MAVEN: An R/Shiny app for compound mechanism of action analysis and visualisation using transcriptomics and compound structure data

Layla Hosseini-Gerami[1], Howard Broughton[2], David Andrew Collier[3], Andreas Bender[1*]

[1]Centre for Molecular Informatics, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, United Kingdom [2]Eli Lilly and Company Centre de Investigacion, Alcobendas, Spain [3]Eli Lilly and Company, Bracknell, United Kingdom

*To whom correspondence should be addressed (ab454@cam.ac.uk)

Abstract

Background

Understanding the Mechanism of Action (MoA) of a compound is an often challenging but equally crucial aspect of drug discovery that can help improve both its efficacy and safety. Computational methods to aid MoA elucidation usually either aim to predict direct drug targets, or attempt to understand modulated downstream pathways or signalling proteins. Such methods usually require extensive coding experience and results are often optimised for further computational processing, making them difficult for wet-lab scientists to perform, interpret and draw hypotheses from.

Results

To address this issue, we in this work present MAVEN (Mechanism of Action Visualisation and Enrichment), an R/Shiny app which allows for GUI-based prediction of drug targets based on chemical structure, combined with causal reasoning based on biological networks and gene expression data. The app computes a systems-level view of the mechanism of action of the input compound. This is visualised as a sub-network linking predicted or known targets to modulated transcription factors *via* inferred signalling proteins. The tool includes a selection of MsigDB gene set collections to perform pathway enrichment on the resulting network, and also allows for custom gene sets to be uploaded by the researcher. MAVEN is hence a user-friendly, flexible tool for researchers without extensive bioinformatics or cheminformatics knowledge to generate interpretable hypotheses of compound Mechanism of Action

Conclusions

MAVEN is available as a fully open-source tool at https://github.com/laylagerami/MAVEN with options to install in a Docker or Singularity container. Full documentation, including a tutorial on example data is available at https://laylagerami.github.io/MAVEN/.

Keywords

Mechanism of Action, Causal Reasoning, Transcriptomics, Shiny, Systems Biology

1 Background

The discovery of the mechanism of action (MoA) of a small molecule, which describes the biochemical interactions a molecule makes to produce a pharmacological effect, is an important aspect of drug discovery for a wide range of reasons, from repurposing to a new indication to anticipating potential side effects and rationalising phenotypic findings[1]. Advances in machine learning techniques, combined with large publicly availably bioactivity databases such as ChEMBL and PubChem, as well high-throughput biological assays such as LINCS L1000 and DRUG-Seq, have contributed to the development of computational methods for generating hypotheses of compound MoA[2]. Two popular approaches include target-based and network-based methods. Target-based methods aim to predict the direct biological target of the compound, and have shown high performance using chemical structure fingerprints as descriptors[3–5]. Network-based methods such as causal reasoning use transcriptomics data along with prior knowledge networks to infer upstream drivers of transcriptional changes, and have been shown to capture biological pathways modulated by drug compounds[6–9].

However, such approaches often require proficiency in programming languages such as R and Python as well as the command-line, and output computer-readable data which can be difficult to convey to non-specialists, which can hinder scientific communication in multi-disciplinary groups. Hence, here we introduce MAVEN, or Mechanism of Action Visualisation and ENrichment, an R/Shiny app which allows users to integrate compound structure-based target prediction with gene expression-based causal reasoning without prior coding experience, and allows for the visualisation and pathway enrichment of the results to obtain a systems-level, easily interpretable view of the mechanism of action of a compound.

2 Implementation

MAVEN was written in the R programming language (v 4.2) using the Shiny application framework, and the source code is available for local installation at https://github.com/laylagerami/MAVEN. To run direct target prediction (which is optional for software functionality) the app also invokes PIDGINv4[4] (https://github.com/BenderGroup/PIDGINv4) Python scripts using a Bash command script called from within R. For causal reasoning over prior knowledge biological networks with CARNIVAL[7] (https://github.com/saezlab/CARNIVAL), it is necessary to install either the free, open-source Cbc solver[10] or the free-for-academic IBM ILOG CPLEX[11]. Installation and configuration instructions for the solvers along with troubleshooting are described in the documentation https://laylagerami.github.io/MAVEN/ along with troubleshooting steps. We also provide an R script to install all packages with the required versions, and a conda .yml file with packages required for running the PIDGINv4 Python scripts. In case a container solution is preferred, build files for Docker and Singularity containers with all required software and environments (including solvers) are provided. The Omnipath[12] prior knowledge network is included along with the package as well as gene expression[13] and compound structure data for lapatinib which is used as an example in the documentation, and will be discussed here in the case study. For pathway enrichment on the predicted signalling network, a collection of MSigDB gene sets has been provided as well as an option to use custom user-uploaded gene set

files. Installation and deployment with the open-source Cbc solver have been tested on the HPC systems at Eli Lilly and Company and AWS in order to ensure compatibility with corporate computational environments.

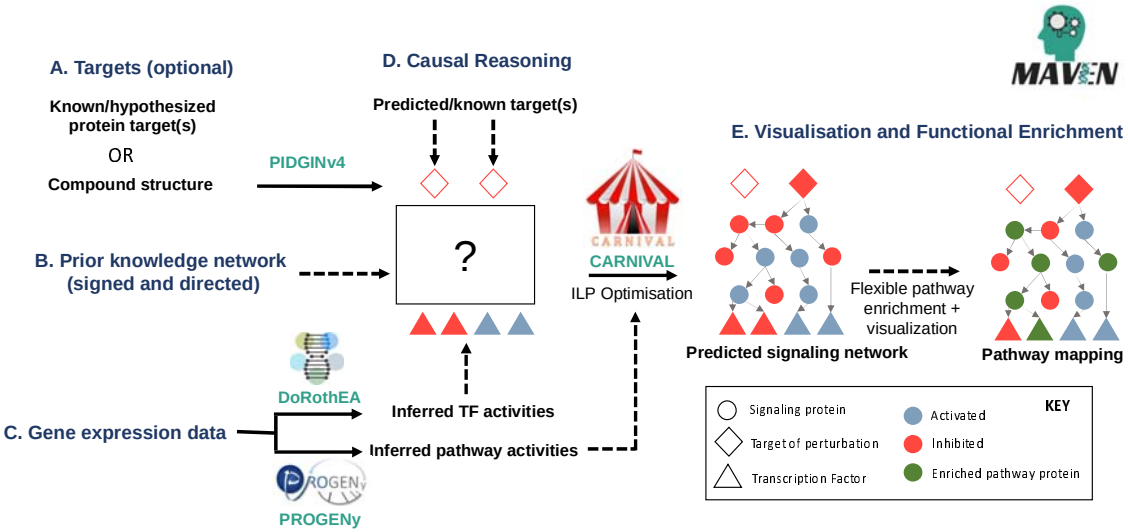

Figure 1: Workflow of analyses and features in MAVEN which requires 3 main inputs: Targets (A, optional) which can be known/hypothesised or predicted from the compound structure with PIDGINv4, a signed and directed prior knowledge network for causal reasoning (B) and gene expression data (C) which is used for the inference of transcription factor and pathway activities with DoRothEA and PROGENy. These are used as input to CARNIVAL for optimisation of the signalling network, connecting the targets to the TFs via intermediate signalling proteins (D). The resulting predicted signalling network is displayed in the GUI and interactive pathway enrichment can be carried out to contextualise the signalling proteins and interpret compound mechanism of action (E)

The overall workflow for MAVEN is depicted in Figure 1. Three inputs are taken; known or hypothesised targets which can be predicted from chemical structure with PIDGINv4[4] or defined a priori (optional) (Figure 1A); a signed and directed (i.e., A activates/inhibits B) prior knowledge network (Figure 1B) for causal reasoning; and compound-induced gene expression data in the form of a summary statistic such as t-values or log2-fold changes (Figure 1C). Signed and directed prior knowledge networks are required to infer causality and function (activation or inhibition), and can be obtained from open source databases e.g., Omnipath[12] (provided), SignaLink[14], SIGNOR[15] . Gene expression data can be from any platform, e.g., microarray, RNA-Seq, and publicly available gene expression data is available for many perturbations in databases such as GEO (https://www.ncbi.nlm.nih.gov/geo/) [16] (provided for the compound lapatinib) and LINCS L1000 (https://clue.io/releases/data-dashboard - Level 5) [2]. The gene expression data is then used to infer transcription factor (TF) activities with DoRothEA[17] and pathway activities with PROGENy[18], which is then used along with the prior knowledge network with CARNIVAL[7] to optimise a subnetwork which captures signalling proteins upstream of transcriptional changes and, if targets are predicted or provided, links them to the targets (Figure 1D). Finally, the subnetwork can be viewed and exported to use in other software such as Cytoscape[19], and we also provide a collection of MSigDB[20] gene sets (or allow for the upload of a custom gene set) for pathway enrichment with over-representation analysis (ORA), the results of which can also be visualised on the network (Figure 1E).

Throughout the workflow, all chosen parameters and command-line options are saved in log files for reproducibility purposes and so that analyses can be re-run programmatically (for information on how to run the tools used in MAVEN, please refer to https://pidginv4.readthedocs.io/en/latest/ and https://github.com/saezlab/transcriptutorial). Additionally, there are help buttons throughout the GUI with more information to aid the user in choosing algorithm parameters, and guidelines on the formatting of data.

The functionalities implemented in MAVEN will now be discussed in more detail.

Target prediction with PIDGINv4

The first data analysis step in MAVEN is target prediction based on compound chemical structure - though this is optional and targets can be manually entered or left out entirely (allowing for flexibility with the choice of use-case, e.g., with non-small-molecule induced gene expression data without a defined molecular target, such as disease data). Targets are used as input to CARNIVAL in a later step, to connect to inferred signalling proteins. Target prediction is implemented in MAVEN by invoking the PIDGINv4 software. PIDGINv4[4] is an open-source target prediction tool trained on ChEMBL[21] (v29) and PubChem[22] data using the scikit-learn[23] Python package, available on GitHub (v4.2). The tool consists of a collection of Random Forest models trained on the chemical structures (ECFP4 fingerprints calculated with the RDKit[24] Python package) of active and inactive compounds against 2,000+ human targets, and Python scripts to generate predictions for query compounds and to search for structurally similar compounds in the model training sets. For target prediction, the user is required to upload a .smi file, and a ChemDoodle[25] widget[26] is embedded in the app GUI to sketch the structure and generate a SMILES file in case the structural SMILES are not known. The user can select various parameters for the target prediction including activity threshold (0.1, 1, 10 or 100µM – default 10), number of cores, and applicability domain filter[27] (to remove low-confidence predictions). Once the user chooses to run the target prediction, a Bash script is invoked which runs the predict.py and sim2train.py PIDGINv4 scripts.

Transcription factor enrichment with DoRothEA and VIPER

To perform causal reasoning on a protein-protein interaction network, the gene expression data must be converted from the “gene-level” to the “protein-level” by inferring upstream transcription factors (TFs) driving the expression changes. DoRothEA (Bioconductor dorothea v1.6.0) consists of a regulon of TF-gene interactions[17]. Each interaction is given a confidence score reflecting the supporting evidence behind it from A (highest confidence, manually curated) to E (lowest confidence, computational predictions). The package is coupled with the VIPER[28] statistical method to infer TF activity from gene expression data, generating normalised enrichment scores for each TF[29]. In-app, the user can select the confidence levels A-E to filter the interactions in the regulon, and a slider for the number of top TFs they want to report and use for causal reasoning analysis. By default, 50

TFs are reported and plotted as a bar-chart in terms of their normalised enrichment score (NES, from -1 to 1). This number is generally a trade-off between coverage and noise which can be examined by adjusting the slider and viewing the NES plot, which updates automatically upon re-calculation. Furthermore, only confidence levels A-C are included by default, but this criterion can be relaxed if more enriched TFs are required. The documentation and help buttons also provide guidance on choosing these parameters.

Pathway activity inference with PROGENy

Pre-weighting proteins on the prior knowledge network has shown to improve the causal reasoning results by CARNIVAL[7]. PROGENy[18] (Bioconductor progeny v1.16.0) is a "footprint" method which infers pathway activities by leveraging a large compendium of publicly available perturbation experiments that yield a common core of Pathway RespOnsive GENes. Based on the pathway footprint genes, PROGENy produces pathway scores for 14 major signalling pathways including WNT, JAK-STAT and PI3K. The pathway scores are converted under-the-hood to protein weights to improve the CARNIVAL optimisation. The user can select the number of most dysregulated genes to include in the PROGENy pathway score calculations (by default 100, but this depends on the number of input genes – for experiments with a higher coverage such as RNA-Seq, this can be increased to e.g., 200 – 500).

Causal reasoning with CARNIVAL

The last analysis tool implemented in MAVEN is causal reasoning to infer dysregulated signalling networks. CARNIVAL[7] (GitHub CARNIVAL v.1.3) is a causal reasoning algorithm based on integer linear programming (ILP) which aims to optimise a subnetwork of signalling proteins contextualising a perturbation of interest. CARNIVAL takes as input dysregulated transcription factors (from DoRothEA) and a prior knowledge network (signed and directed protein-protein interactions), with pre-computed node weights (based on pathway activity scores from PROGENy) to aid in the network optimisation. CARNIVAL generates multiple solutions which are then aggregated to form a consensus network which connects TFs to targets (pre-defined or predicted with PIDGINv4) *via* inferred dysregulated signalling proteins, including their sign (activated or inhibited). If no target is defined then the signalling proteins get connected to a proxy "perturbation" node. The user can choose the runtime (in seconds) and number of cores. To solve the ILP problem a separate solver must be installed - Cbc[10] (v.2.9, free and open source) or IBM ILOG CPLEX[11] optimisation studio (v20.10, free for academic use or a license is required).

Visualisation and functional enrichment

Following causal reasoning with CARNIVAL the consensus network is visualised in the GUI using the visNetwork package (CRAN visNetwork v.2.1.0). To put the inferred signalling network into biological context it is possible to

perform functional enrichment. To this end, 11 MSigDB[20] gene sets collections are included with MAVEN (such as Hallmark[30], GO[31], Reactome[32], Wikipathways[33]). Alternatively, a .gmt file can be uploaded by the user for custom enrichment analysis. Over-representation analysis of the signalling network nodes in the gene sets using the prior knowledge network as background is performed with piano[34] (Bioconductor piano v2.10.1). Following enrichment and tabular display of the results, the user can select a pathway of interest and highlight the participating proteins on the network. The pathway results can also be downloaded. The network .sif file is also saved for further analysis and visualisation in Cytoscape[19] or other software packages.

3 Demonstration and Discussion

To demonstrate the app's utility for generating hypotheses for compound mechanism of action in practice, and to give an overview of the UI and app functionalities, we will now present a case study using the EGFR and ERBB2 inhibitor lapatinib. (We also provide this as a tutorial including in the documentation https://laylagerami.github.io/MAVEN/) The gene expression data comes from lapatinib-treated (1uM, 6h) HER2-positive BT474 breast cancer cells, from a publication by Sun et al[13] (GEO[16] accession GSE129254). The aim of the MAVEN analysis in this case study is to infer a signalling network which captures the cellular response of HER2+ positive cells treated with lapatinib.

The MAVEN UI is split into five tabs; Index (landing page), Data, Targets, Analysis and Visualisation (Figure 2). The landing page provides a summary of the MAVEN workflow and the case study will proceed from the second tab (Data).

Data

Here the gene expression data and prior knowledge are uploaded and stored in local memory for use in the Analysis tab. The user can browse for their files or use the toggle to load the Omnipath network and the lapatinib gene expression data used in this case study (Figure 2). As well as the documentation, there are help buttons throughout the workflow to explain file formats, definitions of parameters, and so on.

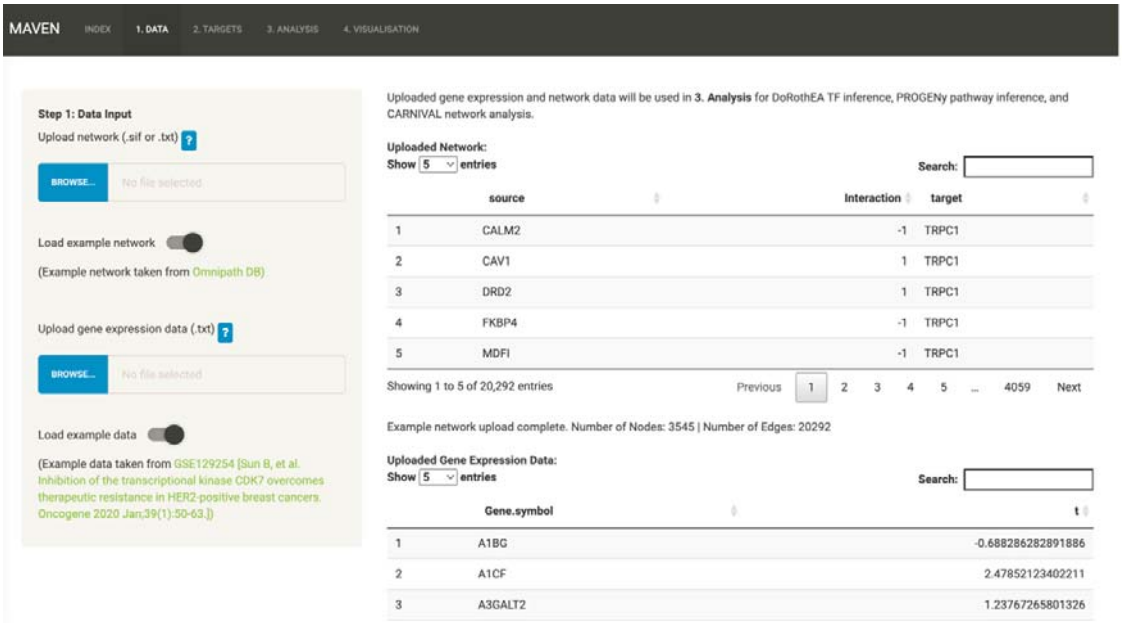

Figure 2: Screenshot of Data tab showing overall layout of MAVEN's GUI and the case study datasets loaded into the app via the "Load example" toggles

After checking that the data is in the correct format (including checking valid HGNC symbols and reporting any invalid symbols), the GUI provides a summary of the uploaded data for the user to check e.g., number of nodes and edges in the network. The user is then prompted to move onto the Targets tab.

Targets

The Targets page is split into four sub-tabs (Figure 3) and is an optional step in the MAVEN workflow. In the first tab (Figure 3A), the user either uploads a SMILES file or sketches their compound to produce a SMILES file. Following successful SMILES upload the compound is displayed as an image for the user to check, which can be seen for the case study with the correctly rendered lapatinib structure. In the second tab, the user is able to select the options for running PIDGIN (Figure 3B). Here, the bioactivity threshold was set to 1µM to correspond with the concentration of lapatinib used to generate the gene expression data. The applicability domain (AD) filter was set to 30, and 20 cores of compute power were used to run the predictions. After choosing the parameters the user is prompted to browse for the location of their PIDGINv4 installation directory, and then a button becomes available to click for running the target prediction analysis (which can be monitored *via* the R console output). Targets can also be defined manually by entering their HGNC symbols in the (D) User-defined targets tab.

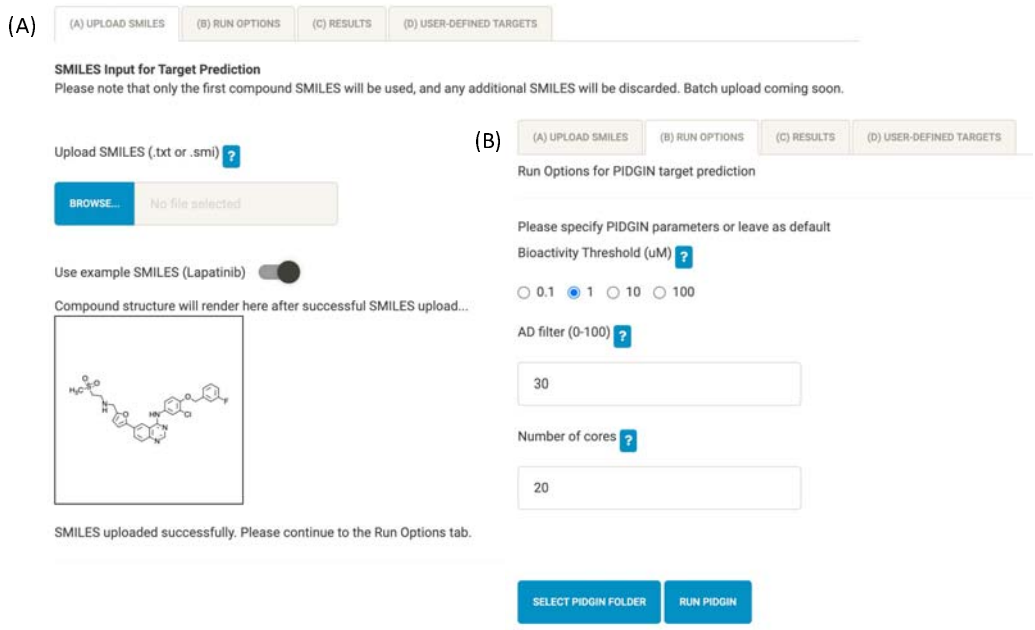

Figure 3: The Targets page contains 4 sub-tabs including SMILES upload (A) and PIDGIN run options (B). Results are displayed in (C) Results, and the user can manually define targets in (D) User-defined targets (not shown)

Once the PIDGIN run is complete, the results are saved and also displayed in the third tab, (C) Results, as a data table (Figure 4), with one row for each target model. The table contains the HGNC symbol, target name, predicted probability of activity, ChEMBL ID of the most structurally similar training compound (nearest neighbour), Tanimoto similarity of the nearest neighbour compared to the query compound computed from ECFP4 fingerprints, and the experimental measurements available for this compound. The target and nearest neighbour are hyperlinked to the UniProt and ChEMBL databases, respectively. It can be seen that many of the highest-predicted targets (ERBB4, EGFR, ERBB2, KCNH2 and PIK3C2B) are experimentally measured targets of lapatinib (CHEMBL554 represents lapatinib, and Tanimoto Coefficients of 1 identity in descriptor space, and in this case also structural space). To recapitulate the mechanism of action of lapatinib in HER2+ positive cells, we take the two known targets in this specific cell type (EGFR and ERBB2) [35] to the causal reasoning analysis stage, the rows are selected in the data table as shown in Figure 4. Targets can be chosen from the PIDGIN output based on the predicted probabilities as well as Tanimoto similarities (the higher the better in both cases), or by consulting the literature references to a wide variety of protein functions listed in their linked UniProt entries (e.g., https://www.uniprot.org/uniprot/P00533 for EGFR). Alternatively, the analysis can be run without targets, and then re-run with selected targets based on these findings to investigate specific target hypotheses. For example, if the final network outputs nodes from a particular signalling pathway, a highly-predicted target upstream of this pathway can be used to refine the final network.

The results table contains the target UniProt IDs, preferred name, and predicted probability of activity (0-1). The next columns show the nearest neighbour to the query compound in each target's training dataset, the Tanimoto similarity (0-1, where 1 is an identical structure) and the nearest neighbour's experimentally measured pChEMBL value (-log10[XC50]).

Show 10 entries Search:

	Gene ID	Name	Predicted Probability	Nearest Neighbour	NN Tanimoto Sim	NN pChEMBL
1153	ERBB4	Receptor protein-tyrosine kinase erbB-4	0.790	CHEMBL554	1.000	7.435
180	EGFR	Epidermal growth factor receptor erbB1	0.784	CHEMBL554	1.000	8.021
229	ERBB2	Receptor protein-tyrosine kinase erbB-2	0.724	CHEMBL554	1.000	7.793
1514	PI4KB	PI4-kinase beta subunit	0.700	CHEMBL554	1.000	6.030
498	ERBB3	Receptor tyrosine-protein kinase erbB-3	0.686	CHEMBL3633928	0.566	6.390
1056	KCNH2	HERG	0.654	CHEMBL554	1.000	6.000
27	PIK3C2B	Phosphatidylinositol-4-phosphate 3-kinase C2 domain-containing beta polypeptide	0.576	CHEMBL554	1.000	6.170
917	EPHB4	Ephrin type-B receptor 4	0.311	CHEMBL3900296	0.419	7.300
1393	MKNK1	MAP kinase-interacting serine/threonine-protein kinase MNK1	0.264	CHEMBL7917	0.318	6.280

Figure 4: Loaded data table of PIDGIN results which appears in tab (C) Results on the Targets page (see Figure 3). The results table includes a row for each target defined by their HGNC symbol (hyperlinked to the UniProt database entry) and preferred names, the predicted probability of activity of the query compound lapatinib, the most structurally similar compound (nearest neighbour - NN) in the model training set (hyperlinked to the ChEMBL database entry), the Tanimoto similarity of the NN compared to lapatinib (where 1 indicates the compounds are exactly the same), and the experimental pChEMBL value of the NN against the target. Here two of the top predicted targets, EGFR and ERBB2, are selected for further analysis to recapitulate lapatinib's MoA in HER2+ cells. In this case, the identical compounds has been retrieved several times from the data (lapatinib has ChEMBL ID CHEMBL554, with a nearest-neighbour Tanimoto Similarity of 1, i.e. identity in descriptor and in this case also structural space)

Analysis

The analysis page is split into three sub-tabs for the three bioinformatics analysis methodologies; DoRothEA, PROGENy and CARNIVAL. The settings for each can be set on the left-hand side of the page (Figure 5). For the case study, DoRothEA (Figure 5A) was run with confidence levels A, B and C and the top 50 enriched TFs have been used for further analysis. For PROGENy (Figure 5B) the top 100 most responsive lapatinib genes (based on the t-values input in the Data stage) were used for the calculations. For CARNIVAL (Figure 5C) it can be seen that the targets selected in the previous step (Figure 4) have populated the CARNIVAL options, and they can be further deselected if required - here, we kept both EGFR and ERBB2 as described above. We set a time limit of 3,600 seconds for the calculations, 20 cores of compute power and used the IBM ILOG CPLEX solver for solving

the ILP problem. This means that the solver will generate as many optimal network solutions as possible with the given time and compute resources, and output the final consensus network. Increasing the time limit or number of cores hence allows the solver to generate more networks, which may be required if no optimal solutions are found.

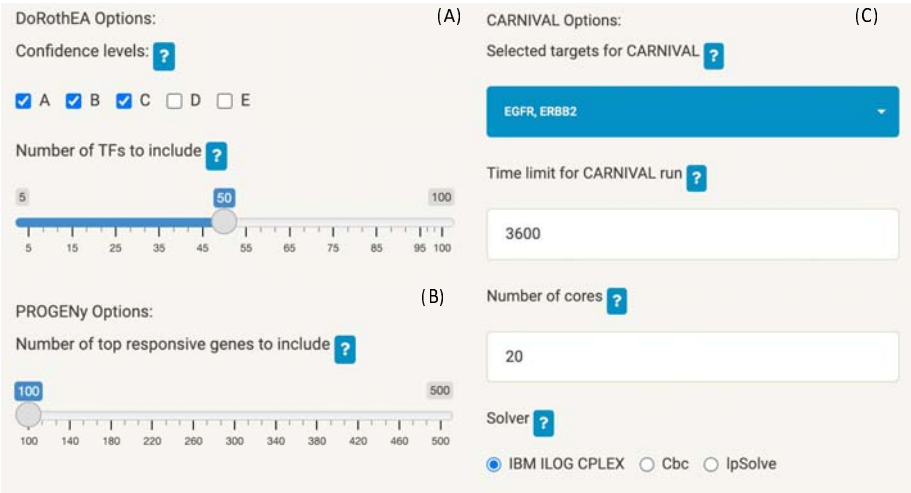

Figure 5: Settings for DoRothEA (A), PROGENy (B) and CARNIVAL (C) which can be interactively set in the MAVEN GUI

Following DoRothEA analysis, the resulting normalised enrichment scores (NES) for each TF are displayed as a bar chart (Figure 6) and corresponding data table with TFs hyperlinked to their corresponding UniProt page. It can be seen from the plot that the top enriched upregulated TF was FOXO3 which is known to be upregulated by lapatinib in HER2+ cells[36], and the top enriched downregulated TF was ESRRA which is known to be degraded by lapatinib in HER2+ cells[37]. Hence, MAVEN is able to generate an easy-to-interpret overview of TFs which are known to be dysregulated by lapatinib in the specific cellular context under investigation.

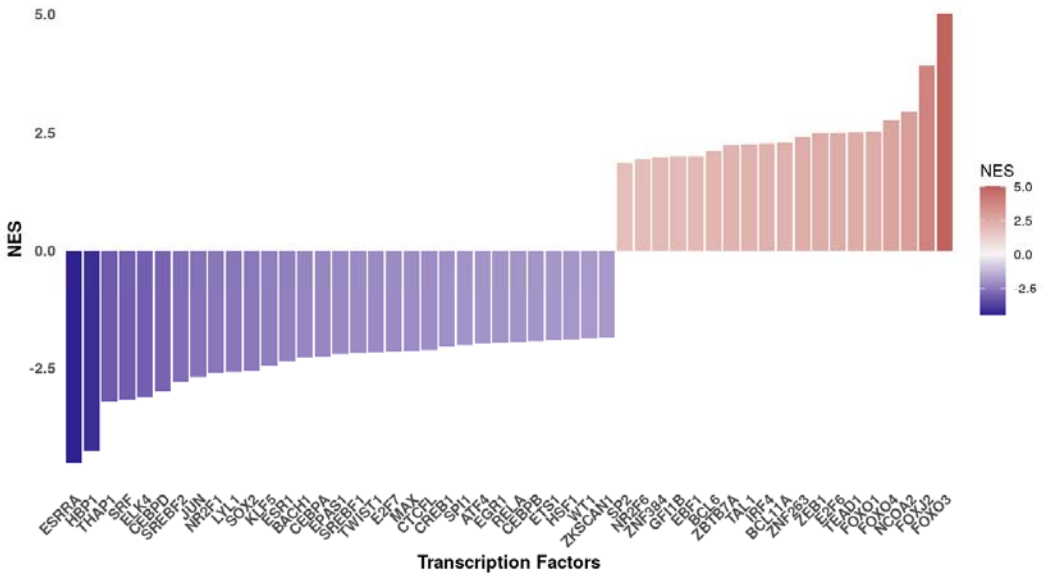

Figure 6: DoRothEA results derived from gene expression data from lapatinib-treated HER2+ BT474 cells expressed as a colour-coded bar plot. NES = Normalised enrichment score. Here, the most enriched upregulated gene (indicated by a positive NES) was FOXO3, and the most enriched downregulated gene (indicated by a negative NES) was ESRRA, which match with the known activity of lapatinib in HER2+ cells

If the slider is adjusted to select a different number of top scoring TFs, the plot and table of results automatically update. The number chosen here is a trade-off between coverage (where selecting a higher number may lead to additional findings) and also noise, where on the other hand a greater number of TFs may not necessarily contribute additional information and instead increase computational time. To aid in this decision, the plot and associated UniProt information for each TF can be consulted to select a number that provides good coverage of different protein functions (i.e., to not solely choose set of proteins in the same family, so that the CARNIVAL analysis can better exploit the prior knowledge network) coupled with prior knowledge/hypotheses on phenotypic findings. The interface help buttons (which can be seen in Figure 5) also provide a guide for selecting theses parameters, from the authors of DoRothEA.

Following PROGENy analysis, the results are visualised in the same way - a bar chart of predicted pathway activity score (from -1 to 1 indicating inhibition and activation) (Figure 7) and a corresponding numerical data table (not shown). In agreement with the results of the analysis, lapatinib is known to inhibit the EGFR[38], MAPK[39] and PI3K[40] pathways in HER2+ cells. The pathway scores are converted to weights on the protein-protein interaction network, which aids optimisation of the signalling subnetwork by CARNIVAL[7]. By default, the top 100 top responsive genes are chosen, but this can be adjusted depending on the coverage of the gene expression experiment – in general, the greater the number of genes measured, the greater the number of top responsive genes (e.g., 200-500 for RNA-Seq experiments). The bar chart will again update upon adjustment of the number of genes, and can be interpreted with regards to the function of each pathway and what would be expected based on what is known about the compound.

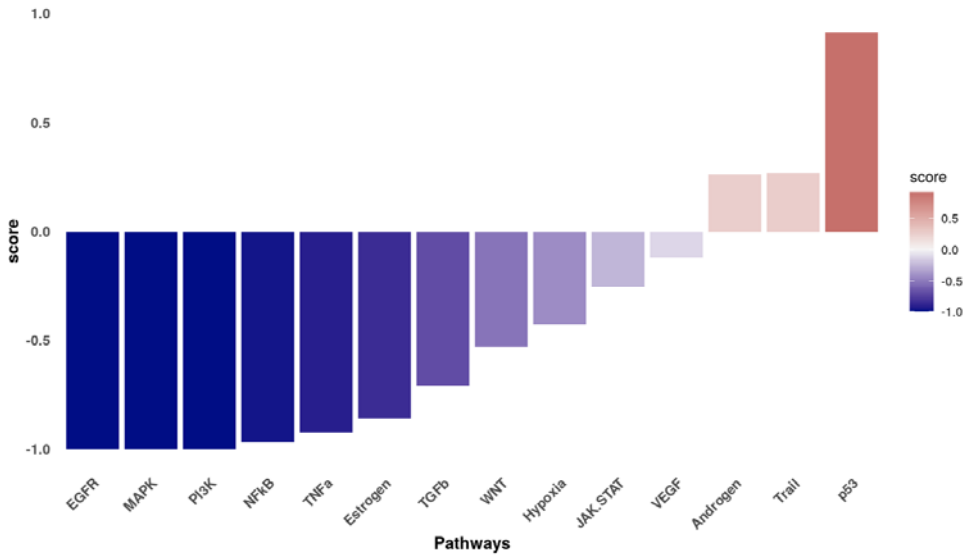

Figure 7: PROGEny results for lapatinib-treated gene expression data in terms of predicted pathway activity score (from -1 to 1 indicating inhibition and activation). It can be seen that known pathways inhibited by lapatinib in HER2+ cells EGFR, MAPK and PI3K, are predicted as inhibited, aiding the optimisation of the CARNIVAL network in the next stage of the analysis.

Visualisation and enrichment analysis

Following DoRothEA and PROGENy analysis the CARNIVAL analysis can be run, taking as input the DoRothEA enriched TFs and pathway weights from PROGENy, as well as the prior knowledge network uploaded in the first Data step. Once complete, the resulting CARNIVAL consensus network is visualised on the visualisation page (Figure 8). Files from previous analysis runs, which are automatically saved, can also be uploaded into the tool for visualisation.

It can be seen that the top layer of the network consists of the two selected targets (ERBB2 and EGFR), the bottom layer consists of the input TFs (e.g., FOXO1, FOXO3, ESR1), and they are connected by signalling proteins with inferred directionality (indicated with blue for up-regulation and red for down-regulation), which along with their interactions have been optimised from the input prior knowledge network. The intermediate node PIK3CA is hovered over, showing that in 100% of the solutions the node was inferred to be downregulated (Down activity: 100) and in none of the solutions it was inferred to be upregulated (Up activity: 0). Interestingly, is known that that PIK3CA mutation causes resistance to lapatinib treatment in HER2+ cancers[41], indicating that it could play a role in its mechanism of action[42]. As well as visualising the resulting network and deriving hypotheses from individual nodes, it is possible to perform pathway enrichment using the network proteins in an over-representation analysis. To illustrate this, we ran the enrichment analysis using the Biocarta gene set and excluded TFs from the enrichment analysis (to avoid the over-enrichment of transcription-related pathways), and the top enriched pathway, HER2 signalling pathway (adjusted p-value = 7.51e-19) is visualised on the network with participating proteins highlighted in green (Figure 8). Hence, CARNIVAL was able to construct a signalling network highly enriched for HER2 signalling, including the signalling proteins MAPKs, RAF1, ESR1, ERBB2,

EGFR, PIK3CA[43] and EP300[44], which are known to be relevant for the primary mechanism of action of lapatinib in HER2+ cancers[43].

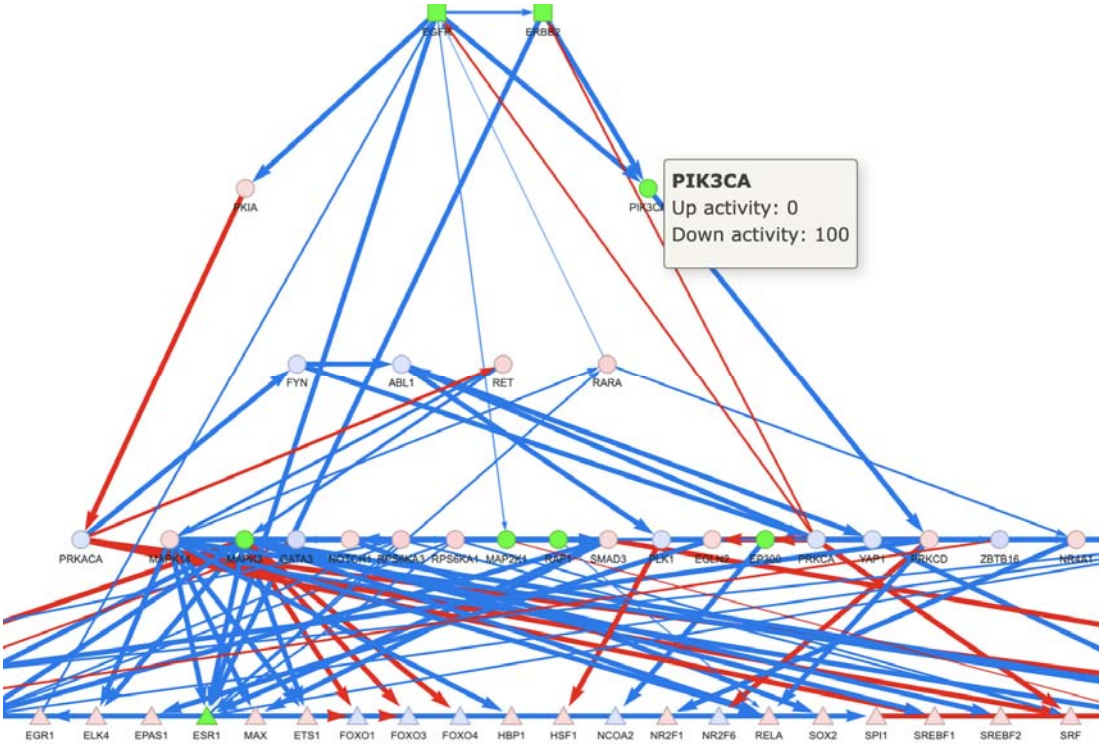

Figure 8: Following CARNIVAL analysis (or re-uploading previous results) the user can carry out pathway enrichment and visualisation on the inferred signalling network. Here, pathway enrichment was carried out on the network derived from lapatinib data with the built-in Biocarta MSigDB gene set, and the HER2 signalling pathway is highlighted on the network (as green nodes).

The enrichment results are also displayed in the GUI as a data table (Figure 9) and if one of the included MSigDB sets was used for the analysis, then they can be clicked through to the entry on the MSigDB website. A .csv file with more information on the enrichment results (e.g., participating proteins, odd's ratio, unadjusted p-value) can also be downloaded.

Network nodes in selected pathway:
MAPK3, MAP2K1, RAF1, ESR1, ERBB2, EGFR, EP300, PIK3CA

Show 10 entries Search:

	Pathway ID	Adjusted p-Value
62	BIOCARTA_EGF_PATHWAY	5.54e-13
164	BIOCARTA_HER2_PATHWAY	1.01e-12
81	BIOCARTA_GH_PATHWAY	6.05e-12
188	BIOCARTA_TCR_PATHWAY	6.05e-12
134	BIOCARTA_PDGF_PATHWAY	8.32e-12
27	BIOCARTA_BIOPEPTIDES_PATHWAY	1.16e-11
204	BIOCARTA_TPO_PATHWAY	2.62e-11
114	BIOCARTA_MAPK_PATHWAY	2.65e-10
50	BIOCARTA_CXCR4_PATHWAY	1.07e-9
165	BIOCARTA_ERK5_PATHWAY	3.12e-9

Showing 1 to 10 of 291 entries Previous 1 2 3 4 5 ... 30 Next

Figure 9: Pathway enrichment results data table following over-representation analysis on the lapatinib-derived signalling network. It can be seen that the HER2 pathway is the top enriched pathway; selecting it lights up participating proteins (MAPK3, MAP2K1, RAF1, ESR, ERBB2, EGFR, EP300, PK3CA) on the signalling network and reports the list of proteins to the user. The full results can also be downloaded as a .csv file.

Case Study Summary

Through the case study, we have demonstrated the ability of the MAVEN R/Shiny app and its constituent tools to produce and report correct target prediction results (predicting the lapatinib targets EGFR and ERBB2), infer both down- and up-regulated transcription factors induced by lapatinib (including FOXO3 and ESRRA), infer pathways known to be modulated by lapatinib (EGFR, MAPK and PI3K), and finally construct and visualise a signalling network which is highly enriched for the HER2 signalling pathway known to be modulated by lapatinib. This demonstrates the detailed insights into compound MoA that can be obtained using MAVEN's user-friendly interface, and without requiring extensive coding knowledge.

Future Development

Future additions to the app will include a batch upload option to analyse multiple compounds at once and compare their results, options to use other causal reasoning algorithms, and the ability to upload and analyse other data types (e.g., phosphoproteomics and metabolomics data). Supplementary files such as the MSigDB gene sets will be continuously updated to reflect any major version changes. Suggestions for new features can also be requested on the GitHub page https://github.com/laylagerami/MAVEN.

4. Conclusions

We have developed an R/Shiny app called MAVEN (Mechanism of Action Visualisation and ENrichment) a novel, feature-rich tool integrating chemical-structure based target prediction with gene expression-based causal reasoning analysis, coupled with visualisation and pathway enrichment analysis. A case study, using the chemical structure of lapatinib and gene expression data derived from lapatinib-treated HER2+ positive cells, has demonstrated the ease of inferring detailed insights into compound MoA using MAVEN.

Availability and requirements

Project name: MAVEN

Project home page: https://github.com/laylagerami/MAVEN

Operating system(s): Platform independent

Programming language: R, Python

Other requirements: R v.4.1 or higher, Python v.2 or higher

License: GNU General Public License

Any restrictions to use by non-academics: A license is needed to use the IBM solver which is only freely available for academics

List of abbreviations

MoA - Mechanism of Action

TF - Transcription Factor

NES - Normalised Enrichment Score

NN - Nearest Neighbour

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The dataset used in the case study is available from GEO (accession number GSE129254).

Competing interests

The authors declare that they have no competing interests.

Funding

LHG would like to thank BBSRC and Eli Lilly (grant code BB/M011194/1) for funding this project.

Authors' contributions

LHG developed the app and wrote the manuscript, HB performed beta testing at Eli Lilly and Company, HB, DAC and AB supervised the project and read and approved the final manuscript.

Acknowledgements

The authors would like to thank Rosa Hernansaiz-Ballesteros and Julio Saez-Rodriguez's research group for beta testing the app and giving feedback. We would also like to thank Jeff Kriske at Eli Lilly and Company for building the Singularity container solution.

References

(1) Trapotsi, M.-A.; Hosseini-Gerami, L.; Bender, A. Computational Analyses of Mechanism of Action (MoA): Data, Methods and Integration. *RSC Chem. Biol.* **2021**. https://doi.org/10.1039/D1CB00069A.

(2) Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K.; Lahr, D. L.; Hirschman, J. E.; Liu, Z.; Donahue, M.; Julian, B.; Khan, M.; Wadden, D.; Smith, I. C.; Lam, D.; Liberzon, A.; Toder, C.; Bagul, M.; Orzechowski, M.; Enache, O. M.; Piccioni, F.; Johnson, S. A.; Lyons, N. J.; Berger, A. H.; Shamji, A. F.; Brooks, A. N.; Vrcic, A.; Flynn, C.; Rosains, J.; Takeda, D. Y.; Hu, R.; Davison, D.; Lamb, J.; Ardlie, K.; Hogstrom, L.; Greenside, P.; Gray, N. S.; Clemons, P. A.; Silver, S.; Wu, X.; Zhao, W.-N.; Read-Button, W.; Wu, X.; Haggarty, S. J.; Ronco, L. V.; Boehm, J. S.; Schreiber, S. L.; Doench, J. G.; Bittker, J. A.; Root, D. E.; Wong, B.; Golub, T. R. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **2017**, *171* (6), 1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

(3) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9* (24), 5441–5451. https://doi.org/10.1039/C8SC00148K.

(4) Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A. Target Prediction Utilising Negative Bioactivity Data Covering Large Chemical Space. *J. Cheminformatics* **2015**, *7* (1), 51. https://doi.org/10.1186/s13321-015-0098-y.

(5) Carracedo-Reboredo, P.; Liñares-Blanco, J.; Rodríguez-Fernández, N.; Cedrón, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A Review on Machine Learning Approaches and Trends in Drug Discovery. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4538–4558. https://doi.org/10.1016/j.csbj.2021.08.011.

(6) Hosseini-Gerami, L.; Collier, D. A.; Laing, E.; Evans, D.; Broughton, H.; Bender, A. Benchmarking Causal Reasoning Algorithms for Gene Expression-Based Compound Mechanism of Action Analysis. January 27, 2022. https://doi.org/10.21203/rs.3.rs-1239049/v1.

(7) Liu, A.; Trairatphisan, P.; Gjerga, E.; Didangelos, A.; Barratt, J.; Saez-Rodriguez, J. From Expression Footprints to Causal Pathways: Contextualizing Large Signaling Networks with CARNIVAL. *Npj Syst. Biol. Appl.* **2019**, *5* (1), 1–10. https://doi.org/10.1038/s41540-019-0118-z.

(8) Enayetallah, A. E.; Ziemek, D.; Leininger, M. T.; Randhawa, R.; Yang, J.; Manion, T. B.; Mather, D. E.; Zavadoski, W. J.; Kuhn, M.; Treadway, J. L.; Etages, S. A. G. des; Gibbs, E. M.; Greene, N.; Steppan, C. M. Modeling the Mechanism of Action of a DGAT1 Inhibitor Using a Causal Reasoning Platform. *PLOS ONE* **2011**, *6* (11), e27009. https://doi.org/10.1371/journal.pone.0027009.

(9) Kumar, R.; Blakemore, S. J.; Ellis, C. E.; Petricoin, E. F.; Pratt, D.; Macoritto, M.; Matthews, A. L.; Loureiro, J. J.; Elliston, K. Causal Reasoning Identifies Mechanisms of Sensitivity for a Novel AKT Kinase Inhibitor, GSK690693. *BMC Genomics* **2010**, *11*, 419. https://doi.org/10.1186/1471-2164-11-419.

(10) Forrest, J.; Ralphs, T.; Vigerske, S.; LouHafer; Kristjansson, B.; jpfasano; EdwinStraver; Lubin, M.; Santos, H. G.; rlougee; Saltzman, M. *Coin-or/Cbc: Version 2.9.9*; Zenodo, 2018. https://doi.org/10.5281/zenodo.1317566.

(11) IBM. *ILOG CPLEX Optimization Studio*. https://www.ibm.com/products/ilog-cplex-optimization-studio (accessed 2019-06-17).

(12) Türei, D.; Korcsmáros, T.; Saez-Rodriguez, J. OmniPath: Guidelines and Gateway for Literature-Curated Signaling Pathway Resources. *Nat. Methods* **2016**, *13* (12), 966–967. https://doi.org/10.1038/nmeth.4077.

(13) Sun, B.; Mason, S.; Wilson, R. C.; Hazard, S. E.; Wang, Y.; Fang, R.; Wang, Q.; Yeh, E. S.; Yang, M.; Roberts, T. M.; Zhao, J. J.; Wang, Q. Inhibition of the Transcriptional Kinase CDK7 Overcomes Therapeutic Resistance in HER2-Positive Breast Cancers. *Oncogene* **2020**, *39* (1), 50–63. https://doi.org/10.1038/s41388-019-0953-9.

(14) Csabai, L.; Fazekas, D.; Kadlecsik, T.; Szalay-Bekő, M.; Bohár, B.; Madgwick, M.; Módos, D.; Ölbei, M.; Gul, L.; Sudhakar, P.; Kubisch, J.; Oyeyemi, O. J.; Liska, O.; Ari, E.; Hotzi, B.; Billes, V. A.; Molnár, E.; Földvári-Nagy, L.; Csályi, K.; Demeter, A.; Pápai, N.; Koltai, M.; Varga, M.; Lenti, K.; Farkas, I. J.; Türei, D.; Csermely, P.; Vellai, T.; Korcsmáros, T. SignaLink3: A Multi-Layered Resource to Uncover Tissue-Specific Signaling Networks. *Nucleic Acids Res.* **2022**, *50* (D1), D701–D709. https://doi.org/10.1093/nar/gkab909.

(15) Licata, L.; Lo Surdo, P.; Iannuccelli, M.; Palma, A.; Micarelli, E.; Perfetto, L.; Peluso, D.; Calderone, A.; Castagnoli, L.; Cesareni, G. SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 Update. *Nucleic Acids Res.* **2020**, *48* (D1), D504–D510. https://doi.org/10.1093/nar/gkz949.

(16) Edgar, R.; Domrachev, M.; Lash, A. E. Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. *Nucleic Acids Res.* **2002**, *30* (1), 207–210. https://doi.org/10.1093/nar/30.1.207.

(17) Garcia-Alonso, L.; Ibrahim, M. M.; Turei, D.; Saez-Rodriguez, J. Benchmark and Integration of Resources for the Estimation of Human Transcription Factor Activities. *bioRxiv* **2018**, 337915. https://doi.org/10.1101/337915.

(18) Schubert, M.; Klinger, B.; Klünemann, M.; Sieber, A.; Uhlitz, F.; Sauer, S.; Garnett, M. J.; Blüthgen, N.; Saez-Rodriguez, J. Perturbation-Response Genes Reveal Signaling Footprints in Cancer Gene Expression. *Nat. Commun.* **2018**, *9* (1). https://doi.org/10.1038/s41467-017-02391-6.

(19) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504. https://doi.org/10.1101/gr.1239303.

(20) Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdóttir, H.; Tamayo, P.; Mesirov, J. P. Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics* **2011**, *27* (12), 1739–1740. https://doi.org/10.1093/bioinformatics/btr260.

(21) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100–D1107. https://doi.org/10.1093/nar/gkr777.

(22) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (Database issue), D1202–D1213. https://doi.org/10.1093/nar/gkv951.

(23) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python. *Mach. Learn. PYTHON* 6.

(24) *RDKit: Open-Source Cheminformatics Software*. https://www.rdkit.org/ (accessed 2020-01-28).

(25) Burger, M. C. ChemDoodle Web Components: HTML5 Toolkit for Chemical Graphics, Interfaces, and Informatics. *J. Cheminformatics* **2015**, *7* (1), 35. https://doi.org/10.1186/s13321-015-0085-3.

(26) zachcp. *Chemdoodle Https://Github.Com/Zachcp/Chemdoodle*; 2022.

(27) Aniceto, N.; Freitas, A. A.; Bender, A.; Ghafourian, T. A Novel Applicability Domain Technique for Mapping Predictive Reliability across the Chemical Space of a QSAR: Reliability-Density Neighbourhood. *J. Cheminformatics* **2016**, *8* (1), 69. https://doi.org/10.1186/s13321-016-0182-y.

(28) Alvarez, M. J.; Shen, Y.; Giorgi, F. M.; Lachmann, A.; Ding, B. B.; Ye, B. H.; Califano, A. Functional Characterization of Somatic Mutations in Cancer Using Network-Based Inference of Protein Activity. *Nat. Genet.* **2016**, *48* (8), 838–847. https://doi.org/10.1038/ng.3593.

(29) Holland, C. H.; Tanevski, J.; Perales-Patón, J.; Gleixner, J.; Kumar, M. P.; Mereu, E.; Joughin, B. A.; Stegle, O.; Lauffenburger, D. A.; Heyn, H.; Szalai, B.; Saez-Rodriguez, J. Robustness and Applicability of Transcription Factor and Pathway Analysis Tools on Single-Cell RNA-Seq Data. *Genome Biol.* **2020**, *21* (1), 36. https://doi.org/10.1186/s13059-020-1949-z.

(30) Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J. P.; Tamayo, P. The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection. *Cell Syst.* **2015**, *1* (6), 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

(31) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000**, *25* (1), 25–29. https://doi.org/10.1038/75556.

(32) Fabregat, A.; Jupe, S.; Matthews, L.; Sidiropoulos, K.; Gillespie, M.; Garapati, P.; Haw, R.; Jassal, B.; Korninger, F.; May, B.; Milacic, M.; Roca, C. D.; Rothfels, K.; Sevilla, C.; Shamovsky, V.; Shorser, S.; Varusai, T.; Viteri, G.; Weiser, J.; Wu, G.; Stein, L.; Hermjakob, H.; D'Eustachio, P. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **2018**, *46* (D1), D649–D655. https://doi.org/10.1093/nar/gkx1132.

(33) Slenter, D. N.; Kutmon, M.; Hanspers, K.; Riutta, A.; Windsor, J.; Nunes, N.; Mélius, J.; Cirillo, E.; Coort, S. L.; Digles, D.; Ehrhart, F.; Giesbertz, P.; Kalafati, M.; Martens, M.; Miller, R.; Nishida, K.; Rieswijk, L.; Waagmeester, A.; Eijssen, L. M. T.; Evelo, C. T.; Pico, A. R.; Willighagen, E. L. WikiPathways: A Multifaceted Pathway Database Bridging Metabolomics to Other Omics Research. *Nucleic Acids Res.* **2018**, *46* (D1), D661–D667. https://doi.org/10.1093/nar/gkx1064.

(34) Väremo, L.; Nielsen, J.; Nookaew, I. Enriching the Gene Set Analysis of Genome-Wide Data by Incorporating Directionality of Gene Expression and Combining Statistical Hypotheses and Methods. *Nucleic Acids Res.* **2013**, *41* (8), 4378–4391. https://doi.org/10.1093/nar/gkt111.

(35) Medina, P. J.; Goodin, S. Lapatinib: A Dual Inhibitor of Human Epidermal Growth Factor Receptor Tyrosine Kinases. *Clin. Ther.* **2008**, *30* (8), 1426–1447. https://doi.org/10.1016/j.clinthera.2008.08.008.
(36) Matkar, S.; Sharma, P.; Gao, S.; Gurung, B.; Katona, B. W.; Liao, J.; Muhammad, A. B.; Kong, X.-C.; Wang, L.; Jin, G.; Dang, C. V.; Hua, X. An Epigenetic Pathway Regulates Sensitivity of Breast Cancer Cells to HER2 Inhibition via FOXO/c-Myc Axis. *Cancer Cell* **2015**, *28* (4), 472–485. https://doi.org/10.1016/j.ccell.2015.09.005.
(37) Deblois, G.; Smith, H. W.; Tam, I. S.; Gravel, S.-P.; Caron, M.; Savage, P.; Labbé, D. P.; Bégin, L. R.; Tremblay, M. L.; Park, M.; Bourque, G.; St-Pierre, J.; Muller, W. J.; Giguère, V. ERRα Mediates Metabolic Adaptations Driving Lapatinib Resistance in Breast Cancer. *Nat. Commun.* **2016**, *7* (1), 12156. https://doi.org/10.1038/ncomms12156.
(38) Imami, K.; Sugiyama, N.; Imamura, H.; Wakabayashi, M.; Tomita, M.; Taniguchi, M.; Ueno, T.; Toi, M.; Ishihama, Y. Temporal Profiling of Lapatinib-Suppressed Phosphorylation Signals in EGFR/HER2 Pathways. *Mol. Cell. Proteomics MCP* **2012**, *11* (12), 1741–1757. https://doi.org/10.1074/mcp.M112.019919.
(39) Estévez, L. G.; Suarez-Gauthier, A.; García, E.; Miró, C.; Calvo, I.; Fernández-Abad, M.; Herrero, M.; Marcos, M.; Márquez, C.; Lopez Ríos, F.; Perea, S.; Hidalgo, M. Molecular Effects of Lapatinib in Patients with HER2 Positive Ductal Carcinoma in Situ. *Breast Cancer Res.* **2014**, *16* (4), R76. https://doi.org/10.1186/bcr3695.
(40) Garrett, J. T.; Olivares, M. G.; Rinehart, C.; Granja-Ingram, N. D.; Sánchez, V.; Chakrabarty, A.; Dave, B.; Cook, R. S.; Pao, W.; McKinely, E.; Manning, H. C.; Chang, J.; Arteaga, C. L. Transcriptional and Posttranslational Up-Regulation of HER3 (ErbB3) Compensates for Inhibition of the HER2 Tyrosine Kinase. *Proc. Natl. Acad. Sci.* **2011**, *108* (12), 5021–5026. https://doi.org/10.1073/pnas.1016140108.
(41) Wang, L.; Zhang, Q.; Zhang, J.; Sun, S.; Guo, H.; Jia, Z.; Wang, B.; Shao, Z.; Wang, Z.; Hu, X. PI3K Pathway Activation Results in Low Efficacy of Both Trastuzumab and Lapatinib. *BMC Cancer* **2011**, *11* (1), 248. https://doi.org/10.1186/1471-2407-11-248.
(42) O'Brien, N.; Wilcox, K.; Mcdermott, M.; Crown, J.; O'Donovan, N.; Slamon, D. Abstract P4-01-06: The Role of the PI3K/AKT, RAS/MAPK and Novel Signaling Pathways in Resistance to HER2-Targeted Therapy. *Cancer Res.* **2011**, *70*, P4-01. https://doi.org/10.1158/0008-5472.SABCS10-P4-01-06.
(43) Vogel, C.; Chan, A.; Gril, B.; Kim, S.-B.; Kurebayashi, J.; Liu, L.; Lu, Y.-S.; Moon, H. Management of ErbB2-Positive Breast Cancer: Insights from Preclinical and Clinical Studies with Lapatinib. *Jpn. J. Clin. Oncol.* **2010**, *40* (11), 999–1013. https://doi.org/10.1093/jjco/hyq084.
(44) Mahmud, Z.; Gomes, A. R.; Lee, H. J.; Aimjongjun, S.; Jiramongkol, Y.; Yao, S.; Zona, S.; Alasiri, G.; Gong, G.; Yagüe, E.; Lam, E. W.-F. EP300 and SIRT1/6 Co-Regulate Lapatinib Sensitivity Via Modulating FOXO3-Acetylation and Activity in Breast Cancer. *Cancers* **2019**, *11* (8), E1067. https://doi.org/10.3390/cancers11081067.