machine-learning (SciML) models (Rackauckas et al., 2020).

A particular advantage of our modeling approach is its complete independence from prior knowledge. This is in contrast, e.g., to other frameworks for interpretable models (Fröhlich et al., 2018) that incorporate prior knowledge by predefining the connections in the network model based on large-scale curation of molecular and biochemical pathways. Such models have been further developed and incorporated with proteome/phosphoproteome data (Schmiester et al., 2020). Although our current modeling framework is completely data driven, prior knowledge of cellular interactions could also be included in the optimization function by adding a term, which rewards the agreement between the inferred and prior-knowledge values for each interaction parameter. However, considering the incompleteness of current pathway knowledge bases and the relatively high prediction accuracy of CellBox, we speculate that rewarding agreement with prior knowledge might give limited improvement in model performance. Besides, it would also be of interest to explore whether one can combine the efficiency advantage described in Fröhlich et al. (2018) via adjoint analysis (Fröhlich et al., 2017) with the advantages of the current implementation of CellBox.

For broad translational applicability, a tantalizing but challenging prospect is to incorporate individual tumor background via genetic perturbations and to propose optimal, personalized combinations of targeted therapeutics. We envision the systems biology approach described here to be broadly applicable to other areas of biology, such as developmental biology or synthetic biology, provided that suitable perturbation-response data become available. Key future challenges are, therefore, the design of experiments for each biological context of interest and the further development of transferable and scalable machine-learning methods.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- METHOD DETAILS
  - Perturbation Dataset Overview
  - Model Configuration
  - Model Training
  - Cross-Validations
  - Sensitivity Analysis
  - Network Interpretation
  - Model Predictions
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Model Performance Analysis
  - Network Analysis

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.11.013.

ACKNOWLEDGMENTS

We thank Aviv Regev, Alexandra Franz, Frank Poelwijk, Laura Kleiman, Nicholas Gauthier, Haozhe Shan, William Yuan, Han Altae-Tran, and members of the Sander and Marks labs for constructive discussions. Funding support from Dana-Farber Cancer Institute, National Human Genome Research Institute (U41HG006623), and National Institute of General Medical Sciences (P41GM103504).

AUTHOR CONTRIBUTIONS

Conceptualization, C. Sander and J.I.; Methodology and Software, B.Y., C. Shen, A.L., and J.I.; Writing - Original Draft, B.Y., C. Shen, and C. Sander; Writing - Review & Editing, B.Y., C. Shen, A.L., A.K., D.S.M., J.I., and C. Sander; Resources, A.K., and C. Sander; Visualization, B.Y. and C. Shen; Funding Acquisition, C. Sander; Supervision, C. Sander.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 17, 2020
Revised: July 13, 2020
Accepted: November 25, 2020
Published: December 28, 2020

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2015). TensorFlow: large-scale machine learning on heterogeneous distributed systems, arXiv https://arxiv.org/abs/1603.04467.

Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell *167*, 1867–1882.e21.

Aldridge, B.B., Saez-Rodriguez, J., Muhlich, J.L., Sorger, P.K., and Lauffenburger, D.A. (2009). Fuzzy logic analysis of kinase pathway crosstalk in TNF/EGF/insulin-induced signaling. PLoS Comp. Biol. *5*, e1000340.

Aoki, K., Yamada, M., Kunida, K., Yasuda, S., and Matsuda, M. (2011). Processive phosphorylation of ERK MAP kinase in mammalian cells. Proc. Natl. Acad. Sci. USA *108*, 12675–12680.

Azmi, A.S., Wang, Z., Philip, P.A., Mohammad, R.M., and Sarkar, F.H. (2010). Proof of concept: network and systems biology approaches aid in the discovery of potent anticancer drug combinations. Mol. Cancer Ther. *9*, 3137–3144.

Babur, O., Demir, E., Gönen, M., Sander, C., and Dogrusoz, U. (2010). Discovering modulators of gene expression. Nucleic Acids Res *38*, 5648–5656.

Bayat Mokhtari, R.B., Homayouni, T.S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., and Yeger, H. (2017). Combination therapy in combating cancer. Oncotarget *8*, 38022–38043, https://doi.org/10.18632/oncotarget.16723.

Baydin, A.G., Pearlmutter, B.A., Radul, A.A., and Siskind, J.M. (2018). Automatic differentiation in machine learning: a survey. J. Mach. Learn. Res. *18*, 1–43.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, G. Montavon, G.B. Orr, and K.R. Müller, eds. (Springer), pp. 437–478, https://doi.org/10.1007/978-3-642-35289-8_26.

Brown, R., Curry, E., Magnani, L., Wilhelm-Benartzi, C.S., and Borley, J. (2014). Poised epigenetic states and acquired drug resistance in cancer. Nat. Rev. Cancer *14*, 747–753.

Bruggeman, F.J., Westerhoff, H.V., Hoek, J.B., and Kholodenko, B.N. (2002). Modular response analysis of cellular regulatory networks. J. Theor. Biol. *218*, 507–520.

Butch, E.R., and Guan, K.L. (1996). Characterization of ERK1 activation site mutants and the effect on recognition by MEK1 and MEK2. J. Biol. Chem. *271*, 4230–4235.

Carreira, S., Goodall, J., Aksan, I., La Rocca, S.A., Galibert, M.D., Denat, L., Larue, L., and Goding, C.R. (2005). Mitf cooperates with Rb1 and activates p21Cip1 expression to regulate cell cycle progression. Nature *433*, 764–769.

Carter, S.L., Brechbühler, C.M., Griffin, M., and Bond, A.T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics *20*, 2242–2250.

Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res *39*, D685–D690.

Chan, T.E., Stumpf, M.P.H., and Babtie, A.C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst *5*, 251–267.e3.

Chandarlapaty, S., Sawai, A., Scaltriti, M., Rodrik-Outmezguine, V., Grbovic-Huezo, O., Serra, V., Majumder, P.K., Baselga, J., and Rosen, N. (2011). AKT inhibition relieves feedback suppression of receptor tyrosine kinase expression and activity. Cancer Cell *19*, 58–71.

Chen, R.T.Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2018). Neural ordinary differential equations, arXiv http://arxiv.org/abs/1806.07366.

Cheng, F., Kovács, I.A., and Barabási, A.L. (2019). Network-based prediction of drug combinations. Nat. Commun. *10*, 1197.

Cheung, H.W., Cowley, G.S., Weir, B.A., Boehm, J.S., Rusin, S., Scott, J.A., East, A., Ali, L.D., Lizotte, P.H., Wong, T.C., et al. (2011). Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc. Natl. Acad. Sci. USA *108*, 12372–12377.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., et al. (2014). The reactome pathway KnowledgeBase. Nucleic Acids Res *42*, D472–D477.

D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics *16*, 707–726.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*, 1853–1866.e17.

Donovan, J.C., Milic, A., and Slingerland, J.M. (2001). Constitutive MEK/MAPK activation leads to p27(Kip1) deregulation and antiestrogen resistance in human breast cancer cells. J. Biol. Chem. *276*, 40888–40895.

Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. *10*, 390.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.

Fitzgerald, J.B., Schoeberl, B., Nielsen, U.B., and Sorger, P.K. (2006). Systems biology and combination therapy in the quest for clinical efficacy. Nat. Chem. Biol. *2*, 458–466.

Frei, A.P., Bava, F.A., Zunder, E.R., Hsieh, E.W., Chen, S.Y., Nolan, G.P., and Gherardini, P.F. (2016). Highly multiplexed simultaneous detection of RNAs and proteins in single cells. Nat. Methods *13*, 269–275.

Fröhlich, F., Kaltenbacher, B., Theis, F.J., and Hasenauer, J. (2017). Scalable parameter estimation for genome-scale biochemical reaction networks. PLoS Comp. Biol. *13*, e1005331.

Fröhlich, F., Kessler, T., Weindl, D., Shadrin, A., Schmiester, L., Hache, H., Muradyan, A., Schütte, M., Lim, J.H., Heinig, M., et al. (2018). Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. Cell Syst *7*, 567–579.e6.

Gardner, T.S., di Bernardo, D., Lorenz, D., and Collins, J.J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. Science *301*, 102–105.

Garraway, L.A., and Jänne, P.A. (2012). Circumventing cancer drug resistance in the era of personalized medicine. Cancer Discov *2*, 214–226.

Gunning, D., and Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. AIMag *40*, 44–58, https://doi.org/10.1609/aimag.v40i2.2850.

Gupta, S., and Davis, R.J. (1994). MAP kinase binds to the NH2-terminal activation domain of c-Myc. FEBS Lett *353*, 281–285.

Hill, S.M., Nesser, N.K., Johnson-Camacho, K., Jeffress, M., Johnson, A., Boniface, C., Spencer, S.E., Lu, Y., Heiser, L.M., Lawrence, Y., et al. (2017). Context specificity in causal signaling networks revealed by phosphoprotein profiling. Cell Syst *4*, 73–83.e10.

Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., and Saltz, J.H. (2016). Patch-based convolutional neural network for whole slide tissue image classification. Proc IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. *2016*, 2424–2433.

Hug, S., Raue, A., Hasenauer, J., Bachmann, J., Klingmüller, U., Timmer, J., and Theis, F.J. (2013). High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling. Math. Biosci. *246*, 293–304.

Kingma, D.P., and Ba, J. (2014). Adam: a method for stochastic optimization, arXiv http://arxiv.org/abs/1412.6980.

Klinger, B., Sieber, A., Fritsche-Guenther, R., Witzel, F., Berry, L., Schumacher, D., Yan, Y., Durek, P., Merchant, M., Schäfer, R., et al. (2013). Network quantification of EGFR signaling unveils potential for targeted combination therapy. Mol. Syst. Biol. *9*, 673.

Korkut, A., Wang, W., Demir, E., Aksoy, B.A., Jing, X., Molinelli, E.J., Babur, Ö., Bemis, D.L., Onur Sumer, S., Solit, D.B., et al. (2015). Perturbation biology nominates upstream-downstream drug combinations in RAF inhibitor resistant melanoma cells. eLife *4*, e04640, https://doi.org/10.7554/eLife.04640.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science *313*, 1929–1935.

Lei, W., Liu, F., and Ness, S.A. (2005). Positive and negative regulation of c-Myb by cyclin D1, cyclin-dependent kinases, and p27 Kip1. Blood *105*, 3855–3861.

Lezon, T.R., Banavar, J.R., Cieplak, M., Maritan, A., and Fedoroff, N.V. (2006). Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. Proc. Natl. Acad. Sci. USA *103*, 19033–19038.

Liang, J., and Liu, R. (2015). Stacked denoising autoencoder and dropout together to prevent overfitting in deep neural network (8th International Congress on Image and Signal Processing (CISP)), pp. 697–701, https://doi.org/10.1109/cisp.2015.7407967.

Lipton, Z.C. (2018). The mythos of model interpretability. Commun. ACM *61*, 36–43, https://doi.org/10.1145/3233231.

Locasale, J.W., and Wolf-Yadlin, A. (2009). Maximum entropy reconstructions of dynamic signaling networks from quantitative proteomics data. PLoS One *4*, e6522.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058.

Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721.

Luna, A., Babur, Ö., Aksoy, B.A., Demir, E., and Sander, C. (2016). PaxtoolsR: pathway analysis in R using pathway commons. Bioinformatics *32*, 1262–1264, https://doi.org/10.1093/bioinformatics/btv733.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2011). Entrez gene: gene-centered information at NCBI. Nucleic Acids Res *39*, D52–D57.

Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S., and Baradaran, B. (2017). The different mechanisms of cancer drug resistance: a brief review. Adv. Pharm. Bull. *7*, 339–348.

McDonald, E.R., 3rd, de Weck, A., Schlabach, M.R., Billy, E., Mavrakis, K.J., Hoffman, G.R., Belur, D., Castelletti, D., Frias, E., Gampa, K., et al. (2017). Project DRIVE: a compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep RNAi screening. Cell *170*, 577–592.e10.

Meinshausen, N., and Bühlmann, P. (2010). Stability selection. J. R. Stat. Soc. B *72*, 417–473, https://doi.org/10.1111/j.1467-9868.2010.00740.x.

Meyer, P.E., Lafitte, F., and Bontempi, G. (2008). minet: A R/bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics *9*, 461.

Mimitou, E.P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., Ouyang, Z., et al. (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. Nat. Methods *16*, 409–412.

Molinelli, E.J., Korkut, A., Wang, W., Miller, M.L., Gauthier, N.P., Jing, X., Kaushik, P., He, Q., Mills, G., Solit, D.B., et al. (2013). Perturbation biology: inferring signaling networks in cellular systems. PLoS Comp. Biol. *9*, e1003290.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. Digit. Signal Process. *73*, 1–15, https://doi.org/10.1016/j.dsp.2017.10.011.

Nelander, S., Wang, W., Nilsson, B., She, Q.B., Pratilas, C., Rosen, N., Gennemark, P., and Sander, C. (2008). Models from experiments: combinatorial drug perturbations of cancer cells. Mol. Syst. Biol. *4*, 216.

Niepel, M., Hafner, M., Duan, Q., Wang, Z., Paull, E.O., Chung, M., Lu, X., Stuart, J.M., Golub, T.R., Subramanian, A., et al. (2017). Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. Nat. Commun. *8*, 1186.

Norman, T.M., Horlbeck, M.A., Replogle, J.M., Ge, A.Y., Xu, A., Jost, M., Gilbert, L.A., and Weissman, J.S. (2019). Exploring genetic interaction manifolds constructed from rich phenotypes. bioRxiv. https://doi.org/10.1101/601096.

Nyman, E., Stein, R.R., Jing, X., Wang, W., Marks, B., Zervantonakis, I.K., Korkut, A., Gauthier, N.P., and Sander, C. (2020). Perturbation biology links temporal protein changes to drug responses in a melanoma cell line. PLoS Comp. Biol. *16*, e1007909.

Osaki, L.H., and Gama, P. (2013). MAPK signaling pathway regulates p27 phosphorylation at threonin 187 as part of the mechanism triggered by early-weaning to induce cell proliferation in rat gastric mucosa. PLoS One *8*, e66651.

Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T.M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods *17*, 147–154.

Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., and Edelman, A. (2020). Universal differential equations for scientific machine learning. arXiv. https://doi.org/10.21203/rs.3.rs-55125/v1.

Ryall, K.A., and Tan, A.C. (2015). Systems biology approaches for advancing the discovery of effective drug combinations. J. Cheminform. *7*, 7.

Schmiester, L., Schälte, Y., Fröhlich, F., Hasenauer, J., and Weindl, D. (2020). Efficient parameterization of large-scale dynamic models based on relative measurements. Bioinformatics *36*, 594–602.

Schraivogel, D., Gschwind, A.R., Milbank, J.H., Leonce, D.R., Jakob, P., Mathur, L., Korbel, J.O., Merten, C.A., Velten, L., and Steinmetz, L.M. (2020). Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat. Methods *17*, 629–635.

Sears, R., Nuckolls, F., Haura, E., Taya, Y., Tamai, K., and Nevins, J.R. (2000). Multiple Ras-dependent phosphorylation pathways regulate Myc protein stability. Genes Dev *14*, 2501–2514, https://doi.org/10.1101/gad.836800.

Şenbabaoğlu, Y., Sümer, S.O., Sánchez-Vega, F., Bemis, D., Ciriello, G., Schultz, N., and Sander, C. (2016). A multi-method approach for proteomic network inference in 11 human cancers. PLoS Comp. Biol. *12*, e1004765.

Städter, P., Schälte, Y., Schmiester, L., Hasenauer, J., and Stapor, P.L. (2020). Benchmarking of numerical integration methods for ODE models of biological systems. biorXiv. https://doi.org/10.1101/2020.09.03.268276.

Stapor, P., Schmiester, L., Wierling, C., Lange, B.M.H., Weindl, D., and Hasenauer, J. (2019). Mini-batch optimization enables training of ODE models on large-scale datasets. biorXiv. https://doi.org/10.1101/859884.

Süli, E., and Mayers, D.F. (2003). An Introduction to Numerical Analysis (Cambridge University Press).

Thelisson, E. (2017). Towards trust, transparency and liability in AI / AS systems. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, pp. 5215–5216, 10.24963/ijcai.2017/767.

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G.B., and Kornblau, S.M. (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol. Cancer Ther. *5*, 2512–2521.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a cancer dependency map. Cell *170*, 564–576.e16.

Vanhaelen, Q., Aliper, A.M., and Zhavoronkov, A. (2017). A comparative review of computational methods for pathway perturbation analysis: dynamical and topological perspectives. Mol. Biosyst. *13*, 1692–1704.

von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res *33*, D433–D437.

Wang, K., Saito, M., Bisikirska, B.C., Alvarez, M.J., Lim, W.K., Rajbhandari, P., Shen, Q., Nemenman, I., Basso, K., Margolin, A.A., et al. (2009). Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nat. Biotechnol. *27*, 829–839.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Wroblewska, A., Dhainaut, M., Ben-Zvi, B., Rose, S.A., Park, E.S., Amir, E.D., Bektesevic, A., Baccarini, A., Merad, M., Rahman, A.H., and Brown, B.D. (2018). Protein barcodes enable high-dimensional single-cell CRISPR screens. Cell *175*, 1141–1155.e16..

Wrzodek, C., Büchel, F., Ruff, M., Dräger, A., and Zell, A. (2013). Precise generation of systems biology models from KEGG pathways. BMC Syst. Biol. *7*, 15.

Yi, S., Lin, S., Li, Y., Zhao, W., Mills, G.B., and Sahni, N. (2017). Functional variomics and network perturbation: connecting genotype to phenotype in cancer. Nat. Rev. Genet. *18*, 395–410.

Zaal, E.A., and Berkers, C.R. (2018). The influence of metabolism on drug response in cancer. Front. Oncol. *8*, 500, https://doi.org/10.3389/fonc.2018.00500.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.

Zou, M., and Conzen, S.D. (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics *21*, 71–79.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Processed data files	Korkut et al., 2015	https://github.com/dfci/CellBox
Software and Algorithms		
TensorFlow	Abadi et al., 2015	v1.15.0
Python	https://www.python.org/	v3.6
R	https://www.r-project.org/	v3.6.1
Belief propagation for perturbation biology	Korkut et al., 2015	https://github.com/korkutlab/pertbio
Pathway Commons	https://www.pathwaycommons.org/	v11
igraph	https://igraph.org/r/	v1.2.4.2
STRING	https://string-db.org/	v11.0
Custom code	This manuscript	https://github.com/dfci/CellBox

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Chris Sander (mathcellbox@gmail.com).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

- The RPPA data used during this study is deposited and publicly available at https://github.com/dfci/CellBox under the data folder.
- All original code is publicly available at https://github.com/dfci/CellBox as release v0.3.1. A shareable, online, interactive environment using Binder is provided with all necessary dependencies pre-installed. The *Quick Start* section provides instructions to quickly try an example script that runs a shortened training process.
- The scripts used to generate the figures reported in this paper are available at https://github.com/dfci/CellBox under the manuscript folder.
- Any additional information required to reproduce this work is available from the Lead Contact.

METHOD DETAILS

Perturbation Dataset Overview

The CellBox models were trained using a perturbation-response dataset of the SK-Mel-133 melanoma cell line (Korkut et al., 2015) (Figure S1). The cells were treated with 12 different single drugs, each at two different concentrations and 66 pairwise combinations of these drugs at IC40 concentrations. 24 h after drug treatment, Reverse Phase Protein Arrays (RPPA) was used to measure the level of 45 proteins and 37 phosphoproteins of interest. Cell- cycle progression, including G1 arrest, G2 arrest, G2/M transition, and S arrest, was measured by flow cytometry. Cell viability was measured 72 h after drug treatment by the resazurin assay. Our experimental data were normalized using median normalization, which is a standard approach to processing RPPA data (Tibes et al., 2006). We used the $\log_2$ ratio of measurements in perturbed conditions over unperturbed conditions as the system variables x_i in the models. The dataset was initialized with 12 drug activity nodes representing the inhibition strengths of different drugs to their targets (Molinelli et al., 2013). The resulting dataset has 89 perturbation conditions and 99 observed nodes (82 protein and phosphoproteins, 5 phenotypes, and 12 drug activity). A more detailed description of the experimental dataset is available in Korkut et al. (Korkut et al., 2015).

Model Configuration

The models were constructed using Python 3.6 and Google TensorFlow (Abadi et al., 2015) (version = 1.15.0). The molecular and phenotypic changes are linked in a unified biological network model using a system of ordinary differential equations

$$\frac{\partial x_i^{\mu}(t)}{\partial t} = \epsilon_i \varphi\left(\sum_{j \neq i} w_{ij} x_j^{\mu}(t) + u_i^{\mu}(t)\right) - \alpha_i x_j^{\mu}(t) \quad \text{(Equation 1)}$$

where $x_i^{\mu}(t)$represents the $\log_2$-normalized relative change of each (phospho)protein or phenotype levels relative to control levels under condition μ. $u_i^{\mu}(t)$ quantifies the strength of the perturbation on target (i). Here the drug effect is assumed to be constant and, therefore, $u(t) = u$ for $t > t_0$ is the perturbation strength determined using the endpoint level change of the primary target of the particular drug (Figure S1B). characterizes the effect of decay, meaning the tendency of protein i to return to the original level before perturbation. The interaction parameters w_{ij} indicate interactions between network node j on network node i, assumed to be a constant property of the pair of molecules in this given cellular setting. We constrain the interaction parameters w_{ij} by disallowing three classes of interactions:

i) ingoing connections for drug nodes (drugs cannot be acted upon by any other node)
ii) outgoing connections for phenotypic nodes (phenotypes cannot act on any other nodes)
iii) self-interaction (nodes cannot act on themselves)

We use a sigmoid function $\varphi(x) = tanh(x)$, to introduce a saturation effect of the interaction term so that it is bounded by the constant value of ε_i. Imposing bounds on the absolute value of the contribution of the interaction term to the derivatives is in part motivated by the fact that we are only modeling a small fraction of cellular components and by the experimental observation of system stability in these experiments. We also tested different alternatives of envelope forms, including a clipped linear function (hard tanh) $\varphi(x) = max(-1, min(1, x))$, symmetric polynomial function $\varphi(x) = \frac{\tilde{x}^n - 1}{\tilde{x}^n + 1}, \tilde{x} = max(0, x)$ (adapted from Hill's equation), and no envelope function $\varphi(x) = x$ (linear) (Figure S5).

The biological network interactions were constructed *de novo* without any prior knowledge input, meaning the network was fully connected with interaction parameters. The interaction parameters w_{ij} were randomly initialized following a normal distribution $\sim N(0.01, 1)$. The other two coefficients α_i and ϵ_i were initialized as 1.0.

Taken together, with this formulation of system dynamics, the effect of each new input perturbation is quantified by the response of downstream protein and phenotypic nodes, and the effect of the perturbation is simulated by propagating the input across the inferred interaction network according to the differential equations. After learning from perturbation-response data, the model can therefore make predictions for responses of perturbations on any experimentally probed node, including nodes not perturbed in the experimental dataset from which the models are derived. The ODE system was numerically solved using Heun's method (Süli and Mayers, 2003) (Equation 2, time steps $N_t = 400$, Figure S2), which is an improved variant of Euler's method. Model performance was evaluated by disagreement between the experimental cell responses and the numerical steady-state levels.

$$\tilde{x}_i(t+h) = x_i(t) + hf(t, x_i(t))$$

$$x_i(t+h) = x_i(t) + \frac{h}{2}\left[f(t, x_i(t)) + f\left(t+h, \tilde{x}_i(t+h)\right)\right] \quad \text{(Equation 2)}$$

where h is the step size, $f(t, x_i(t)) = x'(t)$, $y(t_0) = log(1) = 0$

The loss function $L(w)$ is defined as a weighted sum of prediction error and complexity penalty in order to avoid overfitting. Here a mean squared error (MSE) and an L1-loss regularization term are used, as defined in Equation 3. We have tested different regularizations, including L1, L2, and both combined (elastic net) with various strengths (regularization strengths for L1, L2 term λ_1, λ_2= 0, 0.01, 0.0001, pairwise combination), and found no significant difference in model performance (Figure S2B). The interaction parameters were optimized end-to-end using the Adam optimizer(Kingma and Ba, 2014), with the objective of minimizing the loss function.

$$L(w) = \sum_{\mu}\sum_{i} \|\hat{x}_i(w, t) - x_i^{\mu *}\|_2 + \lambda_1 \|w\|_1 \quad \text{(Equation 3)}$$

$\hat{x}_i^{\mu}(w, t)$ is calculated as the converged value of the numerical simulation of the ODE with the interaction parameters w and defined simulation timestep t. $x_i^{\mu *}$ indicates experimental measurement, which is used as a gold standard in training. The dataset was divided into training, validation, and test sets, in order to optimize parameters, provide an indication for stopping training, and test model performance, respectively. Optimization was conducted with an initial learning rate for the Adam optimizer (lr=0.1) and regularization strength (λ=0.01). It has been shown that a gradually decreasing learning rate is helpful for model convergence (Bengio, 2012). The model training was stopped when the loss function of the validation set does not further decrease for a continuous of 20 iterations (stopping patience).

The model was trained with mini-batching: a random 80% portion of the training set was used to optimize parameters for each iteration, and a fixed batch size of 4 was used in each epoch. Models that failed to converge (MSE for training set > 0.05) were excluded as unsuitable. For a larger dataset, we recommend using a fixed batch size of 8 or 16, as documented in the latest version of the software on GitHub.

Model Training

For initial model training and analysis of model performance, the cell line perturbation-response dataset was randomly partitioned into training, validation, and test set in the proportion of 56% (n = 50 conditions), 14% (n = 12 conditions), and 30% (n = 27 conditions). 1,500 models were generated on 1,500 independently random-partitioned datasets.

The models were examined and categorized into non-oscillating and oscillating solutions based on time derivatives at the final time step of the ODE simulation. The non-oscillating solutions are defined as those with the average absolute value of time derivatives of all nodes and conditions in the training set smaller than δ, i.e. $\frac{1}{m}\sum_{i=1}^{m}\left|\frac{\partial x_i{}^{\mu}(t)}{\partial t}\right| < \delta; \delta = 1e-03$. In each category, twenty models were randomly selected, and each re-trained with the original data partitioning but forty different random seeds, covering all the random processes in training, including parameter initialization and mini-batching sampling (Figure S4). Oscillating solutions comprise about 30 percent of all models. In the following analysis, models that converged to oscillating solutions were excluded.

Under the training scheme using random data partition, we examined the sensitivity to data scaling and to the choice of ODE solver, by applying the following modifications to the data or the model and evaluating the changes of model performance. i) In addition to the $\log_2$ scale of the ratio of measurements in perturbed conditions over unperturbed conditions, we applied other scalings to the raw data to generate model input, including a non-log (linear) scale (Figure S3C). ii) It has been argued that the selection of numerical methods could have a significant influence on model training (Städter et al., 2020). To test such limitations, we tested other numerical ODE solvers, including Euler methods, Midpoint methods, and Runge-Kutta methods (Figures S4E–S4G), in addition to Heun's ODE solver. On the current dataset, these different methods perform similarly.

Cross-Validations

In the single-to-combo task, all single-drug conditions were allocated to the training set (n = 23 conditions), and the combination perturbation conditions were randomly distributed among the validation and test set (n = 53 conditions) in a 20/80 ratio. To evaluate model performance by cross-validation for each drug, the data were partitioned into training (n = 78 conditions) and test (n = 11 conditions) sets, where each test set contains all the drug combination conditions with the particular drug. 20% of training conditions (n = 15) are used as a validation set. The predictions on the test set were averaged over 100 models. In the more stringent leave-one-drug-out task where all perturbation conditions involving one drug were withheld as the test set, the evaluation of model performance was conducted by applying perturbation of the same strength directly on the downstream effector node of each withheld drug activity node (Figure S1B).

The Belief Propagation (BP) models for both single-to-combo and cross-validation prediction were performed as in our earlier publication (https://github.com/korkutlab/pertbio). The predictions on the test set were averaged over 100 models. The deep neural network model (NN) network had 2 densely connected hidden layers (hidden layer H1: 20 neurons, H2: 100 neurons), connecting the parameterized perturbation tensor with the cell response tensor. We used the *tanh* activation function for the hidden layers and the leaky rectified linear unit function (ReLU) for the output layer. The NN models were constructed in the TensorFlow framework in Python and optimized using the same optimization methods (Adam optimizer). The co-expression static model (Co-exp) was constructed in the Python environment using the sklearn (version = 0.21.3, https://scikit-learn.org/stable/) *MultiTaskElasticNetCV* module, which is a linear regression model with L1 and L2 regularizations whose strengths are automatically determined using built-in cross-validation. The model was trained to use the changes in levels of each pair of protein nodes (input) to predict levels for the rest of the proteins and phenotypes (output). To compare the four different predictive models, CellBox, BP, NN, and Co-exp, the paired t-test on two related samples was used to analyze the difference between model performance (test set correlation) and to assign p-values.

Sensitivity Analysis

We conducted a sensitivity analysis of our model to evaluate the robustness of its prediction in response to noise. To model experimental noise, we first applied varying levels of multiplicative Gaussian noise to the input molecular and phenotypic data (Equation 4). The assumption is that the uncertainty and noise in experimental measurements arise around the true values with a scaling factor depending on the experimental approach.

$$x_i(t) = x_i^*(t) * N(1, \sigma_{mul}) \quad \text{(Equation 4)}$$

The scaling factor for each node and each condition was independently drawn from a Gaussian distribution $N(1,\sigma_{mul})$, with a mean of 1 and a standard deviation of σ_{mul}. $x_i^*(t)$ represents the experimental measurements, and $x_i(t)$ represents the values with noise added. For each noise level, we evaluated 15 different training/validation/test partitioning, and each with 5 independent random noise patterns. Model training was performed on noisy training and validation sets, while the model performance evaluation was performed on the original, noise-free test data. For each noise level, the percentage of successful models, defined as those that converged in terms of both MSE and oscillation filters, was recorded.

To further evaluate model performance from a technical perspective, we applied additive Gaussian noise to the input data (Equation 5) and evaluated model performance as above.

$$x_i(t) = x_i^*(t) + N(0, \sigma_{add}) \quad \text{(Equation 5)}$$

We examined model sensitivity to training and validation set size. We reduced the combined size of the training and validation set, from 90% to 10% in steps of 10%, while keeping their relative size constant, 4:1. The remaining data were allocated to the test set. For each training set size, the percentage of successful models, defined as those that converged in terms of both MSE and oscillation filters, was reported.

Network Interpretation

The entire dataset was used to generate 1,000 full successful models, each with an independent data partitioning of training (n = 71 conditions) and validation (n = 18 conditions).

In order to compare the model inferred interactions to those present in prior-knowledge pathway databases, all the proteins and phosphoproteins nodes were identified by their corresponding gene names (Table S1). The interactions were compared against the Pathway Commons (PC) database (current version at https://www.pathwaycommons.org/archives/PC2/v11) using the paxtoolsr (Luna et al., 2016) software. The database was filtered down to direct interactions, which include ''controls expression of'', ''controls phosphorylation of'', ''controls state change of'', ''controls production of'', ''controls transport of'', and ''controls transport of chemical''. All the direct interactions were converted to a directed PC graph using igraph (version = 1.2.4.1, https://igraph.org/r/). A distance was calculated as the length of the shortest path(s) between two components in the PC graph (Table S2, column: distance). The number of the shortest paths (npath) between the two components was also reported. Additional detailed information on protein links in Homo sapiens was obtained from the STRING database, including combined score (string) and individual-channel scores (co-expression, experimental, database, text-mining) (https://stringdb-static.org/download/protein.links.detailed.v11.0/9606.protein.links.detailed.v11.0.txt.gz) (von Mering et al., 2005). The co-citation score was calculated as the number of papers mentioning both components in a customized paper collection aggregating all PubMed papers that refer to at least two and no more than five gene names (Maglott et al., 2011).

Model Predictions

We used the models trained with the full (non-partitioned) dataset to simulate responses of novel, experimentally unobserved, in silico perturbation conditions. These conditions included different doses of single perturbations (five different levels of perturbation strengths $u \in [0, 3]$on all individual (phospho)protein and drug activity nodes within the network and as well as all pairwise combinations ($n_{all} = 94 \times 5 + C_{94}^2 \times 5^2 = 109,745$ conditions). The cell responses were dynamically simulated with u as the input perturbation with the same number of steps as in training ($N_t = 400$). For each perturbation condition, predictions for cell responses were averaged across 1,000 different models (Table S3). Perturbations were nominated as therapeutic candidates by ranking the predicted magnitude of the phenotypic change in terms of cell cycle arrest.

QUANTIFICATION AND STATISTICAL ANALYSIS

Model Performance Analysis

The numbers of models trained for different tasks are reported in the corresponding figure panels or Method Details. To compare the model performance of the four predictive models, CellBox, BP (Belief Propagation), NN (neural network model), and Co-exp (co-expression static model), on the single-to-combo and leave-one-drug cross-validation tasks, the paired t test on two related samples was used to compare the correlation between the predicted values of the models and the observed values on the test set, and to assign p values.

Network Analysis

For each inferred interaction (w_{ij}) between two nodes from the 1,000 full models, a t-score ($\frac{w_{ij}}{s\sqrt{m}}$) (which is effectively a one-sample Student's t test with the null hypothesis that the population mean is zero) was calculated as an indication of the confidence level of obtaining a value different from zero, where w_{ij} is the average interaction strength across models, s is the standard deviation, m is the number of models (m = 1,000) (Table S2).

To quantitatively evaluate whether the inferred networks recapitulate more known interactions in the database than the random networks, the same number of network models (m = 1,000) were generated with interaction parameters randomly drawn from the pool of all (phospho)protein- (phospho)protein interaction parameters in the CellBox-inferred network models. For each of the inferred or random network models, the number of interactions in Pathway Commons (PC) out of the top one-hundred interactions (ranked by absolute interaction strengths) was calculated, and the t test was performed to compare the difference in means between the two groups. To provide an additional comparison with a well-known alternative method for network inference, the partial information decomposition and context (PIDC) algorithm was implemented in the Python environment based on the detailed mathematical description of the algorithm in Chan et al. (Chan et al., 2017). The same number of PIDC networks (m = 1,000, same as for CellBox) were generated based on 1,000 independent random partitions of the entire dataset (80% of the conditions were used to infer PIDC networks). The t test was performed to compare the difference in means between the number of interactions inferred by CellBox network models and PIDC networks consistent with the PC database (Figure S9).

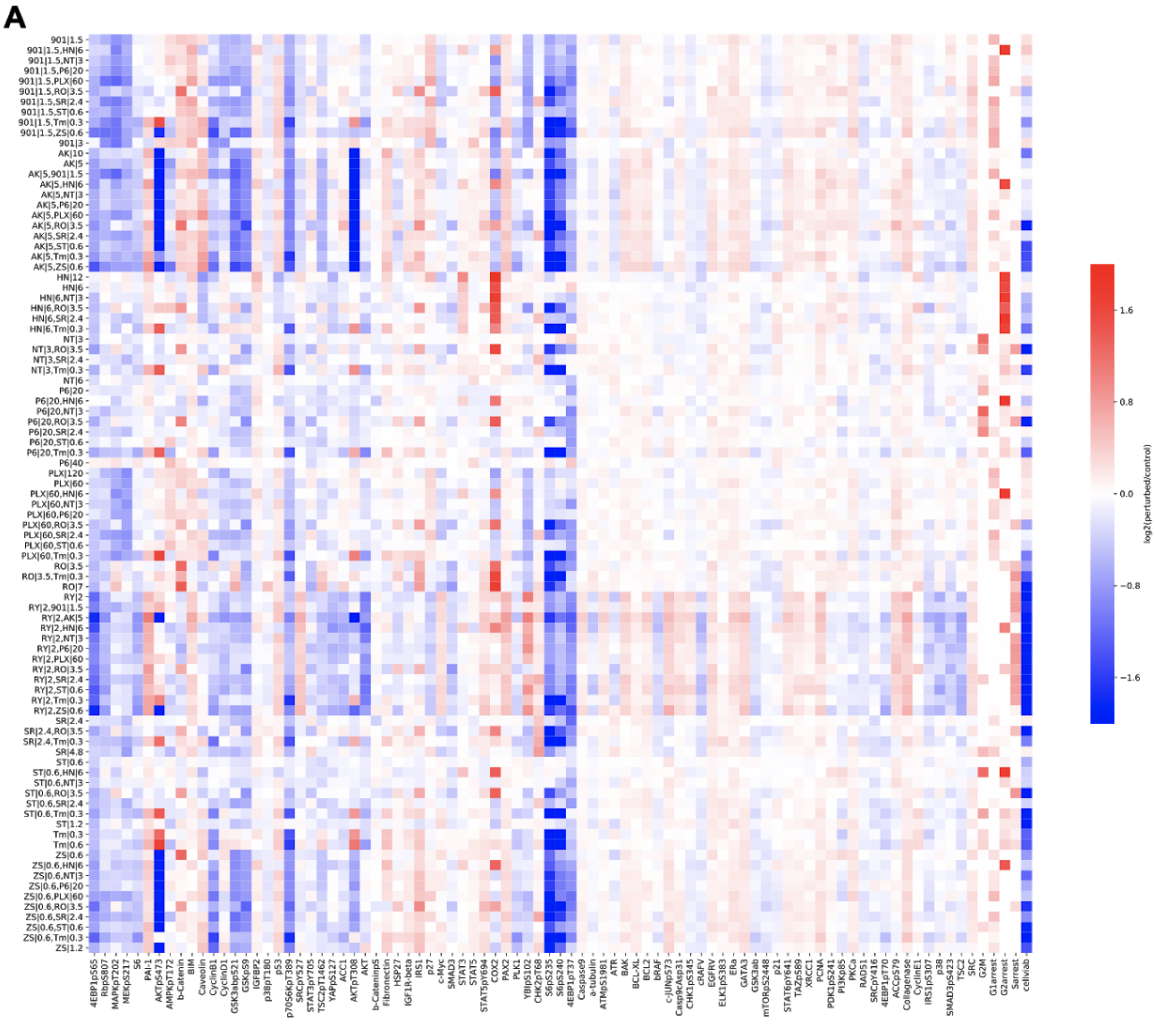
Cell Systems, Volume *12*

Supplemental Information

CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy

Bo Yuan, Ciyue Shen, Augustin Luna, Anil Korkut, Debora S. Marks, John Ingraham, and Chris Sander

A



B

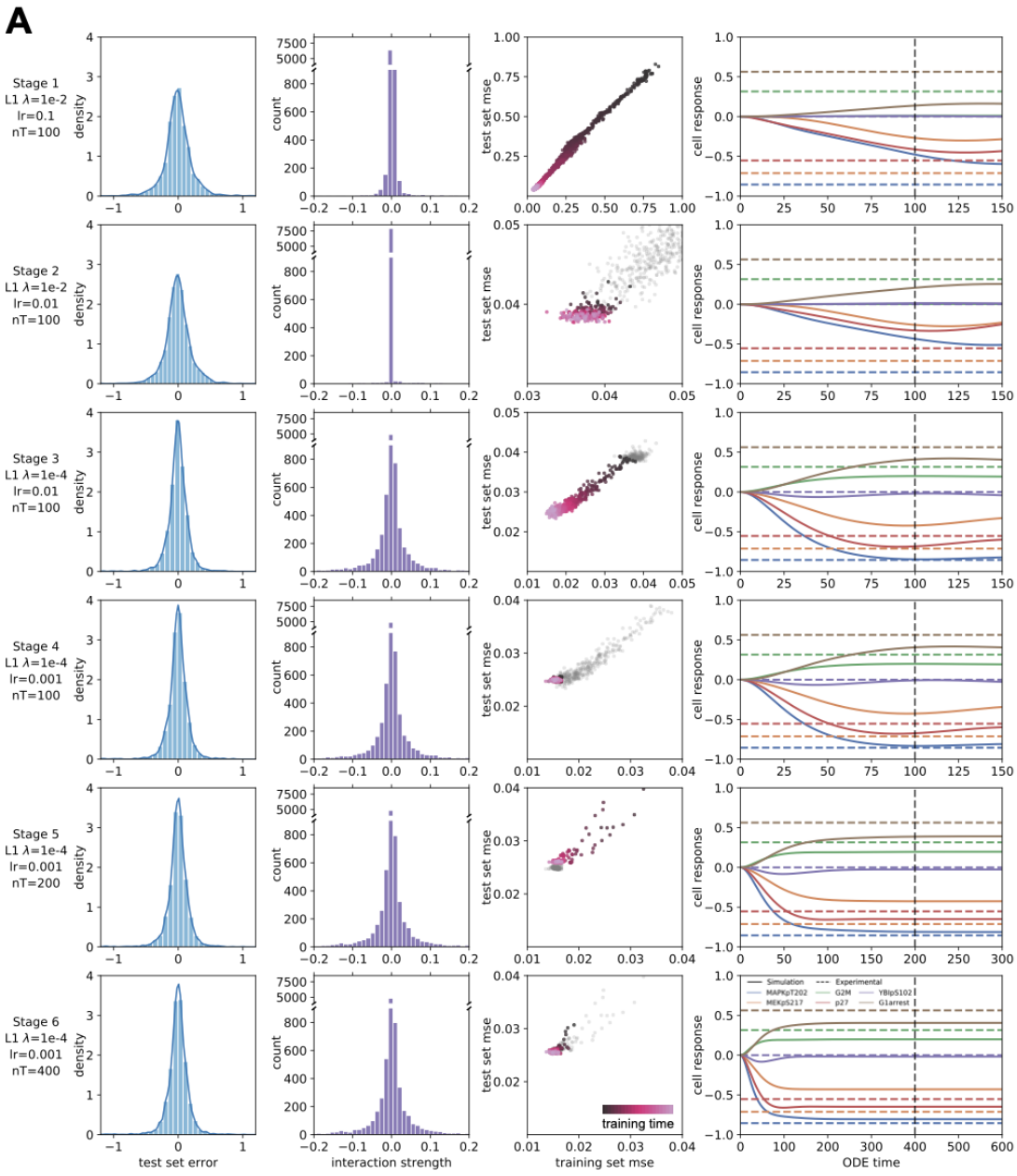
Drug used in perturbation experiments

Abbreviation	Drug name	Target	Downstream effector	Dose 1 (nM)	Dose 2 (nM)
AK	AKTi 1/2	AKT	AKTpS473	5.0	10.0
Tm	Temsirolimus	mTOR	S6pS240	0.3	0.6
ZS	ZSTK474	PI3K	AKTpS473	600	1,200
901	PD0325901	MEK1/2	MAPK1pT202	1.5	3.0
ST	Stattic	STAT3	STAT3pY705	600	1,200
HN	HNHA	HDAC	P27/Kip1	6,000	12,000
PLX	PLX4720	$BRAF^{V600E}$	MAPK1pT202	60	120
RO	RO-31-7549	PKC	S6pS240	3,500	7,000
RY	RY	CDK4	RbpS807	2,000	NA
P6	P6	JAK	GSK3pS21	20	40
SR	SR	SRC	GSK3pS21	2,400	4,800
NT	Nutlin	MDM2	4EBP1pS65	3,000	6,000

Figure S1. An overview of the perturbation dataset.

(A) The perturbation dataset (log_2 transformation of the ratio of measurements in perturbed conditions over unperturbed condition) was visualized in a heatmap with rows

of perturbation conditions (format: drug abbreviation with concentration, single or pair of drugs) and columns of (phospho)proteins or phenotypic measurements. (B) 12 drugs were used in the perturbation experiments, and the table contains for each drug their abbreviations, targets, downstream effectors used in the inference of drug activity nodes, and the two doses used in single-drug perturbation conditions. In combinatorial perturbation conditions, the lower dose of each drug was combined.


A
Stage 1
L1 λ=1e-2
lr=0.1
nT=100
Stage 2
L1 λ=1e-2
lr=0.01
nT=100
Stage 3
L1 λ=1e-4
lr=0.01
nT=100
Stage 4
L1 λ=1e-4
lr=0.001
nT=100
Stage 5
L1 λ=1e-4
lr=0.001
nT=200
Stage 6
L1 λ=1e-4
lr=0.001
nT=400
density
count
test set mse
cell response
training time
Simulation
Experimental
MAPKpT202
G2M
YBlpS102
MEKpS217
p27
G1arrest
test set error
interaction strength
training set mse
ODE time

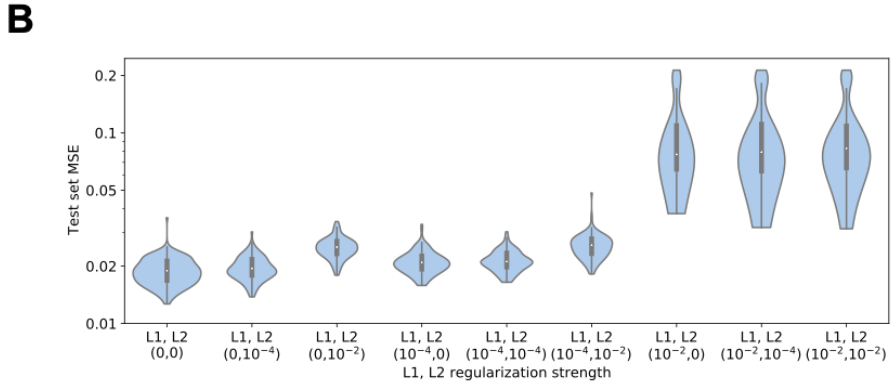

B
Test set MSE
L1, L2 (0,0)
L1, L2 (0,10^-4)
L1, L2 (0,10^-2)
L1, L2 (10^-4,0)
L1, L2 (10^-4,10^-4)
L1, L2 (10^-4,10^-2)
L1, L2 (10^-2,0)
L1, L2 (10^-2,10^-4)
L1, L2 (10^-2,10^-2)
L1, L2 regularization strength

Figure S2. Multi-step fine-tuning of hyperparameters facilitates model training.
During the training of each model, the training loss decreased along training time together with the test loss (Stage 1). The model parameters started to fluctuate around a local optimum after N = 2,000 training iterations. It has been shown that a decreasing learning rate is helpful for model convergence(Bengio, 2012). Decreasing the learning rate to 0.1x allowed the model to escape local minima and continue learning (Stage 2). The loss function stopped decreasing again after (up to) N = 4,000 iterations when the magnitude of the MSE loss was comparable to that of the regularization loss. To further improve training and decrease MSE loss mainly, we decreased the regularization strength by loosening the L1 constraints on the parameters (Stage 3). MSE decreased further while the numerical range of the parameters (interaction strengths) started to increase. Continuous decreasing of the learning rate did not further change the loss (Stage 4). ODE simulation of the model indicated that a steady state had not been fully reached. Therefore, the ODE simulation time was doubled twice (Stage 5 and Stage 6) while the learning rate and regularization were kept the same as Stage 4. The training and testing loss, together with the ODE trajectory, indicated that, at the final stage, the models had converged, optimization made no further improvements, and the ODE simulation has reached a steady state. We then stopped the training and examined the results closely on the test dataset. (A) Models were trained in six individual stages with varying learning rate, regularization strength, and ODE simulation time. As the training proceeds, 1. the distribution of differences between predicted and experimental values in the test set narrowed around zero; 2. the distribution of interaction strengths widened as the L1 regularization was weakened; 3. both the training and test loss decreased; 4. ODE simulation reached a steady state as simulation time increased. (B) Model performance was evaluated on different regularization approaches, including L1, L2, or combined (elastic net) with different regularization strengths (regularization strengths for L1, L2 term λ_1, λ_2 =0, 1e-2, 1e-4, pairwise combination). For a mild level of regularization ($\lambda_1 =$ 1e-4 and $\lambda_2 =$ 1e-4), we observed no significant difference between L1, L2, and elastic net. A stronger L1 regularization compromises the accuracy of model predictions.

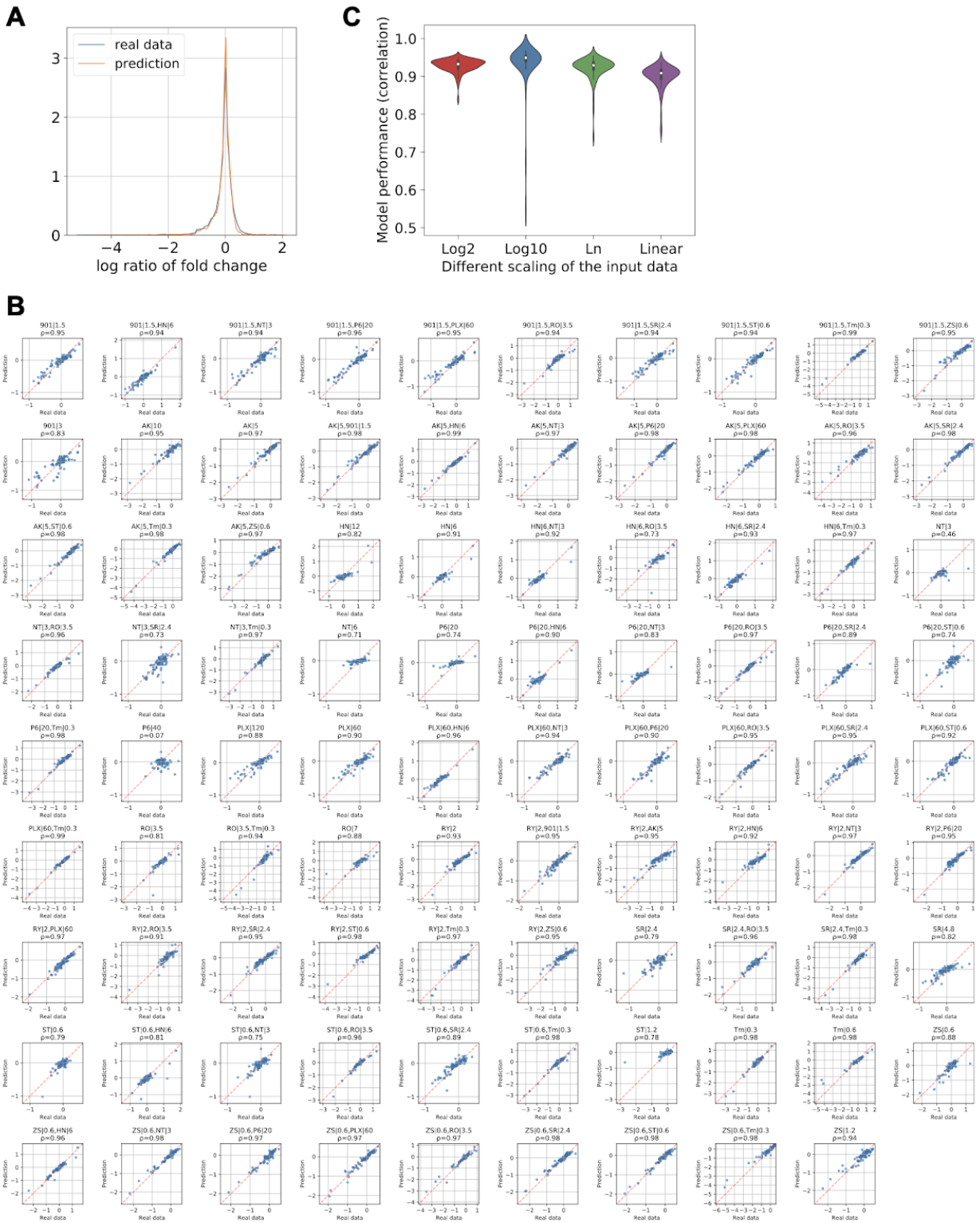

A
real data
prediction
log ratio of fold change
C
Model performance (correlation)
Log2
Log10
Ln
Linear
Different scaling of the input data
B
Prediction
Real data

Figure S3. Correlations between predicted and experimental data were consistently high across different perturbation conditions.

(A) Model prediction and experimental data had a similar range and distribution without skewing and extreme predictions. (B) In addition to overall performance, the model predictions for each perturbation condition were examined. The prediction of cell response under each condition reached similarly high correlation (median Pearson's correlation 0.95) with experimental data. Meshes indicate the range of real data. Models generally performed better for conditions with larger data range. (C) After median normalization of the RPPA data, the $\log_2$ ratio of measurements in perturbed conditions over unperturbed conditions was calculated and used as the input for the model. To evaluate whether the model performance depends on our choice of logarithmic scaling, we tested the model robustness against different scaling, including $\log_{10}$, natural log (ln), and non-log (linear) scaling. The model performance remains similar for different logarithmic bases (Pearson's correlation coefficient for test set $\rho>0.93$). Without logarithmic transformation, the model training still converges and results in a slightly reduced predictivity ($\rho=0.90$). The performance of CellBox is robust against different scaling methods.

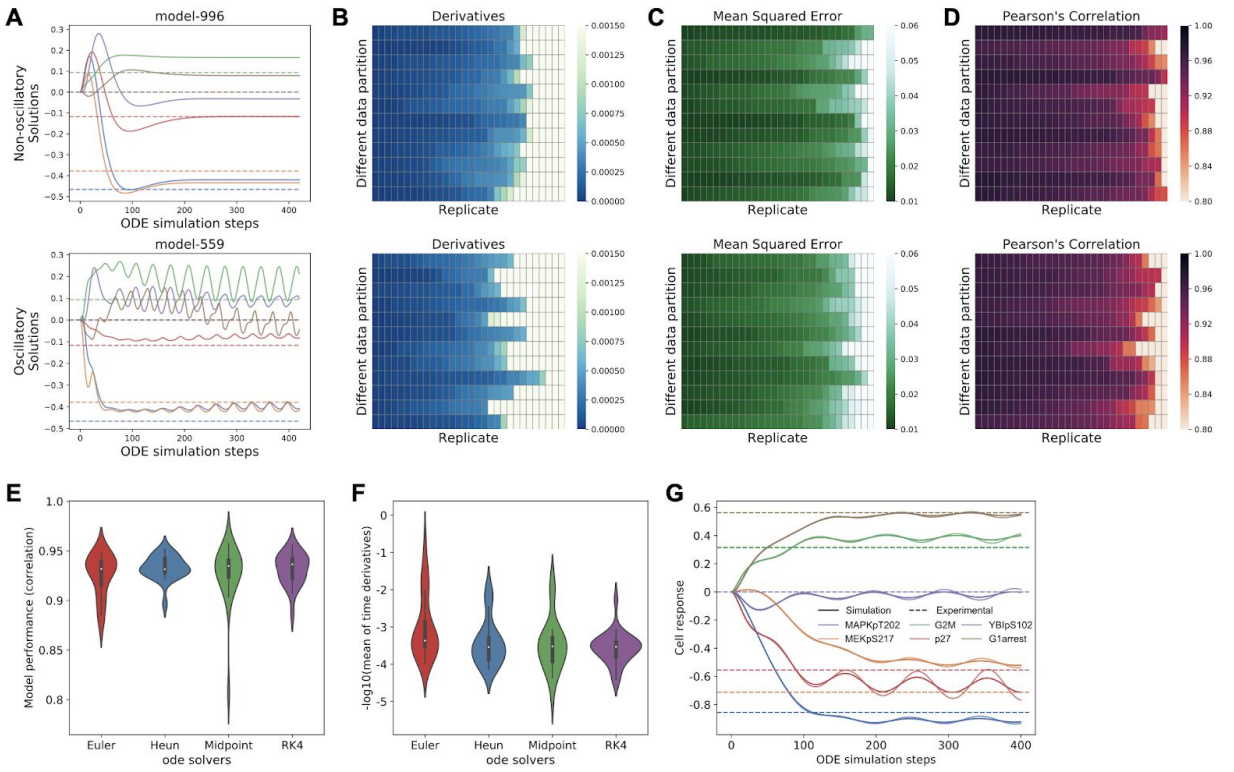


Figure S4. Oscillatory models from stochastic training are independent of data partitioning.

(A) Models were examined and categorized into non-oscillatory and oscillatory solutions based on the ODE simulation trajectories. (B-D) For each data partitioning of training and test set (rows) in the two categories, different seeds for random processes in model training (columns). The models were examined for their performance in terms of average derivatives of each variable at the end of the ODE simulation (B), average mean squared error in the training set (C), and Pearson's correlation between prediction and experimental data (D). Therefore, the solution oscillation and the model convergence are independent of the data partition. (E) The model performance in terms of Pearson's correlation between prediction and experimental data remains similar when different ODE solvers, including Euler, Heun, Midpoint and Runge-Kutta (RK4), were used in CellBox models. (F) Oscillatory solutions (mean of time derivatives threshold $\delta = 1e-03$) exist when all different ODE solvers are used. (G) For a given

solution, all solvers are consistent in terms of whether the solution is oscillatory or not but might differ slightly in terms of the oscillation amplitude.

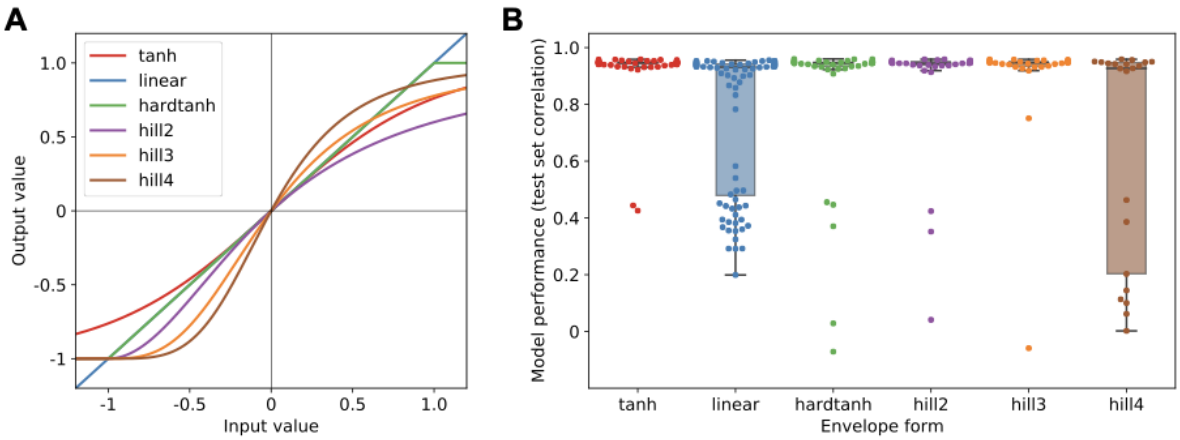


Figure S5. Model performance with different envelope forms

(A) We tested different alternatives of envelope forms, including a hyperbolic *tanh* (used as current default in CellBox), a clipped linear function (hard *tanh*), symmetric sigmoid function (hill of various degrees), and no envelope function (linear). All the other envelope forms model the saturation effect except for the linear function. (B) We observed comparable prediction accuracy between *tanh*, clipped linear, and low order of sigmoid functions, while the training with a linear envelope and high order of sigmoid functions was much less stable, and prediction performance was less accurate on average .

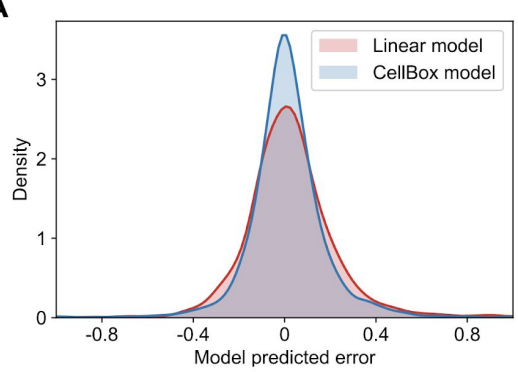
		Belief Propagation	Co-expression	Neural Network	CellBox
Performance	Training time* (CPU only)	30-40 mins	**5 mins**	**5 mins**	10-20 mins
	Training time* (GPU)	N/A	N/A	**<1 min**	4-8 mins
	Correlation (Leave-one-drug-out)	0.80	0.92	0.92	**0.94**
	Correlation (Single-to-combo)	0.69	0.84	0.88	**0.92**
Interpretability	Framework	**Biological mechanism**	Correlation	Correlation	**Biological mechanism**
	Dynamics	**Traceable**	-	-	**Traceable**
Prediction	Cooperative effects (non-linear)	**Yes**	-	**Yes**	**Yes**
	Experimentally untested targets	**Yes**	-	-	**Yes**

* Model training is timed for network size m = 99, using 4 vCPUs and 2GB RAM, with an optional Nvidia P100 GPU.

Figure S6. Comparison of various features between CellBox and other models.

In addition to the comparison of model performance in several training schemes between different models (Figure 3), we provided a more detailed comparison between models in the perspectives of training time and model capabilities and labeled the best features out of all models (bold). Generally, we believe the faster the training is with a similar number of inferred parameters, and the more biological mechanisms the model can provide, the more generalizable and useful the model is.

A



B

Examples of non-additive effects in drug combinations

Combination	Effector	Drug A effect	Drug B effect	Effect if additive	Measured effect	CellBox prediction	Effect
AKTi+PI3Ki	b-Catenin	0.25	1.28	1.51	0.42	0.51	antagonistic
PI3Ki+PKCi	b-Catenin	1.28	1.26	2.52	0.76	1.03	antagonistic
PI3Ki+SRCi	S6pS235	-1.18	-0.22	-1.15	-2.26	-2.25	synergistic
JAKi+PKCi	Cell viability	-0.07	0.10	-0.33	-1.88	-2.07	synergistic
SRCi+PKCi	Cell viability	-0.05	0.01	-0.24	-2.08	-2.36	synergistic

Figure S7. CellBox predicts the non-additive effects of drug combinations.

One of the most important applications of CellBox is to raise candidates of synergistic drug combinations. In order to test model prediction for synergy, we examined the model predictions in the single-to-combo training scheme where only the single-drug perturbation conditions were used to predict cell response to drug combinations. We compared those predictions with additive predictions (linear model) and observed CellBox models could predict both synergistic and antagonistic effects. (A) A systematic comparison of the prediction from CellBox and linear additive models. The CellBox models predict with higher accuracy (smaller prediction errors). The error is defined as the difference between model predictions and experimental responses for each data point. (B) Examples of non-additive effects were chosen from the top ranking list of difference between prediction error from linear model and that from CellBox models, which are effectively the proteins or phenotypes in drug combinations that CellBox predicts accurately while the linear model does not.

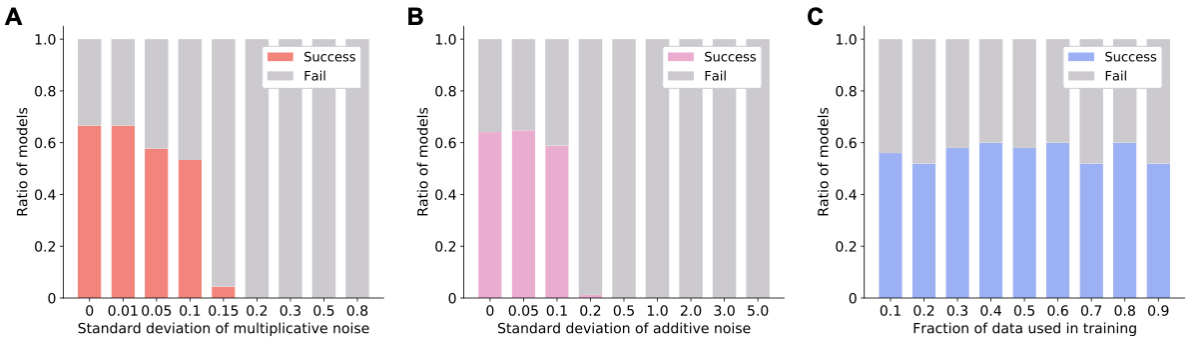


Figure S8. Model convergence against noise and reduced training set

(A-B) The percentage of models that successfully converged, defined by MSE of training set below a threshold of 0.05, decreases as an increased level of multiplicative Gaussian noise (A) or additive Gaussian noise (B) was added into training data. Such a decrease in model accuracy is plausible: as the signal is overwhelmed by the increasing noise level, model performance and stability should decrease in terms of MSE. (C) The percentage of successful models stayed the same as more data was used for model training.

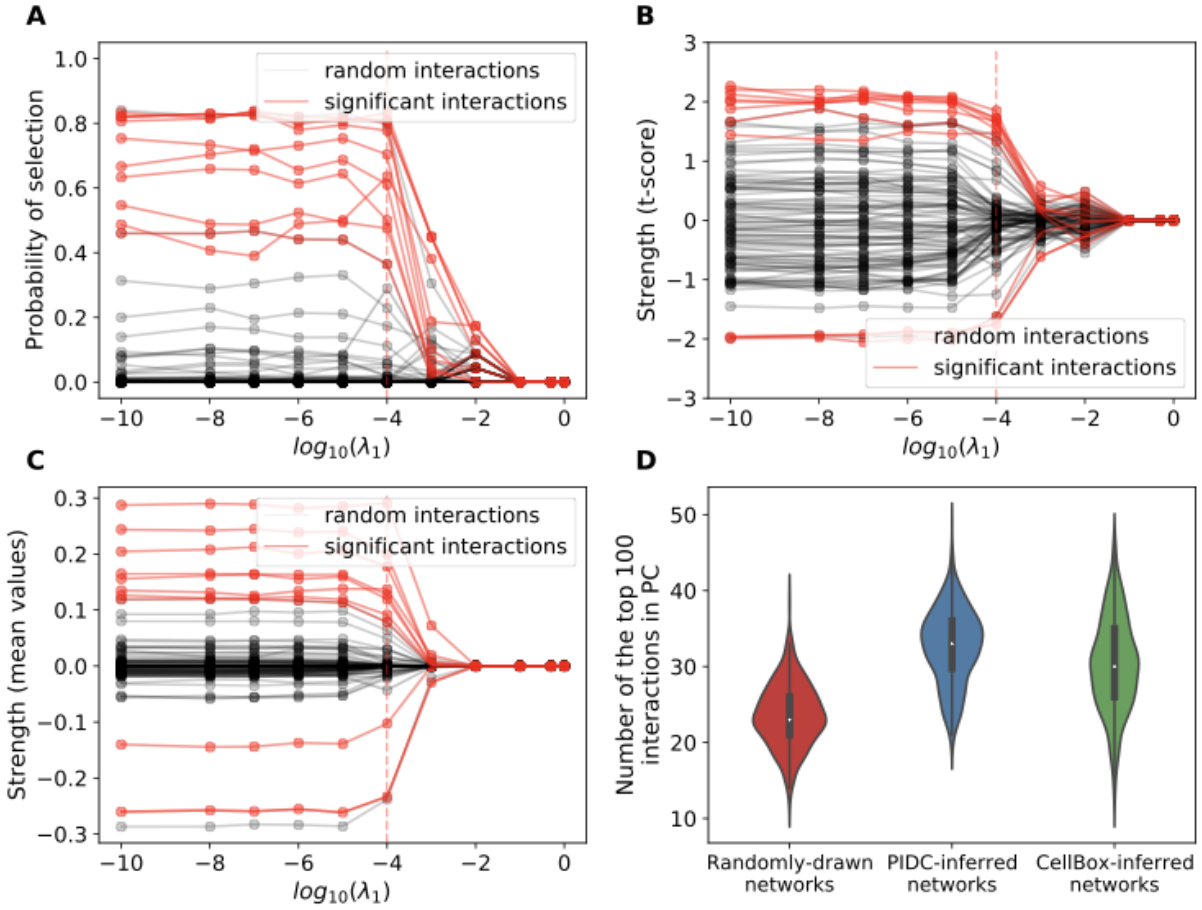


Figure S9. De novo inferred interactions are both robust and significantly consistent with literature derived networks.

(A-C) We used stability selection approaches(Meinshausen and Bühlmann, 2010) to examine network inference stability. We trained CellBox models with different L1 regularization strengths λ_1 and plotted the stability paths, including the probability of selection (A), t-score of the interactions over the models (B), and mean values of the interactions over the models (C), to compare the top 100 interactions (10 out of these are highlighted in red lines) identified by CellBox to 100 randomly chosen interactions from the pool of all possible (phospho)protein- (phospho)protein interactions (grey lines). The CellBox models result in significantly stable parameter inference, compared to random permutations (m>140 models for each λ_1). (D) To quantitatively evaluate our models in the context of prior knowledge, we compare CellBox models to random

models: the same number of network models (m = 1,000) were generated with interaction parameters randomly drawn from the pool of all (phospho)protein-(phospho)protein interaction parameters in the CellBox-inferred network models. For each of the inferred or random network models, the interactions existing in PC were identified out of the top one-hundred interactions (ranked by absolute interaction strengths). Our results find a significant difference in the number of interaction edges consistent with prior knowledge from Pathway Commons between the CellBox-inferred network and a network with random interactions (t-test, p=3e-149, $N_{\text{CellBox-inferred interactions in PC}} - N_{\text{random interactions in PC}} = 7$). We also compared our methods with other network inference methods(Chan, Stumpf and Babtie, 2017). Partial information decomposition and context (PIDC) recovers more prior knowledge interactions in PC compared to CellBox (by t-test, p-value=8.0e-21, $N_{\text{PIDC-inferred interactions in PC}} - N_{\text{CellBox-inferred interactions in PC}} = 2$). Nevertheless, we argue that the slight advantage of PIDC compared to CellBox in this measure is more than compensated for by the crucial ability of the dynamically executable CellBox model to predict cell response to unseen perturbations, which is the primary objective of our approach.