Linköping university | Department of biomedical engineering
Master's thesis, 30 hp | Engineering biology
Spring 2016 | LiTH-IMT/BIT30-A-EX--16/529--SE

A first phosphoproteome-wide mechanistic model of insulin signaling

William Lövfors

Supervisor: Elin Nyman
Examiner: Gunnar Cedersund

Linköping university
SE-581 83 Linköping
013-28 10 00, www.liu.se

Abstract

Diabetes, a disease occuring in almost 200 million people world-wide, have been studied for centuries and have been discovered to occur when cells stop responding to insulin stimuli properly. Recently, the insulin response of adipocytes has been modeled in detail for the core of the signaling network. These mechanistic models are based on time-resolved, high-quality data and can therefore capture the dynamics of insulin signaling in cells from control subjects. These models can also, by changing a single mechanism, describe the insulin response in patients with type 2 diabetes. However, all such mechanistic models developed so far only contains a handful of protein modifications, and there exists no method that can scale such models to satisfactorily account for large-scale recently available mass-spectrometry data, which simultaneously measures tens of thousands of protein modifications. In this thesis, I present a first method that can scale small-scale reliable mechanistic modelling to also work for such high-throughput mass-spectrometry data. More specifically, the method expands an existing core model in an iterative fashion. At each iteration, the states of the (expanded) core model is used as input to new proteins, using a list of possible interactions from prior knowledge databases. If the model simulations and data are in agreement, the new protein will be added to the model. These added proteins are then used as new potential inputs for the next iteration. This process is repeated until no more proteins can be added.

With this method, a phosphoproteome-wide mechanistic model for the cellular insulin response which is in agreement with both time-resolved data and prior knowledge has been achieved. The model was then used to predict the effect of inhibitions and diabetes system-wide. All in all, these results illustrate that the long-held dream of systems biology now is at hand: a time-resolved mechanistic understanding of systems-wide intracellular signaling that allows for computer-based screenings of drug targets across the entire proteome.

Preface

This work has been done as a master's thesis for engineering biology, worth 30 hp, during the spring of 2016 at Linköping university. All work has been done by me, except for the compilation of the lists of interactions. This compilation of lists has been done by Emil Arkstål, and has been much appreciated. I would also like to thank my supervisor and examiner for being highly involved in the project and providing useful support when needed.

Throughout the report, proteins have mostly been referred to by their gene names instead of their protein names, except for when referencing the work of others and in some figures. For this purpose, a translation table have been constructed to help the reader. It can be seen below.

Table 1: **Translation between gene names and protein names**

Gene names	**Protein names**
Insr	IR
Irs1	IRS1
Akt1	PKB
Akt2	PKB
Akt3	PKB
Tbc1d4	AS160
Rps6kb1	S6K
Rps6	S6
Mapk1	Erk1
Mapk3	Erk2

Contents

1. Introduction

Multiple molecules circle around in the blood each day. These are responsible for many essential bodily functions, such as getting the cells to take up energy consumed via food. This is e.g. done via a molecule called insulin, which makes the cells take up glucose. The key steps in this function, as well as a general overview, can be seen in Figure 1.1. It has been shown that when this network stops functioning properly in human adipose (fat) tissue, diabetes (of type 2) occurs in the person [1]. As is clear from the figure, the system is extremely complicated and is occurring over many layers, ranging from intracellular networks to whole-body interactions. In the last 10 years, a gradually evolving mathematical model has sought to bring order to this complexity, by unraveling the role and function of the intracellular network that is activated by insulin in adipocytes. However, that model is still limited in size, describing only a handful of proteins. In this thesis, I present a new approach that allows us to expand such small-scale models to describe all of the thousands of protein modifications that we nowadays can measure in a cell. Let us now consider a little bit more closely the state-of-the-art methods and understandings, to which my new approach then will be added.

1.1 Diabetes

Diabetes is not a modern disease. The oldest known description of diabetes is from ancient (around 1500 BCE) Egyptian papyrus texts, and at approximately the same time physicians in India developed the first clinical test for the disease [2]. Diabetes have therefore most likely been studied for at least 3000 years, with many significant discoveries having been made since then. Sharpey-Schafer proposed the first hypothesis in 1910 that the deficiency of a single molecule (insulin) is the cause of diabetes, which is known today to be the cause behind one of the two sub types of diabetes (type 1) [3]. Diabetes type 1 occurs when the body is unable to produce insulin, and type 2 (which is more common) has to do with that the cells stop responding to insulin stimuli [3]. Furthermore, Clark and Lyons developed the first electronic glucose-sensor in 1962 [4] and more recently the first biosynthetic human insulin dose [3]. In total, research on diabetes have generated 10 Nobel prizes over the last 200 years [3]. However, despite these advances the amount of people living with diabetes are ever increasing. It was

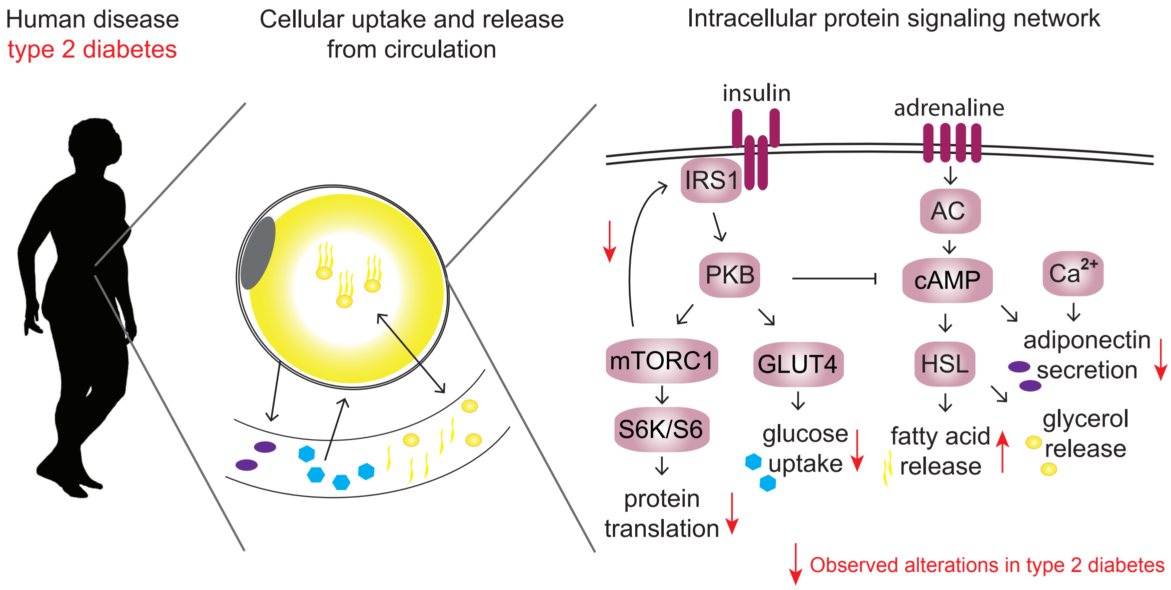


Figure 1.1: **Insulin signaling network and diabetes.** An adipocyte and key signaling proteins are shown in the middle. The key proteins in the signaling network in adipose tissue are shown in purple to the right.

estimated that 171 million people world-wide was living with diabetes in 2000, with an expected increase to 366 million in 2030 [5]. Since diabetes is still on the rise, there is incentive in trying to get more insights into the mechanisms of diabetes in order to come up with new treatments and/or ways to prevent the disease from arising in the first place.

1.2 Omics-data and protein-protein interactions

In order to find out how different diseases work, and why they arise, scientists have for a long time been looking into how various parts of the cells interact with other parts. This is done in an attempt to understand the intrinsic mechanisms in the cells, with the long-term goal of finding the cause and potentially the cure to various diseases. In recent years, new experimental methods have given scientists the ability to measure vast amounts of different components in living systems. One of the first applications for this was the characterization of the human genome, the enormous atlas over every gene in the human cells. However, genes are not the only component of living systems that exists in tens of thousands of different varieties. Other common large-scale, so called omics, areas are the atlas of proteins (proteomics) and phosphorylations of proteins (phosphoproteomics). These different components of the living systems are not static components, isolated from their environment. On the contrary these parts of the systems interact with other parts to such an extent that scientists have started talking about an interactome. The interactome suggests cross-talk between all levels of the living cells. One of the most important types of these interactions is the interaction between proteins, the protein-protein interaction (PPI).

It is known that proteins interact with each other in many different ways. Proteins can e.g bind to each other, and they can make different kinds of modifications of the fundamental protein structure after the protein is translated from mRNA, so called post-translational modifications. Since the protein is made up of a combination of only 20 different amino acids (AAs), adding post-translational modifications can widely increase the range of functionality of the protein [6]. It can also work as a way for cells to reversibly regulate the activity of a protein. One of the most prevalent types of post-translational modifications is the phosphorylation reaction [6]. In this type of reaction, a phosphoryl-group (a phosphate with three attached oxygen atoms) is added to the side chain of one of the AAs in the protein. However, this addition of a phosphoryl-group is not a permanent modification and the phosphoryl-group can therefore be removed. This removal is called a dephosphorylation, and can be seen as the reverse reaction to phosphorylation. In total, phosphorylations can be found on 139,582 sites on 530,264 proteins [6]. Clearly there is a need to organize this knowledge in an structured way. Therefore multiple databases of PPIs have been constructed. However, simply having the information stored in large databases is not enough, to truly understand a living system, the interactions have to be put into a bigger picture. An example of such a bigger picture is a hypothesis of a signaling network.

1.3 Signaling networks

In biological signaling networks, some type of external signal is transduced throughout the network, leading up to some kind of intracellular response. Such networks are very common in biology, to such an extent that cells have developed specific receptors on the outer side of the plasma membrane that are able to differentiate between different types of signals (such as different molecules). When a receptor is activated by an extra-cellular signal, the receptor activates various proteins inside of the cells, which in turn active other proteins. This signaling cascade is propagated within the cell, leading up to a cellular response. An example of this would be the stimulation of cells with insulin, which signals through the insulin receptor, via various proteins and ends up with an increased uptake of glucose into the cell.

To make things more complicated the signal that is transduced through out the network can take different shapes (such as constantly increasing, constantly decreasing, plateauing, or having an overshoot) all depending on the structure of the network. Clearly there is a need to structure this complex system using a systematic approach. This is where the field of systems biology comes into the picture.

1.4 Systems biology

Systems biology is a field where mechanistic modeling, using ordinary differential equations (ODEs), are used to achieve an understanding of how biological networks are constructed. This is done by formulating and testing hypotheses of how the network is constructed. A typical modeling cycle is shown in Figure 1.2, and in detail the knowledge of a biological network available is formulated into equations describing how the different parts of the network interact with each other. By using the mathematical model to predict the outcome of an experiment, which is also performed in vitro or in vivo, the model can be evaluated and possibly rejected. The model (and therefore also the hypothesis) is rejected, if the model prediction does not agree with the experimentally acquired data. Unfortunately, it is not possible to prove that a hypothesis (in the form of a model) is true, it is only possible to reject false hypothesis. Therefore many predictions must be made, and tested, in order for a model to be accepted as true. One such model which has been thoroughly tested is the model over insulin signaling [7].

1.5 Previous work

Previously, a detailed mechanistic model of insulin signaling in adipocytes (referred to as the **core-model**) has been developed by Brännmark et al. [8]. Around the same time the phosphoproteome was measured in 3T3-L1 cells (an adipocyte cell line) using mass-spectrometetry (referred to as the **MS-data**).

1.5.1 The core-model of insulin signaling

In a mechanistic model published in 2013, the main cause of how healthy controls and type 2 diabetic subjects in primary human adipocytes [8] differ were unraveled. This model has then been further extended by Nyman et al.[7] to also include Mapk1/3 (also referred to as ERK) and Elk, and can be seen in Figure 1.3. In this extended model, referred to as the **core-model**, the core proteins and their interactions in insulin signaling have been extensively studied. However, it is still a relatively small model with only around 10 proteins (less than 1% of the entire proteome).

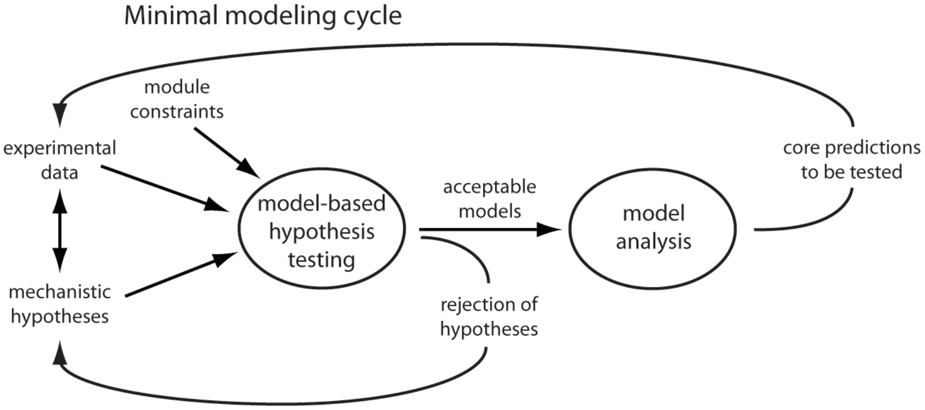


Figure 1.2: **A typical modeling cycle.** Experimental data are collected and used to evaluate a hypothesis. If the data can not be explained by the model, the hypothesis is rejected. If the model can not be rejected, predictions are made and compared against new data. This process is iterated until the model is considered to be valid.

In the core-model the mechanism behind insulin resistance has been unraveled. This was done by using a mathematical model to explain data collected from primary human adipocytes (both from diabetic patients and healthy controls) and measured using western-blot (WB) [8]. When explaining the data, the same model was used for both healthy and diabetic data, only the values of a few parameters were changed [8].

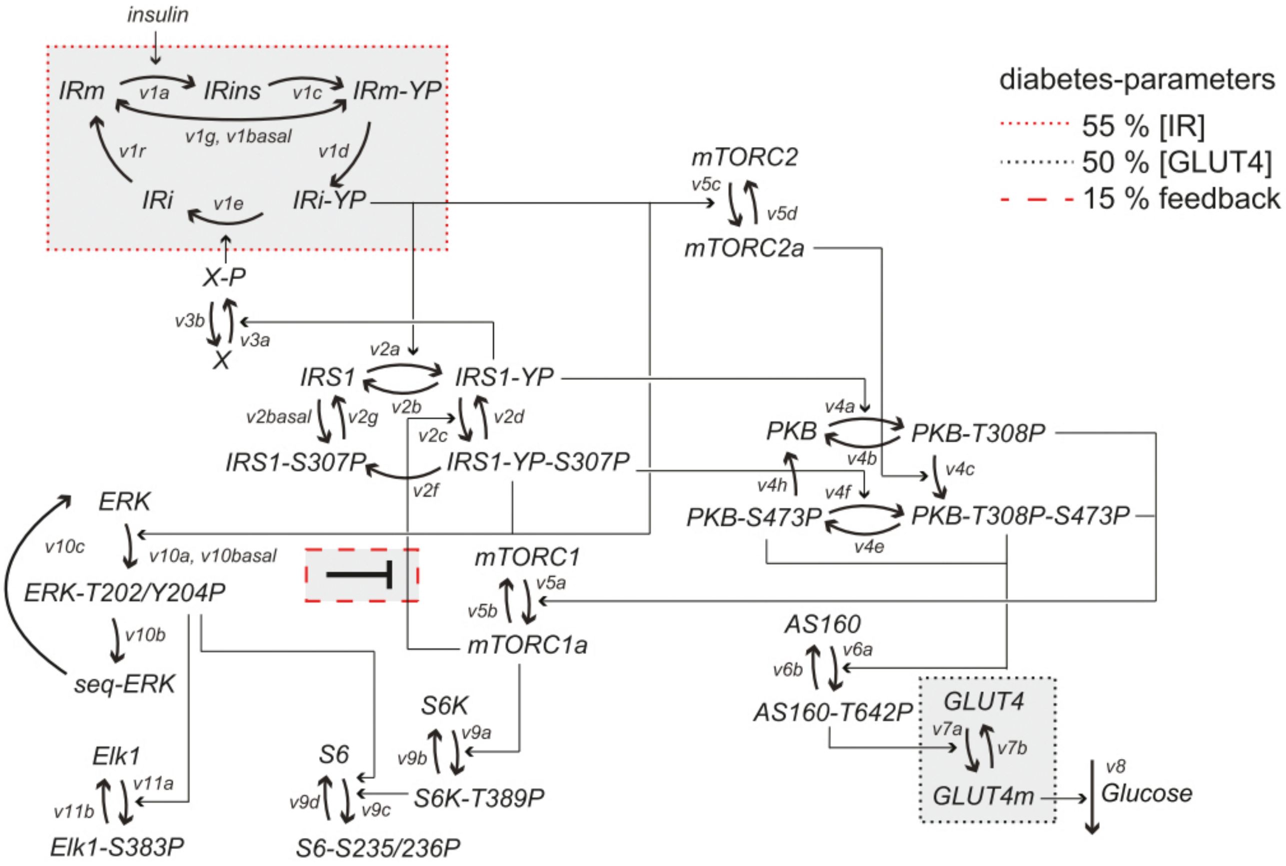


Figure 1.3: **The structure of the core-model of how insulin signaling results in glucose uptake.** In the model, the mechanisms can be seen captured in red, and under ”diabetes-parameters” in the top right corner. By changing only these parameters, the model (with the same model parameters) can explain either healthy or type-2 diabetic data. This research was originally published in The Journal of Biological Chemistry. Nyman et. al. A Single Mechanism Can Explain Network-wide Insulin Resistance in Adipocytes from Obese Patients with Type 2 Diabetes. The Journal of Biological Chemistry. 2014; 289:33215-33230. © the American Society for Biochemistry and Molecular Biology.

1.5.2 The mass-spectrometry data

In an experiment performed by Humphrey et al. in 2013, the entire phosphoproteome of 3T3-L1 cells (an adipocyte cell line) was measured using mass spectrometry [9]. The authors measured the phosphoproteome at 9 different time points (0 sec, 15 sec, 30 sec, 1 min, 2 min, 10 min, 20 min and 60 min), as well as how the phosphoprotome was affected by two different inhibitions (by the chemicals MK22006 or LY294002, referred to as MK and LY)[9]. The inhibitions could either affect Pi3k (LY) or Akt (MK), two proteins involved early in the insulin signaling network. The authors had also created an additional data set from the main data set, where only the sites responding to insulin were present. These will be referred to as ”insulin responding” and ”full” data set.

1.6 Limitations in previous works

In the mechanistic core-model, the complex system which connects insulin stimulation with glucose uptake in primary human adipocytes has been investigated [7]. However, the mechanistic model developed does only account for less than one percent of the total proteins in the cell, and is therefore in need of expansion. The MS-data on the other hand, covers the proteome but has not been used for generating mechanistic models.

The works previously performed (represented in Figure 1.4) have been done with two different purposes, one with the aim of creating a small-scale, detailed, reliable network, and the other with the aim of measuring the proteins involved in the entire network, yet not looking at the individual interactions within the network. However, no work currently exists that has tried to do both: to create a mechanistic model that quantitatively describes the interactions involved in the entire intracellular signaling network that responds to insulin. In fact, no method has ever been presented that allows for such large-scale mechanistic models to be created. Herein, I present a first such method.

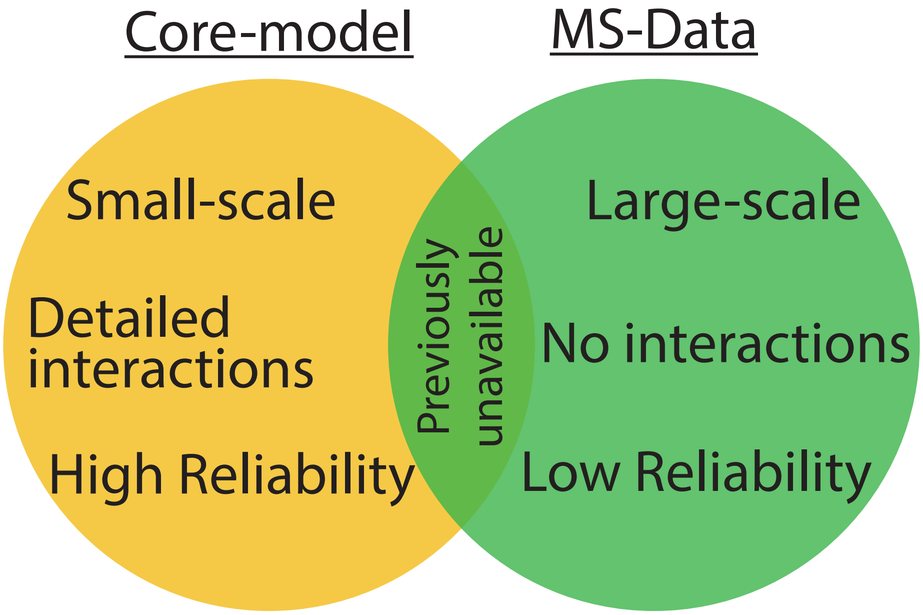


Figure 1.4: **Features of the previous work.** The core-model has been studied in great detail. By collecting high reliability data to compare the model against, a model with much knowledge regarding the interactions in the model has been achieved. However, this high detailed modeling is cumbersome and takes much time, therefore only a few proteins (small-scale) are part of the model. In the MS-data, on the other hand, the entire phosphoproteome has been measured (large-scale). However, no interactions between different proteins are given in the data, and such large-scale data have previously only been analyzed using low reliability statistical methods. In this thesis, the previously unavailable cross-section of the two works has been explored.

1.7 Aim

The aim of this project is to develop a method for combining the detailed small-scale core-model of insulin signaling with the MS-data using a list of interactions from databases, into a large-scale mechanistic model, which is in agreement with both data and prior-knowledge, and that can simulate diabetes in the phosphoproteome.

1.8 Delimitations

The ambitious aim outlined above will have some constraints: 1) only the MS-data will be used to evaluate the suggested interactions, 2) only proteins that can pass a χ^2-test for the agreement between the model and the uninhibited MS-data will be added to the model, 3) the order of tested interactions will be arbitrary, 4) only the core-model of insulin signaling will be expanded, and 5) the core-model will only have more proteins added and will not have any feedbacks.

2. Method

In order to create a mechanistic model for the entire phosphoproteome, new efficient ways of expanding available models had to be found. This was done by adding new proteins to the model that lies downstream of the current model. Using this approach, it was possible to expand the model iteratively.

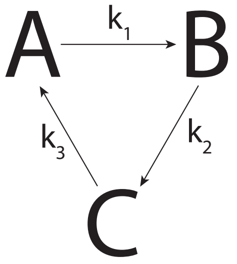
The method essentially consisted of three steps. The first step was to extract the information from the MS-data. The second step was to use databases of the interactome to gain prior knowledge of what protein interactions to add. The third and last step was to add all these, with decreasing quality, prior knowledge interactions, assuming that they also pass a χ^2-test. By using this approach, the model will have the possibility to give the end user a choice of how reliable the model should be, e.g. what reliability on the suggested interactions that should be allowed.

2.1 Mathematical modeling, software and parameter estimation

In order to expand the core-model, the models have to be defined in an explicit way. This was done by using time-resolved ODEs, implemented in the framework of the MATLAB toolbox ”SBToolbox2”. Using ODEs, the relationship between the different proteins and phosphorylations can be defined as equations. In the models, the proteins that changes over time are referred to as ”states” (A, B, C in Figure 2.1), the interactions between them as ”reactions” (the right-hand side in the equations for the states in Figure 2.1), and the rate with which the interactions are carried out are referred to as ”parameters” (k_1, k_2, k_3 in Figure 2.1).

These parameters are not given and needs to be estimated. In detail, the parameters are tuned such that the difference between the model simulation and the measured value is minimized. This optimization of the parameter values are done using the MATLAB implementations of the simulated annealing algorithm, and a local optimization algorithm.

The difference is calculated by taking the sum of squares of the residuals between the model simulation ($\hat{y}_t(p)$) and the data (y_t), normalized by the squared standard error of the mean (SEM) for all time points (t) in the measured



$$
\begin{aligned}
\frac{d}{dt}(A) &= -k_1 \cdot A + k_3 \cdot C \\
\frac{d}{dt}(B) &= -k_2 \cdot B + k_1 \cdot A \\
\frac{d}{dt}(C) &= -k_3 \cdot C + k_2 \cdot B \\
A(0) = a, B(0) &= b, C(0) = c \\
k_1 = d, k_2 &= e, k_3 = f
\end{aligned}
$$

Figure 2.1: **An example of an interaction graph and the corresponding ODEs.** The interaction graph on the left represents a system where A is transformed into B, which is transformed into C, which in turn is transformed back to A. This transformation is done with the rate constants k_1, k_2, k_3 (all having arbitrary values in the equations to the right) and the initial values $A(0), B(0), C(0)$ (also having arbitrary values).

data, as can be seen in Equation 2.1. This measure is commonly referred to as the ”cost” of the model.

$$\sum_t \frac{\left(y_t - \hat{y}_t(p)\right)^2}{SEM_t^2} \tag{2.1}$$

If the cost is low, the agreement with data is high, and a way of determining if the agreement is good enough is to use a χ^2 test. This can be done since the cost is relatable to the values given by the inverse of the χ^2 cumulative distribution function, with degrees of freedom equal to the number of data points. If the cost is larger than the χ^2 value, then the model is rejected, with a certain statistical reliability (often 95 %).

When the mathematical model had been expanded and was in agreement with data, the resulting network was vizualised in Cytoscape[10].

2.2 Data analysis

The first step was to analyze the MS-data since unfortunately the sequences measured in the MS had been mapped against proteins using an already outdated version (of the now defunct) mouse international protein index (IPI) database (v3.68). Therefore, the first step was to remap the sequences against an up-to-date list before the mean values and SEMs could be calculated.

2.2.1 Mapping against unique identifiers

The sequences were mapped to proteins using the UniProt database[11] instead of the IPI database. Firstly, the sequences were mapped against the more reliable (reviewed), database entries from the UniProtKB/Swiss-Prot database. If a sequence did not match any entries from UniProtKB/Swiss-Prot, UniProtKB/TrEMBL (unreviewed) entries was tested next. If an AA-sequence mapped to multiple proteins, the protein suggestions was checked against the names in the original MS-data and if a match was available, that protein was used. Since Uniprot is built up with gene names with only alphanumericals, in combination with that the databases of protein interactions also have the same standardized gene names, these names were used for identification of the proteins.

2.2.2 Extracting mean and standard-error of the mean values

When the sequences had been mapped to up-to-date information, the next step was to calculate the mean and SEM values for each protein. For this, the sites were initially scaled to be as similar to each other as possible. This scaling was done by using the 'Least-squares solution in presence of known covariance' (lscov) function in MATLAB. Once the sites had been scaled against each other, mean and SEM values were calculated for each protein.

The next step was to rescale the mean and SEM values, since the individual sites had been arbitrarily scaled against each other. For this, two new data sets were constructed. In these two new data sets the mean and SEM value had been scaled to either have the first mean value for each protein equal to one, or the maximum mean value for each protein equal to 50. The first scaling was done since the original data was normalized to one. The second scaling was done since due to how the model are constructed, the phosphorylation levels are seen as a percentage of total phosphorylation, and the data was therefore seen as having reached half of the maximum response.

2.3 Lists of interactions

The second step was to formalize the prior knowledge of the PPIs from different databases into a list of known interactions. The databases have slightly different information and intended usage, and therefore the list of known interactions had to be constrained to only a few properties. In detail, the interactions were stored as pairs of proteins with a directed interaction. A directed interaction means that protein A only affects protein B, while an undirected interaction means that both A affects B, and that B affects A. Since the databases have different levels of detail regarding the direction, type and reliability of the interactions, the total list of interactions was separated into three different lists for high, medium and low quality interactions. The databases from which the interactions have been collected are summarized in Table 2.1.

2.3.1 Different quality of the interactions

Table 2.1: **Summary of data bases used.** Eight different databases were used to make the three different lists of interactions. The interactions from the databases were separated into three different lists, called ”high”, ”medium” or ”low”, depending on the quality of the information given about the interactions. Which list a database were used in is listed under ”List”. How many interactions that were used from each database are given under ”Interactions used”.

Database	**List**	**Interactions used**	**Reference**
Pathway Studios	High	10,246	[12]
PhoSigNet	Medium	7,223	[13]
PhosphoSitePlus	Medium	16,354	[14]
IntAct	Medium	3,062	[15]
RegPhos	Medium	6,510	[16]
PhosphoELM	Medium	5,532	[17]
Pathway commons	Medium	11,890	[18]
BioGRID	Low	1,011,962	[19]

In the high quality list, only interactions which fulfilled the criteria of both being explicitly stated that the interaction were either a phosphorylation or a dephosphorylation, and that were mentioned in at least 3 scientific articles were included. For this list, the Pathway studio database was used. This list contains a total of 10,246 interactions.

The medium quality list contains interactions that are explicitly listed as phosphorylations, with no requirement of reliability. For this list, the interactions were collected from seven databases: PhoSigNet, IntAct, Pathway Studio, RegPhos, PhosphoELM and Pathway Commons. This list containes a total of 23,342 interactions not included in the high quality interactions.

The low quality list includes all interactions with no added criteria, i.e. as long as two protein were mentioned together, the interactions were used. In this list are a total of 838,580 interactions not included in the high or medium list, all from BioGRID.

2.4 Achieving agreement between the core-model, list of interactions and MS-data

There were three mayor inconsistencies between the core-model, the list of interactions and the MS-data which had to be addressed. Firstly, the core-model used protein names, while the databases used gene names to identify the proteins. For this reason, the names in the core-model were changed to the same gene names that were used in the databases. Secondly, some proteins in the core-model only had a state for a combination of all isoforms of the protein, while the databases had them listed separately. This was solved by giving the same input from the core-model for all isoforms in the list of interactions. Thirdly, the core-model had combined the proteins in the mammalian target of rapamycin-complex (mTORC) 1 and 2 into two states called mTORC1 and mTORC2, while the databases have the proteins that make up the complexes listed separately. For interactions having any of these proteins involved, the corresponding state for the protein complex was tested as inputs. Also, some proteins were part of both mTORC1 and mTORC2. This was solved by making copies of these interactions, and trying both states from the core-model (mTORC1/mTORC2) as inputs independently. Lastly, the core-model had states for some proteins that corresponded to specific phosphorylation sites, while the databases did not have such detailed information. In these cases, the interaction was copied, and all corresponding states from the core-model were tested as inputs independently. Once the core-model, the list of interactions and MS-data were in agreement, the core-model could be expanded.

2.5 Expanding the model

In order to expand the model, prior knowledge interactions were tested using different types of kinetics. In this step, the suggested interactions were tested partially independently from the rest of the model. More specifically, the simulation of the input from the already available model will be given as an interpolated input to the target protein. This method will be referred to as **pairwise** testing.

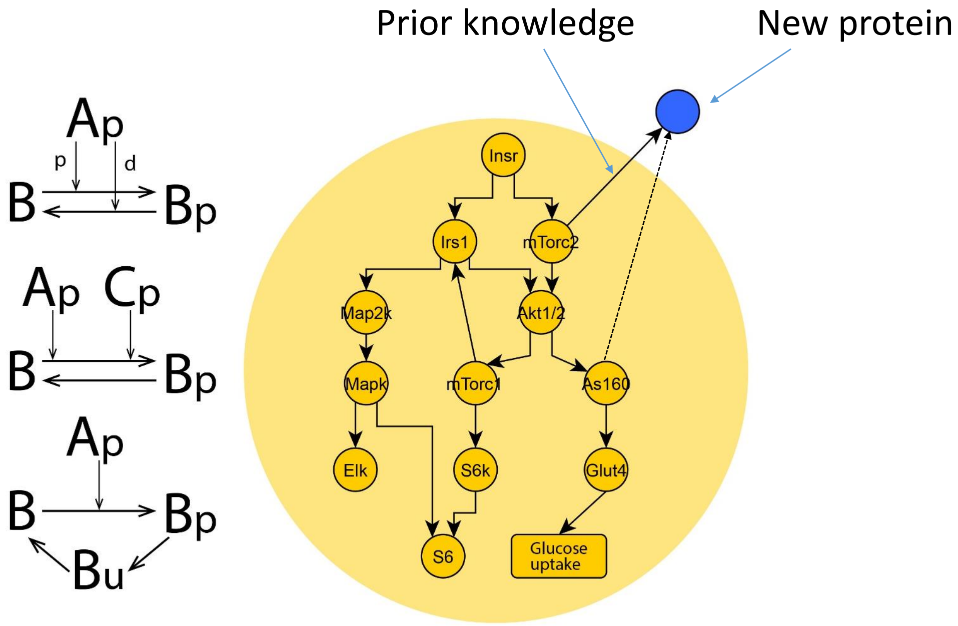


Figure 2.2: **Overview of the pairwise testing.** New proteins (blue dot) are being added to the core-model (yellow) using prior knowledge in suggested interactions (black line). The interaction is tested using four different interaction types: 1) phosphorylation, seen in the first interaction graph, with the input used denoted with ”p”, 2) dephosphorylation, also seen in the first interaction graph but using the input denoted with ”d”, 3) multiple input seen as the solid and dashed line in the figure and second interaction graph and 4) using an additional state seen as the solid line in the figure and third interaction graph.

2.5.1 Pairwise testing

The pairwise model is essentially formed by the following differential equations, where B is the target protein and A_p is the interpolated input. In this example (also represented in Figure 2.2, upper interaction graph), the target protein can be either phosphorylated or unphosphorylated.

$$
\begin{aligned}
\frac{d}{dt}(B) &= -v \\
\frac{d}{dt}(B_p) &= v \\
v &= k_f \cdot A_p \cdot B - k_b \cdot B_p
\end{aligned}
\tag{2.2}
$$

By using an interpolation of a simulation as input to the pairwise interaction, the enormous problem of generating the full model can be broken down into multiple smaller ones. In this way the otherwise impossible problem can be solved in a relatively straight forward manner. In Equation 2.2, v can easily be changed to different model kinetics, as opposed to the mass-action (MA) kinetics used. Two other kinetics would be either Michaelis-Menten (MM) kinetics (v_1 in Equation 2.3), or a dephosphorylation (v_2 in Equation 2.3). Another possibility is that more than one input is required to get the right time dynamics. In v_3, in Equation 2.3 an extra interpolated input is given, called C_p (also represented in Figure 2.2, middle interaction graph). When adding multiple inputs, the number of inputs is contained to only two inputs, since using more than two would unlikely result in the desired time dynamics.

$$\begin{aligned} v_1 &= k_f \cdot A_p \cdot B - k_b \cdot B_p \\ v_2 &= k_f \cdot A_p \cdot B - k_b \cdot B_p \\ v_3 &= k_f \cdot A_p \cdot C_p \cdot B - k_b \cdot B_p \end{aligned} \tag{2.3}$$

Besides using alternative kinetics or multiple inputs, it is possible to allow for the protein to shift between an additional state, referred to as an ”unavailability pool” (U in Equation 2.4) (also represented in Figure 2.2, lower interaction graph). Such an alternative is not unreasonable in biological systems. E.g. proteins are often internalized, or shuttled between the cytosol and the nucleus.

$$\begin{aligned} \frac{d}{dt}(B) &= -v_1 + k_{ub} \cdot U \\ \frac{d}{dt}(B_p) &= v_1 - k_{uf} \cdot B_p \\ \frac{d}{dt}(U) &= -k_{ub} \cdot U + k_{uf} \cdot B_p \\ v &= k_b \cdot B_p - k_f \cdot A_p \cdot B \end{aligned} \tag{2.4}$$

When expanding the model, the different types of interactions are tested in the following order. First, each available input is tested using phosphorylation with MA kinetics, then phosphorylation with MM kinetics and last dephosphorylation. Second, if no inputs are successful in yielding simulation that is in agreement with data, then the input with the lowest cost is tested in combination with one other input. These are tested one by one until a pair is able to yield simulations in agreement with data. Thirdly, if no pair of multiple inputs is able to explain the data, then unavailability pools are tested. If any of these three steps is able to yield a simulation that is in agreement with data, then that interaction is added to the model and no further types of interactions are tested for that specific protein. However, if none of these steps can yield a simulation in agreement with data, then the interaction is not added. That specific interaction is not tested again, but it can be used as one of the multiple inputs if new interactions are added (i.e. when allowing for lower quality interactions).

2.5.2 Adding interactions

In the beginning, only the proteins in the core-model will be used as inputs and only the high quality list of interactions will be used. This essentially means that only the high quality interactions where the suggested input is already in the model will be tested. Since the interactions are tested pairwise, there can be no cross-talk between the suggested proteins. Therefore it is possible to test all new interactions independent of each other. All interactions that pass a χ^2-test will be added to the model. When the model has been expanded one layer, the proteins that have been added to the model can be used as potential inputs for more interactions. In this next iteration, target proteins that previously did not pass the χ^2-test will be tested again using any of the newly added proteins (assuming the interaction is in the prior knowledge list) as input, or a combination of available inputs (including the ones that previously failed). This step will iterate until no more interactions are able to pass the χ^2-test, and no more proteins are added.

When no more interactions can be added to the model, the medium and low quality interactions will be added to the list of interactions. This will be done in two steps, where the medium quality interactions are added first and the process of incrementally expanding the model started again. When no more proteins can be added to the model using the high and medium quality interactions, the low quality interactions will be added to the list of interactions. Once no more proteins can be added to the model, the model has been fully expanded.

2.5.3 Estimating the upper limit on model expansion

An estimate on how large the model could potentially be (ignoring the agreement with data) was determined by removing the interactions where either of the proteins was not present in the MS-Data, or using all possible interactions left. By entering the interactions into Cytoscape, a full model could be acquired. Since the interactions had been collected without any regard to their connection to the original model, some interactions formed small islands not connected to the original model. These were removed, and the remaining proteins were counted. This gives the upper limit of the model expansion. Naturally, by using different levels of quality and different data sets (insulin responding or full) different upper limits were estimated.

2.6 Model validation

Once the model had been expanded, it was validated using the inhibition data. The data contain two types of perturbations (MK and LY), which were predicted using the model. For MK inhibition the effect of Akt on downstream proteins were decreased by reducing the rate of which Akt is phosphorylated at threonine 308. The protein Pi3k was not in the model and could therefore not be inhibited. Instead, the LY inhibition had to be implemented on the input of Irs1 to Akt phosphorylation on threonine 308 and serine 473, since that is where Pi3k effects Akt. The authors of the MS-data had not specified the level of inhibition achieved, and therefore different levels of inhibitions were simulated with the model. These inhibitions ranged from 50 to 96 % of inhibition.

2.7 Simulating diabetes and inhibitions

Once the model had been validated, it could be used to make predictions of diabetes. Since the mechanism of how cells shift from a healthy state to a diabetic state was included in the core-model, diabetes could be simulated for the entire cell using the expanded model. Unfortunately, the MS-data did not contain any measurements of diabetic cells (since it was from a cell-line similar to healthy cells). Therefore, it was not possible to use the diabetic simulation as a way of validating the model, only to make predictions.

3. Results

To be able to generate an expanded model capable of e.g simulating diabetes for a whole cell, the suggested model to expand had to be suitable to use as a core-model, in relation the MS-data. Since the model was used to explain the WB-data, the model must be suitable if the data from both sources were similar. To verify that the model was indeed suitable, the data from the WB-experiment and the MS-experiment were compared.

3.1 The validity of the core-model

In order to check if the data from the two different sources were similar, the MS-data was scaled to be as similar as possible to the WB-data (for each protein), and then plotted together in Figure 3.1. For some comparisons, an average of different sources (such as different isoforms) was used, while sometimes individual phosphorylation sites were used. The reason for this is that the different measurement techniques are able to identify phosphorylations at different precisions.

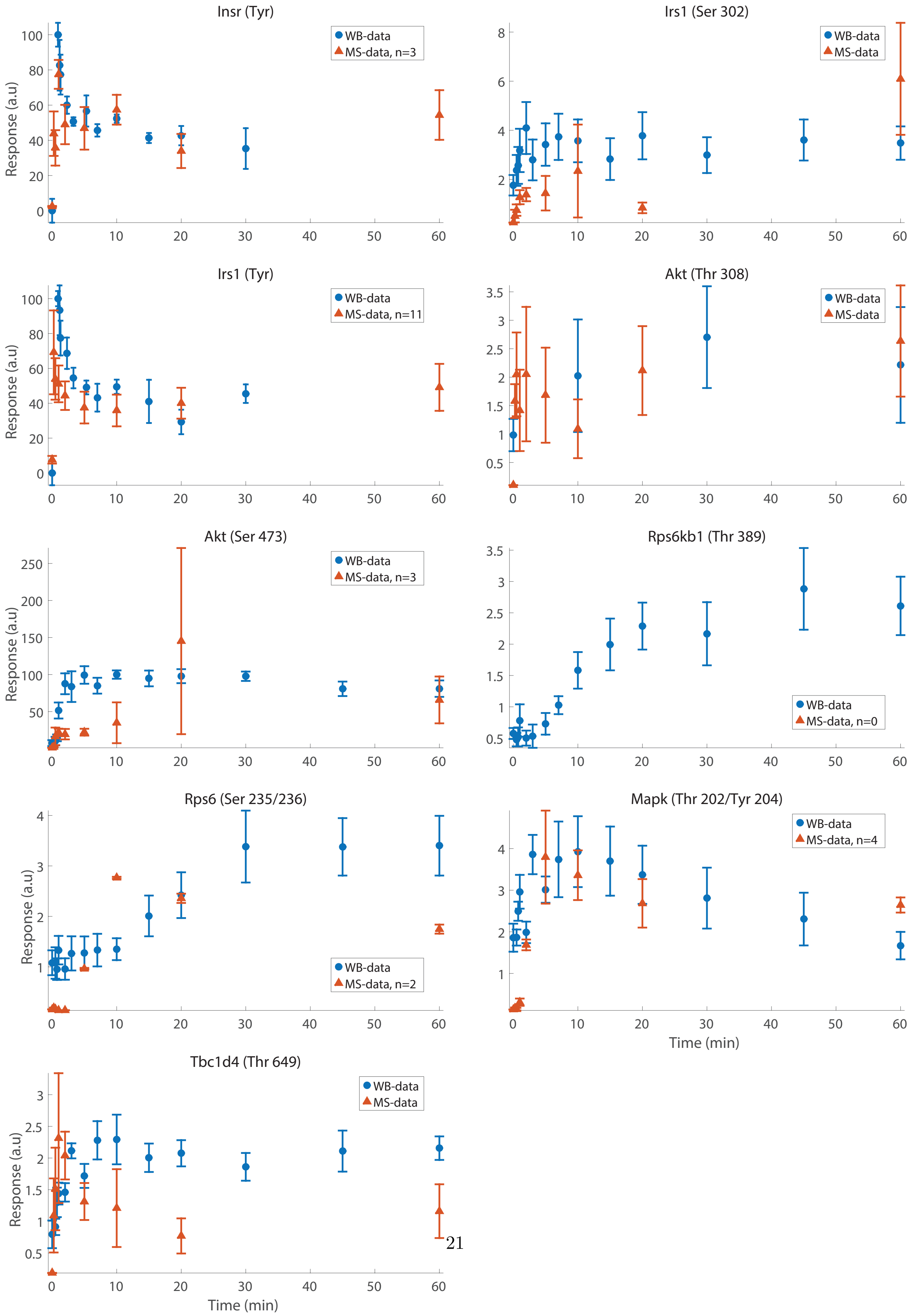


Figure 3.1: (Caption on next page)

Figure 3.1: **Comparision of western blot (WB) data and mass-spectrometry (MS) time-series data.** The measured phosphorylation levels for the different proteins from the WB experiments were compared with the corresponding values from the MS experiment. In the WB experiments, antibodies were used to quantify the amount of phosphorylation [8], and in the MS experiments the sequences of aminoacids (AA) (the phosphorylated site plus 6 AA before and after) were measured directly [9]. For **Insr**, all measurable tyrosine sites (n=3 for the MS-data) were measured in both experiments. For **Irs1** both the serine 302 site and all measurable tyrosine sites (n=7 for the MS-data) were measured with both experiments. For **Akt** two sites were measured: 1) the threonine 308 site and 2) the serine 473 site. In the WB-experiment, both sites were measured for all three different isoforms (Akt1, Akt2 and Akt3) at the same time. In the MS-experiment it was possible to measure the isoforms separately, however all isoforms could only be measured at the serine 473 site (n=3 in the MS-data), while at the threonine 308 site, only the Akt2 isoform could be measured. For **Rps6kb1** only the WB-experiment was able to measure the threonine 389 site (n=0 for the MS-data). For **Rps6** the serine 235 site and the serine 236 site were measured at the same time in the WB-experiment, and separately in the MS-experiment (n=2 in the MS-data). For **Mapk** two isoforms were measured, Mapk1 at threonine 202 and tyrosine 204, and Mapk3 at threonine 185 and tyrosine 187 (n=4 for the MS-data). For **Tbc1d4** the threonine 642 site was measured in both experiments.

For Akt in Figure 3.1, the SEM at 20 minutes is much larger than the other SEMs. This is due to the fact that Akt has different isoforms (Akt1, Akt2, and Akt3) which are measured separately and then combined into an average, and if any of these have been measured incorrectly, the SEM will be large . For Akt, this appears to be the case for certain time-points and isoforms. The individual measurements of the three isoforms can be seen in Figure 3.2. Note that the isoforms do not have the exact same structure, and that the serine 473 phosphorylation therefore has a slight shift in location. However, the sequences surrounding the phosphorylated AA are identical. Akt1 and Akt3 appear to have had some problems in regard to the measurement in the MS-experiment, however Akt2 does not appear to have the same problem. Comparing Akt2 in the MS-data with the WB-data suggests similar time-dynamics.

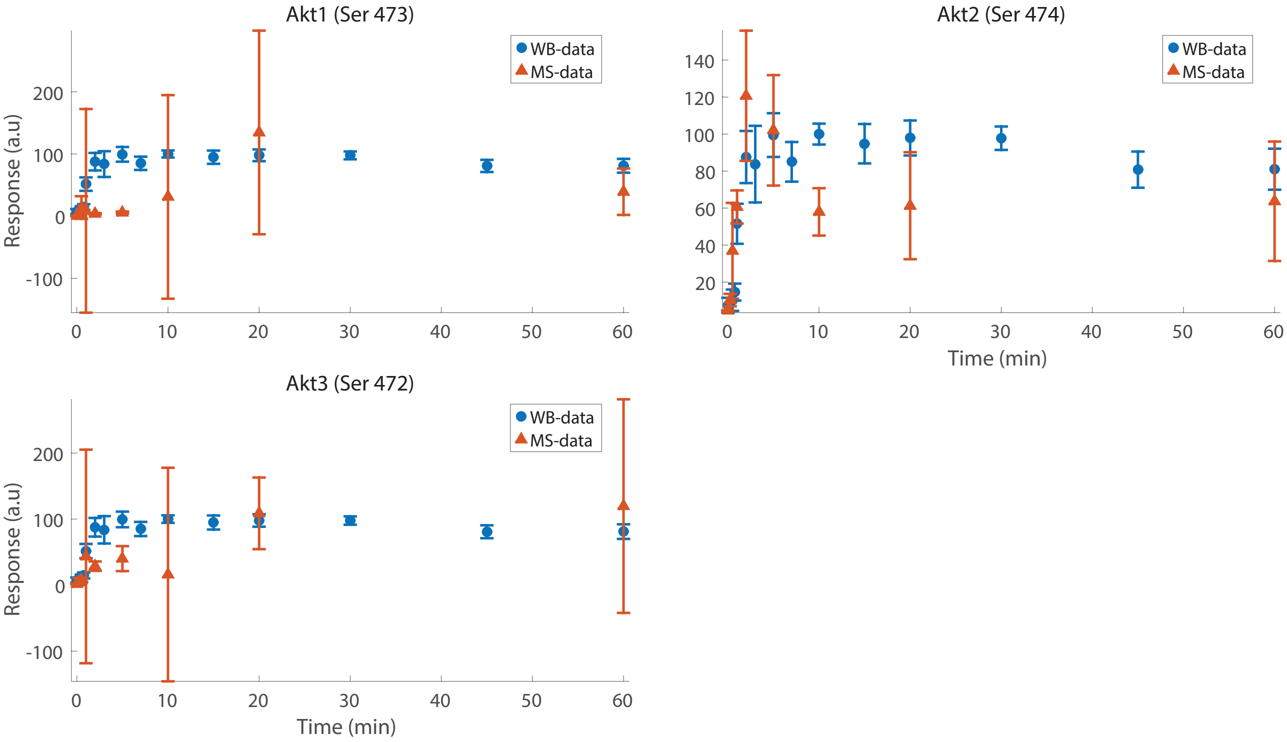


Figure 3.2: **The different isoforms of Akt.** The WB-data (blue) is an average of three isoforms of Akt (Akt1, Akt2 and Akt3), while the MS-data (red) are different for each isoform. Since the different isoforms do not have the exact same AA-sequence, the location of the phosphorylation site is slightly shifted between the isoforms. The first graph shows the comparison between Akt1 in the MS-data, with the WB-data. The second graph shows the comparison between Akt2 in the MS-data and the WB-data. The third graph shows the comparison between Akt3 in the MS-data and the WB-data.

For Rps6kb1, threonine 389 had not been measured in the MS-data. As an alternative, an average of all threonine phosphorylations (four different sites) was used and compared with the WB-data. This comparison can be seen in 3.3. The average of threonine sites in the MS-data appears to be in agreement with the WB-data.

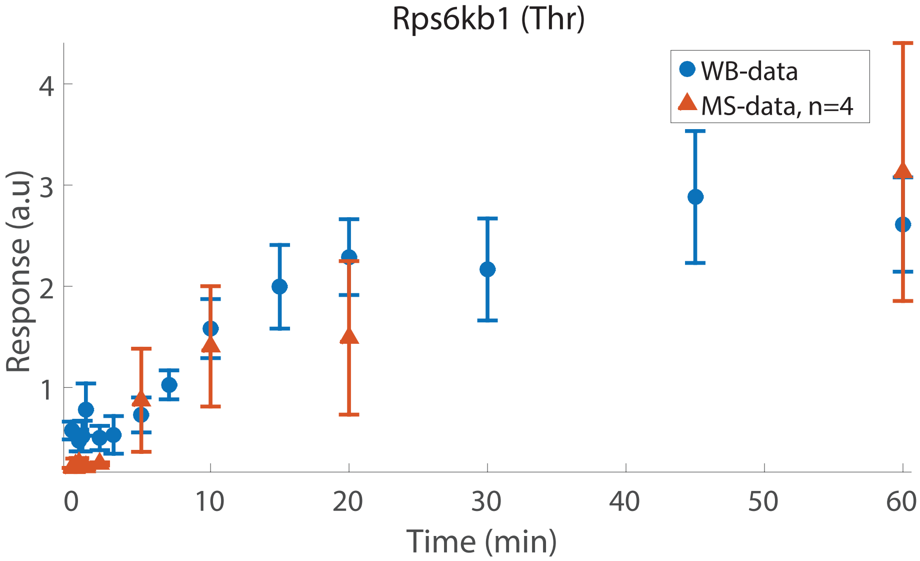


Figure 3.3: **Comparing threonine sites**. The mean of all threonine sites (n=4) in the MS-data (red) was scaled and compared to the threonine 389 site in the WB-Data (blue).

As can be seen in Figure 3.1, the time-dependent response of Rps6 at serine 235/236 in the MS-data has very small SEMS. This is since the different sites respond very similarly. If the different two sites are compared separately, it can be seen that the SEMs should be slightly larger. This comparison can be seen in Figure 3.4. The individual sites do however not respond exactly as the WB-data. Both sites in the MS-data reach maximum response with in 10 minutes from start of insulin stimuli, and then start to drop down again. The sites in the WB-data appears to be slower, and reach their maximum value after 30 minutes and appear to plateau at the maximum response.

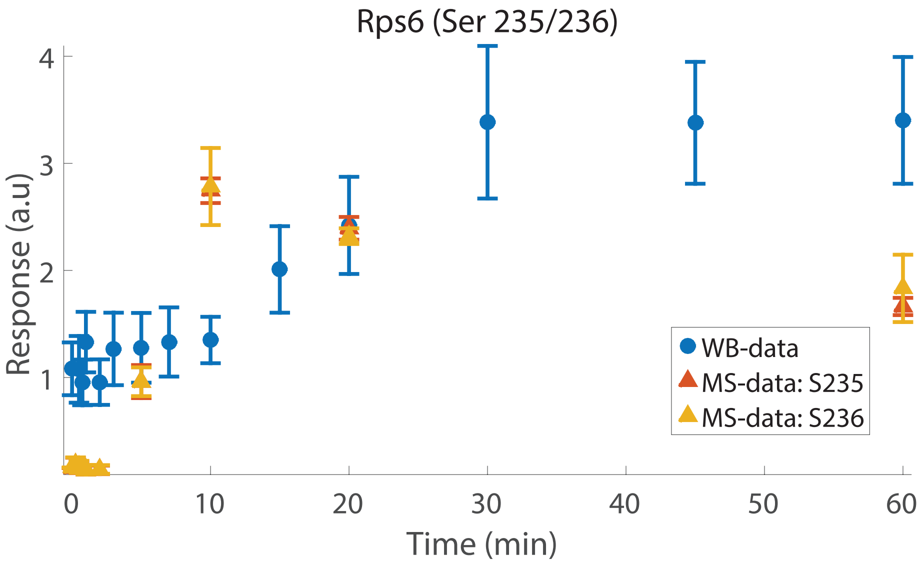


Figure 3.4: **The difference between serine 235/236 in Rps6 in the MS-data**. The WB-data (blue) is an average over the two sites, while the sites are measured independently in the MS-data. The measurements of serine 235 (red) and 236 (yellow) were scaled and compared with the WB-data (blue).

3.2 Collapsing the MS-data

Once the selection of the core-model had been verified, the data was collapsed from measurements of individual sites, into an average over the sites for each individual protein, an example of this can be seen in Figure 3.5. For each protein, the phosphorylation sites were first scaled to be as similar as possible. From these scaled sites, mean and SEM values were calculated. Once the values had been calculated, the mean and SEM where scaled to either begin at 1 or to have a maximum value of 50.

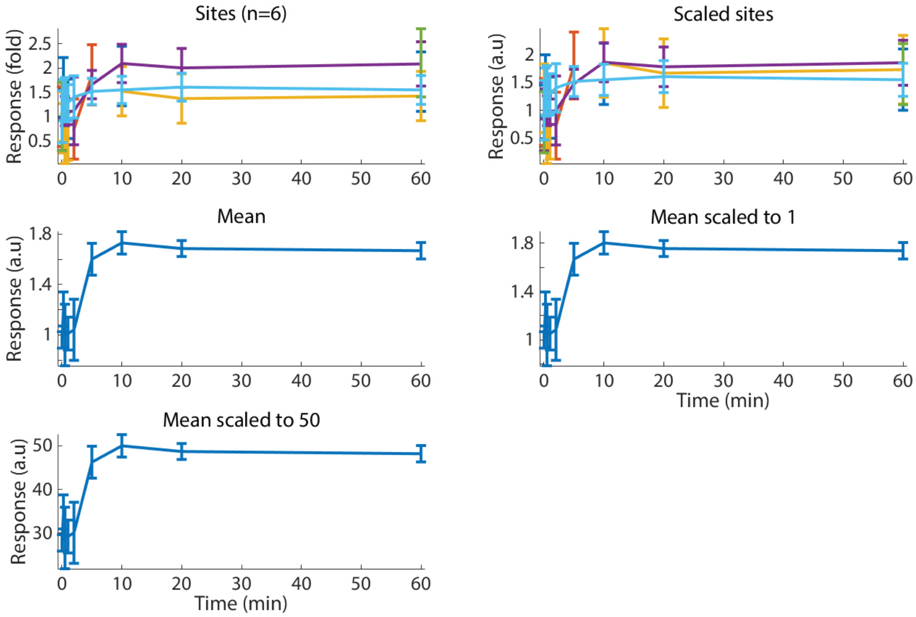


Figure 3.5: **Collapsing of the data.** All sites (n=6) of the protein Ahnak2 were plotted in the same graph. Straight lines have been inserted between each pair of measured values, except if one or more values were missing between two values. The sites were then scaled against one site arbitrarily chosen from the ones with the least amount of missing values (light blue). The mean values and SEM were then calculated for these scaled sites. Then, the mean and SEM values where scaled such that the mean values either begin at one or have the maximum value of 50.

3.3 The first expanded model

When the data had been collapsed into an average of phosphorylation for each protein, all necessary requirements for model expansion were established. By using the core-model, the list of interactions and the collapsed MS-data, the core-model could be expanded into a first system-wide mechanistic model. The first expansion used the MS-data from the insulin responding sites, with no scaling of the mean and SEM. The resulting network can be seen in Figure 3.6. An alternative way of representing the model can be seen in Figure A.1.

In total 162 proteins were added, where 710 unique proteins were present in the data and 522 of these were reachable from the core-model using all possible interactions (if agreement with data were ignored). In total, 31% of all possible proteins were added.

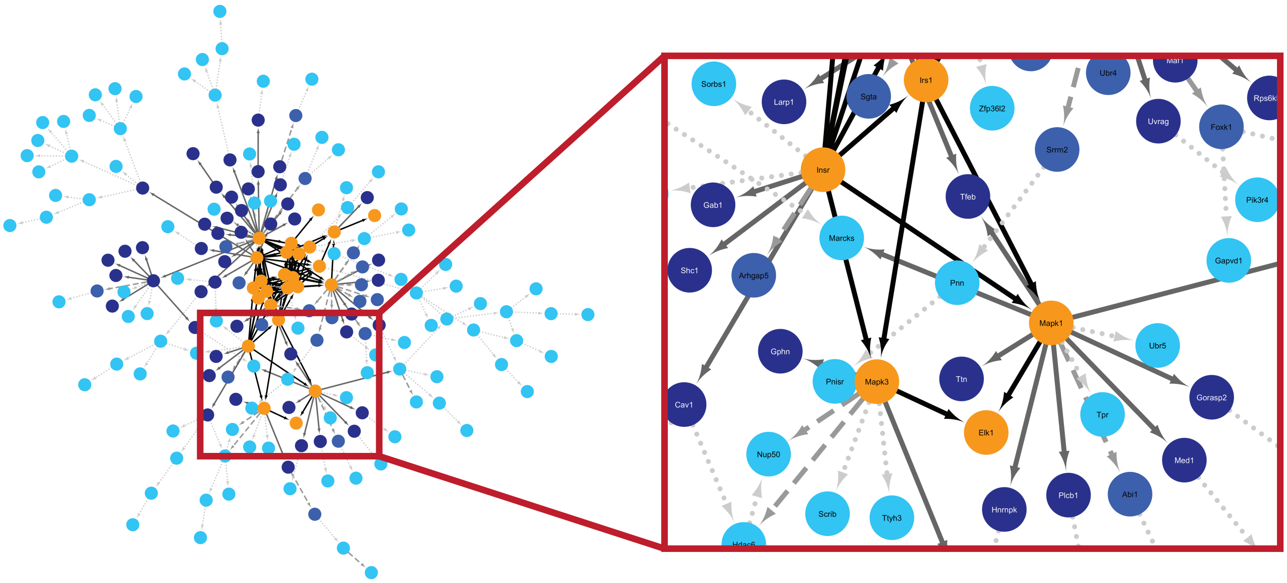


Figure 3.6: **The first expanded model.** The network has been color-coded based on the lowest quality of interactions present in the list when the protein was added. The proteins in core-model were colored yellow, proteins added when only high quality interactions were in the list were colored with a dark blue color, the proteins added when medium were present in the list were colored blue, and the proteins added when low quality were present were colored using a light blue color. Also, the interactions were given different shapes based on their quality. Interactions within the core-model were represented as solid black lines, high quality as solid dark grey lines, medium quality as dashed grey lines and low quality as dotted light grey lines.

3.3.1 Evaluating different scaling

After the first model expansion had been performed, different types of scaling (no scaling, scaling first value to 1, or the maximum value to 50) of the insulin responding data were tested. The resulting networks can be seen in Figure 3.7. As can be seen, the different scalings resulted in different expanded models. The alternative representation of the networks as additions per iteration can be seen in Figure A.2.

Since the different data sets used in the expansion contained the same proteins (only the measured values had been scaled by a constant), the list of interactions were the same and the same core-model was used, the total number of proteins and reachable proteins were the same. In total, 710 unique proteins were present in the data, and 522 were reachable from the core-model when using all possible interactions (ignoring agreement with data). The expansion without scaling added 162 proteins (31%), the expansion with scaling to 1 added

199 (38%) and the expansion with scaling to 50 added 208 (40%). Since expansion when scaling to 50 gave the largest model, this model was selected for further investigation as well as being the selected scaling method for expanding the model with the entire data set.

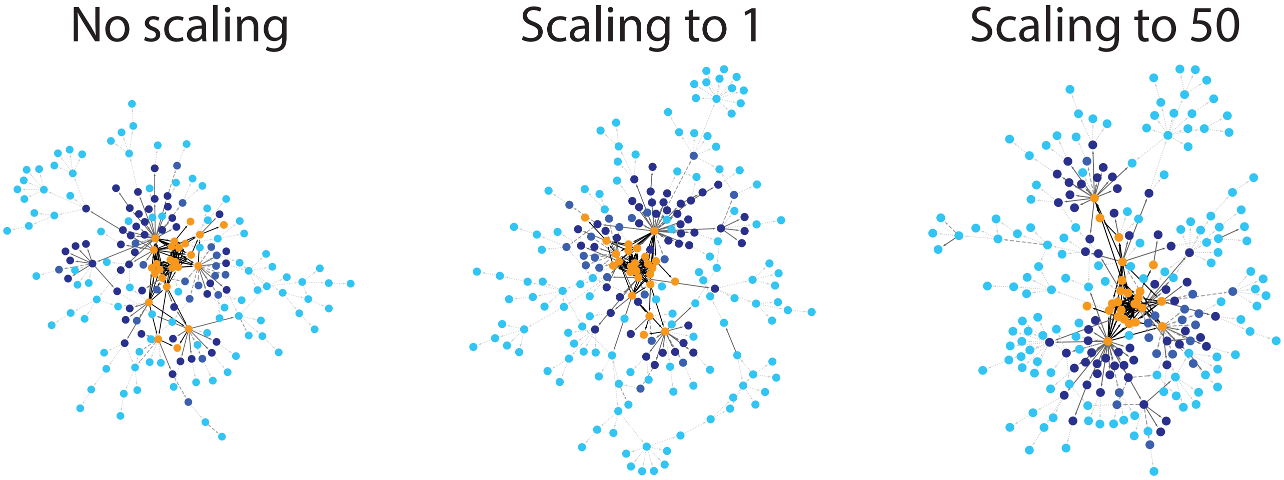


Figure 3.7: **Evaluating different scaling.** The first network is the one that was generated when the core-model was expanded using data which had not been scaled. The second network is the one aquired when the data was scaled in such a way that the first time point had the value of 1. The third network is the one aquired when the data was scaled such that the maximum value was equal to 50. The color coding is the same as in Figure 3.6.

3.3.2 Predicting the effect of inhibitions

To validate the model, two different inhibitions were simulated. The two different inhibitions affect two proteins (Akt and Pi3k) that are at the beginning of the insulin signaling pathway, and the inhibition therefore spread throughout the network. Since the level of inhibition achieved in the experimental measurement was not known, 5 levels of inhibitions were tested (50%, 75% 87%, 93% and 96%). If the model was able to accurately predict how the inhibition would affect the network was measured in two ways. Firstly, if the model was able to accurately predict how the inhibition affects a downstream protein within a range given by the experimental data and secondly, if the model could predict if an inhibition went in the right direction (i.e. if the protein is downregulated or upregulated). Moreover, if the data for the inhibition was inconclusive as to which direction the inhibition was in, the model was assumed to have predicted the direction correctly. The results of the model predictions are summarized in Table 3.1.

Table 3.1: **Predicting different levels of inhibition.** Five different levels of inhibition were tested (given in the heading of the table). Two different types of inhibitions (MK or LY), and two evaluation methods (within SEM or same direction) were chosen. The percentages of proteins predicted correctly are given in the table. Note, MK+LY means if the model is able to predict both inhibitions for the same protein.

	Inhibition (%)				
	50	**75**	**87**	**93**	**96**
MK, SEM (%)	19	20	21	22	21
LY, SEM (%)	16	21	22	24	20
MK+LY, SEM (%)	7	7	8	9	10
MK, direction (%)	68	68	68	69	69
LY, direction (%)	77	79	87	87	87
MK+LY, direction (%)	62	63	65	65	65

The simulated inhibition level of 96% resulted in the best model predictions, where 10% of all proteins were predicted within SEM for both MK and LY inhibition and 65% were predicted in the right direction. The model predictions for 96% are presented in the two correlation plots for MK and LY in Figure 3.8, as well as in a time resolved example (for both MK and LY). All correlation analyses can be seen in Figure A.4 and A.5.

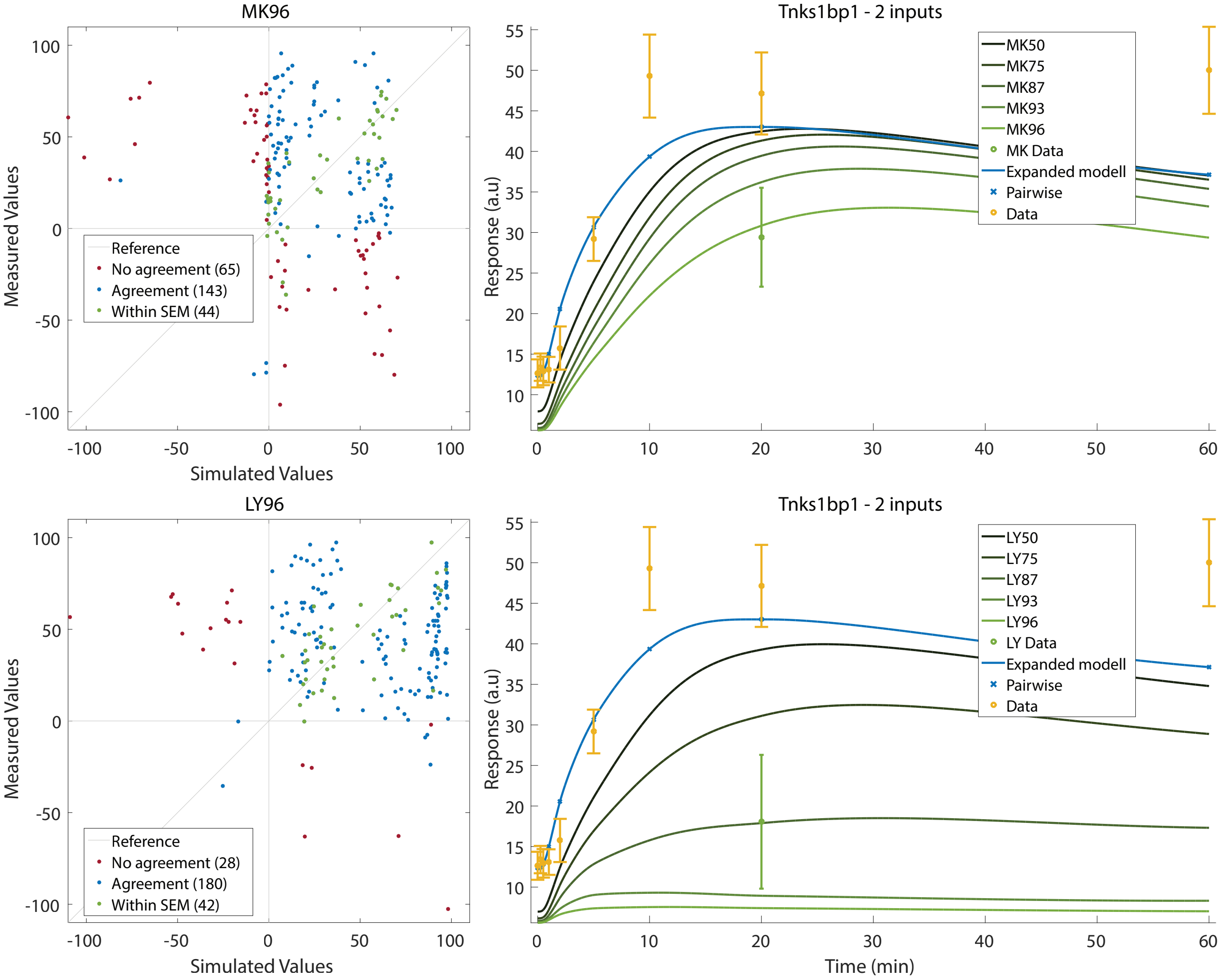


Figure 3.8: **Predicting inhibitions.** The effect of inhibiting with MK or LY with a 96% effectiveness was predicted with the model and compared to experimentally measured values. A correlation plot and a time-resolved example are presented for both MK and LY independently of the other. For the correlation plot, predictions that were within the SEM of the measured values are colored green, predictions that correctly predict the direction are colored blue and incorrect predictions are colored red. In the time-resolved example, all five levels of effectiveness (50%, 75% 87%, 93% and 96%) of the inhibitions are presented as lines in shades of green (lower effectiveness as dark green to higher effectiveness as bright green), as well as the experimentally measured value as green error bars. The uninhibited measurement values are given as yellow error bars and the expanded model prediction of the uninhibited data as a blue line. The values produced when performing pairwise tests are given as 'x', and should be (and were) overlapping with the expanded model simulation.

3.3.3 Giving the model diabetes

Once the inhibitions had been tested, the insulin responding model was used to simulate diabetes. This was done by changing the diabetes mechanism already present in the core-model. Diabetes was then simulated for every protein within the model, of which some examples can be seen in Figure 3.9.

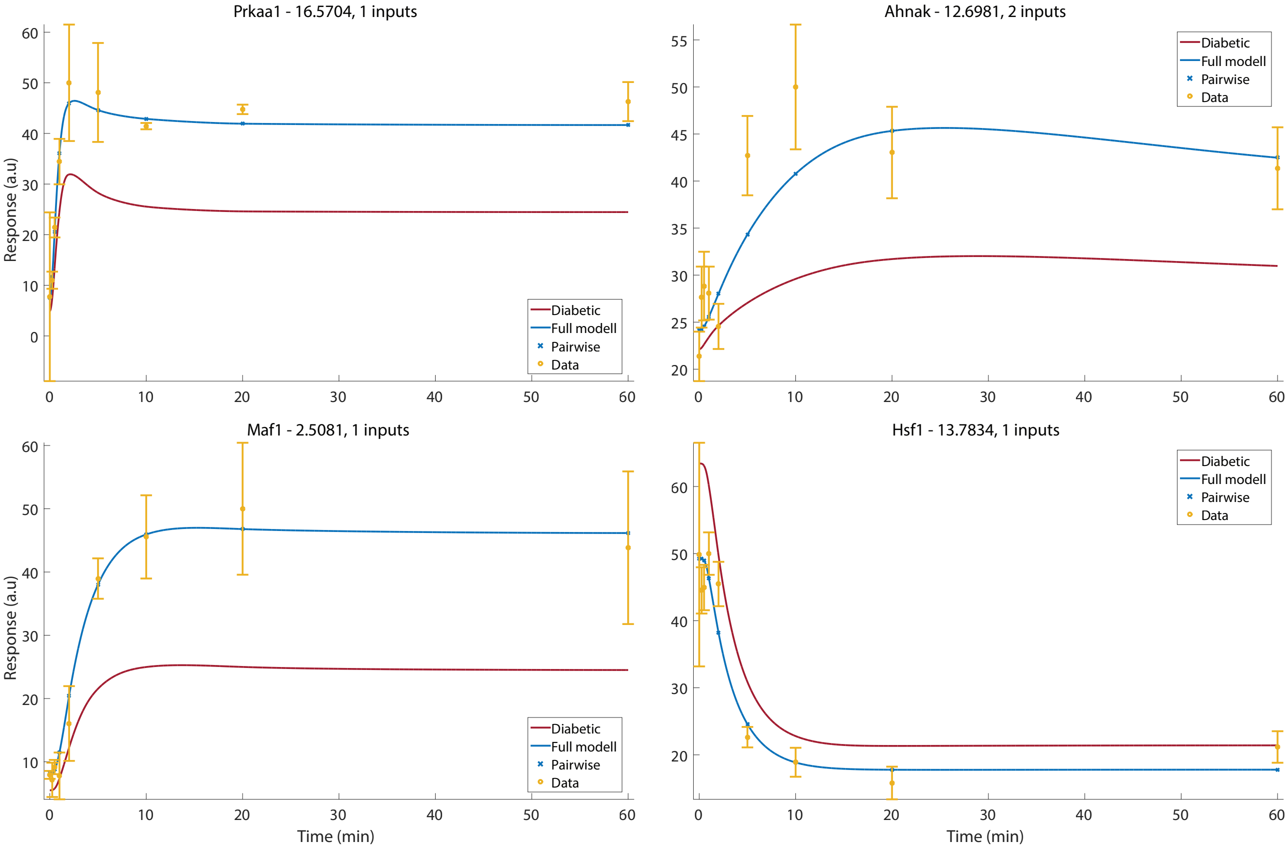


Figure 3.9: **Simulating diabetes.** Some examples from the insulin responding model's simulation of diabetes. In all figures, the experimental data from healthy cells is represented as yellow error bars, the healthy simulations as blue lines and the diabetic simulation as red lines. The values of the simulation when the protein was tested in a pairwise interaction are given as blue 'x' and should (and do) overlap with the healthy simulation.

3.4 Making a large model

Once all key steps had been performed on the smaller, insulin responding data set, the method was applied on the full data set. The resulting model can be seen in Figure 3.10, and the alternative representation in Figure A.3. In this model 2074 proteins where added, out of a total of 3169 of which 2905 were reachable if agreement with data was ignored. The model was able to explain 71% of all reachable proteins.

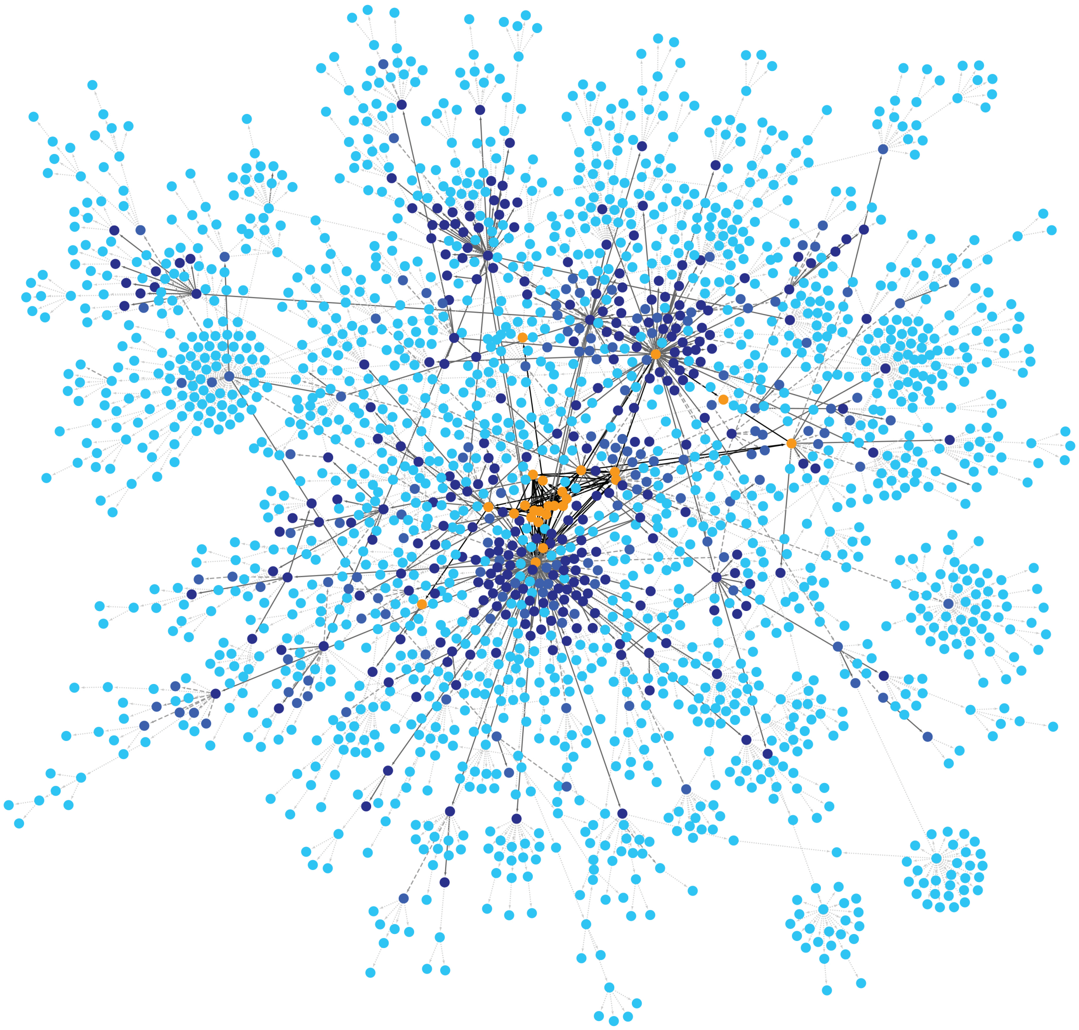


Figure 3.10: **The full model.** A graphical representation of the full model. Color coding is the same as in Figure 3.6.

3.4.1 Predicting inhibitions

As was done with the insulin responding model, the full model was used to predict MK and LY inhibition, with varying amount of levels of achieved inhibition. The results are collected in Table 3.2. The simulated inhibition level of 75% resulted in the best model predictions, where 10% of all proteins were predicted within SEM for both MK and LY inhibition and 65% were predicted in the right direction. The model predictions for 96% is presented in the two correlation plots for MK and LY in Figure 3.11. All correlation analyses can be seen in Figure A.6 and A.7

Table 3.2: **Predicting different levels of inhibition for the full model.** Five different levels of inhibition were tested (given in the heading of the table). Two different types of inhibitions (MK or LY), and two evaluation methods (within SEM or same direction) were chosen. The percentages of proteins predicted correctly are given in the table. Note, MK+LY means if the model is able to predict both inhibitions for the same protein.

	Inhibition (%)				
	50	**75**	**87**	**93**	**96**
MK, SEM (%)	23	23	24	25	24
LY, SEM (%)	27	29	26	25	22
MK+LY, SEM (%)	9	10	9	8	7
MK, direction (%)	76	76	76	76	76
LY, direction (%)	77	78	82	81	81
MK+LY, direction (%)	65	65	66	66	67

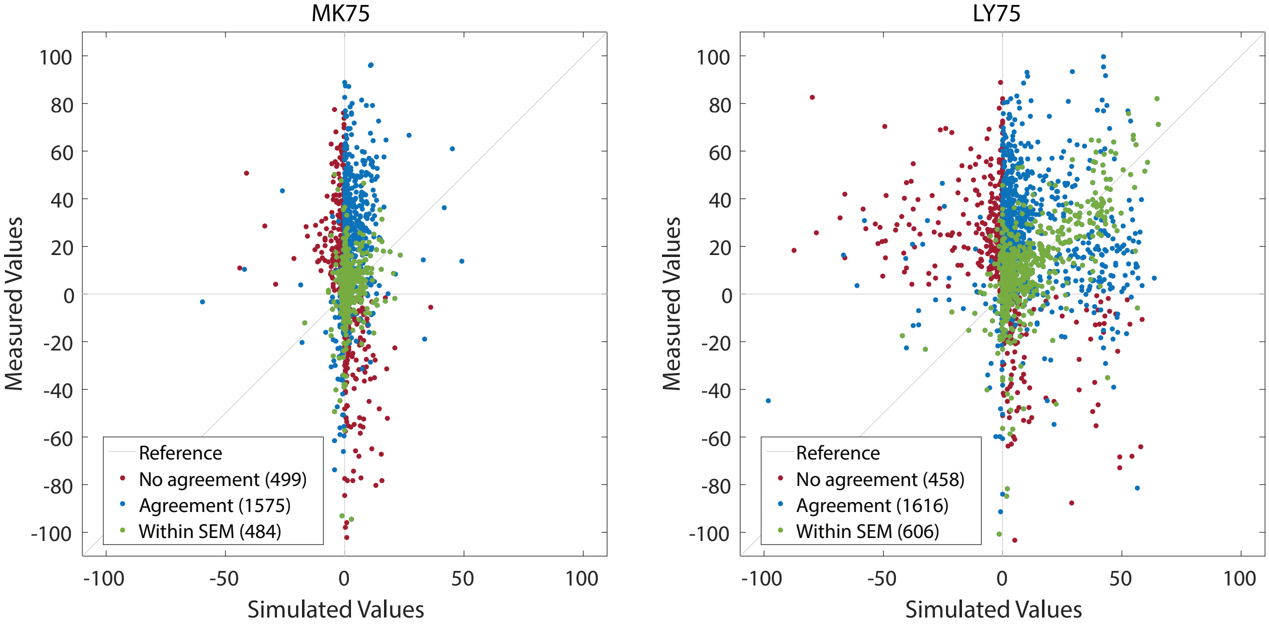


Figure 3.11: **Predicting inhibitions.** The effect of inhibiting with MK or LY with a 75% effectiveness was predicted with the model and compared to experimentally measured values. A correlation plot and a time-resolved example are presented for both MK and LY independently of the other. For the correlation plot, predictions that were within the SEM of the measured values are colored green, predictions that correctly predict the direction are colored blue and incorrect predictions are colored red.

3.4.2 Giving the model diabetes

Once the model had predicted inhibitions, it was used to simulate how diabetes spread throughout the network. Some examples of this can be seen in Figure 3.12

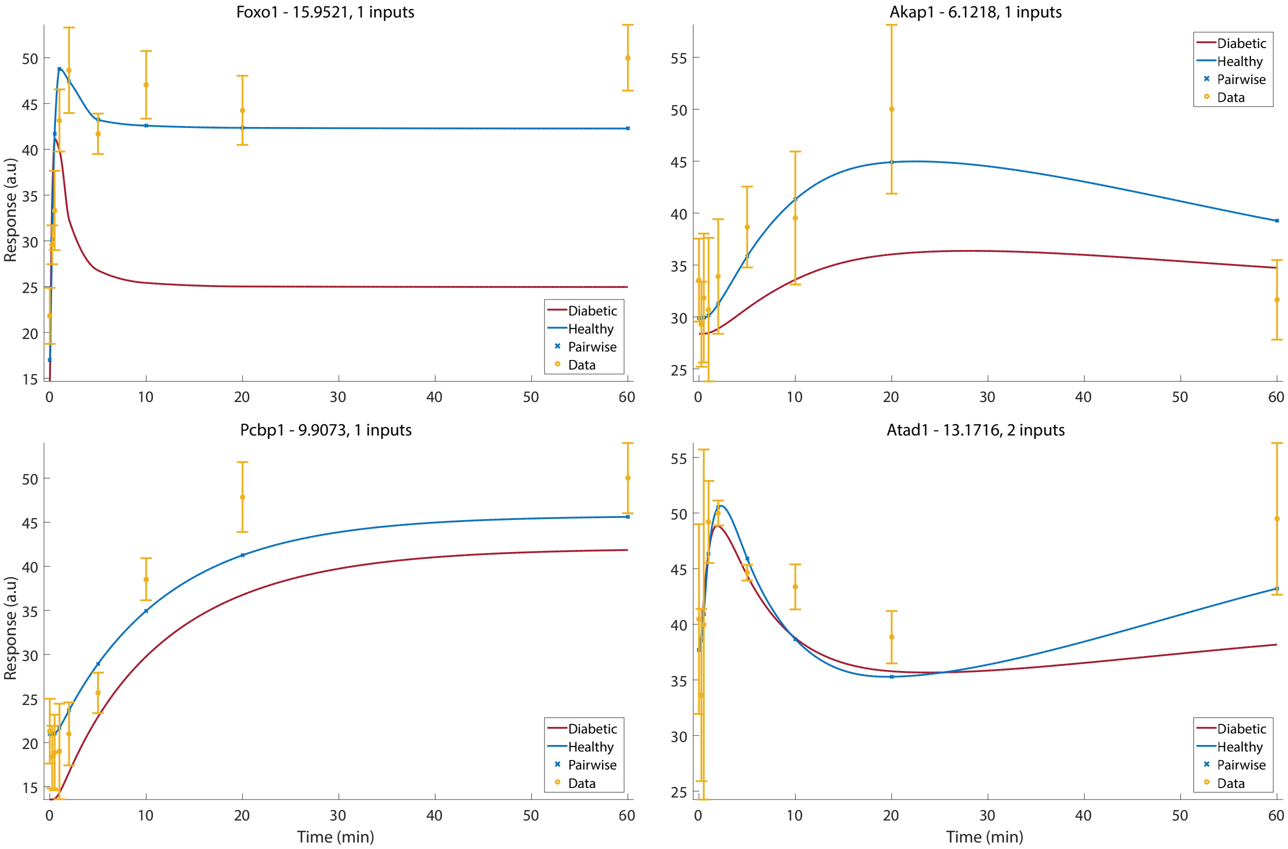


Figure 3.12: **Simulating diabetes.** Some examples from the full model's simulation of diabetes. In all figures, the experimental data from healthy cells is represented as yellow error bars, the healthy simulations as blue lines and the diabetic simulation as red lines. The values of the simulation when the protein was tested in a pairwise interaction are given as blue 'x' and should (and do) overlap with the healthy simulation.

4. Discussion

The aim of this project has been to develop a method to expand small mechanistic models, into large-scale models, while still retaining the mechanistic properties of the model. This aim has been fulfilled. The models constructed are in agreement with the uninhibited data, and the interactions are compatible with prior knowledge. With the work done in this master's thesis, there now exists a model which can simulate 70% of the phosphoproteome, and can be used to make predictions and simulate diabetes. This model, however, is not the only model possible, since some assumptions made throughout the project change how the core-model is expanded.

4.1 Assumptions made and sources of error

The most important assumption made, was that it is reasonable to take the average over different sites of the same protein, and treat it as a single measure, for each protein. It is likely that some sites respond differently in the same protein, despite being given the same input, and this is not captured by the model. If sites respond differently, the result will be a changed mean value and an increase in SEM when collapsing the data, compared to if only sites responding in the same way were collapsed. Another important assumption was that the suggested model is usable as a core-model for the MS-data. This is likely, since most of the time-series behave similarly for both the MS-data and the WB-data, for the compared proteins as can be seen in Figure 3.1 and 3.3. A few proteins, Tbc1d4, Rps6 and Irs1 at threonine 302, did however not behave similarly in the MS-data and WB-data. Despite this, Tbc1d4 and Irs1 at threonine 302 should not be a problem, since Tbc1d4 is never suggested as input in any interaction and Irs1 at threonine 302 is not representative for the general activity of Irs1 and is thus neither used as an input (only the tyrosine phosphorylations of Irs1 are). For Rps6, it is possible that some downstream proteins were not added to the model, since Rps6 might have had the wrong dynamics (assuming the MS-data is right and the WB-data is not), thus possibly resulting in a smaller model.

The list of interactions currently used are not curated to specific tissues, which could be problematic if some interactions are known not to exist in the modeled tissue (in this case, adipose tissue). It has been assumed that the interactions are shared between the different tissues. Under that assumption, less important interactions can simply be ignored by not using them, or by giving them a low rate, in the model if there is not a good agreement between the model simulation and data.

A probable source of error is the fact that proteins are added to the model based on if they can pass a χ^2 test, which will essentially let through some false positives, since we use a 0.95 probability. This means that some interactions that are present in the model, should probably not be there. Another source of error is the fact that the parameters of the model have to be optimized. Since only small models with a few parameters are used when evaluating an interaction, it is likely that the optimization methods used are able to find the best parameter values, but there is no guarantee of this. If an interaction is wrongly rejected due to non-optimal parameters, it is possible that it could have a big impact on the final model structure, since the target protein in that interaction can not be used as an input in any other interaction, thus potentially limiting the size of the final expanded model.

4.2 Use of the model and societal impact

Despite the potential short comings, there are uses for the models created. Since the model has different levels of reliability, depending on which quality of the interactions that have been added, the model can be used for different purposes. For example, if one only uses the most reliable interactions, the model can be used as a visualization tool, as well as for making different predictions. These predictions can be used to design new experiments that are more likely to provide more insights than just using a standard protocol. A typical use could be to know when to make measurements, to capture the important features of the time-response. On the other side of the spectrum, when using the low quality interactions, the model can be used to find new interactions of interest to further investigate, e.g. in research or in the development of new drugs. Since the low quality interactions might not have been studied thoroughly previously, the model can be used as a screening for potential candidates.

It is possible that finding new drug targets can result in reduction of disease symptoms or restore the cells to a healthy state, and therefore lead to reduced suffering on a personal level and a decreased cost for the society as a whole. If the disease can be managed, people will need to be hospitalized less and will be able to work more, which will benefit the economy. Another benefit of having a mathematical model is that one can then perform experiments *in silico*, as opposed to *in vitro* or *in vivo*, potentially reducing the need for animal testing of new drugs.

4.3 Future work

Even though a first model exists that is able to explain most of the data on phosphorylations in insulin signaling, this is not the final model. There are some more steps that can possibly be done in order to expand the model even further. More specifically this can most likely be done by restarting the expansion process with interactions that are not given in the prior knowledge and instead are purely data-driven. This is a reasonable step since it is highly unlikely that all interactions in the entire signaling network have been studied. Therefore by adding interactions based on the shape of the data it is possible to expand the model even further.

4.3.1 Adding data-driven interactions

When all interactions with prior knowledge have been added, purely data-driven interactions could be added. This step would attempt to add interactions that are not in the prior knowledge list, yet can still pass a χ^2-test. These data-driven interactions would be added in a similar way as the first layer proteins. The shape of the data for the proteins not added would be compared with the simulated proteins already in the model. The proteins that appear to be of similar shape would be tested using a pairwise test, and if they pass a χ^2-test, they would be added. These newly added data-driven proteins would then be used as new starting points for addition using the prior knowledge. As before, the high quality interactions would be tested first, and then the lesser qualities will be tested. When all proteins have been added, the model could be fine-tuned by adding plausible extra interactions between proteins in the models.

4.3.2 Fine-tuning the model

When the interactions were added to the model, the new proteins were only allowed to be downstream of proteins that are already present in the model. This could now be adjusted by allowing the newly added proteins to interact with the other proteins already in the model (again assuming the interaction is in the prior knowledge list). Since the current model can (on a quantitative level) already explain all proteins, adding more interactions cannot make the model to no longer be in agreement with data. This is due to that extra interactions can always be set to zero, which will essentially be the same as having the same model as before. Due to this behavior, it is possible to ”tune” the parameters of the model to shift the input from one protein to another. This would allow for end-users of the model to make adjustments to the model so that the model is in agreement with their understanding of the system. The end-user would have a model that 1) is compliant with the prior knowledge of the system, 2) is able to explain the data and 3) is able to generate new insights of the system.

4.3.3 Modeling of individual sites

As mentioned before, it is now assumed that all sites of a single protein behave in the same way when in the same protein. This might not be the case, and it would therefore be better to model the sites individually. At the moment, the list of interactions does not give any information on how phosphorylations on certain sites of the upstream protein affects the activity of the protein. However, such information might be available in the future, and will be beneficial if one wants to model individual sites. An alternative if this information does not arrive is to model each site on its own, and let each site of an upstream protein be used as an input to every downstream protein site. However, the total optimization problem might end up being too big to solve using that approach.

4.3.4 Other systems

This method is not specific for just phosphorylations within insulin signaling in the adipose tissue, and should therefore be possible to use on other types of networks, such as other tissues or other types of post-translational modifications. As long as there exists a core-model, a list of interactions and time-series data, the method can most likely be applied.

4.4 Analysis of work-flow

In the beginning, a brief outline of the project was constructed, where key steps that needed to be performed were noted. However, this proved to not be a realistic estimate of how the work should progress. Analyzing the data, trying out if the pairwise testing and writing the scripts for automatically writing the model files took longer than expected, and therefore it was impossible to test modeling of individual sites and to add data-driven interactions. However, it did not prevent the aim of creating a large-scale mechanistic model, in agreement with prior knowledge and data, from being fulfilled.

If the project would have been done from the start again, then it would probably be a good idea to find a way of modeling the individual sites first, and then try to expand the model. This way, less time would have been spent on collapsing the data into a single measurement for each protein, on analyzing the effect of different scaling and generating multiple models, and more time could have been spent on doing further analysis of the final model.

Another fact that relates to time, is the estimation of the time needed to write the thesis. In the end, it took much longer than expected and things became stressful. A lesson learned is that it would have been much better to write the report in different parts (first introduction, then method, last results and discussion). This way, a better understanding of how much time it takes to write and how many revisions that needs to be done would have been achieved, which would allow for better estimations.

4.5 Conclusions

It is now possible to extend a small-scale mechanistic model, into a large-scale model using only a list of interactions and high-throughput, time-resolved data. For the first time, a truly systems-wide mechanistic model has been achieved. The achieved model is in agreement with data and prior knowledge, and can be used to simulate inhibitions, diabetes and other predictions. The aim of the project was to develop a method of expanding a core-model using large-scale data. This has been achieved.

A. Appendix

A.1 Alternative representations of the models

An alternative way of representing the moodel is by grouping the proteins based on in which iteration they were added. The proteins have been color-coded in the same way as in the main network view (high guality as dark blue, medium as blue and low quality as light blue. This gives an overview of how large each iterations is, and how much larger the model can become by adding interactions of lesser quality.

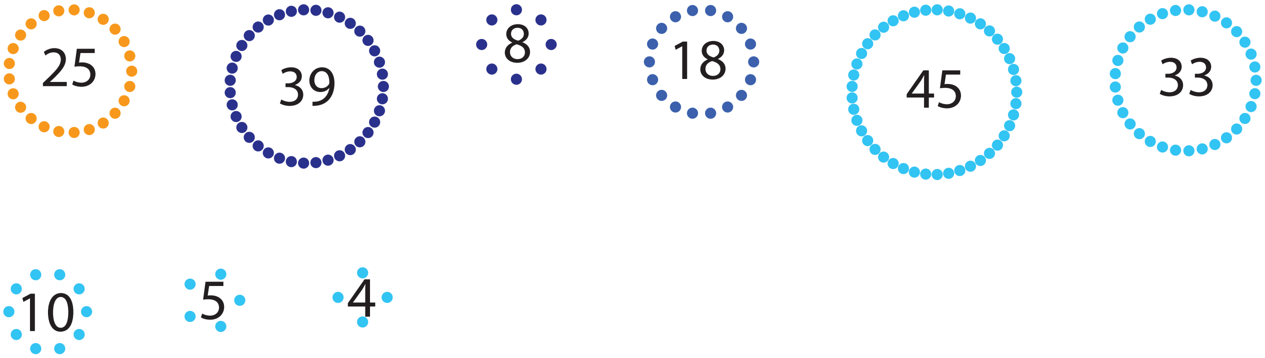


Figure A.1: **Expansion per iteration for the insulin responding, no scale model.**

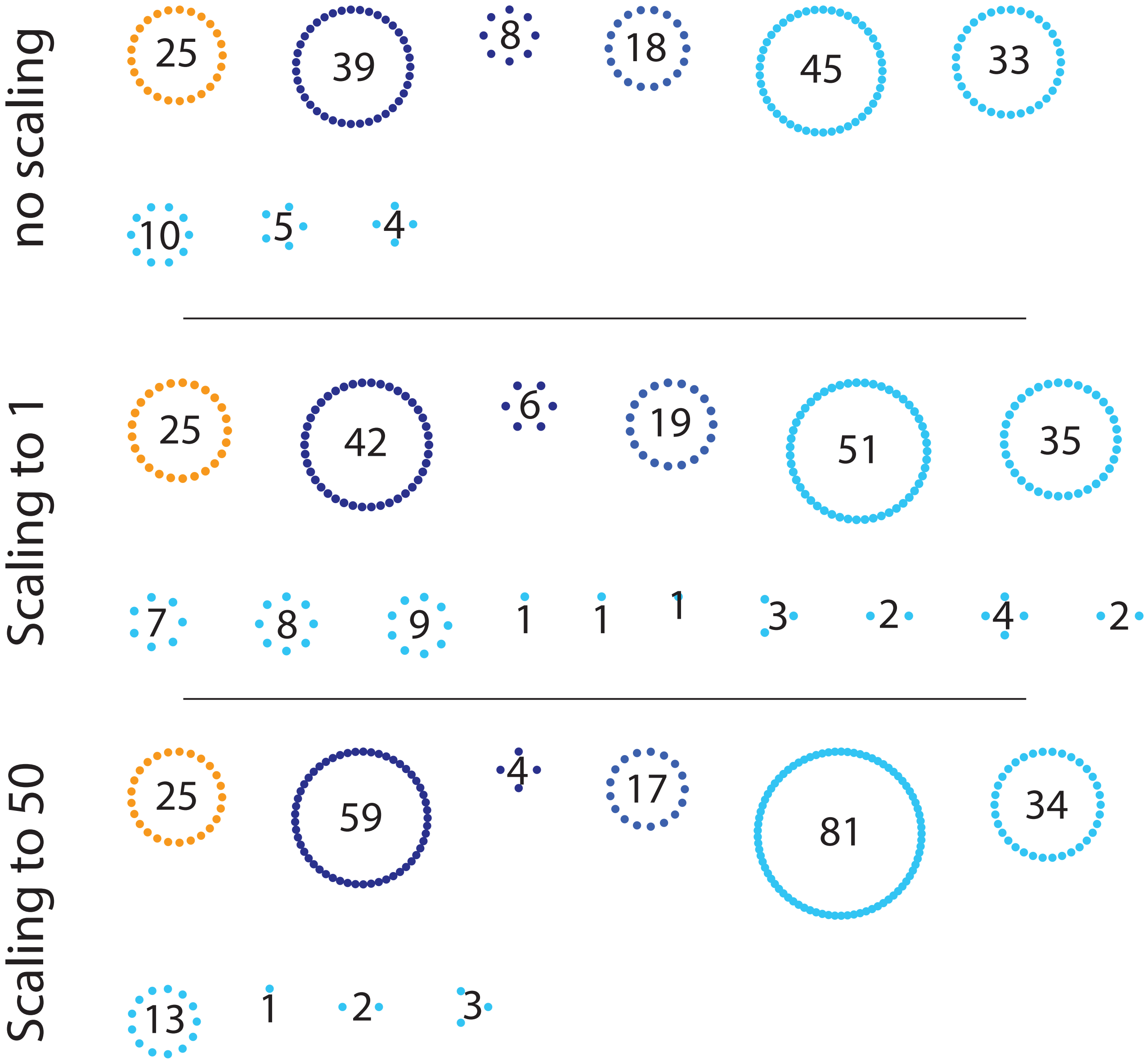


Figure A.2: **Evalutating scaling, visualized as expansion per iterations.** The three acquired models for the different scalings, where the proteins have been grouped based on iteration.

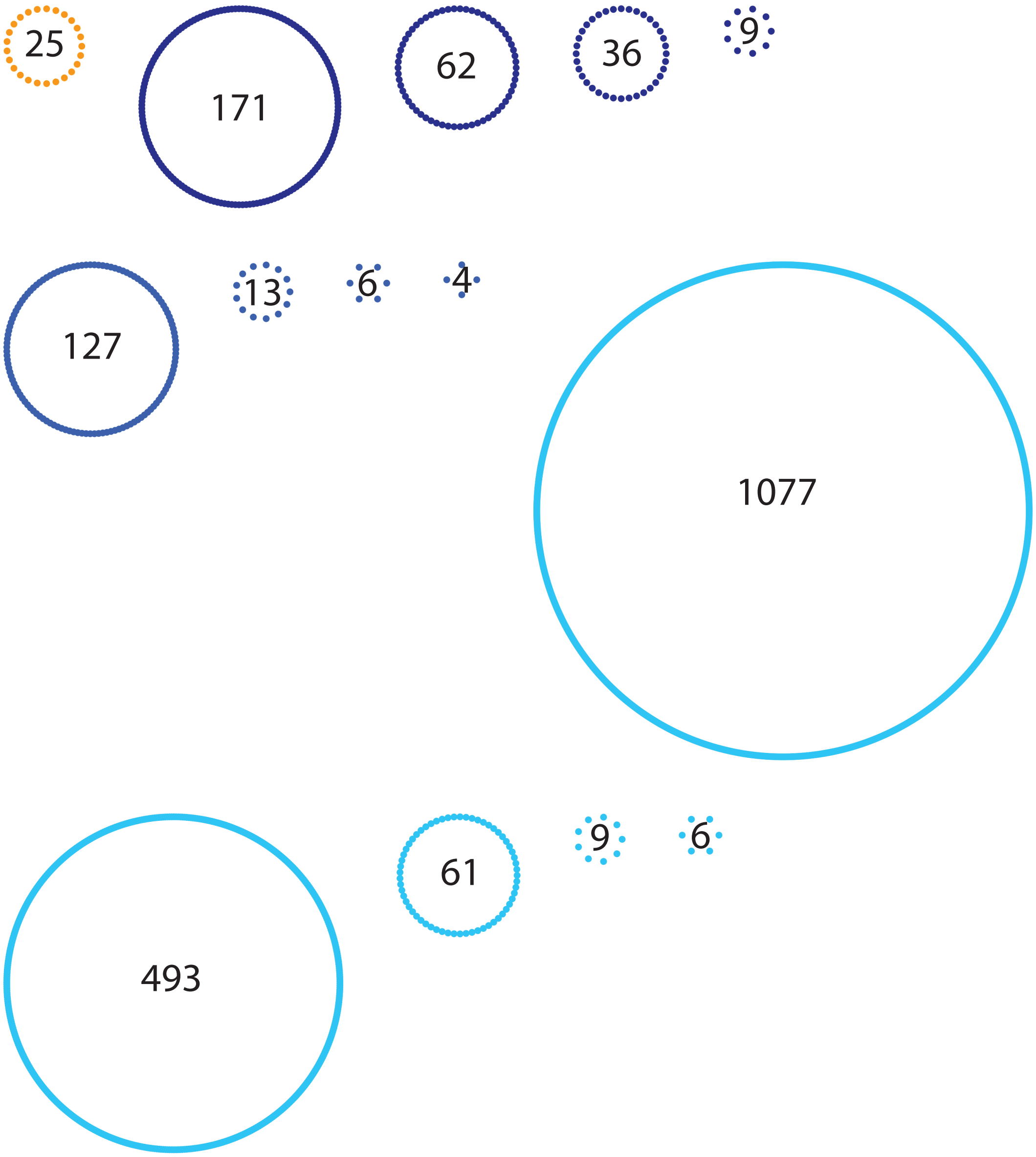


Figure A.3: **Expansion per iteration for the full model, with scaling to 50.**

A.2 Correlation analysis

Five different levels of inhibitions (50%, 75% 87%, 93% and 96%) were predicted and plotted against the corresponding measured values. Predictions that were within the SEM of the measured values are colored green, predictions that correctly predict the direction are colored blue and incorrect predictions are colored red.

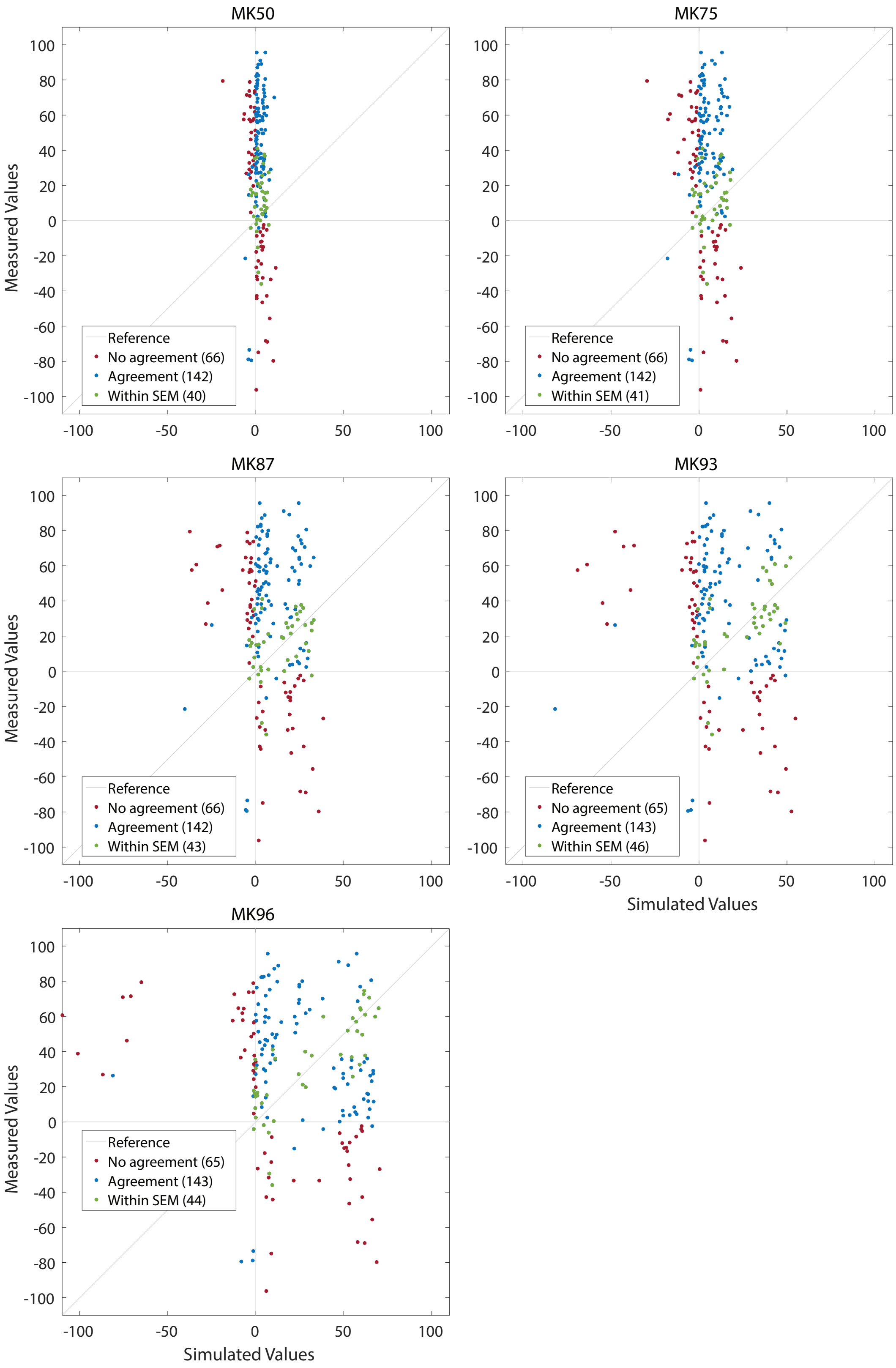


Figure A.4: **Correlation plots for measured vs predicted MK inhibitions for the insulin responding model, with scaling to 50.**

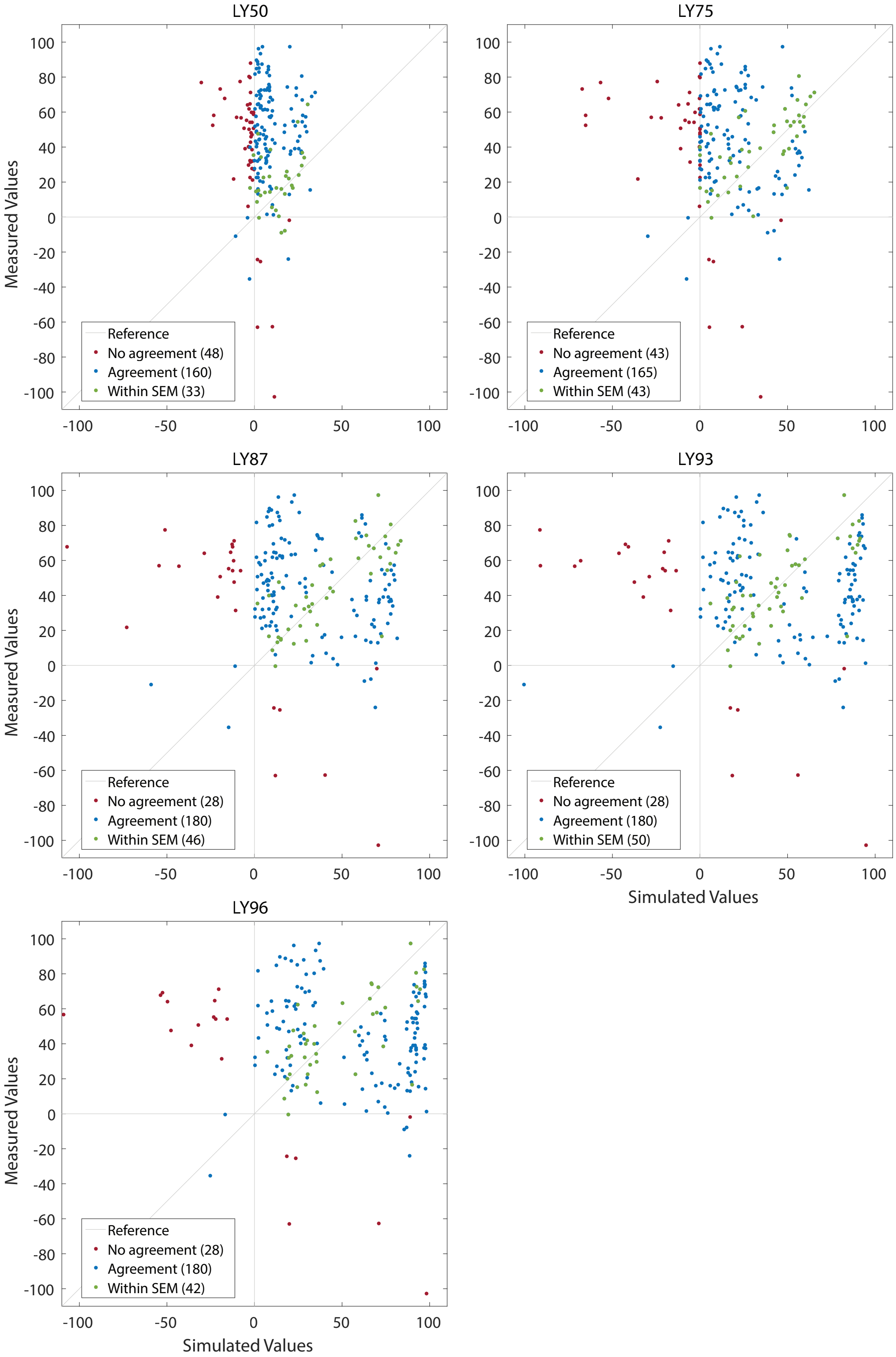


Figure A.5: **Correlation plots for measured vs predicted LY inhibitions for the insulin responding model, with scaling to 50.**

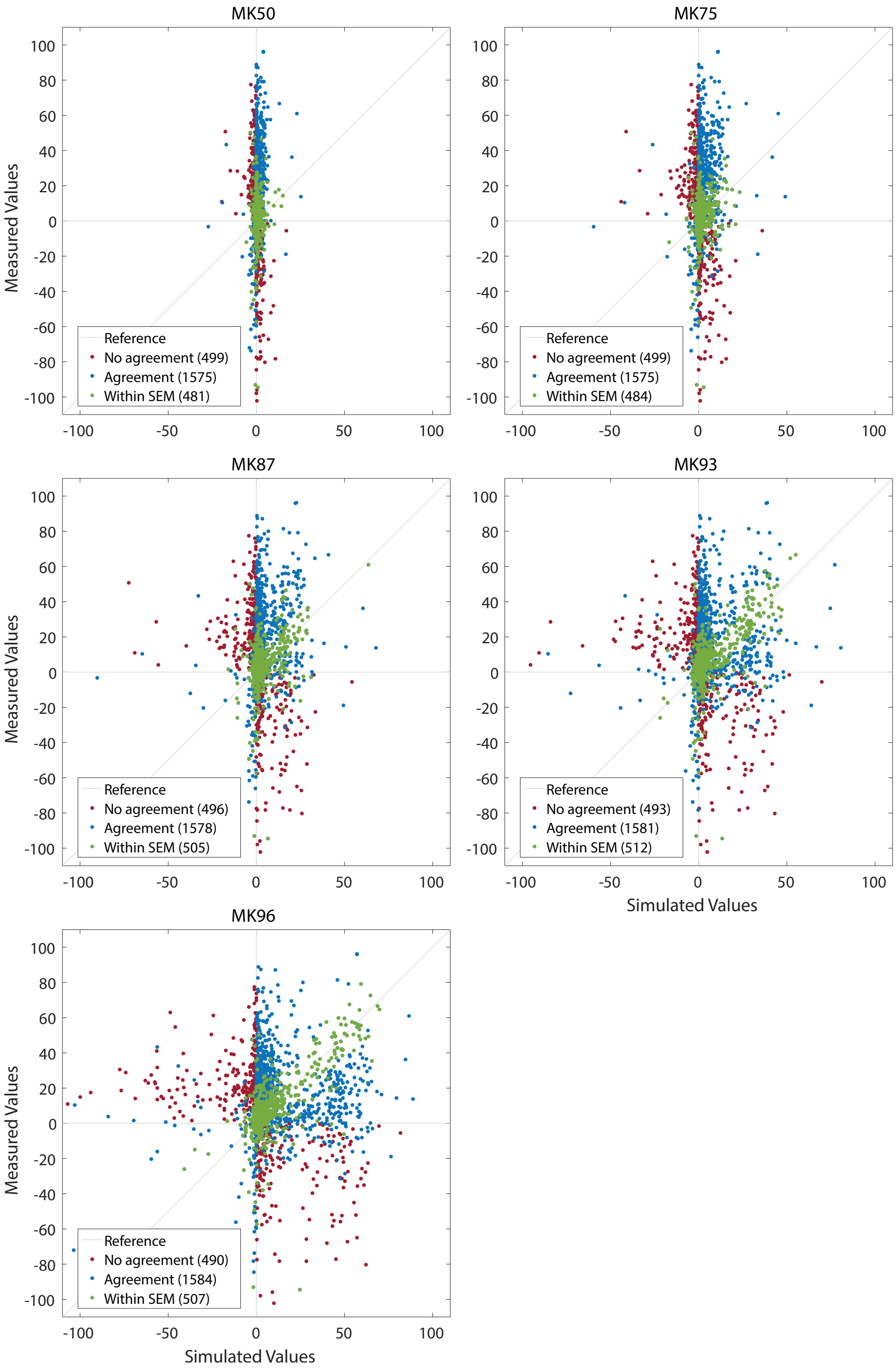


Figure A.6: **Correlation plots for measured vs predicted MK inhibitions for the full model, with scaling to 50.**

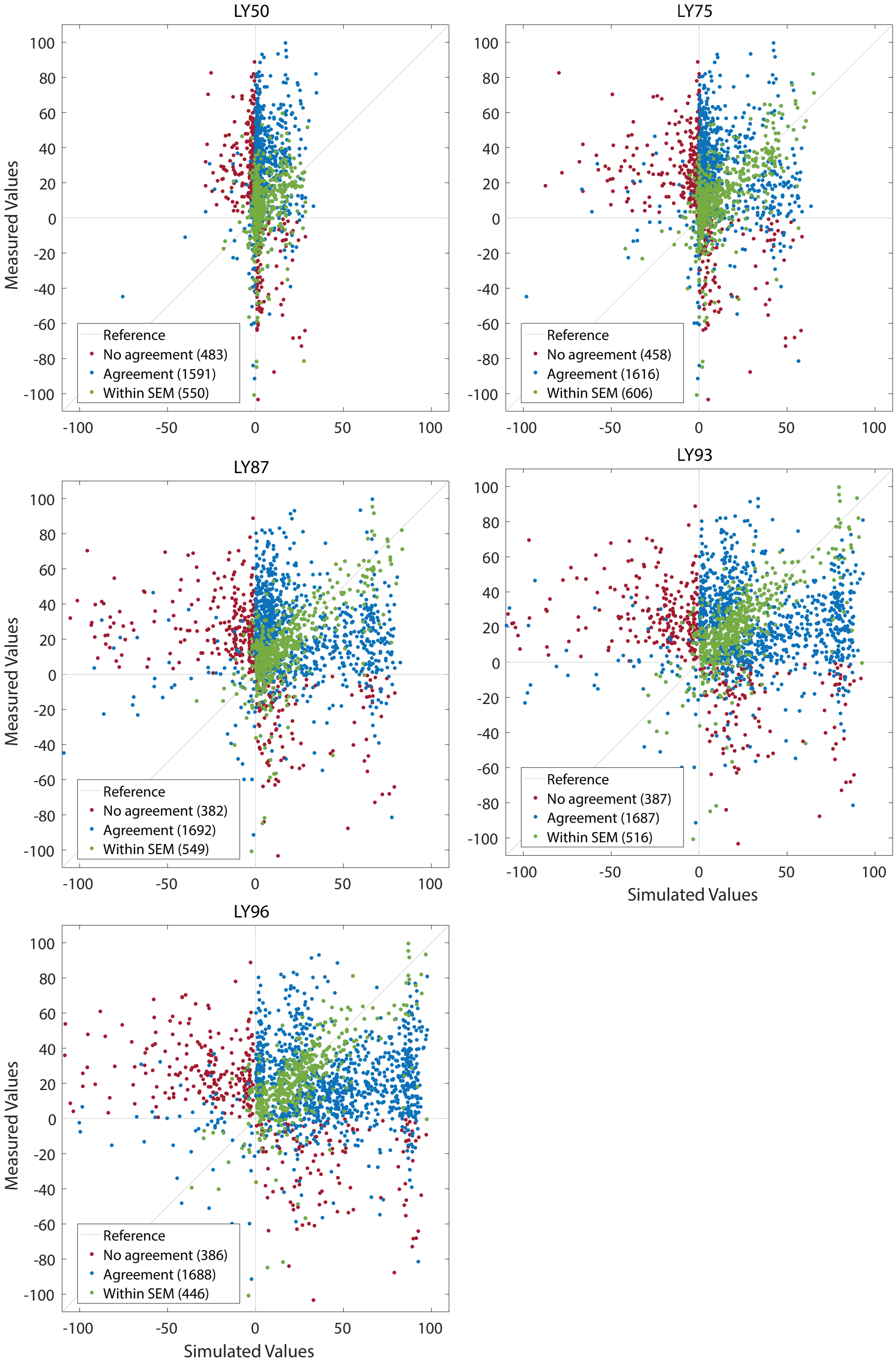


Figure A.7: **Correlation plots for measured vs predicted LY inhibitions for the full model, with scaling to 50.**

Bibliography

1. Guilherme, A., Virbasius, J. V., Puri, V. & Czech, M. P. Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes. *Nature reviews. Molecular cell biology* **9,** 367–77. ISSN: 1471-0080 (2008).
2. Poretsky, L. Principles of diabetes mellitus. *Principles of Diabetes Mellitus,* 1–887. ISSN: 0098-7484 (2010).
3. Polonsky, K. S. *The Past 200 Years in Diabetes* **14,** 1332–1340. ISBN: 1533-4406 (Electronic)\r0028-4793 (Linking). doi:10.1056/NEJMra1110560 (2012).
4. Clark, L. C.; Lyons, C. Electrode Systems for Continuous Monitoring in Cardiovascular Surgery. *Ann. N.Y. Acad. Sci.* **102,** 29–45. ISSN: 00778923 (1962).
5. Wild, S., Roglic, G., Green, A., Sicree, R. & King, H. Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030. *Diabetes Care* **27,** 1047–1053. ISSN: 01495992 (2004).
6. Khoury, G. A., Baliban, R. C. & Floudas, C. A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific reports* **1,** 90. ISSN: 2045-2322 (2011).
7. Nyman, E. *et al.* A single mechanism can explain network-wide insulin resistance in adipocytes from obese patients with type 2 diabetes. *Journal of Biological Chemistry* **289,** 33215–33230. ISSN: 1083351X (2014).
8. Brännmark, C. *et al.* Insulin Signaling in Type 2 Diabetes: EXPERIMENTAL AND MODELING ANALYSES REVEAL MECHANISMS OF INSULIN RESISTANCE IN HUMAN ADIPOCYTES. *Journal of Biological Chemistry* **288,** 9867–9880. ISSN: 0021-9258 (2013).
9. Humphrey, S. J. *et al.* Dynamic adipocyte phosphoproteome reveals that akt directly regulates mTORC2. *Cell Metabolism* **17,** 1009–1020. ISSN: 15504131 (2013).
10. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research* **13,** 2498–2504. ISSN: 10889051 (2003).
11. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research* **43,** D204–12. ISSN: 0305-1048 (2014).

12. Elsevier. *Elsevier Pathway Studio* `https://www.elsevier.com/solutions/pathway-studio-biological-research`. Accessed March 21, 2016.

13. Zhang, M. *et al.* Construction and Deciphering of Human Phosphorylation-Mediated Signaling Transduction Networks. *Journal of Proteome Research* **14,** 2745–2757. ISSN: 15353907 (2015).

14. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Research* **43,** D512–D520. ISSN: 13624962 (2015).

15. Orchard, S. *et al.* The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* **42.** ISSN: 03051048. doi:`10.1093/nar/gkt1115` (2014).

16. Lee, T. Y., Hsu, J. B. K., Chang, W. C. & Huang, H. D. RegPhos: A system to explore the protein kinase-substrate phosphorylation network in humans. *Nucleic Acids Research* **39.** ISSN: 03051048. doi:`10.1093/nar/gkq970` (2011).

17. Dinkel, H. *et al.* Phospho.ELM: A database of phosphorylation sites-update 2011. *Nucleic Acids Research* **39.** ISSN: 03051048. doi:`10.1093/nar/gkq1104` (2011).

18. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research* **39.** ISSN: 03051048. doi:`10.1093/nar/gkq1039` (2011).

19. Tyers, M. *et al.* BioGRID: a general repository for interaction datasets. *Nucl. Acids Res.* **34,** D535–539. ISSN: 1362-4962 (2006).