SUPPLEMENTARY MATERIAL

S1. MIDAS format

Metadata in *DataRail* and SBWiki is based on MIDAS (Minimum Information for Data Analysis in Systems Biology). MIDAS is derived from other minimum-information models, particularly MIACA (Minimum Information about a Cellular Assay), which in turn includes terms from GO (Gene Ontology), with several additions. Most importantly, each MIDAS file has a unique identifier (UID) composed of the following fields: (i) a two-letter data/file-type code (e.g., PD for Primary Data, MD for multiplex data), (ii) a three-letter creator code (typically initials), (iii) an identification number of arbitrary length that is unique across the entire system, and (iv) a free-text suffix that serves as a mnemonic to improve human readability. For example, the primary data discussed in the text might be tagged `MD-LGA-11111-CytoInh17phFI-BLK` (where MD denotes the Multiplex Data file type and LGA (after author L. G. Alexopoulos). The UID provides a convenient and unique "pointer" for referencing data files, and its integration with SBWiki provides a convenient repository for the data and metadata. These UID are generated and catalogued in the SBWiki (Muhlich *et al*., unpublished data).

Each column in a MIDAS file contains a header that begins with a two-letter prefix that describes its function: ID for identifiers, TR for treatments, DA for data acquisition, and DV for data values. MIDAS files can also include the concept of cues, signals, and responses (Gaudet *et al*., 2005): *cues* are biological perturbations to a system (such as the addition of extracellular ligands), *signals* represent the activities of proteins or other biomolecules involved in transducing biological information (activation of an intracellular kinase, for example), and *responses* represent phenotypic changes such as proliferation, cell death or cytokine release. Accordingly, the column headers in a MIDAS files may contain a second (user-selected) level of identification (e.g. headers for columns describing various cytokine treatments might begin with "TR:Cytokine"). When present, these secondary identifiers allow *DataRail*'s importer to identify automatically the dimensions of a new `compendium`.

S2. Handling data from multiple sets of data

DataRail, thanks to its intrinsic flexibility, provides a technical means to handle data consolidation and bundling either at the level of a `project` (defining a `compendium` for each source of data), the level of the `compendium` (using different `arrays` for different sets of data) or even at the level of an `array`. For example, as a complementary experiment to the CSR Compendium presented here, we have measured cell death with the Cytotoxicity Detection Kit LDH using a plate reader. This data can easily be added to an array containing the micro-Elisa data by increasing the number of dimensions.

When the user tries to load a new file and a compendium is already loaded, *DataRail* will ask the user to either (i) replace the compendium with the new data, (ii) add the data as a new array to the current compendium, (iii) add the data as a different compendium to the same project, or (iv) append the data to another array in current compendium. Importantly, the append function works only when the array to be appended is identical in dimensionality and type to the target array. It is also possible to perform a transformation on one or more primary data arrays and subsequently to join them. In this case, the user specifies how multiple dimensions can be collapsed into a single dimension.

S3. DataRail Documentation

More detailed documentation of *DataRail* is available as a separate PDF file and as help files that are accessible from the MATLAB documentation browser. Prior to publication, *DataRail* and its documentation will be available at http://web.mit.edu/goldsipe/www/DataRail/. After publication, this URL will redirect to a more permanent repository (http://code.google.com/p/sbpipeline/).

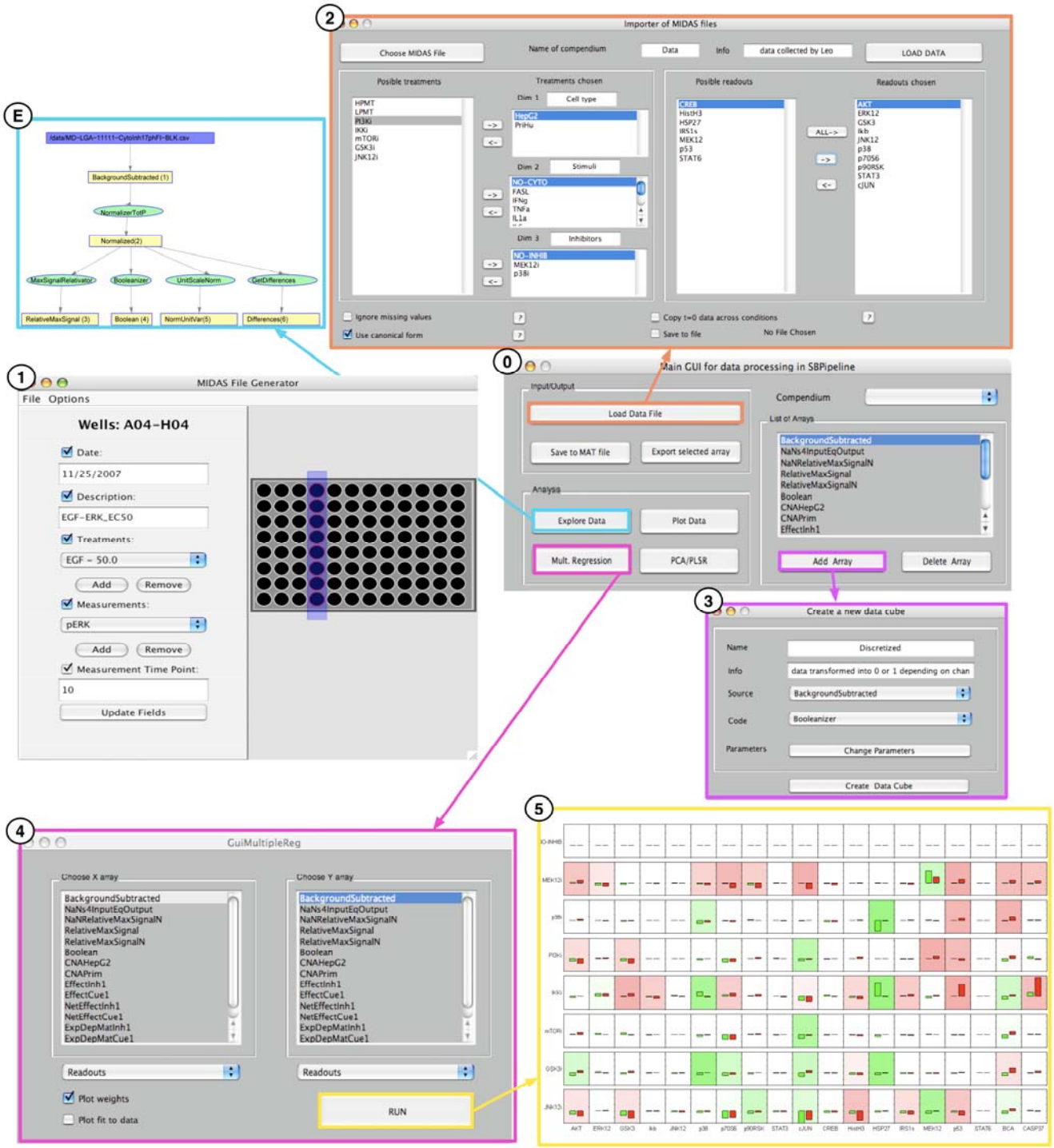

Fig. S1: Screenshots of *DataRail*. From the main GUI (0), different User Interfaces can be called, e.g, to (1) create MIDAS files, (2) create new arrays, (3) set up, (4) analyze, (5) develop a multiple regression model. The GUI for creating MIDAS files (1) is a stand-alone Java application. The numbers correspond to the steps in the workflow in Fig. 1. In addition, it is possible to explore the structure of the `compendium` (E). A function in *DataRail* creates a graph representing all data arrays connected by the functions used to derive them. This visual representation of the provenance is supplemented with the ability to explore the actual values, parameters and code.

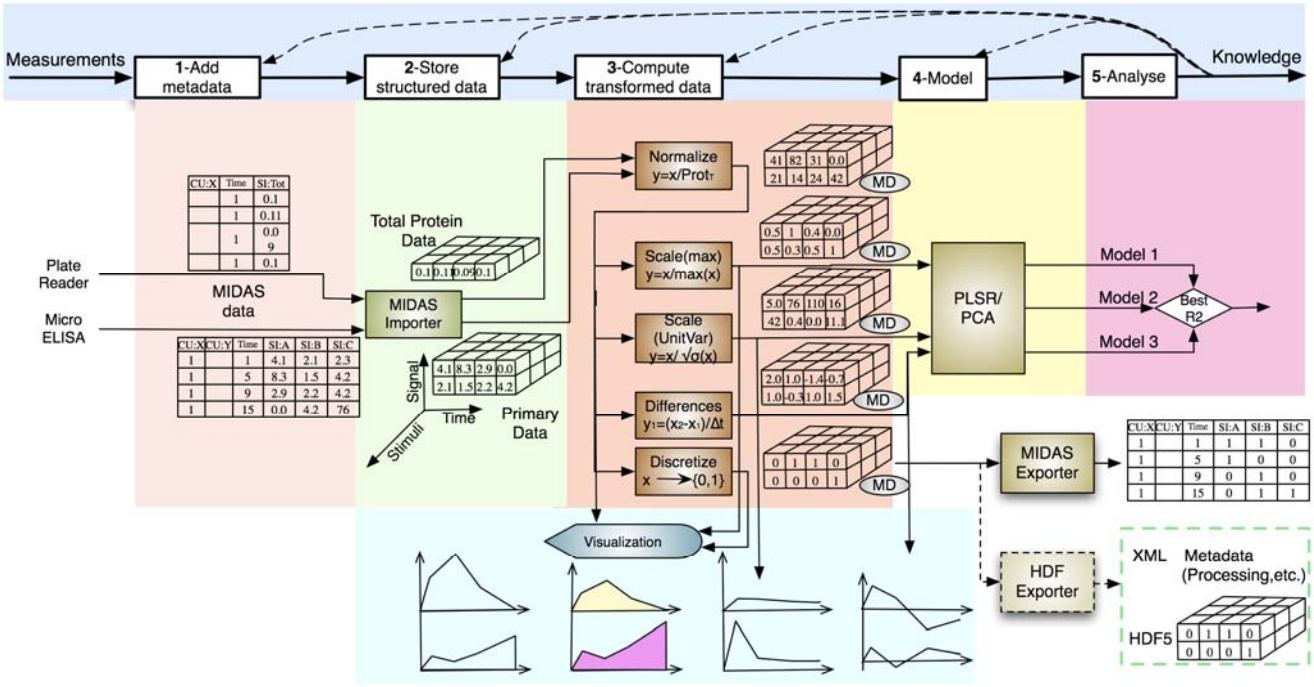

Fig. S2. Working scheme of *DataRail*. Data is stored in multidimensional arrays that can be easily transformed into new cubes using different normalization and metric algorithms. In the example in the figure, a set of data is imported and can be visualized on the fly. Then, a number of transformations are performed: signal relative to total protein and to the maximal signal for the same readout, discretized data, and data relative to the variance for the same time point across conditions (unit-variance scaling). Furthermore, a metric (the differences in the signals across time) is computed. Every data `array` carries all the information (metadata-MD tag in the Figure) concerning the transformations performed to create it. When an array is exported in MIDAS format, this metadata is lost. An export using HDF5 or XML is under preparation that will preserve all the metadata. Plots based on one or more arrays can be easily generated; here, the discretized data is used to characterize the dynamics, and this information is encoded in the coloring of the plots (see also Figure 4 in the main text).

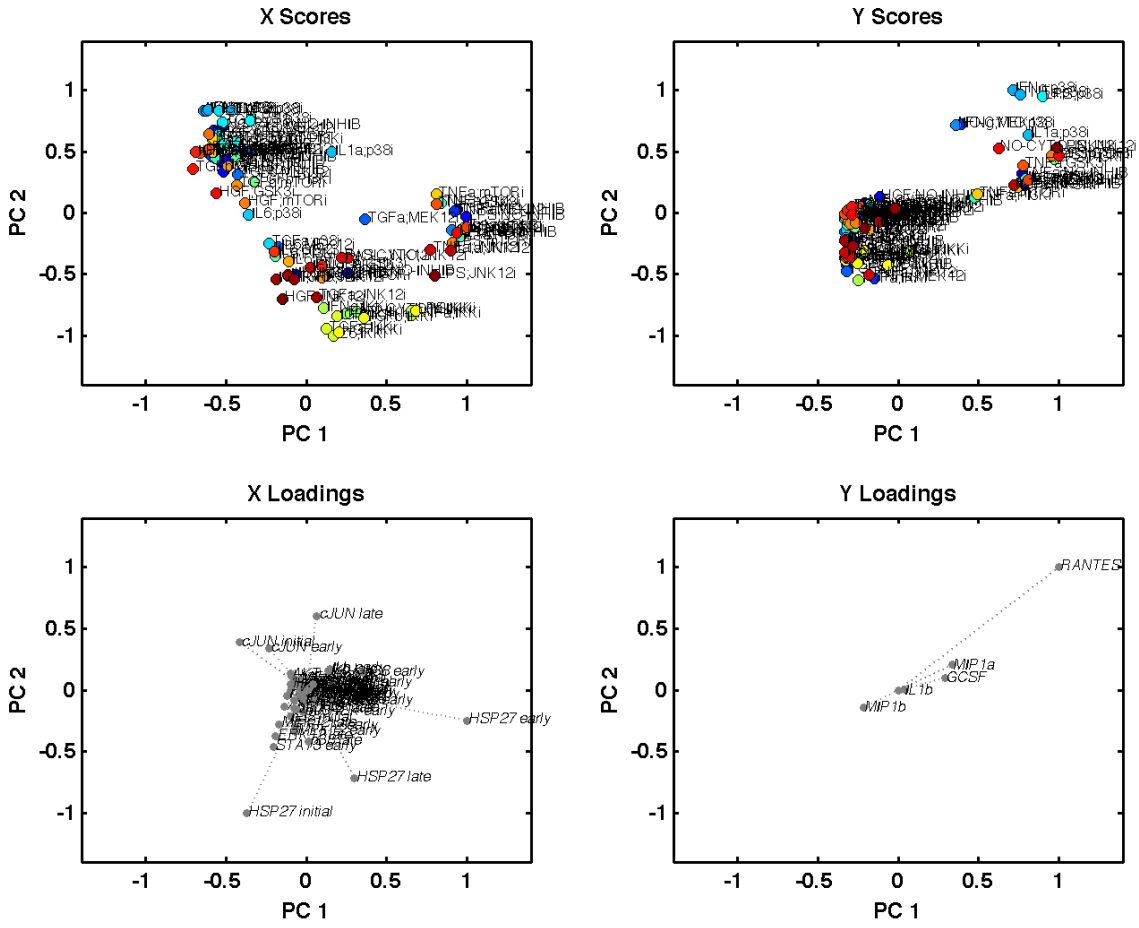

Fig. S3. Scores and loading plots for the best PLSR model. The scores (top plots) and loadings (bottom plots) for the phosphorylation data ("X" inputs, left plots) and cytokine data ("Y" outputs, right plots) indicate how treatment conditions and measurements map onto the two principal components of the PLSR model. Treatment conditions that result in similar input or output measurements cluster together in the scores plots. Phosphorylation measurements that are important for prediction are located farther from the origin in the X loadings plot.

Table S1. Eight compendia that we have assembled from research in our lab and processed with *DataRail*. Abbreviations—cyto: cytokine concentration; DNA: DNA content; inh: small-molecule inhibitors; lig: ligands; lines: cell lines; pp: phosphoprotein concentration; rep: replicates; time: time points; cl: cellular location. The largest data set could be loaded in less than 3 minutes on a desktop computer.

Description	Measurement Types	Data Points	Treatment Dimensions	Data Acquisition Dimension
Liver CSR	bead-based immunoassay, plate reader	36,000	3 time x 11 lig x 8 inh x 2 lines	70 = 50 cyto + 17 pp + 3 phenotypic
Liver toxicity	bead-based immunoassay, plate reader	23,000	5 time x 6 lig x 11 drugs	69 = 50 cyto + 17 pp + 2 phenotypic
Apoptosis	plate reader	15,000	12 time x 6 lig x 8 inh x 2 rep	13 = 9 pp + 4 protein
EGFR	Quantitative microscopy	1,626,084	6 time x 3 cl x 2 channels	1 = 1pp ≈ 2800 cells/well
EGFR	quantitative microscopy	480,000	6 time x 8 lines	10 = 4 pp + 5 protein + DNA ; 1000 cells/well
EGFR	bead-based immunoassay	5,000	6 time x 2 lig x 8 lines	55 = 50 cyto + 5 pp
Macrophages	bead-based immunoassay	24,000	12 time x 15 lig x 2 rep	67 = 50 cyto + 17 pp
Apoptosis	kinase assay, immunoblot, Ab microarray	10,000	13 time x 12 lig x 3-6 rep	19 = 10 pp + 9 protein

Table S2. Example of MIDAS file. The file comprises a small subset of the data of the Liver CSR `compendium`.

TR: HepG2	TR: PriHu	TR: NOCYTO	TR: FASL	TR: IFNg	TR: LPS	TR: NOINH	TR: MEK12i	TR: p38i	TR: PI3Ki	TR: IKKi	TR: mTORi	TR: GSK3i	TR: JNK12i	DA: AKT	DA: ERK12	DA: GSK3	DV: AKT	DV: ERK12	DV: GSK3
	1	1				1								0	0	0	5578	275	1123
	1	1					1							0	0	0	4544	89	905
	1	1						1						0	0	0	5108	240	945
	1	1							1					0	0	0	2796	212	601
	1	1								1				0	0	0	4409	266	931
	1	1									1			0	0	0	4269	256	1101
	1	1										1		0	0	0	4215	280	1020
	1	1											1	0	0	0	3336	78	384
	1	1				1								30	30	30	7100	189	950
	1	1					1							30	30	30	5942	39	517
	1	1						1						30	30	30	6296	136	541
	1	1							1					30	30	30	4011	107	549
	1	1								1				30	30	30	6862	141	804
	1	1									1			30	30	30	6402	160	957
	1	1										1		30	30	30	5485	193	720
	1	1											1	30	30	30	4720	58	340
	1		1			1								30	30	30	4551	239	788
	1		1				1							30	30	30	3799	106	1085
	1		1					1						30	30	30	4577	165	909

Table S3. Ranking of models. The normalizations and scaling for the best 60 models are shown.

Rank	R^2	XCube	XProcess	YCube	YProcess
1	0.5224	max scale	mean&variance	primary	AUC, mean&variance
2	0.5134	primary	mean activ., mean center	primary	AUC, mean center
3	0.5129	primary	mean activ., mean center	primary	mean center
4	0.5123	primary	mean activ., mean center	primary	mean activ., mean center
5	0.5116	max scale	mean activ., mean center	primary	AUC, mean center
6	0.5106	max scale	mean activ., mean center	primary	mean activ., mean center
7	0.5097	max scale	mean activ., mean center	primary	mean center
8	0.5067	primary	mean&variance	primary	AUC, mean&variance
9	0.4991	primary	slope, mean center	primary	AUC, mean&variance
10	0.4952	primary	mean activ., mean center	primary	AUC, mean&variance
11	0.4864	max scale	mean&variance	max scale	AUC, mean&variance
12	0.4862	max scale	slope, mean center	primary	AUC, mean&variance
13	0.4852	primary	slope, mean center	max scale	AUC, mean&variance
14	0.4841	primary	mean activ., mean center	max scale	AUC, mean center
15	0.4826	primary	mean activ., mean center	max scale	mean activ., mean center
16	0.4818	max scale	mean activ., mean center	max scale	AUC, mean center
17	0.4804	max scale	mean activ., mean center	max scale	mean activ., mean center
18	0.4797	primary	mean activ., mean center	max scale	AUC, mean&variance
19	0.4795	primary	mean activ., mean center	max scale	mean center
20	0.4761	max scale	mean activ., mean center	max scale	mean center
21	0.4707	primary	mean&variance	max scale	AUC, mean&variance
22	0.4675	max scale	mean&variance	primary	mean activ., mean center
23	0.4671	max scale	mean&variance	primary	AUC, mean center
24	0.459	max scale	slope, mean center	max scale	AUC, mean&variance
25	0.4554	max scale	slope, mean center	primary	mean center
26	0.4544	max scale	slope, mean center	primary	AUC, mean center
27	0.4535	max scale	slope, mean center	primary	mean activ., mean center

28	0.4466	max scale	mean activ., mean&variance	max scale	AUC, mean&variance
29	0.4459	max scale	mean activ., mean&variance	primary	AUC, mean&variance
30	0.4453	max scale	mean activ., mean center	primary	AUC, mean&variance
31	0.4418	primary	slope, mean center	primary	mean center
32	0.4368	max scale	mean&variance	max scale	AUC, mean center
33	0.4322	max scale	slope, mean center	max scale	AUC, mean center
34	0.431	max scale	slope, mean center	max scale	mean activ., mean center
35	0.4308	max scale	mean&variance	max scale	mean center
36	0.4296	max scale	slope, mean center	max scale	mean center
37	0.4284	max scale	mean activ., mean center	max scale	AUC, mean&variance
38	0.4249	primary	mean center	primary	mean activ., mean center
39	0.4243	primary	mean&variance	primary	mean activ., mean center
40	0.4211	primary	mean center	primary	AUC, mean&variance
41	0.4204	primary	mean&variance	primary	mean center
42	0.4195	max scale	mean activ., variance sc.	max scale	AUC, mean&variance
43	0.4171	primary	slope, mean center	max scale	mean center
44	0.4163	max scale	mean activ., variance sc.	primary	AUC, mean&variance
45	0.4137	primary	mean activ., mean&variance	max scale	AUC, mean&variance
46	0.413	max scale	slope, mean&variance	primary	AUC, mean&variance
47	0.406	max scale	mean center	primary	AUC, mean center
48	0.4053	max scale	mean center	primary	mean center
49	0.4007	primary	mean center	max scale	mean center
50	0.3996	max scale	slope, mean&variance	max scale	AUC, mean&variance
51	0.3918	primary	mean center	max scale	AUC, mean&variance
52	0.3915	primary	slope, mean&variance	primary	AUC, mean&variance
53	0.3885	primary	mean&variance	max scale	mean center
54	0.3882	max scale	mean activ., mean center	primary	slope, mean center
55	0.3865	primary	mean activ., variance sc.	max scale	AUC, mean&variance
56	0.38	primary	mean activ., mean center	primary	slope, mean center
57	0.3754	primary	mean activ., variance sc.	primary	AUC, mean&variance
58	0.3733	primary	slope, mean&variance	max scale	AUC, mean&variance
59	0.3683	max scale	slope, variance sc.	primary	AUC, mean&variance
60	0.3659	max scale	mean&variance	primary	slope, mean center

Table S4. Comparison of metrics and scaling methods. The mean R^2 value is listed for all models that use a particular metric or scaling method. Note that negative R^2 values occur in poor models that do not include a constant offset.

Model Class	**Mean R^2**
primary X	-0.31
max scale X	-0.32
primary Y	-0.42
max scale Y	-0.22
mean center X	-0.41
variance sc X	-0.04
mean&variance X	-0.41
time series X	-0.37
AUC X	-0.27
slope X	-0.27
mean activ. X	-0.35
mean center Y	0.27
variance sc Y	-1.38
mean&variance Y	0.23
time series Y	-0.5
AUC Y	-0.64
slope Y	0.01
mean activ. Y	-0.1