ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΣΥΜΠΕΡΙΦΟΡΑΣ ΚΥΤΤΑΡΩΝ ΜΕΣΩ ΑΝΑΛΥΣΗΣ ΣΗΜΑΤΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΙΩΑΝΝΗ Ν. ΜΕΛΑ
Διπλωματούχου Μηχανολόγου Μηχανικού Ε.Μ.Π

ΕΠΙΒΛΕΠΩΝ:
ΛΕΩΝΙΔΑΣ ΑΛΕΞΟΠΟΥΛΟΣ
Επικ. Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ, Φεβρουάριος 2014

Η έγκριση της διδακτορικής διατριβής απο την Ανώτατη Σχολή Μηχανολόγων Μηχανικών του Ε.Μ. Πολυτεχνείου δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα (Ν. 5343/1932, Άρθρο 202)

Πρόλογος

Η εν λόγω διδακτορική διατριβή πραγματεύεται την μοντελοποίηση ενδοκυτταρικών σηματοδοτικών μονοπατιών με σκοπό την κατανόηση της λειτουργίας και συμπεριφοράς βιολογικών συστημάτων σε περίπλοκες ασθένειες. Η διδακτορική διατριβή διεξήχθει στο εργαστήριο του Επικ. Καθηγητή Λεωνίδα Αλεξόπουλου στη σχολή Μηχ. Μηχ. Ε.Μ.Π., τομέας Μηχανολογικών Κατασκευών και Αυτομάτου Ελέγχου. Ενώ ο συγγραφέας συνεργάστηκε καθ΄ όλη τη διάρκεια της διατριβής του με ξένα πανεπιστήμια στα οποία φιλοξενήθηκε κατά διαστήματα συμπεριλαμβανομένων: του Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής Cambridge, UK, του Massachusetts Institute of Technology, Department of Mechanical Engineering, Massachusetts, USA, και του Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany.

Ο συγγραφέας θέλει να ευχαριστήσει για την βοήθειά τους, τους ακόλουθους ανθρώπους. Απο το ΕΜΠ, σχολή Μηχανολόγων Μηχανικών, τον Καθ. Χριστόφορο Προβατίδη και τον Καθ. Ιωάννη Αντωνιάδη για τις συμβουλές τους και πολλές χρήσιμες συζητήσεις. Από το πανεπιστήμιο του Aachen τον Καθ. Αλέξανδρο Μητσό για την ενεργή βοήθειά του σε όλη την διάρκεια των διδακτορικών σπουδών του υποψηφίου και την πολύ επικοδομητική μας συνεργασία. Από το European Bioinformatics Institute τον Dr. Julio Saez-Rodriguez για την φιλοξενία του κατά την επίσκεψη του υποψήφιου στο γκρούπ του το καλοκαίρι του 2010 και την ενεργή καθοδήγησή του εκείνο το διάστημα. Από το Massachusetts Institute of Technology, department of Biological Engineering, τον Prof. Douglas A. Lauffenburger για την φιλοξενία του κατά την επίσκεψη του υποψήφιου στο γκρούπ του τα καλοκαίρια του 2011 και 2012 και την ενεργή καθοδήγησή του εκείνο το διάστημα. Από το Max Planck Institute, Department for complex technical systems, τον Dr. Steffen Klamt για την φιλοξενία του κατά την επίσκεψη του υποψήφιου στο γκρούπ του το φθηνόπωρο του 2012 και την ενεργή καθοδήγησή του εκείνο το διάστημα. Τον Καθ. Σωκράτη Τσαγγάρη, τον Καθ. Νικόλαο Χρόνη, τον Καθ. Αριστοτέλη Χατζηιωάννου και τον Καθ. Φραγκίσκο Κολίση, μέλη της επταμελούς επιτροπής του υποψηφίου για την εξέταση και επικοδομητική κριτική της διατριβής του. Ολα τα μέλη της ομάδας συστημικής βιολογίας και εμβιομηχανικής του Καθ. Λεωνίδα Αλεξόπουλου στο ΕΜΠ, σχολή Μηχανολόγων Μηχανικών, από το καλοκαίρι του 2008 έως το καλοκαίρι του 2013 για την πολύ επικοδομητική συνεργασία μας, και με ειδική αναφορά στην Αικατερίνη Δ. Χαιρακάκη, στον Δημήτρη Μεσσήνη, στην Δανάη Κυρλή-Φλώρου, στον Θεόδωρο Σακελλαρόπουλο, στον Ευάγγελο Ζυμπελούδη και στην Ελισάβετ Χατζοπούλου για την στενή μας συνεργασία. Και τέλος τον Καθ. Λεωνίδα Αλεξόπουλο για την καθοδήγησή του επί πέντε χρόνια και που έδωσε το ερέθισμα για την βαθύτερη κατανόηση των βιολογικών συστημάτων που κινητοποίησε τον υποψήφιο σε όλη τη διάρκεια των διδακτορικών του σπουδών.

Περιεχόμενα

Περίληψη

Η εν λόγω διδακτορική διατριβή πραγματεύεται την μοντελοποίηση ενδοκυτταρικών σηματοδοτικών μονοπατιών με σκοπό την κατανόηση της λειτουργίας και συμπεριφοράς βιολογικών συστημάτων σε περίπλοκες ασθένειες.

Τα σηματοδοτικά μονοπάτια απεικονίζουν αλληλεπιδράσεις μεταξύ πρωτεϊνών και περιγράφουν πώς τα κύτταρα αποκρίνονται σε ερεθίσματα του εξωτερικού τους περιβάλλοντος. Τα μονοπάτια αυτά είναι διαθέσιμα στην βιβλιογραφια, σε διαδικτυακές βάσεις δεδομένων. Τα τελευταία χρόνια η διεθνής κοινότητα κάνει προσπάθειες να τα μοντελοποιήσει, υιοθετώντας μεθοδολογίες από την θεωρία συστημάτων, προς την δημιουργία εκτελέσιμων μοντέλων που θα δίνουν την δυνατότητα προσομοίωσης σημαντικών κυτταρικών διεργασιών. Η μοντελοποίηση σηματοδοτικών μονοπατιών αποτελεί κύριο ενδιαφέρον της βιολογίας συστημάτων και αναμένεται να βελτιώσει τις διαδικασίες ανάπτυξης φαρμάκων, όπως αυτές εφαρμόζονται τώρα στην φαρμακευτική βιομηχανία.

Στην παρούσα διδακτορική διατριβή, ο υποψήφιος διδάκτορας εφαρμόζει μεθόδους Ακέραιου Γραμμικού (και μή γραμμικού) Προγραμματιμού (Integer Linear Programming - ILP, Non Linear Programming - NLP) για την μοντελοποίηση ενδοκυτταρικών σηματοδοτικών μονοπατιών και την εκπαίδευση των εν λόγω μοντέλων σε πειραματικά δεδομένα με σκοπό την πιστή απεικόνιση των σηματοδοτικών μηχανισμών στις υπο εξέταση κυτταρικές σειρές. Πιο συγκεκριμένα, κατασκευάζονται μοντέλα για την ασθένεια της οστεοαρθρίτιδας και τον καρκίνο του ήπατος. Τα πειραματικά δεδομένα που χρησιμοποιήθηκαν μετρήθηκαν με την τεχνολογία xMAP της Luminex ενώ οι αλγόριθμοι μοντελοποίησης, που αποτελούν το κύριο μέρος της διδακτορικής διατριβής, σχεδιάστηκαν με άξονα την λογική Bool και επιλύθηκαν με μεθόδους Ακέραιου Γραμμικού (και μή-Γραμμικού) Προγραμματισμού. Αποτελέσματα της έρευνας αυτής δημοσιεύτηκαν σε έγκριτα επιστημονικά περιοδικά και συνέδρια, και οι μέθοδοι μοντελοποίησης που περιγράφονται έλαβαν δύο βραβεία σε διεθνή συνέδρια.

Εκτεταμένη Περίληψη

Εισαγωγή

Η εν λόγω διδακτορική διατριβή πραγματεύεται την μοντελοποίηση ενδοκυτταρικών σηματοδοτικών μονοπατιών με σκοπό την κατανόηση της λειτουργίας και συμπεριφοράς βιολογικών συστημάτων σε περίπλοκες ασθένειες. Η μελέτη των σηματοδοτικών μονοπατιών είναι ένα απο τα πεδία έρευνας της μοριακής βιολογίας που έχει αλλάξει δραστικά τα τελευταία χρόνια με την εμφάνιση της Συστημικής Βιολογίας (ΣΒ). Τα σηματοδοτικά μονοπάτια περιγράφουν πώς το κύτταρο αποκρίνεται σε ερεθίσματα του εξωτερικού του περιβάλλοντος.

Τα κύτταρα υπάρχουν σε ένα πολύ περίπλοκο βιοχημικό περιβάλλον. Αντιλαμβάνονται το περιβάλλον αυτό μέσω εξειδικευμένων μορίων στην κυτταρική τους μεμβράνη, ονόματι υποδοχείς. Οι υποδοχείς δεσμεύουν τα ερεθίσματα του εξωκυτταρικού περιβάλλοντος (κάθε ερέθισμα έχει εξειδικευμένο υποδοχέα στο οποίο μπορεί να προσδεθεί). Κατόπιν πρόσδεσης του ερεθίσματος στον αντίστοιχο υποδοχέα, το ενδοκυτταρικό τμήμα του υποδοχέα αλλάζει την στερεοχημική του διάταξη (ενεργοποιείται) και άλλες ενδοκυτταρικές πρωτεϊνες μπορούν να προσδεθούν πάνω του, οι οποίες με την σειρά τους αλλάζουν την στερεοχημική τους διάταξη (ενεργοποιούνται / φωσφορυλιώνονται). Με αυτόν τον τρόπο διαμορφώνεται η σηματοδοτική διαδικασία και κάθε πρωτεϊνη ενεργοποιεί (φωσφορυλιώνει) την κάτοθί της στο μονοπάτι. Ο τρόπος με τον οποίον η μία πρωτεϊνη αλληλεπιδρά με τις υπόλοιπες είναι ιδιαίτερα περίπλοκος και εξαρτάται απο την στερεοχημική της διάταξη και πολλές άλλες βιοχημικές ιδιότητες των εμπλεκόμενων μορίων. Τέλος, κάποιες από τις πρωτεϊνες στο σηματοδοτικό μονοπάτι έχουν την ιδιότητα να περνάνε στον πυρήνα του κυττάρου και να εκκινούν την έκφραση γονιδίων (μεταγραφικοί παράγοντες), ρυθμίζοντας έτσι την κυτταρική συμπεριφορά. Ενδεικτικές κυτταρικές συμπεριφορές είναι: **(i)** ο κυτταρικός διπλασιασμός, **(ii)** κυτταρικός θάνατος (απόπτωση) **(iii)** έκλυση εξωκυτταρικών πρωτεϊνων (κυτοκινών) οι οποίες θα αποτελέσουν ερεθίσματα για άλλα κύτταρα (εξωκυτταρική σηματοδότηση) και **(iv)** κυτταρική μετανάστευση.

Μοντελοποίηση σηματοδοτικών μονοπατιών

Τα σηματοδοτικά μονοπάτια αποτελούνται από την συννένωση μεμονομένων πρωτεϊνικών αλληλεπιδράσεων οι οποίες αναγνωρίζονται μέσω πρωτεομικών πειραμάτων. Τα τελευταία δέκα χρόνια περισσότερες από 200.000 αλληλεπιδράσεις μεταξύ πρωτεϊνών έχουν αναγνωριστεί και είναι διαθέσιμες σε διαδικτυακές βιβλιοθήκες, ενώ η ΣΒ αποσκοπεί στην μοντελοποίηση των δικτύων αυτών με χρήση μαθηματικών φορμαλισμών για την ποσοτικοποίηση της μετάδοσης σήματος από την μία πρωτεϊνη στην άλλη και την κατασκευή εκτελέσιμων μοντέλων. Προσομοιώνοντας τα μοντέλα αυτά μπορούμε να αποσαφηνίσουμε τους μηχανισμούς κυτταρικής απόκρισης σε ερεθίσματα του εξωκυτταρικού περιβάλλοντος και να εξάγουμε συμπεράσματα για την σημαντικότητα του κάθε κόμβου (πρωτεϊνης) στο δίκτυο. Οι πιο δημοφιλείς φορμαλισμοί περιλαμβάνουν την Boolean λογική, την fuzzy λογική και τις συνήθεις διαφορικές εξισώσεις. Στην Boolean λογική οι πρωτεϊνες μπορούν να πάρουν μόνο δύο τιμές (ON/OFF) και η συνδεσμολογία μεταξύ τους μοντελοποιείται με χρήση λογικών πυλών. Εν συνεχεία επιβάλλονται οριακές συνθήκες στους κόμβους όπου εισάγεται το εξωτερικό ερέθισμα και η ροή του σήματος προσομοιώνεται από την μια πρωτεϊνη στην άλλη χρησιμοποιώντας την Boolean λογική.

Η πιστότητα των υπολογιστικών μοντέλων στα πραγματικά βιολογικά συστήματα εξαρτάται σε μεγάλο βαθμό από την ακρίβεια των δικτύων που χρησιμοποιήθηκαν σαν βάση. Οι αντιδράσεις που περιλαμβάνονται στα δίκτυα αυτά μετρήθηκαν πειραματικά είτε μέσω Yeast-two-hybrid είτε μέσω φασματογραφίας μάζας. Ωστόσο οι μέθοδοι αυτές χρησιμοποιούν purified protein και τα αποτελέσματα τους δεν αναφέρονται σε κάποιο συγκεκριμένο τύπο κυττάρων. Ενώ διαφορετικοί κυτταρικοί τύποι έχουν διαφορετικούς σηματοδοτικούς μηχανισμούς (επομένως και διαφορετικά μονοπάτια) αναλόγως με την λειτουργία που επιτελούν. Επομένως, για την κατασκευή σηματοδοτικών

μονοπατιών εξειδικευμένων για τον υπο εξέταση κυτταρικό τύπο εκτός από τα δίκτυα που υπάρχουν διαθέσιμα σε βιβλιοθήκες, πειραματικά δεδομένα πρέπει επίσης να χρησιμοποιηθούν που θα αιχμαλωτίζουν το βιολογικό υπόβαθρο των εν λόγω κυττάρων. Εν συνεχεία τα πειραματικά δεδομένα θα πρέπει να συνδυαστούν με τα βιβλιογραφικά δίκτυα μέσω αλγορίθμων βελτιστοποίησης που θα εκπαιδεύσουν τα υπολογιστικά μοντέλα στα πειραματικά δεδομένα, καταλήγοντας σε μοντέλα που πολύ στενά αιχμαλωτίζουν τους σηματοδοτικούς μηχανισμούς των υπο εξέταση κυττάρων.

Βελτιστοποίηση σηματοδοτικών μονοπατιών με χρήση μεθόδων Ακέραιου Γραμμικού Προγραμματισμού

Η διδακτορική διατριβή αυτή πραγματεύεται την βελτιστοποίηση σηματοδοτικών μονοπατιών με χρήση μεθόδων regular optimization, δηλαδή Ακέραιου γραμμικού ή μη γραμμικού προγραμματισμού. Κατα την διαδικασία αυτή οι κανόνες της Boolean λογικής κωδικοποιούνται με την μορφή περιορισμών ενώ ελαχιστοποιείται η ασυμφωνία με πειραματικά δεδομένα. Οι μεθοδολογίες που δημοσιεύτηκαν σαν αποτέλεσματα της εν λόγω διατριβής υπολογίζουν global minimum της αντικειμενικής συνάρτησης και απαιτούν τάξεις μεγέθους λιγότερο υπολογιστικό χρόνο απο συγγενικές μεθόδους, καθιστόντας δυνατή την κατασκευή σηματοδοτικών μονοπατιών που αντιπροσωπεύουν πιστά τα πειραματικά δεδομένα και που περιλαμβάνουν εκατοντάδες πρωτεϊνες.

Στα επόμενα κεφάλαια της διατριβής τρείς διαφορετικές μεθοδολογίες παρουσιάζονται με λεπτομέρεια

1. Μέθοδος Ακέραιου Γραμμικού Προγραμματισμού για την μοντελοποίηση σηματοδοτικών μονοπατιών με χρήση Boolean λογικής.

2. Μέθοδος Μη Γραμμικού Προγραμματισμού για την μοντελοποίηση σηματοδοτικών μονοπατιών με χρήση fuzzy λογικής.

3. Μέθοδος Ακέραιου Γραμμικού Προγραμματισμού για την μοντελοποίηση σηματοδοτικών μονοπατιών ως άκυκλους γράφους και την εξάλειψη ασυμφωνιών με πειραματικά δεδομένα μέσω πλήθους διαφορετικών στρατηγικών

Οι μέθοδοι αυτοί χρησιμοποιούνται για την αντιμετώπιση προβλημάτων συνδεδεμένων με τον ηπατικό καρκίνο και την οστεοαρθρίτιδα.

Αποτελέσματα

Βελτιστοποίηση εκτεταμένου δικτύου μεταγωγής σήματος σε φυσιολογικά ηπατοκύτταρα

Στην ενότητα 2.3 κατασκευάζουμε εκτεταμένο δικτύου μεταγωγής σήματος με βάση φωσφοπρωτεομικά δεδομένα. Το εν λόγω δίκτυο περιγράφει τους σηματοδοτικούς μηχανισμούς φυσιολογικών ηπατοκυττάρων και την απόκρισή τους σε 81 ερεθίσματα του εξωκυτταρικού περιβάλλοντος (κυττοκίνες). Η έρευνα αυτή δημοσιεύτηκε από τους Melas et al. [46] και πραγματοποιήθηκε σε συνεργασία με τον Αλέξανδρο Μητσό (την στιγμή της συγκεκριμένης δημοσίευσης επίκουρο καθηγητή στο τμήμα μηχανολόγων μηχανικών του MIT, Cambridge, MA, USA, την παρούση στιγμή καθηγητή στο RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany) και τον Julio Saez-Rodriguez, group leader στο European Bioinformatics Institute (EBI), Cambridge, UK. Αυτή είναι από τις πρώτες απόπειρες για την κατασκευή σηματοδοτικού δικτύου, εξειδικευμένου για ένα τύπο κυττάρων και προσφέρει την βαθύτερη κατανόηση των σηματοδοτικών μηχανισμών των φυσιολογικών ηπατοκυττάρων.

Πιο συγκεκριμένα, η προτεινόμενη προσέγγιση αποτελείται από τα εξής βήματα (δείτε επίσης σχήμα 2.2): (i) Έλεγχος των 81 πιο σημαντικών κυττοκινών για τον υπο εξέταση κυτταρικό

τύπο, (ii) Επιλογή των πιο ενεργών κυττοκινών, (iii) εισαγωγή τους σε συνδυαστικά πειράματα με σκοπό την δημιουργία ενός συνόλου φωσφοπρωτεομικών δεδομένων που αντιπροσωπεύουν τους σηματοδοτικούς μηχανισμούς των φυσιολογικών ηπατοκυττάρων, (iv) βελτιστοποίηση ενός εκτεταμένου βιβλιογραφικού δικτύου στα φωσφοπρωτεομικά δεδομένα με χρήση της μεθόδου Ακέραιου Γραμμικού Προγραμματισμού που περιγράφεται στην ενότητα 2.2.

Αναγνώριση των επιδράσεων αντικαρκινικών φαρμάκων στο δίκτυο σηματοδότησης καρκινικών ηπατοκυττάρων

Στην ενότητα 2.4 χρησιμοποιούμε την μεθοδολογία που περιγράφεται στην ενότητα 2.2 για την αναγνώριση των επιδράσεων 4 αντικαρκινικών φαρμάκων στο δίκτυο σηματοδότησης καρκινικών ηπατοκυττάρων. Η έρευνα αυτή δημοσιεύτηκε από τους Mitsos et al. [23] και πραγματοποιήθηκε σε συνεργασία με τον Αλέξανδρο Μητσό. Η προτεινόμενη μεθοδολογία οδήγησε στην ανακάλυψη παρενεργειών (off target effects) για τα υπο εξέταση φάρμακα που μέχρι εκείνη την στιγμή ήταν άγνωστες, συνδράμοντας έτσι στην κατανόηση του τρόπου δράσης των εν λόγω φαρμάκων και των χαρακτηριστικών που τα καθιστούν αποτελεσματικά ή όχι.

Πιο συγκεκριμένα, η προτεινόμενη προσέγγιση αποτελείται από τα εξής βήματα: (i) Εκτέλεση φωσφοπρωτεομικών πειραμάτων για την ποσοτικοποίηση των σηματοδοτικών μηχανισμών καρκινικών ηπατοκυττάρων. (ii) Κατασκεύη σηματοδοτικού μοντέλου, εξειδικευμένου για τα καρκινικά ηπατοκύτταρα. (iii) Εκτέλεση δεύτερης σειράς πειραμάτων για την μέτρηση των επιδράσεων των υπο εξέταση φαρμάκων στα καρκινικά ηπατοκύτταρα. (iv) Κατασκευή σηματοδοτικών μοντέλων για το εκάστοτε φάρμακο. Οι επιδράσεις του φαρμάκου εκφράζονται σαν αντιδράσεις που έχουν μπλοκαριστεί από την εισαγωγή του. Για την εκτέλεση των φωσφοπρωτεομικών πειραμάτων χρησιμοποιείται η τεχνολογία xMAP της Luminex [12], και μετρώνται τα επίπεδα ενεργοποίησης 13 φωσφοπρτεϊνων σε περισσότερους από 50 συνδυασμούς κυττοκινών και των 4 φαρμάκων. Για την κατασκεύη σηματοδοτικών μονοπατιών χρησιμοποιείται η μέθοδος Ακέραιου Γραμμικού Προγραμματισμού που περιγράφηκε στην ενότητα 2.2, η οποία συνδυάζει τα πρωτεομικά δεδομένα με βιβλιογραφικά σηματοδοτικά δίκτυα και αφαιρεί τις αντιδράσεις εκείνες απο το δίκτυο που αντικρούουν τα δεδομένα. Τα ακόλουθα φάρμακα χρησιμοποιήθηκαν: (1) Lapatinib (αναστολέας του (EGFR/ErbB-2)) [59], (2) Erlotinib (Αναστολέας της κινάσης του EGFR [60]), (3) Gefitinib (Αναστολέας της κινάσης του EGFR [61]), (4) Sorafenib (Αναστολέας του RAF [62]).

Κατασκευή εκτεταμένων σηματοδοτικών μονοπατιών ώστε να περιλαμβάνουν ενδοκυτταρική και εξωκυτταρική σηματοδότηση

Στην ενότητα 2.5 επεκτείνουμε την μεθοδολογία που περιγράφηκε στην ενότητα 2.2 για την κατασκευή εκτεταμένων σηματοδοτικών δικτύων που θα περιλαμβάνουν ενδοκυτταρική και εξωκυτταρική σηματοδότηση. Θα εκκινούν δηλαδή από τα εισαχθέντα ερεθίσματα, θα προχωρούν στο φωσφοπρωτεομικό επίπεδο και την ενεργοποίηση μεταγραφικών παραγόντων και θα τερματίζουν με την έκκριση κυτοκινών στο εξωκυτταρικό περιβάλλον. Η έρευνα αυτή δημοσιεύτηκε από τους Melas et al. [51] και πραγματοποιήθηκε σε συνεργασία με τον Αλέξανδρο Μητσό. Σαν εφαρμογή κατασκευάστηκαν τα εκτεταμένα σηματοδοτικά μονοπάτια σε φυσιολογικά και καρκινικά ηπατοκύτταρα και αναγνωρίστηκαν οι μεταξύ τους διαφορές.

Βάση της προτεινόμενης μεθοδολογίας αποτελεί η μέτρηση φωσφοπρωτεϊνων και εκκρινόμενων κυτοκινών κάτω από τις ίδιες πειραματικές συνθήκες. Εν συνεχεία, χρησιμοποιήθηκε σαν βάση ο αλγόριθμος Ακέραιου Γραμμικού Προγραμματισμού που περιγράφηκε στην ενότητα 2.2, για την περιγραφή της σηματοδοτικής διαδικασίας στο φωσφοπρωτεομικό επίπεδο, ενώ αλγόριθμος γραμμικής παρεκβολής συσχέτισε τις μετρούμενες φωσφοπρωτεΐνες με τις εκκρινόμενες κυτοκίνες. Το εκτεταμένο σηματοδοτικό δίκτυο αντιπροσωπεύει τους σηματοδοτικούς μηχανισμούς της υπο εξέταση κυτταρικής σειράς και στα δύο επίπεδα.

Μέθοδος Μη Γραμμικου Προγραμματισμού για την ποσοτική μοντελοποίηση σηματοδοτικών δικτύων

Στην ενότητα 3 εισάγουμε μέθοδο Μη Γραμμικού Προγραμματισμού για την ποσοτική μοντελοποίηση σηματοδοτικών δικτύων. Η έρευνα αυτή δημοσιεύτηκε από τους Mitsos et al., [86] και πραγματοποιήθηκε σε συνεργασία με τον Αλέξανδρο Μητσό και τον Douglas A. Lauffenburger (head of the Biological Engineering department of MIT, Cambridge, MA, USA). Σε αντίθεση με προηγούμενες μεθόδους η εφαρμογή μεθόδων Μη Γραμμικου Προγραμματισμού επιταχύνει σημαντικά την βελτιστοποίηση σηματοδοτικών δικτύων σε πρωτεομικά δεδομένα επιτρέποντας έτσι την δημιουργία εκτεταμένων μοντέλων που περιγράφουν πιο πιστά τους σηματοδοτικούς μηχανισμούς του υπο εξέταση κυτταρικού τύπου.

Πιο συγκεκριμένα, η προτεινόμενη μέθοδος υιοθετεί την fuzzy [10] λογική για την ποσοτική μοντελοποίηση της μεταγωγής σήματος από την μια πρωτεΐνη στην άλλη μέσα στο δίκτυο, και εν συνεχεία εισάγει μια αντικειμενική συνάρτηση που αντιπροσωπεύει την ασυμφωνία μεταξύ πειραματικών δεδομένων και υπολογιστικού μοντέλου και μη γραμμικούς περιορισμούς με σκοπό την κατασκευή μοντέλου που περιγράφει πιστά τους σηματοδοτικούς μηχανισμούς του υπο εξέταση κυτταρικού τύπου. Κατα την fuzzy λογική χρησιμοποιείται μια συνάρτηση μεταφοράς για την μοντελοποίηση της μεταγωγής σήματος της μορφής

$$f(x) = a(p^n + 1)\frac{x^n}{x^n + p^n} \tag{1}$$

Όπου, με x αντιπροσωπεύεται η ενεργοποίηση του αντιδρώντος, n ο εκθέτης Hill, p είναι μια σταθερά που ορίζει το σημείο καμπής της συνάρτησης, το a είναι συντελεστής στάθμισης και με $f(x)$ αντιπροσωπεύεται η ενεργοποίηση του προϊόντος της εν λόγω αντίδρασης. Στόχος της προτεινόμενης μεθόδου είναι δοθέντων φωσφοπρωτεομικών δεδομένων να υπολογιστούν οι παράμετροι a, n και p ώστε να ελαχιστοποιηθεί η ασυμφωνία μεταξύ υπολογιστικού μοντέλου και πειραματικών δεδομένων.

Μοντελοποίηση σηματοδοτικών δικτύων με την μορφή άκυκλων γράφων και ελαχιστοποίηση των ασυμφωνιών τους με πειραματικά δεδομένα μέσω αλγορίθμου Ακέραιου Γραμμικού Προγραμματισμού

Στην ενότητα 4 εισάγουμε μέθοδο Ακέραιου Γραμμικού Προγραμματισμού για την μοντελοποίηση σηματοδοτικών δικτύων με την μορφή άκυκλων γράφων και την ελαχιστοποίηση των ασυμφωνιών τους με πειραματικά δεδομένα. Η έρευνα αυτή κατατέθηκε προς δημοσίευση από τους Melas et al., τον Ιανουάριο του 2013 στο έγκριτο περιοδικό PLoS Computational Biology, και πραγματοποιήθηκε σε συνεργασία με τον Dr. Steffen Klamt (group leader in the Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany).

Πιο συγκεκριμένα, σε αντίθεση με προηγούμενες μεθόδους όπου υιοθετούν έναν αυστηρό μαθηματικό φορμαλισμό για την μοντελοποίηση της μεταγωγής σήματος από την μια πρωτεΐνη στην άλλη μέσα στο δίκτυο (Boolean, fuzzy λογική, συνήθεις διαφορικές εξισώσεις), η προτεινόμενη μέθοδος μοντελοποιεί το σηματοδοτικό δίκτυο σαν προσημασμένο άκυκλο γράφο και η μόνη παραδοχή που επιβάλει στην σηματοδοτική διαδικασία είναι τα πρόσημα της ενεργοποίησης στους κόμβους να είναι συμβατά με τα πειραματικά δεδομένα. Σε αντίθετη περίπτωση εισάγει στρατηγικές για την αφαίρεση των ασυμφωνιών αυτών, που περιλαμβάνουν την αφαίρεση αντιδράσεων από το δίκτυο, η την διόρθωση της τιμής ενεργοποίησης συγκεκριμένων κόμβων. Σαν εφαρμογή εξετάζουμε το σηματοδοτικό δίκτυο που δημοσιεύτηκε από τους Samaga et al. [32] αντιπαραβάλλοντας το με τα δεδομένα των Alexopoulos et al. [12], και συγκρίνουμε την λύση μας με δημοσιευμένες λύσεις που στηρίζονται στην Boolean λογική. Ο κώδικας που υλοποιεί την συγκεκριμένη μέθοδο είναι διαθέσιμος για το κοινό με την εμπορική ονομασία *SigNetTrainer*

(http://www.mpi-magdeburg.mpg.de/projects/cna/etcdownloads.html). Για μια λεπτομερή περιγραφή του λογισμικού κοιτάξτε το παράρτημα Α.

Αναγνώριση σηματοδοτικών μονοπατιών που σχετίζονται με την κλινική αποτελεσματικότητα φαρμάκων στον ηπατικό καρκίνο χρησιμοποιώντας φωσφοπρωτεομικά, γενομικά και κλινικά δεδομένα

Στην ενότητα 5 προτείνουμε μια καινοτόμα μέθοδο για την αναγνώριση σηματοδοτικών μονοπατιών που σχετίζονται με την κλινική αποτελεσματικότητα φαρμάκων στον ηπατικό καρκίνο χρησιμοποιώντας φωσφοπρωτεομικά, γενομικά και κλινικά δεδομένα. Η έρευνα αυτή κατατέθηκε προς δημοσίευση από τους Melas et al., τον Φεβρουάριο του 2013, και πραγματοποιήθηκε σε συνεργασία με τον Prof. Douglas A. Lauffenburger.

Πιο συγκεκριμένα εξετάζουμε τις επιδράσεις στο φωσφοπρωτεομικό επίπεδο 3 καρκινικών ηπατικών σειρών (HUH7, HEPG2, HEP3B), 8 φαρμάκων για μη εγχειρήσιμο ηπατικό καρκίνο, και στην συνέχεια χρησιμοποιούμε αλγόριθμο μηχανικής μάθησης (SVM) για την εξαγωγή εκείνων των επιδράσεων που είναι ενδεικτικές της κλινικής αποτελεσματικότητας των εν λόγω φαρμάκων. Εξετάζουμε τα παρακάτω φάρμακα:

- Lapatinib, αναστολέας του epidermal growth factor receptor (EGFR), tyrosine kinase 1 and 2 (Her2/Neu). Δεν αποδείχθηκε αποτελεσματικός κατά του μη εγχειρήσιμου ηπατικού καρκίνου σε κλινικές δοκιμές φάσης 2.
- Gefitinib, επίσης αναστολέας του EGFR ινηβιτορ. Δεν αποδείχθηκε αποτελεσματικός κατά του μη εγχειρήσιμου ηπατικού καρκίνου σε κλινικές δοκιμές φάσης 2.
- Sorafenib, μικρός αναστολέας του VEGFR, PDGFR και του Raf. Έχει εγκριθεί για την αγωγή μη εγχειρήσιμου ηπατικού καρκίνου [62].
- Erlotinib, επίσης αναστολέας του EGFR. Έχει εγκριθεί για την αγωγή μη εγχειρήσιμου ηπατικού καρκίνου [135, 136].
- Vandetanib, ανταγωνιστής του VEGFR και EGFR. Δεν αποδείχθηκε αποτελεσματικός κατά του μη εγχειρήσιμου ηπατικού καρκίνου σε κλινικές δοκιμές φάσης 3.
- Sunitinib, αναστολέας του PDGFR και VEGFR. Δεν αποδείχθηκε αποτελεσματικός κατά του μη εγχειρήσιμου ηπατικού καρκίνου σε κλινικές δοκιμές φάσης 3.
- Dasatinib, αναστολέας του BCR/ABL και Src. Βρίσκεται ακόμα σε δοκιμές τύπου 2 για μη εγχειρήσιμο ηπατικό καρκίνο.
- Bortezomib, αναστολέας πρωτεασών. Βρίσκεται ακόμα σε δοκιμές τύπου 2 για μη εγχειρήσιμο ηπατικό καρκίνο [133].

Τα συγκεκριμένα φάρμακα εισάγουμε στις τρεις υπό εξέταση κυτταρικές σειρές σε συνδυασμό με 6 ερεθίσματα (κυτοκίνες): IL1β, TGFα, Heregulin (HER), Insulin (INS), IL6 and TNFα και τρεις εξειδικευμένους αναστολείς: MEKi, PI3Ki, cMETi, ενώ μετράμε την ενεργοποίηση 16 φωσφοπρωτεϊνικών σημάτων. Στην συνέχεια κατασκευάζονται σηματοδοτικά μονοπάτια με χρήση αλγορίθμου Ακέραιου Γραμμικού Προγραμματισμού προς αναγνώριση των αντιδράσεων που μπλοκαρίστηκαν από τα εν λόγω φάρμακα (κατα αντιστοιχία με την μέθοδο που περιγράφηκε στην ενότητα 2.4), και τέλος, εφαρμόζεται αλγόριθμος SVM για την εξαγωγή των επιδράσεων εκείνων που είναι ενδεικτικές της κλινικής αποτελεσματικότητας των 8 αντικαρκινικών φαρμάκων.

Συμπεράσματα

Σε αυτή τη διδακτορική διατριβή ο συγγραφέας εισήγαγε μια νέα κατηγορία μεθόδων για την μοντελοποίηση ενδοκυτταρικών σηματοδοτικών μονοπατιών, με βάση τον Ακέραιο Γραμμικό και μη-Γραμμικό Προγραμματισμό (ILP/NLP). Η χρήση των εν λόγω μεθόδων επιτρέπει την μοντελοποίηση πολύ μεγαλύτερων δικτύων από ότι ήταν δυνατό πρώτα, και την κατασκεύη υπολογιστικών μοντέλων που αναπαριστούν πιστά τους σηματοδοτικούς μηχανισμούς της υπο εξέταση κυτταρικής σειράς.

Πιο συγκεκριμένα, αναπτύχθηκαν 3 διαφορετικοί φορμαλισμοί: **(i)** Ένας φορμαλισμός ILP που μοντελοποιεί τα δίκτυα μεταγωγής σήματος με χρήση της Boolean λογικής και ελαχιστοποιεί την ασυμφωνία με πειραματικά δεδομένα αφαιρώντας ακμές από το δίκτυο, που φαίνεται να μην είναι λειτουργικές στον υπο εξέταση κυτταρικό τύπο. **(ii)** Ένας αλγόριθμος NLP που μοντελοποιεί το σηματοδοτικό δίκτυο με χρήση λογικής fuzzy και προσδιορίζει την βέλτιστη τιμή των παραμέτρων ώστε να ελαχιστοποιηθεί η ασυμφωνία με πειραματικά δεδομένα. **(iii)** Αλγόριθμος ILP που μοντελοποιεί το δίκτυο ως προσημασμένο άκυκλο γράφο και βελτιστοποιεί την τοπολογία του ώστε να ελαχιστοποιηθεί η ασυμφωνία με πειραματικά δεδομένα.

Επιπρόσθετα, εξετάστηκαν 5 διαφορετικές εφαρμογές: **(i)** Αναγνώριση των επιδράσεων αντικαρκινικών φαρμάκων στο δίκτυο σηματοδότησης καρκινικών ηπατοκυττάρων. **(ii)** Βελτιστοποίηση εκτεταμένου δικτύου μεταγωγής σήματος σε φυσιολογικά ηπατοκύτταρα. **(iii)** Κατασκευή εκτεταμένων σηματοδοτικών μονοπατιών ώστε να περιλαμβάνουν ενδοκυτταρική και εξωκυτταρική σηματοδότηση. **(iv)** Μοντελοποίηση σηματοδοτικών δικτύων στα χονδροκύτταρα και αναγνώριση καινούριων κυττοκινών που μπορεί να ευθύνονται για την αποδόμηση του αρθρικού χόνδρου σε παθολογικές περιπτώσεις. **(v)** Αναγνώριση σηματοδοτικών μονοπατιών που σχετίζονται με την κλινική αποτελεσματικότητα φαρμάκων στον ηπατικό καρκίνο χρησιμοποιώντας φωσφοπρωτεομικά, γενομικά και κλινικά δεδομένα.

Abstract

Modeling signal transduction pathways is of the utmost importance in understanding how cells respond to environmental perturbations. Signaling pathways consist of a set of protein-protein interactions, identified via high throughput proteomic experiments and made available through on line pathway databases. Over the past few years a range of approaches have been introduced to model these networks in an attempt to gain insight into the cells function and uncover the etiology underlying complex disease. Aim of this work is the development of a novel class of methodologies that model signal transduction networks as logic models, and using regular optimization formulations (Integer Linear Programming (ILP) and Non Linear Programming (NLP) formulations) cross reference them with high throughput phosphoproteomic data to construct predictive models of the signaling mechanisms of the interrogated cell type.

Chapter 1

Background

This chapter aims to provide a non-technical introduction to the modeling of signal transduction networks based on high throughput proteomic data and prior knowledge of protein connectivity. The significance of signal transduction networks is discussed, especially with respect to understanding the etiology of complex disease such as cancer and osteoarthritis. Moreover, modeling approaches that aim to construct executable models of the cells signaling mechanisms are presented in brief. The construction of proteomic datasets using the Luminex platform is also presented, together with a sample dataset. Finally, a case study of training a prior knowledge network to proteomic data using a Genetic algorithm approach is illustrated. The reader may skip this chapter if he is familiar with these concepts.

1.1 On Systems Biology

Systems Biology (SB) is defined as the application of dynamical systems theory to molecular biology [1]. SB aims at the study of interactions between the components of biological systems, and how these interactions give rise to the function and behavior of that system. SB became a necessity because of the vast amount of data currently being generated (including proteomic, genomic and metabolomic data) and the difficulty in interpreting this data in such a manner that biologically relevant insight can be gained for the interrogated system, that being a specific cell type of interest, tissue, organ and so forth.

1.2 Modeling of signaling pathways

1.2.1 Background

The study of signaling pathways is one of the fields SB has drastically revolutionized. Signaling pathways describe how cells respond to environmental perturbations.

Cells exist in a very complex biochemical microenvironment; they perceive their environment through a family of transmembrane proteins (proteins located on the cell's membrane with a part of them inside the cell and a part outside the cell) called receptors. The part of the receptor placed outside the cell is formed in such a manner that factors of the cell's environment (hereafter called stimuli) can bind to it. Upon binding of the stimulus, the receptor changes its 3-d structure (because of forces of electrostatic nature), i.e. becomes active. Due to changes in its structure intracellular proteins (proteins inside the cell) can bind to it and become activated (i.e. phosphorylated) as well. These proteins phosphorylate other proteins and so forth. Thus, a signaling pathway forms, starting at the receptor level and propagating from one protein to the next reaches the nucleus [2] (see figure (1.1) for an illustration of a simplified signaling pathway).

Based on the phosphorylation patterns of certain protein kinases called Transcription Factors (TFs) the cell expresses different genes and/or modifies its response and phenotype. Cellular responses include **(i)** proliferation (i.e. cell multiplication), **(ii)** apoptosis (i.e. cell death), **(iii)** expression of proteins and **(iv)** release of cytokines, chemokines, hormones and other stimuli in the cell's environment to communicate with other cells and alter their response (i.e. extracellular signaling).

Because of the significance in understanding the cells signaling machinery, identification and modeling of signal transduction pathways is of the main targets of SB [3].

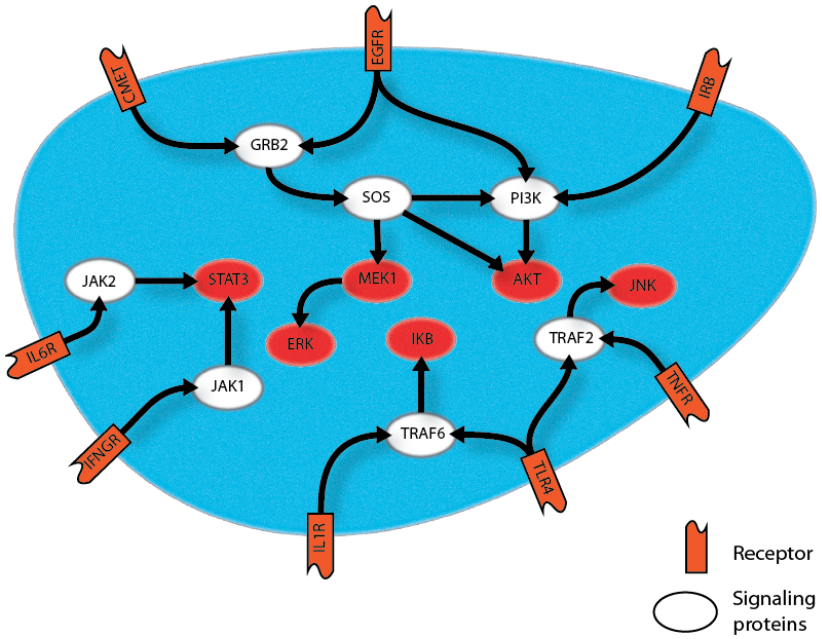


Figure 1.1: Illustration of a simplified signaling pathway

1.2.2 Construction of signaling pathways

Identification of protein interactions (which proteins interact with each other) is carried out with proteomic experiments. Typically, Yeast-two-hybrid (Y2H) and Mass Spectrometry (MS) is used to identify interacting protein pairs [4]. Over the past decade over 100,000 protein-protein interactions have been identified using these two platforms and made available to the public through protein-protein interaction (PPI) databases [5] (see also figure (1.2) for a consensus network of 7 online pathway databases merged together). Understanding the structure of these networks will shed light into the cells' response to factors of its chemical microenvironment, and provide powerful insight into the etiology underlying complex disease [6].

1.2.3 Modeling of signaling pathways

A wide range of methods have been developed to model signal transduction pathways [8]. These employ different mathematical formalisms to describe how signal propagates from one node to the next and construct an executable model of the cells signaling machinery. By simulating the model, conclusions can be drawn for the importance of each node and and its role in the propagation of the signal. Amongst the most widely used formalisms are the various forms of logic modeling [9] and ordinary differential equations (ODEs) [8].

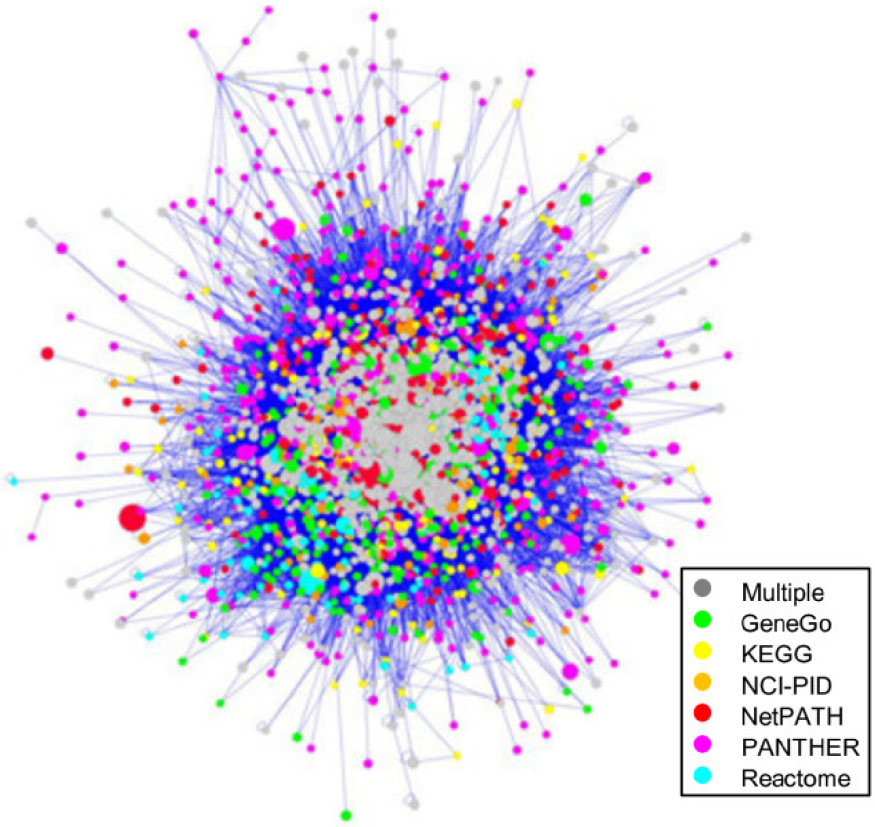


Figure 1.2: A consensus network of 7 online pathway databases merged together, as presented in [7].

In ODEs signaling reactions are modeled as coupled differential equations. ODE networks represent the rates of production and consumption of individual biomolecular species, $d[X_i]/dt$, in terms of mass action kinetics — an empirical law stating that rates of a reaction are proportional to the concentrations of the reacting species. Each biochemical transformation is therefore represented by an elementary reaction with forward and reverse rate constants.

Logic modeling includes two different formalisms, Boolean modeling and constrained fuzzy logic. In Boolean modeling, signal transduction is modeled using the rules of Boolean logic. Protein nodes assume only binary values $\{0, 1\}$, denoting the activation (or not) of the corresponding signaling molecule, and signal is propagated from the receptor level to downstream nodes using a combination of OR and AND gates. In constrained fuzzy logic proteins assume real values and a transfer function is introduced to propagate the signal from one protein to the next [10]. A set of parameters in the transfer function defines its behavior (see figure (1.3)).

1.2.4 Optimization of signaling pathways to high throughput phosphoproteomic experiments

motivation

ODEs and logic modeling succeed in the construction of executable models of the cells signaling mechanisms, however, their predictive power depends on the accuracy of the signaling pathway's connectivity. Protein interactions included in the signal transduction networks are typically

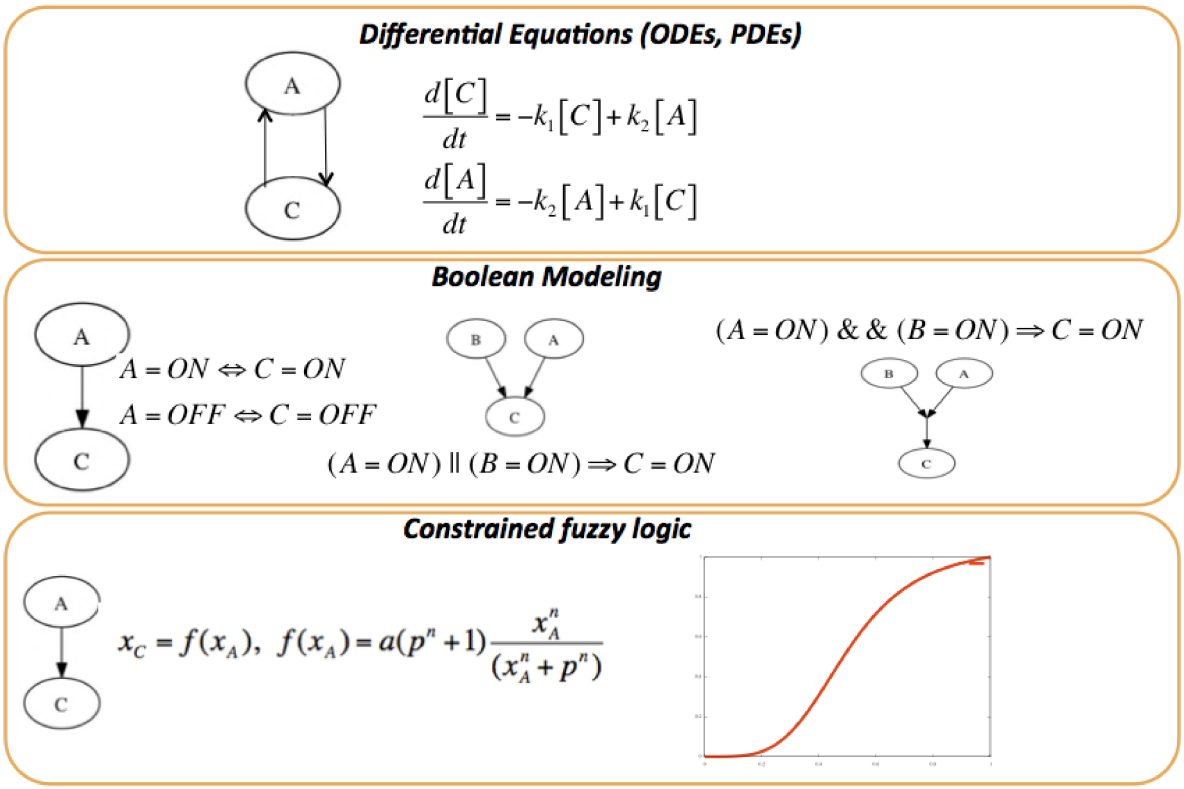


Figure 1.3: Overview of the most widely used modeling approaches for signal transduction networks.

identified experimentally via Yeast-two-hybrid and Mass Spectrometry. These platforms identify interacting pairs of proteins; whether two proteins interact, depends on their amino acid sequence and their folding in 3-d space (Tertiary structure) [2]. Thus two proteins will interact if their tertiary structure enables them to, irrelevant of the interrogated biological system. However, not all proteins are present in all cell types. Different cell types may express different proteins based on their wider role in the tissue. Consequently, different cell types may demonstrate different signaling topologies.

To construct predictive models of the cell's signaling mechanisms, apart from the signaling topology (as obtained from on line pathway databases) experimental data must be measured, to take into account the biological context of the interrogated cell type. Subsequently, experimental data must be combined with the prior knowledge of protein connectivity and the mathematical formulation that models signal transduction, and within an optimization framework, construct a model that closely replicates the cells signaling mechanisms.

Construction of high throughput phosphoproteomic datasets

Luminex xMAP technology is of the most novel, versatile platforms used to construct high throughput phosphoproteomic datasets, with the scope of constraining logic models of signal transduction pathways [11]. The experimental procedure is the following:

Cells are plated on 96-well plates, in every well a different combination of stimuli is introduced to perturb the cells and key phosphoprotein signals are measured [12] (see figure (1.4) for an illustration of the experimental setup). Not all of the stimuli introduced in these experiments

activate signaling pathways, there is a class of stimuli called inhibitors that are small molecules that diffuse through the cells membrane and block signal transduction in certain proteins. Thus, MEKi (MEK inhibitor) blocks signal transduction downstream the MEK protein. Sample data are plotted in figure (1.5) [12]. In the dataset presented in figure (1.5) the 17 key phosphorylation signals are measured in 3 separate time points, the unstimulated state (time=0), 30 minutes after stimulation (early time point) and 3 hours after stimulation (late time point).

The high throughput phosphoproteomic data presented in figure (1.5) essentially capture the response of the interrogated cell type to the 8 stimuli and 8 inhibitors it was perturbed with. Cross referencing this data with the signaling pathway topologies obtained from on-line databases, typically using an optimization based approach, will provide us with a model that closely replicates the cells signaling mechanisms [13].

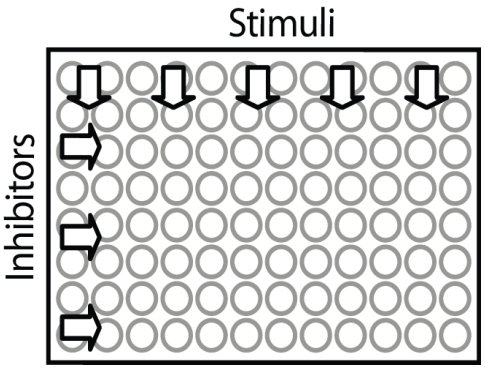


Figure 1.4: Illustration of the experimental setup.

Optimization of signaling pathways to phosphoproteomic data – a Genetic Algorithm (GA) approach

In the work published by Saez-Rodriguez et al. in [13], a GA was used to optimize a signal transduction network to high throughput phosphoproteomic data by Alexopoulos et al. [12] (the sample data featured in the previous paragraph), in an attempt to construct a predictive model of the signaling mechanisms of cancer liver cells.

Hepatocellular carcinoma (HCC) (i.e. liver cancer) is correlated with alterations in the signaling pathways of the liver [14]. Uncovering these alterations will lead to a deeper understanding of the disease and provide potential targets for novel therapeutic interventions.

In [13], a signal transduction network was put together based on literature citations of protein interactions and Boolean logic was used to model how signal propagates from one node to the next. As discussed in a previous paragraph, the signaling topology, being generic in nature and lacking biological context for the specific cell type, would not be able to capture the signaling response of liver cells. Thus, the phosphoproteomic dataset published in [12] (briefly discussed in the previous paragraph) was used in combination to the Boolean model, within an optimization framework, to prune the signaling topology by removing reactions that appear not to be functional in the interrogated cell type. The experimental data consist of multi-combinatorial treatments of 8 stimuli (activate certain signaling pathways) and 8 inhibitors (block signaling pathways in certain proteins) on cancer liver cells, while measuring the activation level of 17 key phosphoprotein signals at 3 time points (unstimulated state, time=30mins, time=3hours). The data is shown in figure (1.5).

Before applying the optimization procedure, simulation of the Boolean model under the same

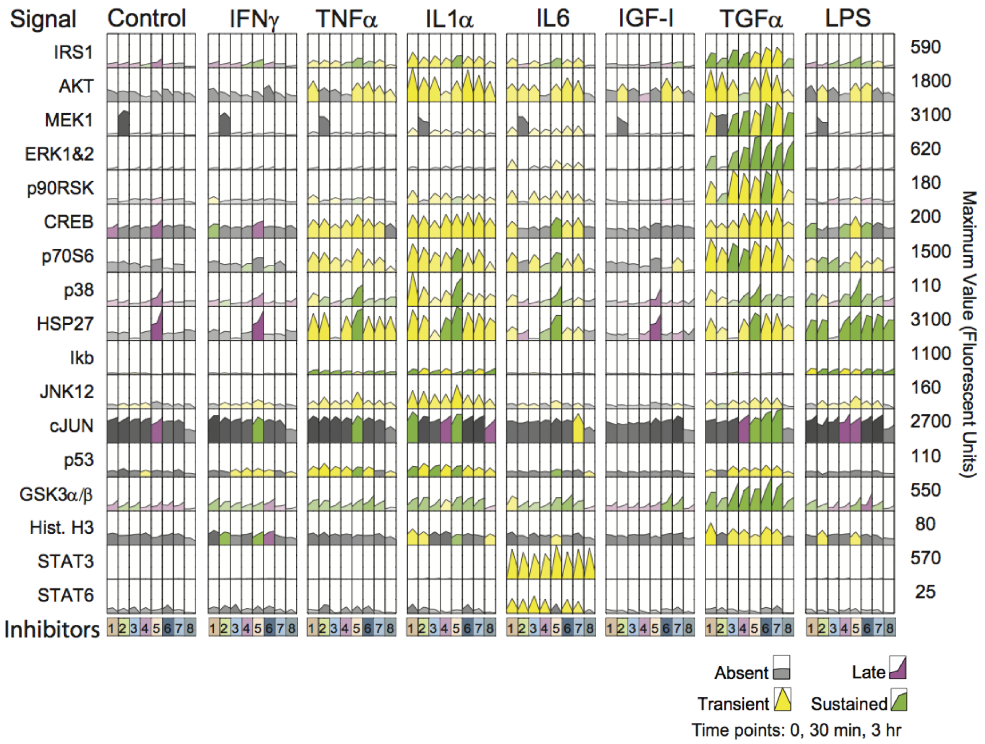


Figure 1.5: Sample experimental dataset published in [12]. Rows correspond to the various signals, the main columns correspond to the stimuli used and the sub columns correspond to the different inhibitors introduced together with the stimuli. The numbers on the right correspond to the maximum fluorescent intensity (in fluorescence units) for every signal. There is a total of 64 experiments in this plot (8 stimuli $\times$ 8 inhibitors). Every subplot represents the time course of the signal from the unstimulated state (time=0) to the 3 hours after stimulation. Based on the trends of the signal (first increase then decrease, increase then stay active, etc) the qualitative response of the proteins is described as "absent", "late", "transient", "sustained".

treatments present in the experimental dataset and subsequent comparison of the predictions with the corresponding measurements reveals a mismatch (i.e. fitness error). The Boolean model may be refined by optimizing its topology to obtain a better fit of the data. The optimization of the network's topology is implemented via a GA that runs the model iteratively and after each iteration evaluates the fitness error and removes reactions that appear not to be functional, to minimize this error. The optimized topology includes only a subset of the original set of interactions (only the ones that appear to be functional based on the data at hand) and more closely captures the signaling response of liver cells.

Optimization of signaling pathways to phosphoproteomic data via an ILP formulation

The GA based approach, presented above, successfully trained a Boolean model of the signal transduction pathway of transformed liver cells to high throughput phosphoproteomic data. The rules of Boolean logic were encoded in an executable model, run iteratively under the GA and gradually refined until its topology fitted the data at hand. However, this black box approach to optimizing the network's topology is far from efficient and more importantly, the GA cannot

guarantee a global minimum of the objective function is obtained.

This thesis revolves around the formulation of the network optimization problem, as described in [13], as a regular optimization problem (Integer Linear Program (ILP) or Non-Linear Program (NLP)). The rules of Boolean logic are encoded as constraints of the formulation and the mismatch of model predictions with the proteomic data is the objective function to be minimized. The published methodologies outperform previous attempts and pave the way for tackling large scale problems capturing the cells signaling mechanisms on a systems level.

In the following chapters of this thesis, three formulations are presented:

1. An ILP formulation to model signal transduction as a Boolean model, trained to proteomic data so it best captures the signaling response of the interrogated cell type.

2. An NLP formulation to model signal transduction as a constrained fuzzy logic model.

3. An ILP formulation to model signal transduction as an interaction graph and subsequently remove inconsistencies with experimental data using a variety of intervention strategies.

These formulations are used to tackle problems related to hepatocellular carcinoma (i.e. liver cancer) and osteoarthritis.

1.3 Phosphoprotein signaling and liver cancer

Hepatocellular Carcinoma (HCC) (i.e. liver cancer) is one of the leading causes of death worldwide [15]. Traditionally, the etiology of the disease is attributed to genetic alterations that accumulate during chronic inflammation of the liver. Mutations are found in several important genes including p73, p53, Rb, APC, DLC-1 (deleted in liver cancer), p16 , PTEN, IGF-2, BRCA2, SOCS-1, Smad2 and Smad4, β-catenin, c-myc, and cyclin D1 [16, 17, 14]. Moreover, as in other cancers, HCC is characterized by an imbalance in growth promoting signals; these are signaling pathways related to cell proliferation (e.g. MEK-ERK pathway). These alterations in signaling pathways give cancer cells a survival advantage against normal cells and cancer tumors are formed.

In more detail, the major pathways implicated in HCC are **(i)** the RAF/MEK/ERK pathway, **(ii)** the PI3K/AKT/mTOR pathway, **(iii)** the WNT/b-catenin pathway, **(iv)** the insulin-like growth factor (IGF1) pathway, **(v)** the hepatocyte growth factor (HGF)/c-MET pathway. and **(vi)** growth factor-regulated angiogenic (VEGF) signaling pathway [18].

RAF/MEK/ERK pathway

The RAF/MEK/ERK pathway regulates crucial cellular processes, including proliferation, differentiation, angiogenesis and survival. It lies downstream of various growth factor receptors such as the EGFR, IGFR, c-MET and VEGFR and the GRB2/SHC/SOS pathway. Within the RAF/MEK/ERK pathway, RAF is activated by upstream kinase RAS and activates MEK's two isoforms MEK1 and MEK2 which are responsible for phosphorylating and activating ERK1 and ERK2. ERK1/2 regulate cellular activity by acting on more than 100 substrates in the cytoplasm and nucleus, including indirect inducers of gene expression, transcription factors and cell cycle-related kinases. In HCC, the RAF/MEK/ERK pathway is over-activated, suggesting a possible role for this pathway in tumor formation.

PI3K/AKT/mTOR pathway

In PI3K/AKT/mTOR pathway, PI3K is activated upon binding of IGF and EGF to their receptors. PI3K subsequently produces the lipid second messenger PIP3b, which in turn activates

AKT. AKT phosphorylates several proteins, such as mTOR and BCL-2-associated death promoter (BAD). The activation of mTOR increases cellular proliferation, and inactivation of BAD decreases apoptosis and increases cell survival. In HCC PI3K/AKT/MTOR pathway is over-activated and in similar fashion to the RAF/MEK/ERK pathway enhances tumor formation.

WNT/b-catenin pathway

The abnormal regulation of the transcription factor β-catenin, a key component of the WNT signaling pathway is a major event in the development of HCC. In the WNT/b-catenin signaling pathway, WNTs (a family of soluble cysteine-rich glycoprotein ligands) bind to their receptors inducing activation of DSH which prevents phosphorylation of b-catenin and its subsequent ubiquitination and proteasomal degradation. The increasing concentration of b-catenin in the cytoplasm results in its translocation in the nucleus where it induces the transcription of genes implicated in cell proliferation (e.g. MYC, CJUN, CYCD1) thus promoting tumor formation.

IGF1, HGF/c-MET and EGFR pathways

The IGF1, HGF/c-MET and EGFR pathways are all over-expressed in HCC, activating pathways such as the RAF/MEK/ERK and PI3K/AKT/mTOR pathways, implicated in cell proliferation and inducing tumor formation and growth. Regarding the IGF pathway, it is activated by binding of the ligands IGF-1 and IGF-2 to IGFR. IGFR regulates key cellular processes including proliferation, motility and inhibition of apoptosis. Ligand binding to the IGFR triggers rapid receptor autophosphorylation, followed by phosphorylation of intracellular targets (mainly insulin receptor substrates 1-4), which in turn initiates downstream cellular effectors, ultimately leading to activation of PI3K, protein kinase B and the RAF/MEK/ERK pathway. Regarding the HGF pathway, it is activated by binding of HGF (a multifunctional cytokine) to tyrosine kinase receptor c-MET. c-MET regulates tissue regeneration, cell proliferation, migration, survival, branching morphogenesis and angiogenesis. HGF-induced activation of c-MET leads to phosphorylation of GRB2 and GAB1, which then activate PI3K and ERK. These in turn promote cell proliferation enhancing tumor formation. The EGFR pathway has similar role in the progression of HCC. EGFR regulates key processes such as proliferation and survival by activating the PI3K/AKT/mTOR and RAF/MEK/ERK pathways and is over-activated in HCC leading to tumor formation and growth.

VEGFR pathway

VEGF is one of the growth factors, most intensely studied in HCC. VEGF binds to the VEGFR and regulates angiogenesis, a key process for tumor growth. Without increased vascularization the tumor cannot obtain the required nutrients and thus, its growth is inhibited. In vivo studies have validated the significance of VEGF in HCC. Experiments on human tumor xenografts in immunodeficient mice showed that neutralization of VEGF inhibited tumor growth and decreased blood vessel density in a variety of tumor types.

1.4 Chondrocytes signaling in osteoarthritis (OA)

Osteoarthritis, a debilitating joint disease, is characterized by an imbalance of competing pro-growth (anabolic) and pro-destructive (catabolic) signals in articular cartilage. It is well established that certain pro-inflammatory cytokines such as IL1 and TNF induce cartilage degeneration and loss of its mechanical integrity mainly through the release in the extracellular space of catabolic effectors such as aggrecanases and matrix metalloproteinases (MMPs) [19, 20, 21, 22]. Regarding the IL1 pathway, binding of IL1a or IL1b to IL1R, enables an adaptor protein, MyD88, to procure kinases IRAK-1 and -2 to activate further intracellular proteins. IRAK-1 and -2 stim-

ulate the response of the NF-kB inducing kinase (NIK) to activate the IKK-b complex, which finally induces the NF-kB protein to effect gene transcription and express MMPs. Regarding the TNF pathway, it becomes activated upon binding of TNF to TNFR inducing the activation of TRAF2/5 and FADD through TRADD (TNF-R1 associated death domain protein). The TRAF protein activates two main signaling pathways, MAPK and NF-kB inducing the expression of MMPs degrading cartilage tissue.

Chapter 2

An Integer Linear Programming (ILP) formulation for the optimization of signal transduction networks to phosphoproteomic data

In this work published in [23], an Integer Linear Programming (ILP) formulation is introduced to optimize signal transduction networks to high throughput phosphoproteomic data and construct predictive models of the signaling mechanisms of the interrogated cell type. This work was carried out in collaboration with Alexander Mitsos (at the time when this work was published an assistant professor in the Mechanical Engineering Department of MIT, Cambridge, MA, USA, currently a professor in RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany). In contrast to previously published approaches, the ILP formulation guarantees to return a global minimum of the objective function (=mismatch between model predictions and experimental measurements) and speeds up the runtime significantly, thus allowing the interrogation of more complex signaling pathways and accompanying phosphoproteomic datasets. Two case studies were addressed with the current formulation, (i) the optimization of a large scale signal transduction network, downstream of 80 receptors of interest and (ii) the identification of drug effects in terms of topology alterations of the signaling pathway. Additionally, the formulation was extended to account for indirect correlations between phosphoprotein signals and cellular response. As a case study, a signaling pathway for liver cells was constructed initiating at the receptors level, moving downstream to the measured signals and predicting the secretion of cytokines in the supernatant.

2.1 Introduction on the modeling of signaling pathways

Signaling pathways are of the utmost importance for understanding cellular function and predicting cellular response to perturbations [3, 4]. Recent advancements in text mining and the construction of Protein-Protein Interaction (PPI) networks have led to large databases of signaling pathways, showing how proteins interact with each other [24, 25, 26, 27, 28]. However, compilation and visualization of protein connectivity in signaling networks is just the first step towards understanding the cell's signaling mechanisms. The modeling and analysis of these networks either at the connectivity level or down at the level of signal transduction mechanics between nodes is a crucial next step towards the construction of functional models, predictive of the cell's biology.

A variety of methods have been proposed for this task, each adopting a different perspective on the nature of the included reactions [8] and focusing on different properties of the signaling network. Two wide classes of network analysis can be distinguished: i) Topological analysis of the signaling network [29, 30] that extracts insight into the cells' function by investigating the structural characteristics of the signaling network (e.g., feedback loops, strongly connected components). ii) Network identification, which identifies the network structure (i.e. connectivity of signaling species), or reaction parameters that define the mechanics of signal transduction from one node to the next. Typically a mathematical formalism is adopted to model how signal transduction takes place and an executable model is constructed by combining this formalism with a prior knowledge network (PKN) that serves as a scaffold. By simulating the model under different node and reaction parameters, conclusions can be drawn for the importance of each node and reaction on the propagation of the signal. Amongst the most widely used formalisms are the various forms of logic modeling [31, 9, 32, 33, 34] and ordinary differential equations (ODEs) [35, 36, 37, 38]. In certain cases, the initial model is trained to signaling data via an optimization approach [13, 10] to compute the values of model parameters that better fit the data at hand, or a sensitivity analysis approach is used [39] to compute the influence of model parameters to the overall response of the model. The incorporation of signaling data allows the construction of cell-specific, tissue-specific, or disease-specific pathways [12].

The selection of the modeling approach, and subsequently of the optimization procedure, is very close related to the availability of data and biological question at hand. For example, if time course data are available and the dynamics of signaling reactions are of interest, then an ODE-based approach may be suitable, especially if the interrogated signaling network is small in size. To this end significant work has been published on parameter estimation of ODE-based models using a wide spectrum of methods including general purpose optimization methods (gradient based algorithms, stochastic search algorithms, branch and bound strategies, geometric programming, Dynamic Flux estimation and others) [40]. However, large scale signaling networks cannot easily be addressed within an ODE framework because of excessive CPU times and lack of proper constrain of the association-dissociation constants. If data are available for large pathways but on a single time point, then logic based modeling (Boolean or fuzzy logic, simulated at a 'pseudo steady-state') can be used to identify the structure of the signaling pathway.

In Boolean modeling, signal transduction is modeled using the rules of Boolean logic [41, 42, 43, 44, 45]. Protein nodes assume only binary values $\{0, 1\}$, denoting the activation (or not) of the corresponding signaling molecule, and signal is propagated from the receptor level to downstream nodes using a combination of OR and AND gates. In [13] an approach was introduced to compress a protein network and convert it into Boolean models that are trained against signaling data. In the approach, implemented in the tool CellNOpt, reactions that appear to contradict the data are removed from the PKN, and thus measurement-prediction mismatch is minimized. In CellNOpt a Genetic Algorithm (GA) was used to prune the pathway by identifying and removing the contradicting reactions. The GA offered a robust and flexible optimization framework and managed to uncover structural differences between normal and cancer liver cell types.

In the present study, the optimization problem was formulated as an Integer Linear Program (ILP) and was solved through CPLEX (ILOG CPLEX 9.0,) and GUROBI (Gurobi Optimization, Inc., http://www.gurobi.com/) via GAMS(http://www.gams. com/). In contrast to GA, the ILP formulation guaranteed global optimality and required a fraction of the CPU time needed by the GA. The computational efficiency of the ILP formulation allows the rapid optimization of large scale signaling networks, offering a systems wide view of the signaling network in the interrogated cell type [46].

2.2 ILP formulation

Here, we describe how the Boolean model described in [13] can be reformulated as an ILP. A pathway is defined as a set of reactions $i = 1, ..., n_r$ and species $j = 1, ..., n_s$. Each reaction has three corresponding index sets, namely the index set of signaling molecules R_i, inhibitors I_i, and "products" P_i ("product" can also correspond to the phosphorylation level of the protein). These sets are all subsets of the species index set ($R_i, I_i, P_i \subset \{1, ..., n_s\}$). Typically, these subsets have very small cardinality (few species), e.g., $\mid R_i \mid = 0, 1, 2; \mid I_i \mid = 0, 1; \mid P_i \mid = 1, 2; \mid R_i \mid + \mid I_i \mid = 1, 2$. A reaction takes place if and only if all reagents and no inhibitors are present. If a reaction takes place, all products are formed. Note that reactions without products as well as reactions with neither reagents nor inhibitors will be excluded here.

While typically the set of species is known, the set of reactions is not known. Rather, only a superset of potential reactions is postulated. The goal of the proposed formulation is to find an optimal (in some sense) set of reactions out of such a superset. To that extent binary variables y_i are introduced, indicating if a reaction is possible or not ($y_i = 0$ connection not present, $y_i = 1$ connection present).

A set of experiments is performed, indexed by the superscript $k = 1, ..., n_e$. In each experiment a subset of species is introduced to the system and another subset is excluded from the system. These are summarized by the index sets $M^{k,1}$ and $M^{k,0}$ respectively (two for each experiment). In the proposed formulation, constants are introduced for all such species, respectively $x_j^k = 1$ and $x_j^k = 0$. In the following it will be assumed that these species do not appear as products in any reaction; this assumption is not limiting, since in the experiments performed only extracellular species and inhibitors are manipulated. In the experiments a third subset of the species is measured (index set $M^{k,2}$) and for the remaining species no information is available. In the proposed formulation for each of the experiments and each such species a binary decision variable $x_j^k \in \{0, 1\}$ is introduced indicating if the species j is present ($x_j^k = 1$) or not ($x_j^k = 0$) in the experiment k according to the model predictions. The last group of variables z_i^k introduced indicate if reaction i will take place ($z_i^k = 1$) or not ($z_i^k = 0$) in the experiment k according to the model predictions.

For the case that a species is measured, the measurement is defined as $x_j^{k,m}$. For Boolean measurements $x_j^{k,m} \in \{0, 1\}$; otherwise $x_j^{k,m} \in [0, 1]$. The primary objective function is formed aiming to minimize the weighted error between model predictions and measurements $\sum_{j,k} a_j^k \mid x_j^k - x_j^{k,m} \mid$. The absolute value is reformulated as $x_j^{k,m} + (1 - 2x_j^{k,m})x_j^k$. It can be easily verified that for binary x_j^k and for $x_j^{k,m} \in \{0, 1\}$ this reformulation is valid:

1. $x_j^k = 0$:
   $x_j^{k,m} + (1 - 2x_j^{k,m})x_j^k = x_j^{k,m} + (1 - 2x_j^{k,m})0 = x_j^{k,m} = \mid x_j^{k,m} \mid = \mid x_j^{k,m} - x_j^k \mid .$

2. $x_j^k = 1$:
   $x_j^{k,m} + (1 - 2x_j^{k,m})x_j^k = x_j^{k,m} + (1 - 2x_j^{k,m})1 = 1 - x_j^{k,m} = \mid 1 - x_j^{k,m} \mid = \mid x_j^k - x_j^{k,m} \mid .$

Note also that alternative norms, such as least-squares errors, could be also used. The resulting optimization problem would still be an ILP, since the objective function involves only integer variables. For instance for the least-square error objective function the following linear reformulation is valid:

$$(x_j^k - x_j^{k,m})^2 = (x_j^k)^2 - (2\, x_j^k\, x_j^{k,m}) + (x_j^{k,m})^2 = (x_j^k) - (2\, x_j^k\, x_j^{k,m}) + (x_j^{k,m})^2$$

The secondary objective is to minimize the weighted number of possible reactions $\sum_i \beta_i y_i$. In multi-objective optimization typically the concept of Pareto-optimal or non-inferior solution is introduced, i.e., a set of decision variable values, such that if one tries to improve one objective, another will be degraded. The set of Pareto points forms the Pareto-optimal curve. Here, however, the primary objective is considered much more important than the secondary objective. Therefore, a single Pareto-optimal point is obtained, by first minimizing the primary objective and then the secondary objective by requiring that the former (more important) objectives are not worsened.

The ILP proposed can be summarized as:

$$min \sum_{k=1}^{n_e} \sum_{j \in M^{k,2}} a_j^k (x_j^{k,m} + (1 - 2\, x_j^{k,m}) x_j^k); \quad \sum_{i=1}^{n_r} \beta_i\, y_i \tag{2.1}$$

s.t.

$$\sum_{i=1}^{n_r} a_i^l\, y_i \leq b^l, \quad l = 1, ..., n_e \tag{2.2}$$

$$z_i^k \leq y_i; \quad i = 1, ..., n_r, \quad k = 1, ..., n_e. \tag{2.3}$$

$$z_i^k \leq x_j^k; \quad i = 1, ..., n_r, \quad j \in R_i \quad k = 1, ..., n_e. \tag{2.4}$$

$$z_i^k \leq 1 - x_j^k; \quad i = 1, ..., n_r, \quad j \in I_i \quad k = 1, ..., n_e. \tag{2.5}$$

$$z_i^k \geq y_i + \sum_{j \in R_i} (x_j^k - 1) - \sum_{j \in I_i} (x_j^k), \quad i = 1, ..., n_r, \quad k = 1, ..., n_e. \tag{2.6}$$

$$x_j^k \geq z_i^k; \quad i = 1, ..., n_r, \quad j \in P_i \quad k = 1, ..., n_e. \tag{2.7}$$

$$x_j^k \leq \sum_{i=1,...,n_r : j \in P_i} z_i^k, \quad j = 1, ..., n_s, \quad k = 1, ..., n_e. \tag{2.8}$$

$$x_j^k = 0, \quad k = 1, ..., n_e \quad j \in M^{k,0} \tag{2.9}$$

$$x_j^k = 1, \quad k = 1, ..., n_e \quad j \in M^{k,1} \tag{2.10}$$

$$X \in \{0,1\}^{n_e \times n_s}, \quad y \in \{0,1\}^{n_r}, \quad Z \in \{0,1\}^{n_e \times n_r} \tag{2.11}$$

where the objectives are separated by a semi-colon. Note that for the elements of the matrices X and Z, the row index (experiment) is indicated as superscript, and the column index (species and reactions respectively) is indicated as subscript. In formulation (2.1)–(2.11) for the manipulated species binary decision variables along with the constraints (2.9) and (2.10) are introduced. This simplifies notation. In the implementation, these variables are replaced by constants. Alternatively the preprocessor of the optimization solver can be used to exclude these trivial variables.

In the following the reasoning for the formulation is given. The first set of constraints, i.e., (2.2) allow the modeler to limit the combinations of connectivities considered. For instance, suppose that two reagents $R1$, $R2$ form a product P, but it is not known if both reagents (AND) or either (OR) are required. This can be modeled as three potential reactions

$$r1: \; R1 + R2 \rightarrow P$$

$$r2: \; R1 \rightarrow P$$

$$r3: \; R2 \rightarrow P$$

with the additional constraint that $r1$ excludes $r2$ and $r3$, which can be modeled as two linear inequalities:

$$y_{r1} + y_{r2} \leq 1$$

$$y_{r1} + y_{r3} \leq 1$$

The constraints (2.3) indicate that a reaction can only take place if it is possible ($y_i = 1$). This can be seen easily, since $y_i = 0$, gives $z_i^k \leq 0$ and together with $z_i^k \in \{0, 1\}$ we obtain $z_i^k = 0$. Similarly, the constraints (2.4) and (2.5) ensure respectively that a reaction can only take place if all reagents and no inhibitors are present. If for instance a reagent is absent, $z_i^k = 0$ is enforced, and the other constraints are redundant. On the other hand, the constraints (2.6) enforce that if a reaction is possible, all reagents are present, and no inhibitors are present, then the reaction will take place ($z_i^k = 1$).

The constraints (2.7) ensure that a species will be formed if some reaction in which it is a product occurs. Note that multiple reactions can give the same species; mathematically this will result in redundant constraints. In contrast, the constraints (2.8) enforce that a species will not be present if all reactions in which it appears as a product do not occur. Recall that manipulated species are not considered as products in reactions. Note also, that it would be possible to combine the constraints (2.7) into a single constraint for each species, e.g.,

$$x_j^k \geq \frac{\sum_{i=1,...,n_r; j \in P_i} z_i^k}{\sum_{i=1,...,n_r; j \in P_i} 1}, \quad j = 1, ..., n_s, \quad k = 1, ..., n_e$$

but this would result in weaker LP-relaxations. Also the reformulation of x_j^k to $[0, 1]$ would no longer be exact.

Construction of toy model and performance assessment

To validate the performance of the optimization algorithm a toy model consisting of 29 nodes and 35 reactions is constructed. The toy model includes only 5 stimuli (TGFA, BTC, NRG1, IL1B, IL1A) and 3 signals (MAP2K1, AKT, IKB) (see figure 2.1) and serves to better illustrate the difference between positive and negative size weights β_i, as well as the execution of cross-validation studies. The main topological features include the activation of MAP2K1 and AKT from TGFA, BTC and NRG1 via a number of alternative pathways, the activation of AKT from IL1A and IL1B via SRC, and the activation of IKB from TGFA, BTC, NRG1 via MAP3K8 and from IL1A, IL1B via TRAF6. An accompanying dataset is also constructed, consisting of single treatments of the above mentioned stimuli in a total of 6 experimental conditions (including the no-inhibitor experiment). The ILP prunes the initial topology to best fit the dataset at hand by minimizing the objective function 2.1.

The toy model is optimized using two different settings: (i) small positive reaction weights (β_i) enforcing the minimization of the pathway size together with the experiments-topology mismatch (see figure 2.1B). (ii) small negative reaction weights (β_i) enforcing the maximization of the pathway size while minimizing the experiments-topology mismatch. This setting results in the superset of possible solutions (see figure 2.1C). It becomes apparent that both solutions share the same basic connectivity patterns, TGFA, BTC and NRG1 activate MAP2K1 and AKT, IL1A and IL1B activate IKB. Their difference lies in the fact that the superset of possible solutions includes all paths fitting these patterns while pathway in figure 2.1B includes only the shortest. No unresolved fitness error is observed (for both solutions).

A better assessment of how the ILP formulation performs, is obtained by omitting datapoints from the initial dataset and monitoring the remaining fitness error. 3 subsets of the initial dataset were constructed by omitting measurements in a random manner. The toy model was

then trained and the unresolved fitness error was plotted (red background in figure 2.1E). In the first of these plots, although 5 of the original 18 datapoints were left out of the dataset no experiments-topology mismatch is observed in the solution. The reason for such a solid performance under missing measurements can be traced to the high degree of overlap between the pathways. In the toy model TGFA, BTC and NRG1 signal via almost identical pathways (same for IL1A and IL1B) creating many internal replicates, allowing the removal of a substantial part of the dataset without affecting the goodness of fit of the solution. As we gradually omit more measurements i) internal replicates are no longer present, ii) subsets of the topology become non-observable and are removed altogether, iii) and the solution fails to fit the latent signals.

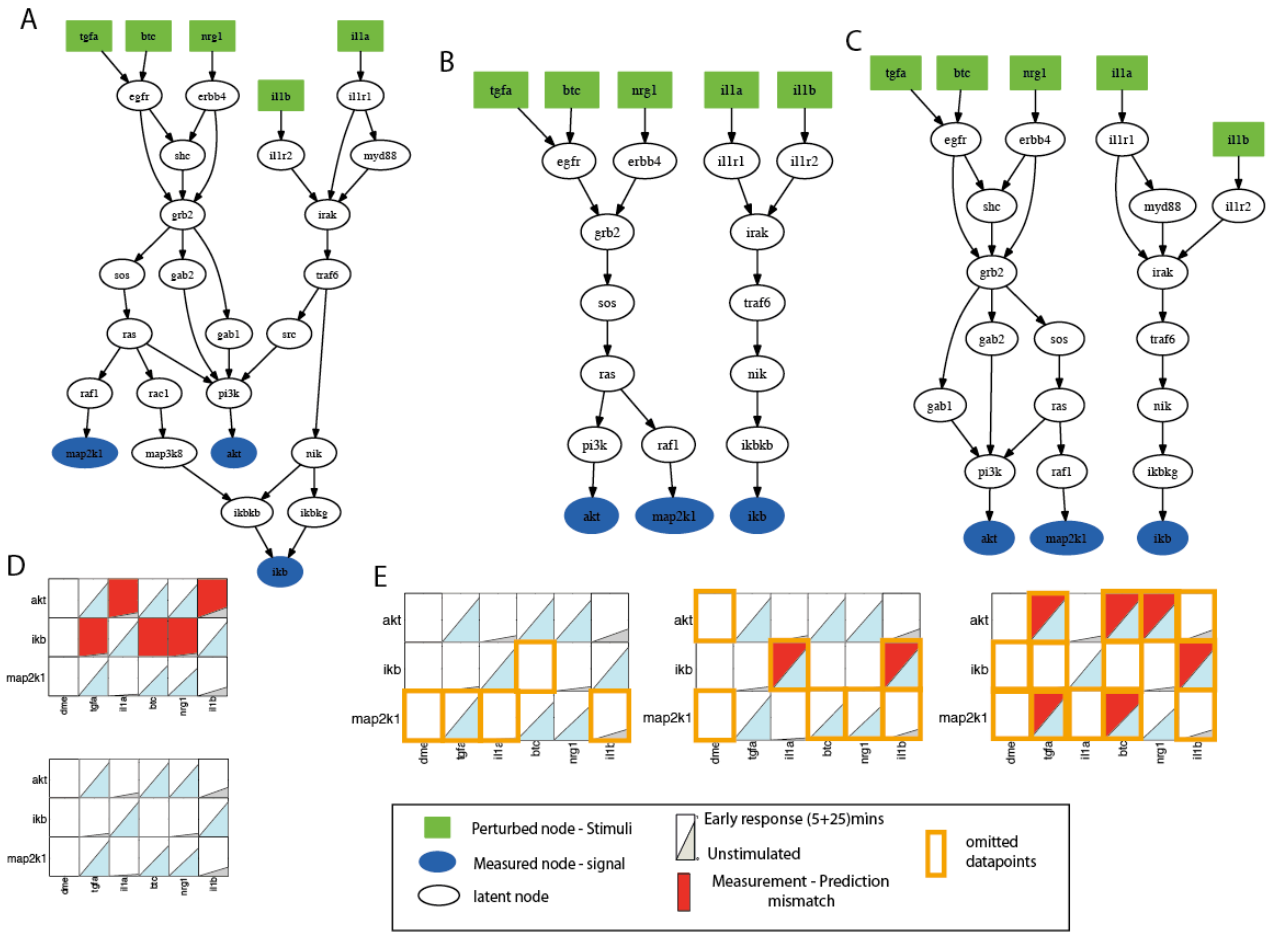


Figure 2.1: Toy model – performance assessment. (a) The toy model consists of 5 stimuli, 3 measured signals and a total of 29 nodes and 35 reactions. (b) The toy model is optimized by the ILP using two different settings, first using a small positive reactions weight, resulting in a minimum-size solution, and second (c) using a small negative reactions weight, resulting in a maximum-size solution. (d) Topology – prediction mismatch of the original and optimized pathway. (e) Topology - prediction mismatch after optimizing the toy model with subsets of the original dataset.

Equivalent reformulation as MILP

The formulation presented above can be reformulated an MILP in the following manner:

Relaxation of Z.
We will argue that relaxing the Z variables from binary to continuous gives an exact reformulation. It suffices to show that constraints (2.3)-(2.8) together with $X \in \{0,1\}^{n_e \times n_s}$ and $y \in \{0,1\}^{n_r}$ imply $Z \in \{0,1\}^{n_e \times n_r}$.

Theorem 1. *Replacing $Z \in \{0,1\}^{n_e \times n_r}$ by $Z \in [0,1]^{n_e \times n_r}$ is an exact reformulation, in the sense that any feasible point in the new program is also feasible in the original program.*

Note that Theorem 1 is a special case of Theorem 2. Nevertheless it is given separately, because it does not require the assumption of acyclical graphs. Moreover, its proof is much simpler, and is used in the proof of Theorem 2.

proof. Take any $X \in \{0,1\}^{n_e \times n_s}$, $y \in \{0,1\}^{n_r}$ and $Z \in [0,1]^{n_e \times n_r}$ that satisfies constraints (2.1)-(2.11). Take any $i = 1, ..., n_r$ and any $k = 1, ..., n_e$. We consider two cases depending on the value of y_i^k.

1. $y_i^k = 0$. From (2.3) we directly obtain $z_i^k \leq 0$ and therefore $z_i^k = 0$.
2. $y_i^k = 1$. We consider two sub cases:
   - If for some $j \in R_i$ we have $x_j^k = 0$ (a reagent is missing), then $z_i^k \leq 0$ from (2.4) and therefore $z_i^k = 0$. Similarly, if for some $j \in I_i$ we have $x_j^k \leq 1$ (an inhibitor is present), then from (2.5) $z_i^k \leq 0$ and therefore $z_i^k = 0$.
   - If for all $j \in R_i$ we have $x_j^k = 1$ (all reagents are present) and for all $j \in I_i$ we have $x_j^k = 0$ (all inhibitors absent), then from (2.6) we obtain $z_i^k \geq 1$ and therefore $z_i^k = 1$. Since the choice of i and k was arbitrary we have shown $Z \in \{0,1\}^{n_e \times n_r}$.

Relaxation of non input x_j^k.
For the case that no loops are present in the pathway, we will argue that we can also use $x_j^k \in [0,1]$ for all species but the input species. In typical pathways the majority of species are noninput species. The formal definition of input species is:

Definition 1 (input species). *Species j that are not products in any reaction i.e., $T = \{j \in \{1, ..., n_s\} : j \notin \cup_{i=1}^{n_r} P_i\}$* are termed *input species.*

Theorem 2. *Suppose that the pathway proposed contains no loops. In (2.1) - (2.11) replacing $Z \in \{0,1\}^{n_e \times n_r}$ by $Z \in [0,1]^{n_e \times n_r}$ and $x_j^k \in \{0,1\}$ by $x_j^k \in [0,1]$ for all $j \notin T$ (for all non-input species) is an exact reformulation, in the sense that any feasible point in the new program is also feasible in the original program.*

Note that input species cannot be relaxed, for otherwise $z_i^k \in \{0,1\}$ would not be ensured. The proof idea is that because the potential pathway form a directed graph, we can proceed from the "top" to the "bottom". In doing so we establish that both x_j^k and z_i^k are forced to be integer.

proof. Take any $X \in [0,1]^{n_e \times n_s}$, $y \in \{0,1\}^{n_r}$ and $Z \in [0,1]^{n_e \times n_r}$ that satisfies the constraints (2.1) - (2.11) and that also satisfies $x_j^k \in \{0,1\}$ for all $j \in T$ (all input species are binary).

In the proof of theorem 1 we have established that if for a given reaction i and experiment k we have $x_j^k \in \{0,1\}$ for all $j \in R_i \cup I_i$ (all reagents and inhibitors are binary), then we also obtain $z_i^k \in \{0,1\}$.

Take $k \in \{1, ..., n_e\}$ (an arbitrary experiment) and $j \in \{1, ..., n_s\}$ (an arbitrary species) We will argue that if $z_i^k \in \{0, 1\}$ for all $i \in \{1, ..., n_r\} : j \in P_i$ (for all reactions for which the species is a product) then $x_j^k \in \{0, 1\}$. There are essentially two cases:

1. If for some $i \in \{1, ..., n_r\} : j \in P_i$ we have $z_i^k = 1$ then by (2.7) we obtain $x_j^k \geq 1$ and therefore $x_j^k = 1$.

2. If for all $i \in \{1, ..., n_r\} : j \in P_i$ we have $z_i^k = 0$ then by (2.8) we obtain $x_j^k \leq 0$ and therefore $x_j^k = 0$.

It is clear that in the absence of loops the above two arguments propagate through the pathway. From an arbitrary species $j \in \{1, ..., n_s\}$ we can traverse the graph in reverse direction and reach the input species in a finite number of steps (a reverse path). Due to the absence of loops, each species depends only on the species which are "further up" in the pathway.

2.3 optimization of a large scale signal transduction network

Herein, the ILP formulation described in section 2.2 is used to construct a large scale signal transduction network downstream 81 receptors of interest based on literature citations of protein interactions and high throughput phosphoproteomic data. This work, published in [46], was carried out in collaboration with Alexander Mitsos (at the time when this work was published an assistant professor in the Mechanical Engineering Department of MIT, Cambridge, MA, USA, currently a professor in RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany) and Julio Saez-Rodriguez, group leader at the European Bioinformatics Institute (EBI), Cambridge, UK. This is amongst the first attempts to construct cell-type specific signal transduction pathways of this size that enables the system level understanding of the cells signaling machinery.

Abstract

Construction of large and cell-specific signaling pathways is essential to understand information processing under normal and pathological conditions. On this front, gene-based approaches offer the advantage of large pathway exploration whereas phosphoproteomic approaches offer a more reliable view of pathway activities but are applicable to small pathway sizes. In this section, we demonstrate an experimentally adaptive approach to construct large signaling pathways from phosphoproteomic data within a 3-day time frame. Our approach—taking advantage of the fast turnaround time of the xMAP technology—is carried out in four steps: (i) screen optimal pathway inducers, (ii) select the responsive ones, (iii) combine them in a combinatorial fashion to construct a phosphoproteomic dataset, and (iv) optimize a reduced generic pathway via an Integer Linear Programming formulation. As a case study, we uncover novel players and their corresponding pathways in primary human hepatocytes by interrogating the signal transduction downstream of 81 receptors of interest and constructing a detailed model for the responsive part of the network comprising 177 species (of which 14 are measured) and 365 interactions.

2.3.1 Introduction on the modeling of large scale signaling pathways

Recent advancements in high throughput technologies are changing the focus of modern biology from the study of individual genes, proteins, or pathways into studying biological systems as a whole [47]. Technologies, such as DNA or protein microarrays for gene expression or protein concentration, Mass Spectrometry (MS) for proteomics, and Yeast two- hybrid (Y2H) screen for Protein–Protein Interactions (PPIs), have led to the construction of vast datasets, addressing cellular behavior on both genomic and proteomic levels. Compilation of all these heterogeneous data into predictive, functional models of cellular processes cannot be tackled by reductionist approaches [1]. Instead, markup languages, data annotation, and advanced mathematical modeling are used to make this integration feasible. These models are usually depicted in computable pathway maps and used for gaining a deeper understanding of cell's machinery as well as for identifying pathway alterations that take place in complex diseases and cannot be regarded as mere malfunction of a single constituent, that being a gene or a protein.

Both gene- and protein-based approaches endeavor the construction of large signaling pathways. A major advantage of the genomic approaches in the pathway construction is the unbiased exploration of the whole genome at the transcriptional level [48]. However, inconsistencies between protein abundance and corresponding mRNA levels in mammalian cells [49] reveal that gene expression alone might not be able to fully explain the function of signaling pathways and thus, proteomic measurements are frequently used for validating gene expression data. On this front, measurement of protein phosphorylation state is considered one of the most reliable

surrogates for studying the activity of signaling pathways [4]. However, in contrast to gene expression data, proteomic and phosphoproteomic technologies are far from capable to cover the whole proteome level [11].

Proteomic technologies applied today can be divided into two main categories, the ones that make no a priori assumption about the sample's protein content (e.g. MS), and affinity- based methods (e.g. protein arrays, xMAP technology, aptamers) that rely on antibodies or aptamers to detect a predetermined set of targets. MS and affinity-based methods serve different needs in studying signal transduction networks. MS supports the measurement of thousands of signals, however, it is limited to screening cells under few experimental conditions because of the semi-automated procedures in sample handling and the extensive data processing times. Complementary to MS approaches, reverse phase protein arrays and xMAP technology are confined by a predetermined number of measured proteins but are capable of measuring thousands of samples in a single day. This advantage—if coupled to assay automation and automated data gathering—opens the road for studying cellular pathways in an adaptive fashion in a short time frame.

In this work, we propose a new method for constructing large topologies based on optimization of canonical pathways with phosphoproteomic data. On the experimental front we propose an adaptive approach that is based on the fast turnaround time of the xMAP technology. Cells are plated on 96-well plates and a 4-step adaptive procedure is employed that includes: (1) ligand screening, (2) a bimodal model for stimuli selection, (3) design, implementation, and acquisition of phosphoproteomic data from combinatorial treatments with selected stimuli in the same batch of cells, and (4) an ILP formulation for pathway construction that guarantees a global optimal solution in a computationally efficient manner. As a case study, we interrogate the signal transduction network of primary human hepatocytes down- stream of 81 receptors, which includes all major players of liver homeostasis but also unknown and less known ligands.

The 4-step approach was implemented in 3-days time frame. In day 1 the cells are plated. In day 2 the cells are treated with 81 stimuli, phosphoproteomic data acquired, optimal ligand candidates selected, and new experimental conditions designed. In day 3 combinatorial treatments for the responsive ligands are performed, data acquired, and the ILP optimization is employed. As a result, a detailed, cell-specific signaling path- way of primary human hepatocytes, incorporating 15 ligands, 177 species and 365 reactions, is constructed within 3 days. To our knowledge, our approach is a first attempt to construct large signaling topologies in an adaptive manner by combining high throughput proteomic measurements with state of the art optimization algorithms. The 4-step procedure is presented in Figure 2.2 and detailed in the following paragraphs.

2.3.2 Results

Ligand screening

Primary human hepatocytes were isolated and plated in 96-well plates using standard methods as briefly described in Materials and Methods (day 1) [12]. A library of 81 stimuli was put together that consist of all major cytokines, chemokines, and other known and unknown (i.e. randomly selected) activators of hepatocyte physiology. Instead of testing the concentration of each individual stimulus (i.e., by performing a dose response curve and measuring the phosphorylation levels of several downstream proteins) we decided to employ a semi-automated text mining approach that searches the literature and builds a histogram of treated concentrations of a particular stimulus. The concentration histogram provides the user with the most commonly used concentrations found in the literature and enables him to choose a high concentration level.

The stimuli library (see the complete names in Table 2.1) consists of prototypical players in

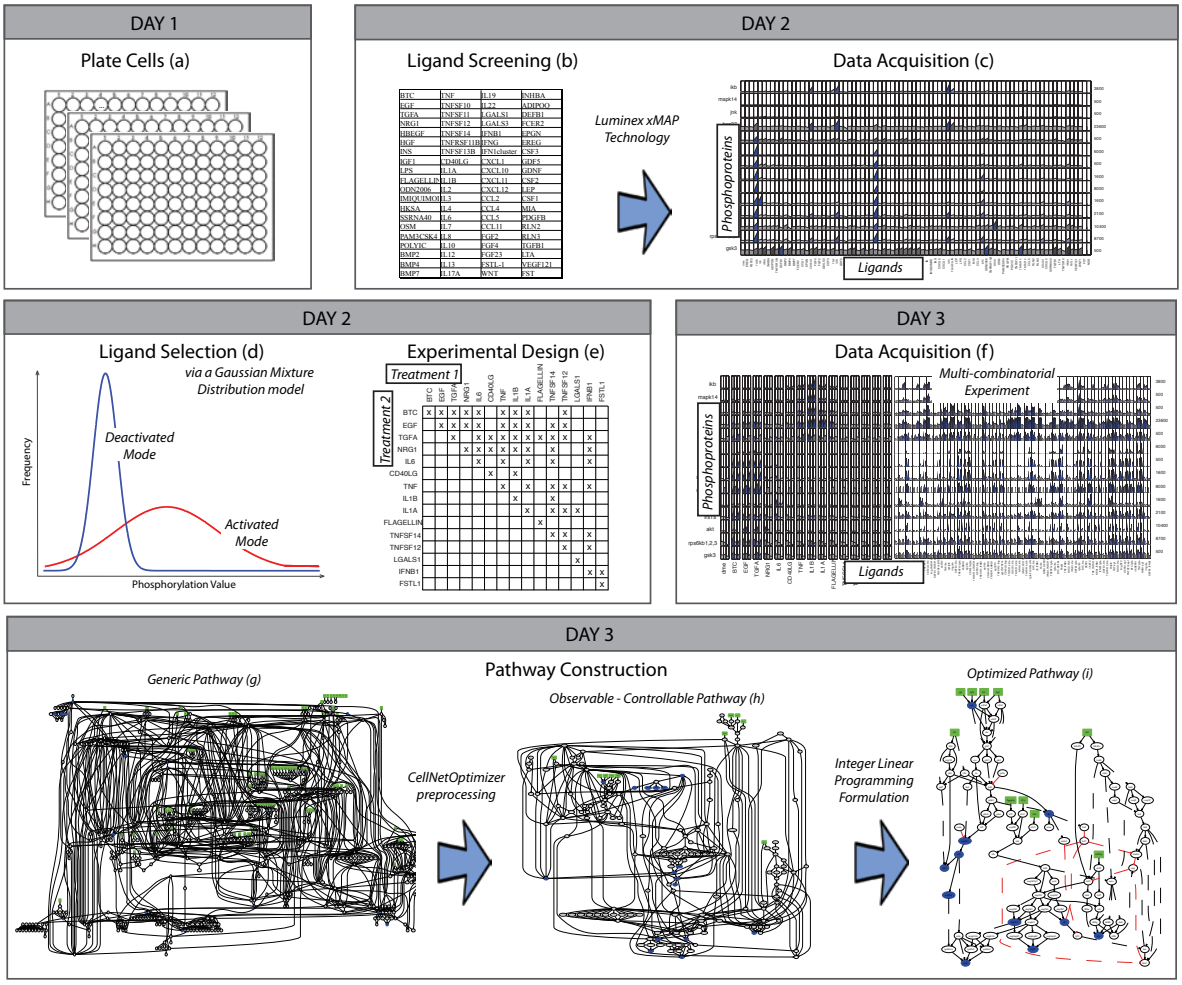


Figure 2.2: Experimental and computational workflow: (a) cells are plated in 96 well plates. (b) Cells are treated with 81 stimuli. (c) For each stimulus, the phosphorylation activity of 14 intracellular signals is measured at 5 and 25 minutes post stimulus. (d) Optimal inducers are selected based on a Gaussian mixture model. (e) The selected inducers are used in a combinatorial fashion. (f) Second phosphoproteomic experiment is performed and data acquired. (g) The generic pathway is pre-processed via CellNetOptimizer and its (h) observable/controllable part is conserved for optimization. (i) The pathway is optimized via an Integer Linear Programming formulation.

the medical literature such as the EGFR pathway (BTC, EGF, TGFA, NRG1, HBEGF), HGF pathway (HGF), Insulin pathway (INS, IGF1), Toll Like Receptor Stimuli (LPS, FLAGELLIN, ODN2006, IMIQUIMOD, HKSA, SSRNA40, OSM, PAM3CSK4, POLYIC), the BMP family (BMP2, BMP4, BMP7), the TNF family (TNF, TNFSF10, TNFSF11, TNFSF12, TNFSF14, TNFRSF11B, TNFSF13B, CD40LG), several inflammatory and pleiotropic interleukins (IL1A, IL1B, IL2, IL3, IL4, IL6, IL7, IL8, IL10, IL12, IL13, IL17A, IL19, IL22), lectins (LGALS1, LGALS3), interferons (IFNB1, IFNG, IFN1 cluster), inflammatory and other chemokines (CXCL1, CXCL10, CXCL11, CXCL12, CCL2, CCL4, CCL5, CCL11), the fibroblast growth factor family (FGF2, FGF4, FGF23), Follistatin-like 1 (FSTL-1), WNT signaling (WNT), as well as several other inflammatory or pleiotropic activators: INHBA, ADIPOQ, DEFB1, FCER2, EPGN,

EREG, CSF3, GDF5, GDNF, CSF2, LEP, CSF1, MIA, PDGFB, RLN2, RLN3, TGFB1, LTA, VEGF121, FST, NOG. High-throughput bead-based ELISA experiments using semi-automated xMAP technology (Luminex, Texas, USA) are performed to measure the "in vitro average phosphorylation level" of 14 key phosphoproteins at 5 and 25 minutes (IKB, JNK, HSP27, CREB, ERK, STAT3, IRS1S, AKT, GSK3, MAPK14 (i.e. p38), RPS6KA1 (i.e. p90RSK), MAP2K1 (i.e. MEK1), RSP6KB1/2/3 (i.e. p70S6K), EGFR). The selection of the above-mentioned signals is based on previous results [12] and assay availability at the time of this study. Data were acquired and analyzed the same day and plotted as shown in Figure 2.3 using the DataRail software [50].

Acronym	**Full name (HUGO Nomeclature)**		**Acronym**	**Full name HUGO nomeclature**
BTC	Betacellulin	hspace10pt	IL22	Interleukin 22
EGF	epidermal growth factor	hspace10pt	LGALS1	lectin, galactoside-binding, soluble, 1
TGFA	transforming growth factor, alpha	hspace10pt	LGALS3	lectin, galactoside-binding, soluble, 3
NRG1	neuregulin 1	hspace10pt	IFNB1	interferon, beta 1, fibroblast
HBEGF	heparin-binding EGF-like growth factor	hspace10pt	IFNG	interferon, gamma
HGF	hepatocyte growth factor	hspace10pt	IFN1cluster	interferon, type 1, cluster
INS	Insulin	hspace10pt	CXCL1	chemokine (C-X-C motif) ligand 1
IGF1	insulin-like growth factor 1	hspace10pt	CXCL10	chemokine (C-X-C motif) ligand 10
LPS	interferon regulatory factor 6	hspace10pt	CXCL11	chemokine (C-X-C motif) ligand 11
FLAGELLIN	component of the bacterial flagellar filament	hspace10pt	CXCL12	chemokine (C-X-C motif) ligand 12
ODN2006	synthetic oligonucleotides	hspace10pt	CCL2	chemokine (C-C motif) ligand 2
IMIQUIMOD	imidazoquinoline amine analogue to guanosine	hspace10pt	CCL4	chemokine (C-C motif) ligand 4
HKSA	preparation of Listeria monocytogenes	hspace10pt	CCL5	chemokine (C-C motif) ligand 5
SSRNA40	20-mer single-stranded RNA oligo	hspace10pt	CCL11	chemokine (C-C motif) ligand 11
OSM	oncostatin M	hspace10pt	FGF2	fibroblast growth factor 2 (basic)
PAM3CSK4	synthetic tripalmitoylated lipopeptide	hspace10pt	FGF4	fibroblast growth factor 4
POLYIC	synthetic analog of double-stranded RNA	hspace10pt	FGF23	fibroblast growth factor 23
BMP2	bone morphogenetic protein 2	hspace10pt	FSTL1	follistatin-like 1
BMP4	bone morphogenetic protein 4	hspace10pt	WNT	wingless-type MMTV integration site family, mem.1
BMP7	bone morphogenetic protein 7	hspace10pt	INHBA	inhibin, beta A
TNF	tumor necrosis factor	hspace10pt	ADIPOQ	adiponectin, C1Q and collagen domain containing
TNFSF10	tumor necrosis factor lig. superfamily, 10	hspace10pt	DEFB1	defensin, beta 1
TNFSF11	tumor necrosis factor lig. superfamily, 11	hspace10pt	FCER2	Fc fragment of IgE, low affinity II, receptor for (CD23)
TNFSF12	tumor necrosis factor lig. superfamily, 12	hspace10pt	EPGN	epithelial mitogen homolog
TNFSF14	tumor necrosis factor lig. superfamily, 14	hspace10pt	EREG	Epiregulin
TNFRSF11B	tumor necrosis factor rec. superfamily, 11b	hspace10pt	CSF3	colony stimulating factor 3 (granulocyte)
TNFSF13B	tumor necrosis factor lig. superfamily, 13b	hspace10pt	GDF5	growth differentiation factor 5
CD40LG	CD40 ligand	hspace10pt	GDNF	glial cell derived neurotrophic factor
IL1A	Interleukin 1 alpha	hspace10pt	CSF2	colony stimulating factor 2
IL1B	Interleukin 1 beta	hspace10pt	LEP	Leptin
IL2	Interleukin 2	hspace10pt	CSF1	colony stimulating factor 1 (macrophage)
IL3	Interleukin 3	hspace10pt	MIA	melanoma inhibitory activity
IL4	Interleukin 4	hspace10pt	PDGFB	platelet-derived growth factor alpha polypeptide
IL6	Interleukin 6	hspace10pt	RLN2	relaxin 2
IL7	Interleukin 7	hspace10pt	RLN3	relaxin 3
IL8	Interleukin 8	hspace10pt	TGFB1	transforming growth factor, beta 1
IL10	Interleukin 10	hspace10pt	LTA	lymphotoxin alpha (TNF superfamily, member 1)
IL12	Interleukin 12	hspace10pt	VEGF121	vascular endothelial growth factor
IL13	Interleukin 13	hspace10pt	FST	follistatin
IL17A	Interleukin 17 alpha	hspace10pt	NOG	noggin
IL19	Interleukin 19	hspace10pt		

Table 2.1: Stimuli abbreviations. For cytokines, the Hugo nomenclature was been used.

Ligand selection

66 out of the 81 inducers did not lead to a significant activation of a key phosphoprotein (Figure 2.3a). A logic-based simulation of the generic pathway (see Figure 2.4) showed that in 56 out of the 66 unresponsive stimuli, hepatocytes should have responded since at least one of the 14 key phosphoproteins are downstream of the ligand in the generic topology. For 10 out of the 66 unresponsive stimuli, we cannot observe their activity because we miss measurements of all downstream signals (see the following discussion on observability). In order to select all active

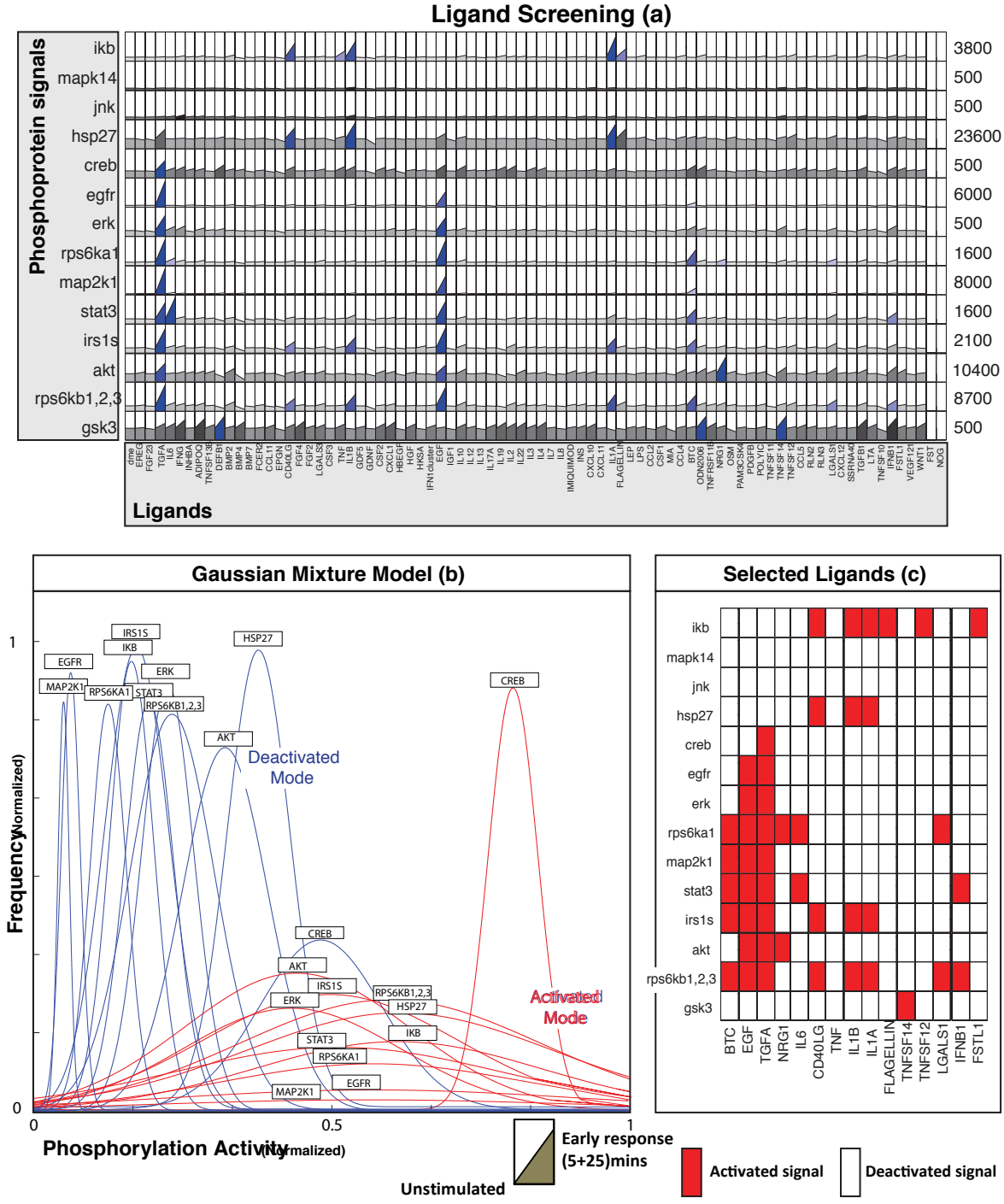


Figure 2.3: Ligand selection: (a) 81 stimuli are screened and data acquired. (b) A Gaussian mixture model is used to classify signals between activated and inactivated. (c) Ligands that activated at least one of the measured key phosphoprotein signals are selected.

ligands, data acquired from the first step were normalized and a Gaussian Mixture Distribution (GMD) model was fitted to discretize data to 0 or 1 (OFF or ON respectively) as detailed in the

next paragraph and plotted in Figure 2.3b. The algorithm leads to the selection of 15 stimuli based on their strong or marginal activation of at least one key phosphoprotein, termed "active responders" (Figure 2.3c).

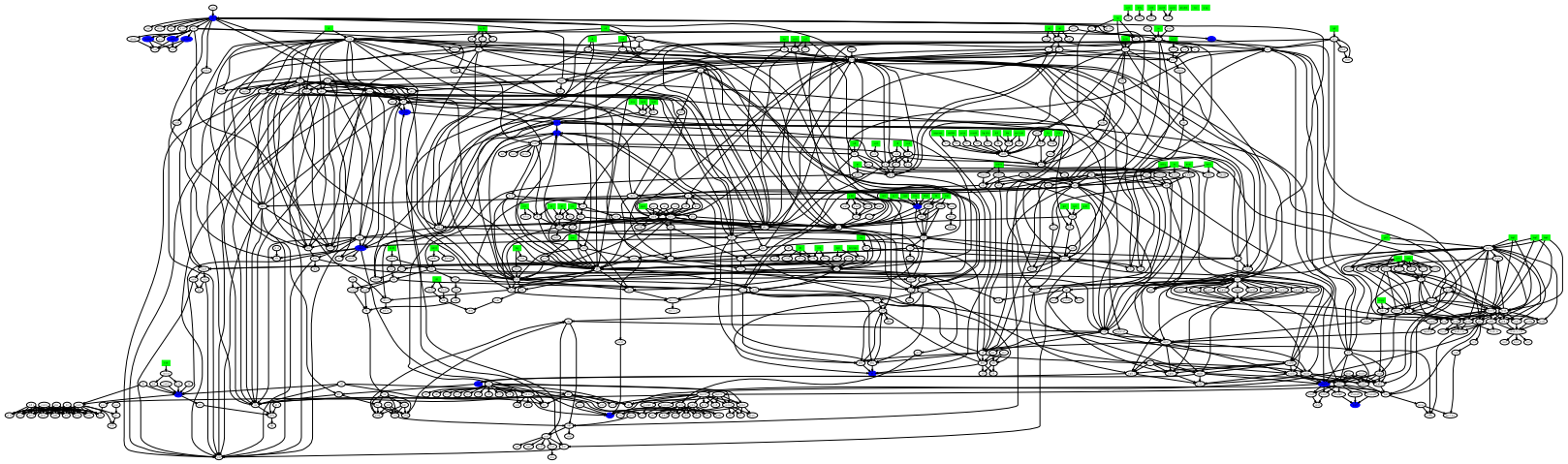


Figure 2.4: Canonical pathway constructed from literature. Numbers 533 nodes, 1064 reactions and serves as starting point for the analysis described in this section.

Primary human hepatocytes responded to IL6, FLAGELLIN, TNF, IFNB1, TGFA, TNFSF14, TNFSF12, IL1A, EGF, IL1B, NRG1, BTC, CD40LG, DEFB1 and LGALS1. As positive control observations, TGFA, EGF, NRG1 and BTC being EGFR ligands activated signals related to pro-growth pathways, namely AKT, MAP2K1 (MEK), ERK, RPS6KB1,2,3 and RPS6KA1. IL1A, IL1B and TNF activated signals lying in inflammatory pathways like IKB and HSP27; IL6 activated STAT3. Moreover, less known players were identified: CD40LG, a member of the TNF superfamily, had a significant effect on IKB, HSP27, IRS1S and RPS6KB1,2,3; FLAGELLIN, a TLR5 ligand activated IKB and HSP27; LGALS1 (also referred as GALECTIN-1) activated RPS6KA1, RPS6KB1,2,3 and had a medium size effect on STAT3, IKB and IRS1S; TNFSF12 and TNFSF14 also belonging to the TNF superfamily yielded medium size responses from RPS6KB1,2,3, IKB, IRS1S, HSP27 GSK3 and AKT. This set of ligands together with IFNB1 which in spite of having medium effects on very few of the signals (STAT3, RPS6KB1,2,3) is known to play a major role in liver homeostasis, were included in the follow-up experiment.

Gaussian bimodal distribution for ligand selection and data normalization

The Gaussian Mixture Distribution (GMD) was used at the core of the ligand selection procedure. The measured phosphorylation values of each of the 14 signals were used to fit a GMD consisting of 2 modes, the deactivated and the activated mode. Subsequently for each measured phosphorylation value, the probability distribution function (PDF) was evaluated respective to both modes: if the PDF respective to the first mode (deactivated) was greater than the PDF respective to the second mode (activated) then the signal was considered to be ON (otherwise OFF). This way the experimental dataset was discretized to 0 and 1 (OFF and ON). Only stimuli that activated at least one of the signals were included in the follow-up steps of the proposed methodology. The Statistics Toolbox of Matlab was used for this analysis and more specifically the `gmdistribution.fit()` and `pdf()` functions.

A different normalization procedure was used for the combinatorial dataset, since the large number of experiments performed with a strong activator and in the presence of inhibitors led to very dispersed distributions that complicated data discretization. The normalization scheme used is similar to the one introduced in [13]. The fold change increase of the signal (before and after stimulation) is evaluated and passed through a hill function filter to be scaled from 0 to 1. In more detail:

$$\hat{x}_j^{k,m} = \frac{(\frac{x_j^{k,m}\,|_{t=1}}{x_j^{k,m}\,|_{t=0}})^n}{p^n + (\frac{x_j^{k,m}\,|_{t=1}}{x_j^{k,m}\,|_{t=0}})^n} \tag{2.12}$$

where,

- $\hat{x}_j^{k,m}$ is the normalized measured value of species j in experiment k,
- $x_j^{k,m}\,|_{t=0}$ is the unstimulated measured value of species j in experiment k,
- $x_j^{k,m}\,|_{t=1}$ is the stimulated measured value of species j in experiment k,
- n is the hill coefficient, herein $n = 4$,
- p is a user defined threshold representing the fold change increase beyond which the signal is considered activated. In the analysis presented herein $p = 2$, implying a 2-fold increase or greater must take place for a signal to be considered activated.

Design and execution of the combinatorial experiment

In the second phase of the experimental procedure, the 15 stimuli, found to have strong effects on the 14 key phosphoproteins in primary hepatocytes, are introduced in a combinatorial manner in order to build the training dataset that will constrain the generic topology. The execution of the combinatorial experiment is taking place on the 3rd day of the procedure with the same batch of primary human hepatocytes plated in day one and stored in the incubator for one additional day to be used in this 2nd phase of experiments.

As mentioned in previous studies [51], the construction of optimal topologies entails biases due to (i) experimental design, (ii) the signals that are selected to be measured, and (iii) the number of perturbations via inhibitors imposed in the network. Finding an optimal experimental design for maximally constraining a generic topology is not a trivial task. Pathway controllability and observability, assay cost limitations, assay performance, reagent performance, and experimental compatibility with 12 $\times$ 8 well plate layout should be taken into consideration. Based on previous assessment of model sensitivity to changes in experimental design [51], in this study we created a dataset that is experimentally feasible and includes all single treatments, the majority of combination of two together, and the presence or absence of two selective inhibitors, MEKi and PI3Ki. Although combinations of more than two ligands may uncover more complex cross-talks and synergistic effects, if done rigorously, the number of experiments needed to include all the different combinations would increase significantly, consequently increasing both the experimental and computational costs. Instead of incorporating more combinatorial patterns we chose to increase the number of interrogated ligands, since the main purpose of this section is the broad, qualitative, less detailed study of the signaling network in human hepatocytes. The dataset is plotted in figure 2.5.

The complexity of the combinatorial dataset in Figure 2.5 offers little potential for manual inspection. Nevertheless, a few basic trends can be identified. MEKi and PI3Ki have blocked their nominal targets: the activation of ERK is clearly inhibited by MEKi under any treatment and PI3Ki blocks the activation of AKT under any treatment. The single treatments included in the second dataset are grouped together and aligned with the respective treatments from the first dataset, so a direct comparison can be made: as positive controls, all the pro-growth stimuli have activated AKT, MAP2K1 (MEK), ERK, RPS6KB1,2,3 and RPS6KA1, as in the first dataset and the pro-inflammatory ligands activated MAPK14, HSP27 and JNK. Despite the fact that qualitatively most ligands performed the same in the 1st and 2nd dataset, few quantitative differences can also be observed: (i) the JNK and MAPK14 signals clearly respond to IL1A, IL1B and TNF in contrast to the first dataset that shows marginal—if any—JNK and MAPK14 activation. (ii) TGFA, EGF and BTC show strong activation in both phases but with some increased activation levels in some of the signals in the second phase (MAPK14, JNK, HSP27). From our experience, mismatches in measurements of this kind are common even when both experiments are performed under the exact same conditions, with the same batch of reagents, and on the same batch of cells from the same donor. What is important for the optimization process is the encapsulation of the clear trends of the signals, since small levels of noise are countered by internal replicates of the dataset and the normalization procedure.

Clustering

Clustering algorithms can be utilized to provide powerful insight when such a large panel of ligands is screened. By evaluating the Euclidean distance (in 14) amongst the 81 stimuli, we identify the ones that exhibit similar profiles to prototypical anabolic or catabolic behavior. The second dataset is clustered hierarchically respective to both axes (stimuli and signals).

Clustering is performed using TM4 MeV (http://www.tm4.org/mev/). Data are reconfigured

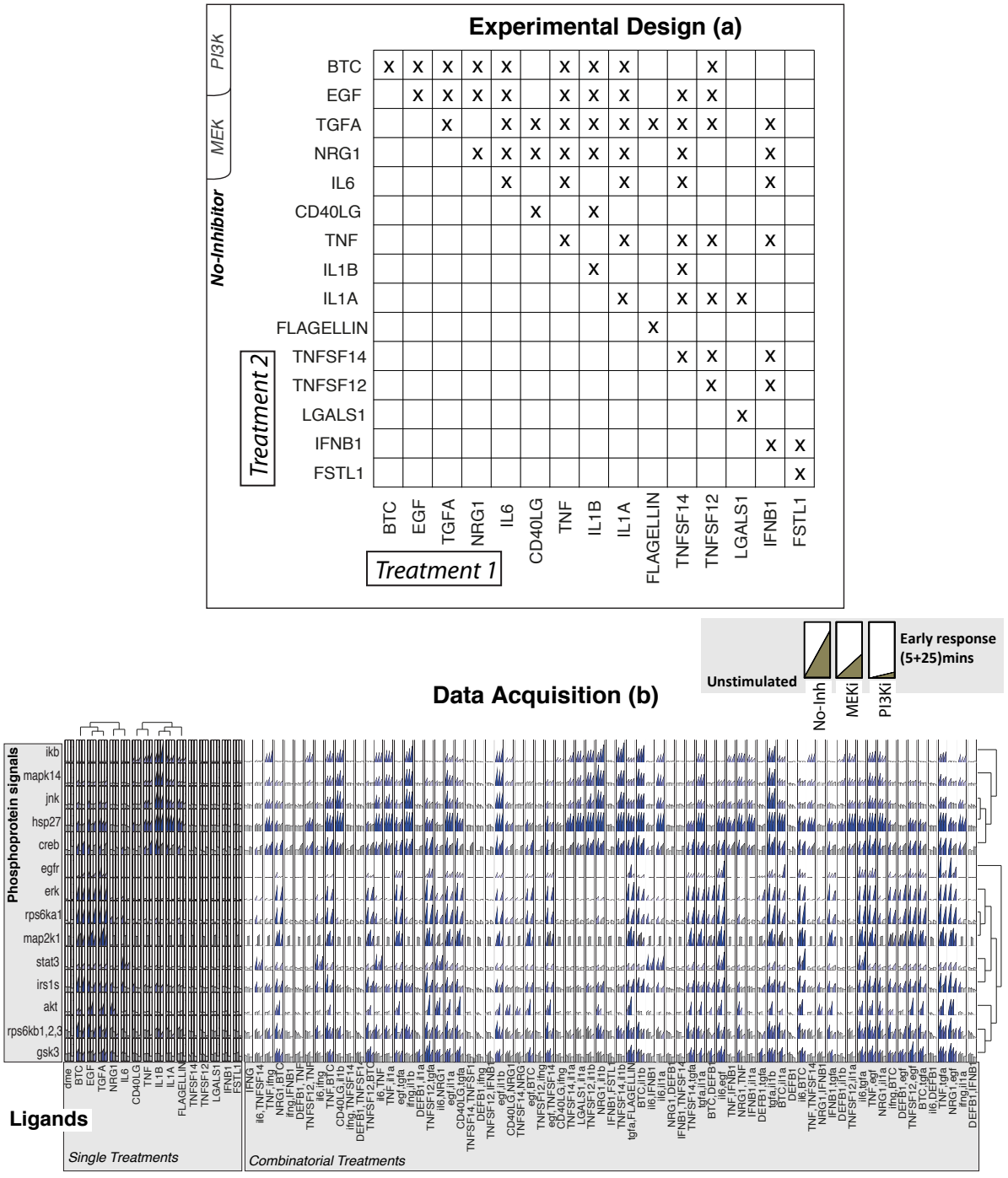


Figure 2.5: Combinatorial experiment: (a) design of the combinatorial experiment. The illustrated stimuli treatments are repeated 3 times: without any inhibitor present, with a MEK inhibitor (MEKi), and with a PI3K inhibitor (PI3Ki). (b) Phosphoproteomic data are acquired acquisition and plotted in DataRail.

into a 2-d matrix and imported to MeV. Hierarchical clustering is performed respective to both axes according to the Euclidean distance and rows and columns were reordered automatically. Other metrics were also used, but offered little insight beyond what had already been acquired using the Euclidean distance, therefore are not discussed more. As far as the stimuli axes are concerned, they seem to be separated into 2 main clusters, the 1st cluster includes IL6, BTC, EGF, TGFA and NRG1, the 2nd cluster includes CD40LG, TNF, IL1B, IL1A and FLAGELLIN, while IFNB1, FSTL1, LGALS1, TNFSF12, TNFSF14 yield marginal response, therefore are left out of all clusters and the analysis that follows. In a similar fashion signals are separated into 2 main clusters. The first one consists of EGFR, ERK, MAP2K1, RPS6KA1, STAT3, IRS1S, AKT, RPS6KB1/2/3 and GSK3. The second cluster consists of IKB, MAPK14, HSP27 and CREB. Interestingly, the ligands of the first cluster share some pro-growth qualities by activating signals in the first cluster (pro-growth signals), whereas the second stimuli cluster consisting of pro-inflammatory stimuli activates mostly the second signal cluster (pro-inflammatory signals) (see also Figure 2.5). Such a procedure reveals the inflammatory role of FLAGGELIN and CD40LG, two less-known stimuli that cluster closely to prototypical inflammatory ones. For a clustering analysis of the full combinatorial dataset see Figure 2.6.

Construction of generic pathway

The generic pathway presented in Figure 2.2 (and in more detail in Figure 2.4) is constructed downstream of 81 receptors of interest and in the neighborhood of 14 measured proteins. Several online databases were queried, but most of the reactions were obtained from KEGG (http://www.genome.jp/kegg/) and Ingenuity (http://www.ingenuity.com/). Conflicting reports on protein interactions were handled by manual search of PubMed database and selection of the most cited alternative. Addressing the problem of combining pathways from various databases, the HUGO Gene Nomenclature Committee database (http://www.genenames.org) was used to ensure standardized protein names. The pathway was built and visualized using Graphviz (http://www.graphviz.org/) following manual curation. The resulting topology is made of more than 500 species and 1000+ reactions (see Figure 2.2).

Pathway pre-processing: controllability, observability and feedback loops

Observability, controllability

Out of the 81 stimuli interrogated originally, only 15 were found to have strong effects on any of the key phosphorylation signals. Thus, a large portion of the generic pathway can be removed before the optimization procedure that is either non- perturbed (by the 15 cytokines) or non-observable [13]. Using Warshall's algorithm for transitive closure, as implemented in CellNetOptimizer (CellNOpt) toolbox [13], the connectivity of every node to the 14 signals and 15 cytokines is examined. If a pathway exists leading from at least a perturbed receptor to an arbitrary node A, then A is considered controllable. If a pathway exists leading from A to at least a signal, then A is considered observable. The observable and controllable subset of the generic topology is plotted in Figure 2.2 (and in more detail in Figure 2.7) numbers 177 species and 365 reactions and is the network to be optimized by the ILP. Note that the reduction of the network also results in a significant reduction of the optimization formulation which makes it computationally more tractable.

feedback loops

The generic topology numbers a large amount of feedback loops that have to be identified before the optimization procedure. The presence of a feedback loop implies that species can affect their own activation state. ODEs (or other frameworks for dynamic modeling) support this kind of

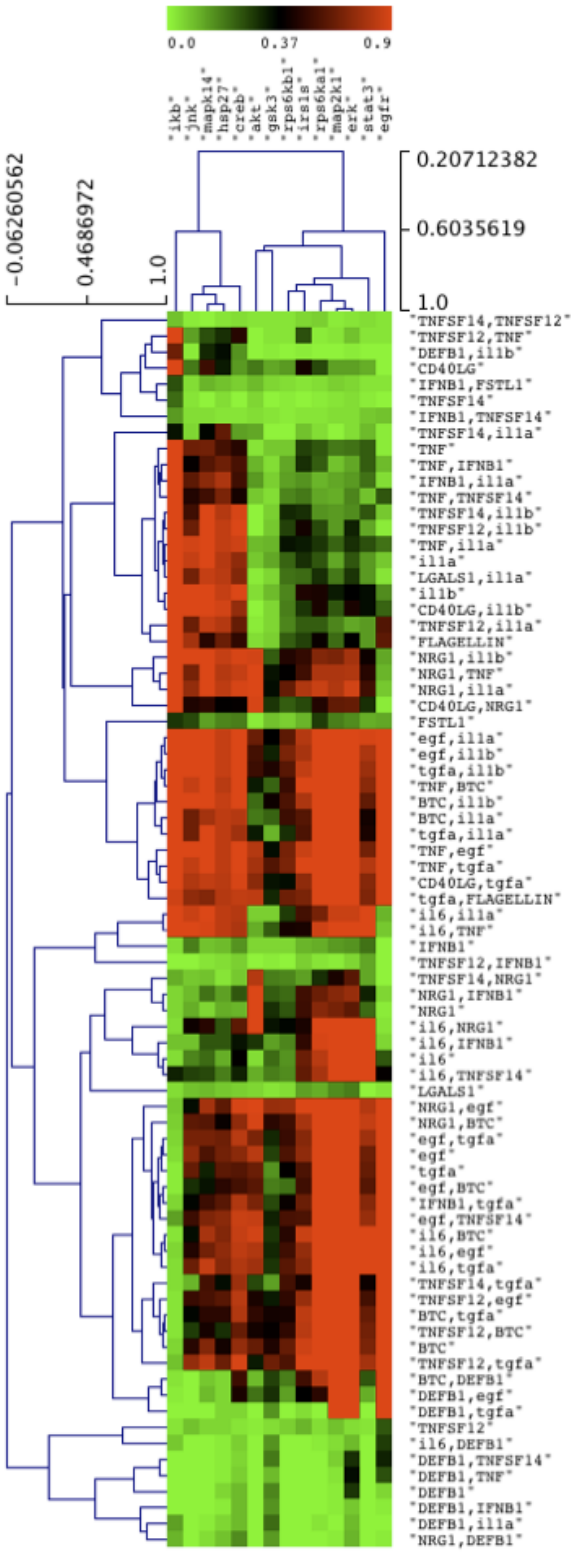


Figure 2.6: Hierarchical clustering of the full combinatorial dataset.

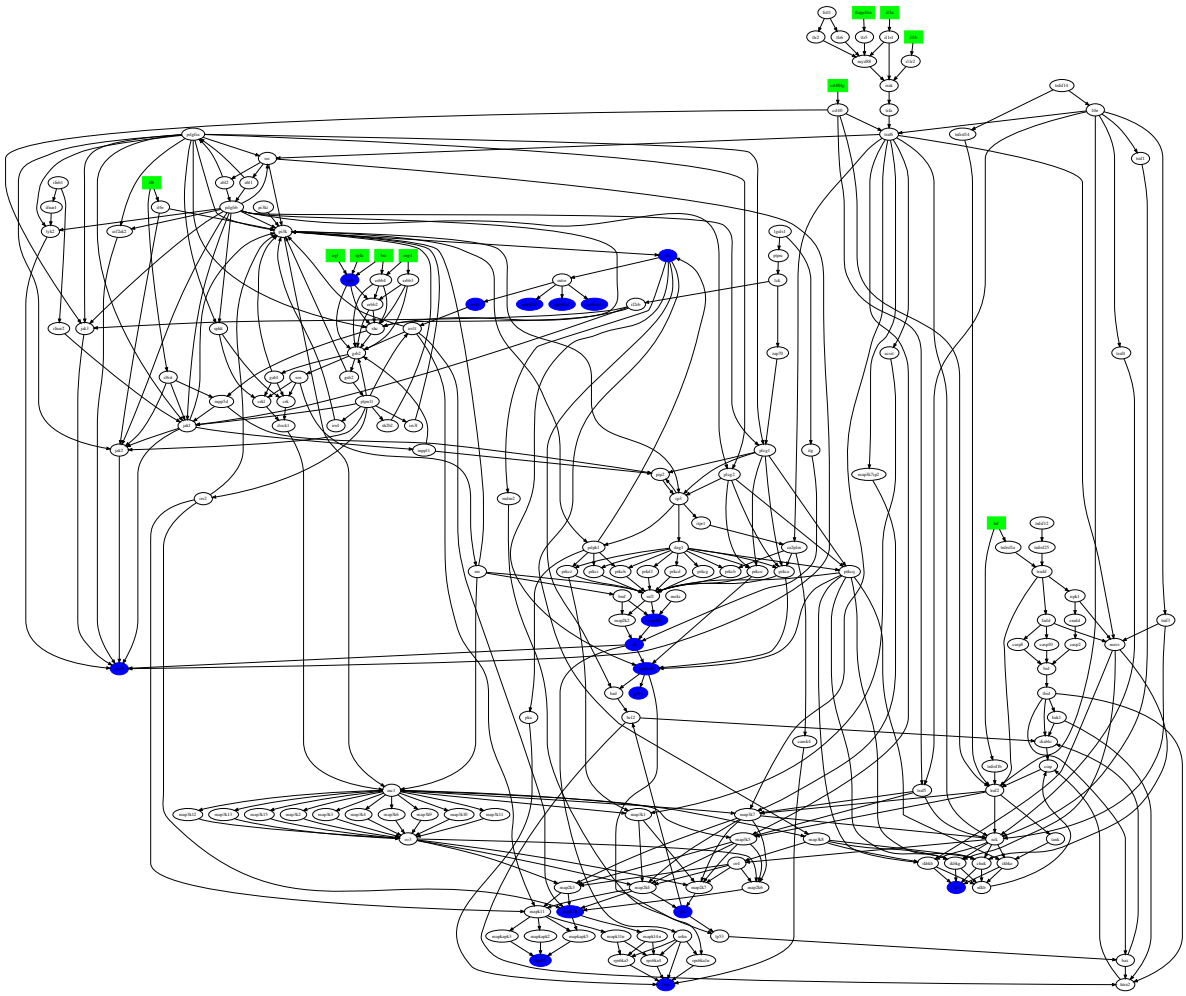


Figure 2.7: Observable-controllable part of the canonical pathway. Numbers 177 species, 365 reactions and is the pathway to be optimized by the ILP algorithm.

dependencies, since the feedback occurs in a subsequent time point. The Boolean approach used herein utilizes only "early response data" and thus assumes only a feed-forward approach where feedback loops are not considered. Thus, the feedback loops present in the generic topology are identified using a Depth-First Search algorithm and additional constraints are crafted to prohibit the ILP from including them in the solution

Without loss of generality we assume a positive feedback loop numbering n species.

$$A_1 \rightarrow A_2 \rightarrow \ldots \rightarrow A_n \rightarrow A_1$$

Equations (2.4) to (2.4) in section 2.2, here repeated for consistency, are:

$$z_i^k \leq x_j^k, \quad i = 1, ..., n_r, \quad k = 1, ..., n_e, \quad j \in R_i$$

$$z_i^k \leq 1 - x_j^k, \quad i = 1, ..., n_r, \quad k = 1, ..., n_e, \quad j \in I_i$$

$$z_i^k \geq y_i + \sum (x_j^k - 1) - \sum x_j^k, \quad i = 1, ..., n_r, \quad k = 1, ..., n_e$$

$$x_j^k \geq z_i^k, \quad i = 1, ..., n_r, \quad k = 1, ..., n_e, \quad j \in P_i$$

where, R is the set of reactants in reaction i, I is the set of inhibitors in reaction i, P is the set of products in reaction i, z_i^k expresses the activation of reaction i in experiment k (assumes only Boolean values)

(2.4), (2.5) and (2.6) imply that reaction i will take place if at least a reactant is present and no inhibitors are present, (2.7) implies that if reaction i takes place all products are formed. If a loop occurs in the initial topology $A_1 \rightarrow A_2 \rightarrow \ldots \rightarrow A_n \rightarrow A_1$, then starting from an arbitrary species in the loop A1 and setting, $x_{A1} = 1$, (2.4),(2.5),(2.6)$=>z_{An->A1} = 1$, (2.7)$=>x_{An} = 1$, $x_{An-1} = 1, \ldots, x_{A2} = 1, x_{A1} = 1$. In other words, species A1 activates itself which is not desirable.

To prevent that, all feedback loops were identified beforehand using a custom depth-first search (DFS) algorithm. Each loop is characterized by n_L reactions with indices in the index set $L = \{1, ..., n\}$ For each such loop the following constraint was added prohibiting the ILP from including the loop in the solution:

$$\sum_{i \in L} y_i < n_L \tag{2.13}$$

The optimizer is thus forced to delete at least one of the reactions in the loop.

Pathway optimization

The optimization procedure is built around an Integer Linear Programming formulation, that receives as input a generic topology and a phosphoproteomic dataset, and by altering the topology that contradicts experimental observations delivers a signaling pathway that fits the experimental dataset with the least possible discrepancies [23]. Pathway optimization relies on the minimization of a two term objective function, the first term penalizes the experiments-topology mismatch and the second term either penalizes or rewards the size of the pathway [13, 23]. Penalty weights are adjusted by assuming very small values so that the goodness of fit is always prioritized over the size of the pathway. In contrast to our previous procedures [23] that utilize the minimization of pathway size as a method to narrow down the solution pool, we consider additional pathway solutions of larger size that may bear strong biological significance and should be conserved. Therefore, in the current analysis we apply the ILP formulation under 2 different settings: first by using the standard procedure with small positive size weights, we minimize the size of the pathway; second by using small negative reaction weights we maximize the size of the pathway while retaining the prioritization over the goodness of fit. The dual-settings approach we adopt aims to obtain a superset of all possible pathway solutions, i.e., pathways with more nodes and edges but with the same match of the experimental data. Additionally, we provide an indicative solution with the minimum number of reactions. The ILP formulation was solved by GUROBI (Gurobi library version 3.0.1. Houston, Texas: Gurobi Optimization, Inc., http://www.gurobi.com/) through GAMS and the optimization results are illustrated in Figure 2.8. Dashed lines correspond to the "negative weights"-solution and the bold ones to the minimum-size solution. For a better assessment of the ILP performance and differences between positive and negative size weights see also section 2.2.

Out of the original 365 reactions present in the reduced canonical pathway, 204 were removed by the ILP formulation (Figure 2.8) for the maximum-size solution (superset has 161 reactions—Figure 2.8, "dashed lines"). In the minimum-size solution only 53 reactions are included. As positive control observations, the modular activation patterns described previously are conserved. The pro-growth stimuli, namely EGF, TGFA, NRG1, BTC and IL6, signal through similar, partially overlapping pathways and activate the first signal cluster consisting of EGFR, AKT, MAP2K1, ERK, STAT3 and RPS6KA1. On the other hand TNF, CD40LG, IL1A, IL1B and FLAGELLIN signal through TRAF6, TRAF2 and the MAPKs and activate

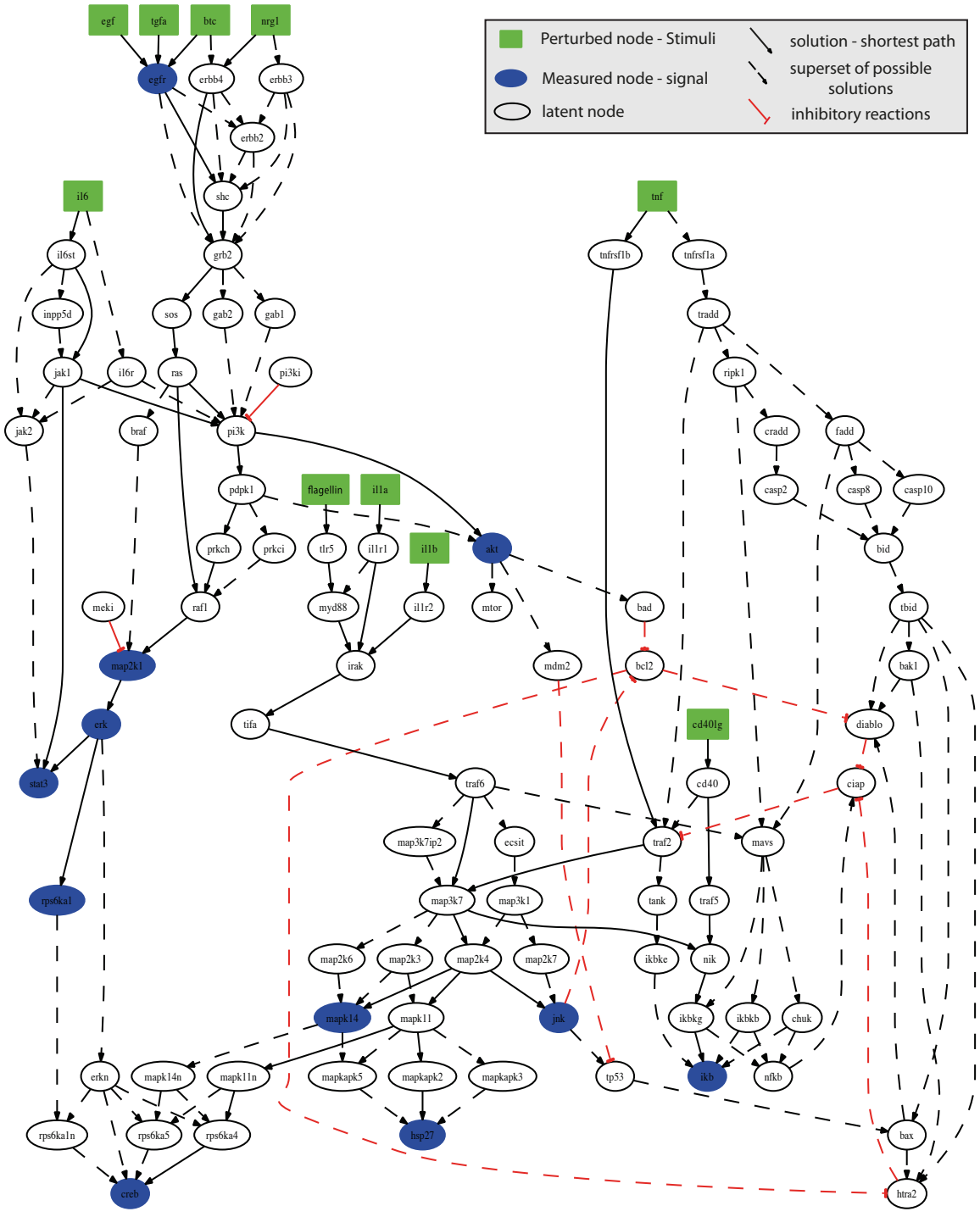


Figure 2.8: Optimized pathway: the optimized pathway consists of the reactions and nodes conserved by the ILP formulation.

MAPK14, IKB, JNK, HSP27 and CREB. RPS6KB1/2/3, GSK3 and IRS1S, although seemed to respond to pro-growth stimuli, are removed from the pathway, since their basal level is relatively

high implying a low signal to noise ratio and causing the fold change to drop below 2.0 (that is the threshold for considering a signal activated). In more detail, concerning the EGFR pathway: NRG1, BTC, EGF and TGFA signal via GRB2 to RAS and then to (i) PI3K - AKT and to (ii) RAF - MAP2K1 - ERK - RPS6KA - CREB, ERK - STAT3. Another branch downstream of IL6 signals via JAK1 to (i) PI3K - AKT and to (ii) STAT3. Concerning the pro-inflammatory ligands (IL1A, IL1B, FLAGELLIN, CD40LG and TNF): IL1A, IL1B and FLAGELLIN signal through TRAF6 to MAP3K7 and then to (i) NIK - IKB, and to (ii) MAP2K4 - JNK, MAP2K4 - MAPK14, MAP2K4 - MAPK11 - HSP27, MAPK11 - RPS6KA4 - CREB. CD40LG signals through TRAF5 to NIK and activates IKB. TNF signals through TRAF2 to MAP3K7 and activates MAPK14, CREB, HSP27, JNK and IKB. It is clear that minimum-size solution (bold lines) and superset of possible solutions (dashed lines) serve the exact same functionality in terms of connecting stimuli to signals. Their sole difference is that the superset of possible solutions includes, apart from the shortest, all alternative paths for achieving that connectivity.

The optimization procedure caused the fitness error to drop from 31% to 7%, validating that the canonical pathway alone could not capture the signal transduction mechanisms of the specific cell type. The ILP formulation managed to fit the basic trends of the experimental dataset and construct a hepatocyte- specific signaling pathway. See following paragraph for a more detailed inspection of fitness error, before and after the optimization procedure.

Inspection of fitness error before and after the optimization procedure

The experiments-topology mismatch is illustrated in Figure 2.9, before and after optimization procedure. The ILP formulation has decreased the fitness error from 31% to 7%. As illustrated in Figure 2.9, there are 2 main areas where the error has been removed drastically : (i) measurement of pro-inflammatory signals (IKB, MAPK14, JNK, HSP27, IKB, CREB) under pro-growth treatments (BTC, EGF, TGFA, NRG1, IL6) and (ii) measurement of pro-growth signals (ERK, RPS6KA1, MAP2K1, AKT, IRS1S, STAT3) under pro-inflammatory treatments (TNF, IL1A, IL1B, CD40LG, FLAGELLIN). Unresolved fitness error is identified mostly under FLAGELLIN, because of partial activation of MAPK14, JNK, HSP27 and CREB. According to the generic topology FLAGELLIN pathway overlaps with the rest of the pro-inflammatory stimuli (IL1A, IL1B, TNF and CD40LG) that fully activate the above mentioned signals. Therefore, minimization of the objective function implies fitting IL1A, IL1B, TNF, CD40LG by including paths to MAPK14, JNK, HSP27 and CREB, and misfitting FLAGELLIN. The same can be observed for IL6 induced AKT activation.

2.3.3 Conclusions

In this section, we propose an approach capable of interrogating signal transduction downstream of 81 receptors of interest and subsequently, constructing a detailed view of the signaling pathway for the responsive part of the network, numbering 15 receptors, 177 species and 365 reactions, all within a 3-day time frame. In contrast to MS-based techniques, the fast turnaround time of the xMAP technology allowed an adaptive approach that learns from a functional phosphoproteomic screen to define the most appropriate combinatorial experiments for pathway construction. The 3-day approach is performed on the same batch of cultured primary cells, thus reducing cell differentiation and eliminating donor-to-donor variability. On the computational front, the ILP formulation solved by Gurobi, a state of the art commercial solver, has successfully negotiated the optimization of a canonical pathway to best fit the functional characteristics of the interrogated cell line.

In contrast to most approaches applied today that focus on the well-studied and the most-cited pathways, in this project we rely on a functional screen of a large library of stimuli to select

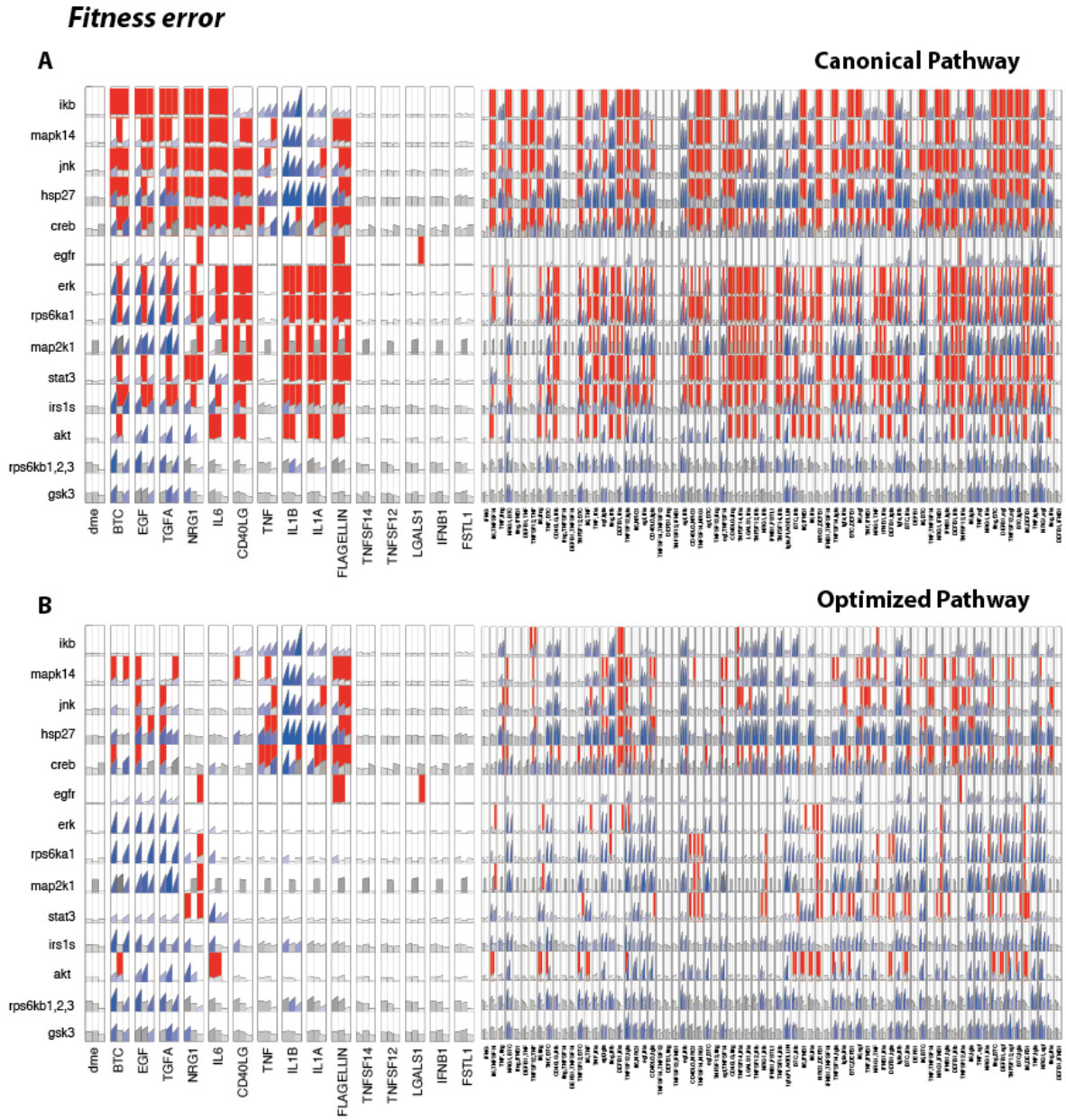


Figure 2.9: Fitness error before and after the optimization procedure. (a) Fitness error before the optimization procedure. (b) Fitness error after the optimization procedure

the most appropriate inducers of hepatocyte intracellular activity. As expected, most ligands we found correspond to well-known and highly cited "prototypical" players for liver homeostasis but surprisingly, we also uncover poorly and unreported stimuli that induced strong intracellular responses. Clustering analysis combined with literature search was able to uncover the potential role of those stimuli due to their functional similarity with prototypical ones. On the other hand, our literature-unbiased approach for ligand selection cannot identify all active players for 4 main reasons: (i) it is based on key measured signals that might not belong to the downstream cascade of the activator (e.g., WNT signaling could not be detected due to the absence of a beta-Catenin signal); (ii) the phosphorylation events are below the limits of detection of the

assay; (iii) the activation occurs in a different time point than the one measured (i.e. not at 5 and 25 minutes). (iv) Finally, since ligands are screened based on their individual effects on the measured proteins, synergistic effects cannot be captured at this point. Implying that players may be left out of the following steps of the analysis with strong, yet unidentified, synergistic effects. In fact, components showing only synergistic effects are lost at this point. As more phosphoprotein assays are added in the future and more time points are selected, more activators can be identified.

Based on our ligand screening approach we identified two under-reported compounds: Flagellin and CD40LG. Flagellin, a TLR5 inducer, was found to activate pro-inflammatory pathways, and function—as the clustering analysis revealed— in a similar way to TNF, IL1A and IL1B; enforcing the hypothesis that hepatocytes assume an active role in immuno-surveillance mechanisms. Moreover, CD40LG, a member of the TNF superfamily previously reported to be expressed in T-cells and affect B-cells and endothelial cells, is identified as having a clear effect on primary hepatocytes by activating the NFkB pathway. CD40LG induced activation of IKB and P38/ HSP27 cluster closely resembling the prototypical TNF, IL1B and IL1A inflammatory players, validating its role in liver inflammation. Even though the role of CD40L in hepatocyte physiology is not clear from the literature, a single study by Zhou et al. [52] identifies CD40 as a driving factor in the pathogenesis of fulminant hepatitis via its upregulation in Kupffer cells and hepatocytes. Together with the present results, CD40L should be considered a major inflammatory player in liver physiology. Apart from the identification of novel players, the high throughput screening revealed clustering of the stimuli in two main groups, the first one consisting of EGF, TGFA, NRG1, BTC and IL6, sharing some pro-growth qualities and the second one consisting of TNF, IL1A, IL1B, FLAGELLIN and CD40LG activating mostly pro-inflammatory pathways (IKB, MAPK14, HSP27, JNK). As expected, this modularity has led to the construction of partially overlapping pathways, namely, EGF, TGFA, BTC and NRG1 all signal through the EGFR and GRB2 adaptor protein, while IL1A, IL1B and FLAGELLIN signal through the IL1R and IRAK.

The ILP formulation conserved a subset of the reactions found in the canonical pathway, shaping a descriptive model of how signal propagates in primary hepatocytes. However, the construction of Boolean models of signaling pathways from phosphoproteomic data entails several validation issues related to optimal experimental design, model sensitivity to the generic topology, and calibration of the weights of the two objective terms. The major limitations in pathway optimization rely on two main factors: pathway controllability (e.g., how many stimuli and downstream inhibitors can be used), pathway observability, and experimental design. Amongst all, the number of the key phosphoprotein signals is of utmost importance for increased pathway observability. Surprisingly, just 14 phosphoprotein signals used in this study were sufficient to give a pathway coverage equal to 68.5% of the generic (685 out of the 1000 reactions can be observed from the generic topology when all stimuli are used). Further development of high throughput assays will enable us to increase the pathway coverage but also increase the quality of results since more than one signal in the same pathway increase the reliability of the pathway topology.

Construction of predictive models for the intracellular signaling cascades is the cornerstone for understanding cellular behavior. Here we presented an integrative approach to construct large scale signaling pathways, based on an adaptive collection of high throughput phosphoproteomic data and a priori knowledge of protein connectivity. State of the art optimization algorithms were coupled to an ILP optimization formulation to prune a canonical pathway of 177 nodes and 365 reactions to best fit the dataset at hand. Our approach sheds light into the complex signaling mechanisms of primary human hepatocytes and constitutes a proof-of-principle for construction of large pathways based on combinatorial data and a limited number of measured

signals.

2.4 Identification of drug effects via pathway alterations

In this work published in [23], we describe an unbiased, phosphoproteomic-based approach to identify drug effects by monitoring drug-induced topology alterations. This work was carried out in collaboration with Alexander Mitsos (at the time when this work was published an assistant professor in the Mechanical Engineering Department of MIT, Cambridge, MA, USA, currently a professor in RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany). The proposed approach uses an Integer Linear Programming (ILP) formulation to identify the effects of drugs on the cells signal transduction network. As a case study, 4 drugs for unresectable HCC were interrogated and previously unpublished off target effects were identified.

Abstract

Understanding the mechanisms of cell function and drug action is a major endeavor in the pharmaceutical industry. Drug effects are governed by the intrinsic properties of the drug (i.e., selectivity and potency) and the specific signaling transduction network of the host (i.e., normal vs. diseased cells). Here, we describe an unbiased, phosphoproteomic-based approach to identify drug effects by monitoring drug-induced topology alterations. With our proposed method, drug effects are investigated under diverse stimulations of the signaling network. Starting with a generic pathway made of logical gates, we build a cell-type specific map by constraining it to fit 13 key phopshoprotein signals under 55 experimental conditions. Fitting is performed via an Integer Linear Program (ILP) formulation and solution by standard ILP solvers; a procedure that drastically outperforms previous fitting schemes. Then, knowing the cell's topology, we monitor the same key phosphoprotein signals under the presence of drug and we re-optimize the specific map to reveal drug-induced topology alterations. To prove our case, we make a topology for the hepatocytic cell-line HepG2 and we evaluate the effects of 4 drugs: 3 selective inhibitors for the Epidermal Growth Factor Receptor (EGFR) and a non-selective drug. We confirm effects easily predictable from the drugs' main target (i.e., EGFR inhibitors blocks the EGFR pathway) but we also uncover unanticipated effects due to either drug promiscuity or the cell's specific topology. An interesting finding is that the selective EGFR inhibitor Gefitinib inhibits signaling downstream the Interleukin-1alpha (IL1a) pathway; an effect that cannot be extracted from binding affinity-based approaches. Our method represents an unbiased approach to identify drug effects on small to medium size pathways which is scalable to larger topologies with any type of signaling interventions (small molecules, RNAi, etc). The method can reveal drug effects on pathways, the cornerstone for identifying mechanisms of drug's efficacy.

2.4.1 Introduction on the characterization of drug's mode of action

Target-based drug discovery is a predominant focus of the pharmaceutical industry. The primary objective is to selectively target protein(s) within diseased cells in order to ameliorate an undesired phenotype, e.g., unrestrained cell proliferation or inflammatory cytokine release. Ideally, other pathways within the diseased cells, as well as similar phenotypes in other cell types, should remain unaffected by the therapeutic approach. However, despite the plethora of new potential targets emerged from the sequencing of the human genome, rather few have proven effective in the clinic [53]. A major limitation is the inability to understand the mechanisms or drug actions either due to the complex signal transduction networks of cells or due to the complicated profile of drug potency and selectivity.

Finding drug's targets is traditionally based on high-throughput in vitro assays using recombinant enzymes or protein fragments [54]. The main goal is to characterize the drug's biochemical activity (binding affinities that describe potency and selectivity) and depict them in drug-interaction maps [55]. In most cases, once the target(s) is known, the in vivo effect

on the signaling pathway is validated by measuring the drug's efficiency to inhibit the activity (usually measured as phosphorylation level [56]) of the downstream protein. However, beyond that measurement, little is know on how the rest of the signaling network is affected. In addition, in vivo drug effects can hardly be calculated from in vitro assays for several reasons: most kinase inhibitors are promiscuous [57], there is discrepancy between in vivo and in vitro binding affinities of drugs [58], and there is an additional discrepancy between in vivo binding affinities and in vivo inhibitor activity for the phosphorylation of downstream signals.

Here, we describe a significantly different approach to identify drug effects where drugs are evaluated by the alterations they cause on signaling pathways. Instead of identifying binding partners, we monitor pathway alterations by following key phosphorylation events under several treatments with cytokines. The workflow is presented in Figure 2.10. On the experimental front, using bead-based multiplexed assays [12], we measure 13 key phosphorylation events under more than 50 different conditions generated by the combinatorial treatment of stimuli and selective inhibitors. Based on the signaling response and an a-priori set of possible reactions (i.e. generic pathway), we create a cell-type specific pathway using an efficient optimization formulation known as Integer Linear Programming (ILP). This approach builds upon the Boolean optimization approach proposed in [13]. The ILP is solved using standard commercial software packages to guaranteed global optimality (within a user-defined, numerically small tolerance). To evaluate drug effects, we subject the cells with the same stimuli in the presence of drugs and we tract the alterations of the same key phosphorylation events. Then, we reapply the ILP formulation without a-priori assumption of the drug target, and we monitor the changes in the pathway topology with and without drug presence. To demonstrate our approach, we construct a generic map and optimize it to fit the phosphoproteomic data of the transformed hepatocytic cell lines HepG2. Then, we identify the effects of four drugs: the dual EGFR/ErbB-2 inhibitor Lapatinib [59], two potent EGFR kinase inhibitors Erlotinib [60] and Gefitinib [61], and the "dirty" Raf kinase inhibitor Sorafenib [62]. When our method is applied on those 4 drugs we find their main target effect and we also uncover several unknown but equally active off-target effects. In the case of Gefitinib, we find a surprising inhibition of cJUN in the IL1a pathway.

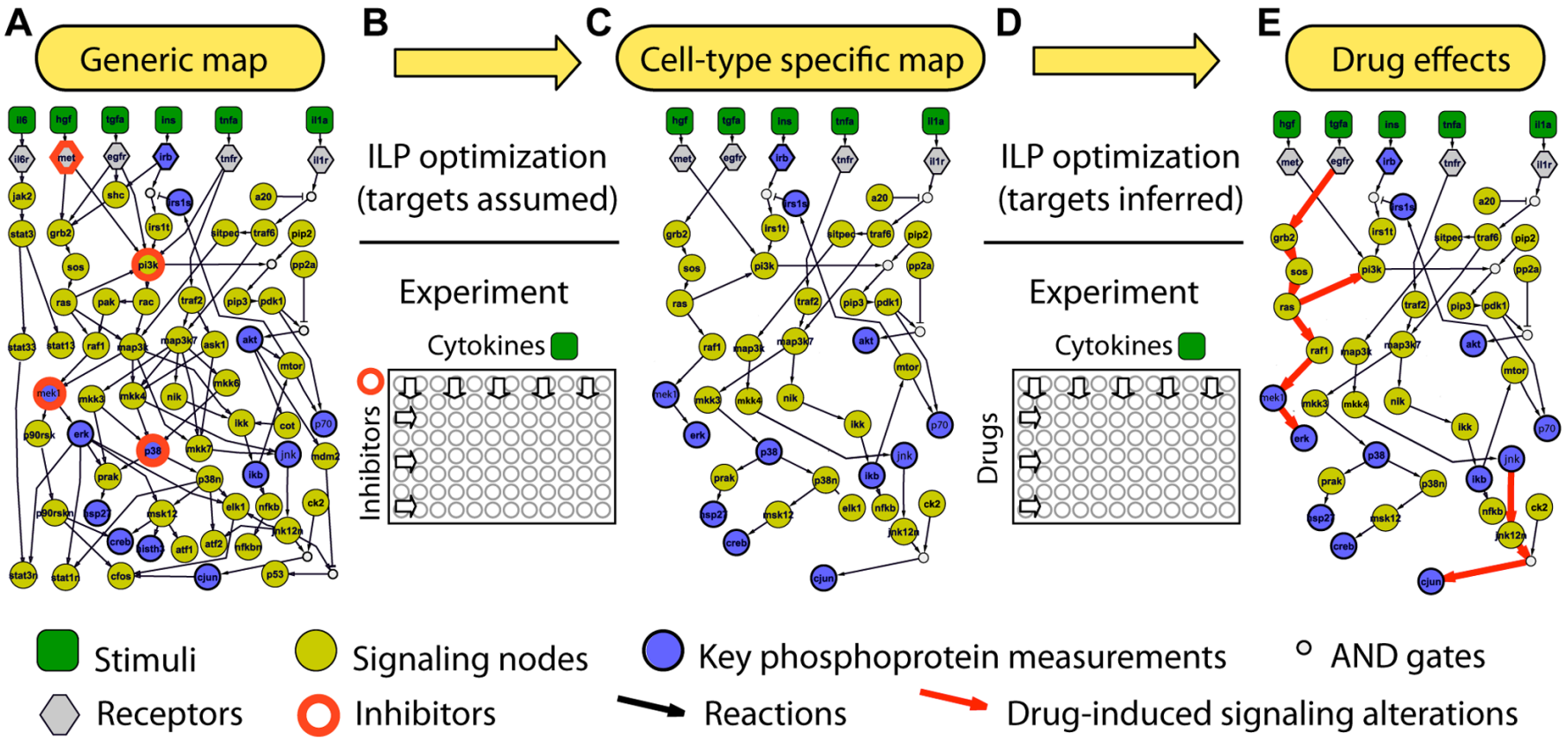


Figure 2.10: Experimental and computational workflow to assess drug effects. (A) A Boolean generic map is assempled from pathway databases and includes stimuli (green squares), key measured phosphoproteins (brown circles), and the neighboring proteins (yellow circles). (B) Cells are treated with a combination of cytokines and selective inhibitors (red circles) of known effects and an ILP formulation is used to fit the data to the Boolean pathway. (C) A cell-type specific pathway is constructed. (D) Cells are treated with a combination of cytokines and drugs –their effects are assumed unknown- and ILP is used for the second time to fit the drug-induced phosphorylation data. (E) Alterations of the the cell-type specific topology reveals drug effects (red arrows).

In contrast to previously developed techniques, our method is based on the actual effect on phosphorylation events carefully spread into the signaling network. Theoretically, it can be applied on any type of intracellular perturbations such as ATP-based and allosteric kinase inhibitors, RNAi, shRNA etc. On the computational front, our ILP-based approach performs faster and more efficient than current algorithms for pathway optimization [13] and can identify the main drug effects as well as unknown off-target effects in areas of pathways constrained between the activated receptors and the measured phosphorylated proteins. Our fast and unbiased characterization of modes of drug actions can shed a light into the potential mechanisms drug's efficacy and toxicity.

2.4.2 Results

Construction of phosphoproteomic datasets

High-throughput bead-based ELISA-type experiments using xMAP technology (Luminex, Texas, USA) are performed as briefly described in the Materials and Methods section and in [12]. We create two datasets: one for the construction of cell-type specific topology and another for the identification of the mechanisms of drug actions. To do that, HepG2s are stimulated in 10 different ways with combinatorial treatments with a diverse set of 5 ligands (TNFa, IL1a, HGF, INS, TGFa, and no stimuli) and either 4 highly selective inhibitors (PI3K, MEK, p38, cMET, and no inhibitor) or 4 commercial drugs (EGFR inhibitors Lapatinib, Erlotinib and Gefitinib, and the "dirty" inhibitor Sorafenib) (Figure 2.10b and 2.10d). For the purpose of this section, we refer to "inhibitors" as the compounds for which we know the target and we use them in a concentration capable to block ,95% of the downstream protein. Conversely, we refer to "drugs" as the compounds for which we assume no a-priori knowledge of their target. For each combination of cytokine and drug/inhibitor we collect cell lysates at 5 and 25 minutes. The two time points are pooled together in 1:1 ratio and the mixed lysates are used as an indicator of the "average early signaling response". For each treatment we measure 13 protein phosphorylations that we consider "key protein activities" (raw data in Figure 2.11). The key phosphorylation signals are chosen based on the availability of the reagents and quality controls performed at the early phases of the experimental setup [12]. The raw data (arbitrary fluorescent intensities) are normalized to fit logic models as described in [13] using a non-linear transformation that converts raw data into values between 0 and 1 where 1 corresponds to the fully activated state and 0 to no-activation. It has to be noted that logic-transformed data depends on what should be considered "protein activation" (transformed value 0.5), a criterion that is embedded in the transformation function and accounts for signal-to-noise limits, saturation of the detection scheme, and eliminates biases that could have been introduced by the variability of antibody affinities [13].

Generic pathway assembly and visualization

The generic pathway map is constructed in the neighborhood of the 5 stimuli and the 13 measurements. The ubiquitous presence of conflicting reports on pathway maps and alternative protein names makes this step a highly nontrivial one. We explored several pathway databases including STKE, Pathway Interaction Database, KEGG, Pathway Commons, Ingenuity, and Pathway Studio [63, 64]. Our limited intracellular protein coverage makes impractical the reduction of very large pathway datasets such as those found in Pathway Commons. Here, we create the initial topology from the union of canonical pathways found in Ingenuity (Redwood City, California) with subsequent manual curation.

A detailed description of Boolean representation of pathways can be found elsewhere [13, 32,

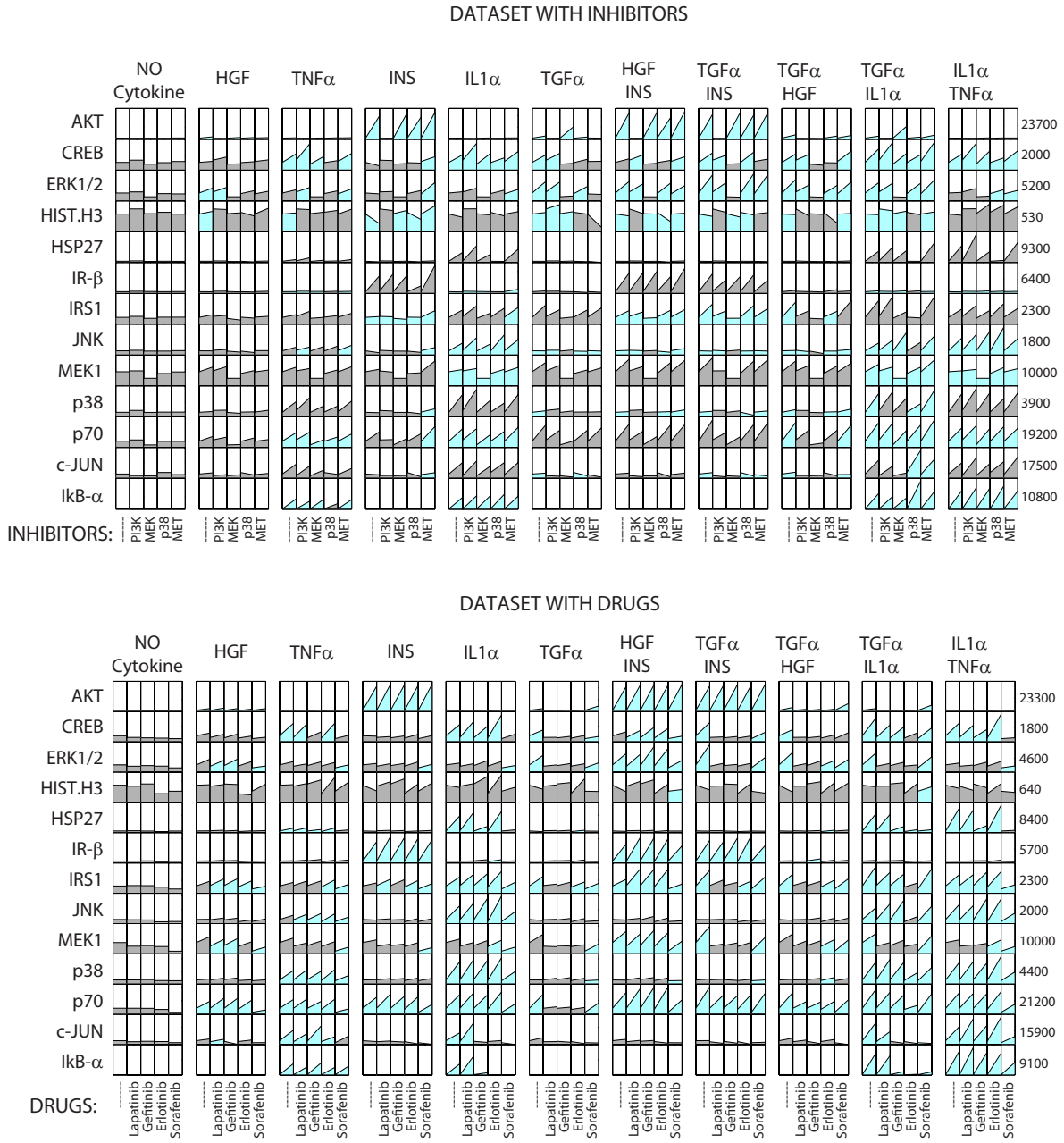


Figure 2.11: Raw data for the construction of the cell-type specific map and the evaluation of the drug effects. The signals in the Y-axis correspond to the measurements of the phosphorylated residues. Each column corresponds to cytokine or cytokine mix and each sub-column to the presence of an inhibitor or drug. The numbers to the left are the maximum values across all treatments measured as arbitrary fluorescent intensities.

33]. In this section as opposed to [13], the connectivity in our pathway (Figure 2.12, left panel) is represented with OR gates and only few connections (represented with small black circles in

Figure 2.12) require an AND gate. We are therefore not comparing OR vs. AND gates, but rather assuming our pathways to be "causal" graphs, and since there are a few AND gates we refer to it as Boolean model.

Construction of cell-type specific pathway via ILP formulation

The formulation for the optimal pathway identification is a 0–1 Integer Linear Program, i.e., an optimization problem with binary variables and linear constraints (see section 2.2). The optimizer picks values for the decision variables, such that the logical constraints are satisfied and the objective(s) optimized. The primary objective is to find an optimal pathway, i.e., a pathway that best describes a set of phosphoproteomic data under a given model (e.g. Boolean). A secondary objective is that the pathway is as small as possible, i.e., has as few connections as possible, such that the best-possible fit of the experiments is maintained. It is shown that some of the binary variables can be relaxed to continuous, without changing the feasible set.

The ILP is solved with the state-of-the-art commercial code (CPLEX) that guarantees minimal error between experimental data and the Boolean topology. The goodness of fit (percent error as described in section 2.2) was decreased from 36.7% on the generic map to 8.3% on the optimized map (Figure 2.12). The main source of error is the inability of TGFa to activate the IRS1_s (serine residue of IRS1) (see the red background on the IRS1 row at the bottom panel of Figure 2.12). This is a result of the infeasibility of the generic pathway to satisfy the activation of IRS1_s in a TGFa/IL1a-dependant but HGF/ INS-independent manner: TGFa activation of IRS1_s requires mTOR activation via AKT which the optimization algorithm removes to satisfy the inactivation IRS1_s by INS that shares the same path with TGFa. This example highlights the importance of multi-perturbations to better constrain the optimization formulation.

Figure 2.12 shows the optimized topology of HepG2s. Our ILP formulation uses two subsequently - imposed objective functions to remove reactions that do not fit the experimental data. During the optimization of the first objective the ILP formulation (A) keeps reactions that lead to phosphorylations of the key proteins and (B) removes reactions that lead to false protein activations. An example of the first case is the Insulin (INS)-induced AKT activation that is maintained via the INS $\rightarrow$ IRb $\rightarrow$ IRS1t $\rightarrow$ PI3K $\rightarrow$ PIP3 $\rightarrow$ PDK1 $\rightarrow$ AKT path (see INS to AKT path in Figure 2.12). An example of a removed reaction is the TNFR $\rightarrow$ PI3K reaction which is removed because there is no TNFa induced AKT activation (see TNFR $\rightarrow$ PI3K $\rightarrow$ ... $\rightarrow$ AKT in Figure 2.12). During the optimization of the secondary objective (see section 2.2), several reactions with no evidence of their existence (no downstream measurements, or no stimuli) are removed. In this step, the overall goodness of fit is not improved, but the size of the topology is reduced. To illustrate this case, we add to the initial topology the receptor IL6R but the associated stimulus IL6 is not introduced on the experiments. After the secondary optimization, all downstream reactions of IL6 are removed because no data are present (see reaction arrows downstream of IL6 in Figure 2.12). Similarly, all reactions downstream of the bottom-of-the-network key proteins are removed (e.g. CJUN $\rightarrow$ CFOS reaction in Figure 2.12). All those reactions might be present in reality and could have been kept if the secondary objective was not present. Here, we apply the secondary objective and follow a network trimming which removes all reactions that might be present in the cell but due to the lack of measured signals or experimental conditions cannot be verified. The resulting network is significantly smaller but contains only elements for which there are solid experimental evidence that explain the topology.

To validate our model, we also examine three scenarios where we remove 20% of our experimental data, and then we try to predict them. Specifically, we create three training datasets, each time by removing all cases where one inhibitor is present (either MEKi, PI3Ki, or p38i) and then we calculate how well our ILP- optimized map can predict each of the inhibitor cases (see

Figure 2.13). For the MEKi, PI3Ki, and p38i scenarios the goodness of fit is 8.22%, 9.46%, 7.05% respectively and our ILP-formulation converges on the same or slightly less optimal solutions compared to the solutions obtained when the whole dataset is used for training (4.47%, 7.76%, and 7.05% respectively) - See Figure 2.13. Note that the errors given refer only to the subset considered in each case, not the entire dataset. More extensive validations for Boolean-type models on similar phospho-proteomic dataset can also be found in Saez-Rodriguez et al. [13].

Comparison with genetic algorithm

In order to compare the ILP algorithm with the previously published genetic algorithm (GA) we use the same initial topology and the same normalized dataset [13]. The two algorithms reached almost identical results (see Figure 2.14). For the ILP, the computational requirements are manageable, in the order of a few seconds (14.3 seconds for this example) on a Quad Core Intel Xeon Processor E5405 (2.00GHz,2X6M L2,1333) running Linux 2.6.25.20 (using only one core). In comparison, the same optimization problem using GA requires approximately 1 hour on a similar power computer. The optimal pathway furnished by the ILP matches all but 98 out of 880 experimental data, as opposed to 110 mismatches in the topology furnished by the GA. It has to be noted that GA does not provide termination criteria, and it is conceivable that after even larger CPU times the GA would have achieved the same fit as the ILP. In contrast the deterministic solution of the ILP guarantees that an optimal fit (not necessarily unique) has been identified within a user-specified tolerance (1023 in our case). In addition to the guaranteed optimal solution, commercial ILP solvers are fast, robust and reliable. Note that open-source ILP solvers also exist, but in our experience are not yet adequate. Note also that for larger network topologies, the differences in CPU time will become even more dramatic, rendering the GA intractable.

The notable differences between the proposed method and the method used in [13] is mainly due to fundamental algorithmic differences: the technology behind deterministic ILP solvers (branch-and-bound, branch-and-cut) is more sophisticated than genetic algorithms, it employs the inherent linearity of the problem, and makes use of the good scalability of linear programs (sub-problems in branch-and-bound tree). In contrast, GA treats the model as a black-box and does not exploit the problem structure. Another point is that herein we used a well-established commercial solver, whereas Saez-Rodriguez et al. [13] used their own implementation of GA. Commercial deterministic ILP solvers, such as CPLEX, rely on several decades of research and development, and have extremely powerful features such as pre-processors and node selection heuristics. Thus, they typically become the default choice for ILPs.

Identifying drug effects via drug-induced topology alterations

For the identification of the drug effects we make use of the second dataset in HepG2s where drugs are applied together with the same set of ligands. In this case, the ILP formulation is being used with the HepG2 specific topology (topology obtained from the previous step) and not the generic map. We also do not impose inhibitor constrains the way we do for pathway optimization (e.g., PI3K inhibitor blocks the signal downstream of PI3K) but we let the optimization algorithm decide which reaction(s) should be removed in order to fit the drug-induced data.

The effect of Lapatinib (Figure 2.15a), the most selective and specific EGFR inhibitor [65], is the complete removal of the downstream reactions of the TGFa branch: TGFa $\rightarrow$ GRB2 $\rightarrow$ SOS $\rightarrow$ RAS $\rightarrow$ PI3K and RAS $\rightarrow$ RAF1 $\rightarrow$ MEK1/2 $\rightarrow$ ERK1/2. This resulted from the fact that Lapatinib blocks the TGFa induced MEK1/2, ERK1/2, and AKT phospho-signals (Figure 2.15e). Note that the PI3K $\rightarrow$ ... $\rightarrow$ AKT branch is not removed because it is being used by

the HGF and INS path for the activation of AKT that cannot be blocked by Lapatinib (Figure 2.15e).

Gefitinib, an EGFR tyrosine kinase inhibitor, alters the topology in a very similar pattern as Lapatinib, but, interestingly enough, it also results in the removal of the JNK → c-JUN branch (Figure 2.15b). Closer examination of the raw data (Figure 2.15f) shows a potent inhibition of IL1a- and (IL1a+TGFa)-induced cJUN activity upon Gefitinib treatment. To follow up this interesting off-target effect, we did a dose-response experiment where Gefitinib shows that it can reduce the activation of cJUN signal induced by the IL1a stimuli (Figure 2.15i). We believe that the inhibition of cJUN is not due to the binding of Gefitinib in the upstream molecule JNK but a collective effect of signaling inhibitions in several species that take part in the path between IL1a and cJUN. For this reason, a fitting with a typical dose response curve has been avoided and a simple linear equation has been used instead (Figure 2.15i). Erlotinib, another EGFR inhibitor, has the same effects as Gefitinib (Figure 2.15c) but at the same time shows an effect in the TRAF6 → MAP3K7 reaction. This effects is probably because IkB-a is inhibited in an IL1a -dependent but TNFa-independent manner (see IkB-a signals upon IL1a and TNFa stimuli in Figure 2.11); the only way for the ILP to satisfy this behavior is to remove the transmission of signal before the merging of TNFa and IL1a paths which can be done through the TRAF6 → MAP3K reaction.

The "dirty" Raf inhibitor Sorafenib shows a very different profile: it also blocks the JNK → c-JUN branch (Figure 2.15d) and in addition affects the p38 path (see complete HSP27 inhibition upon IL1a treatment in Figure 2.15h). An interesting observation is that network optimization does not remove the RAF → ERK1/2 reaction despite the fact that RAF is the main target of Sorafenib. Close inspection of the data shows that Sorafenib reduces but does not block the MEK1 phosphorylation (see MEK phosphorylation in Figure 2.15h). This is in agreement with previous published results where Sorafenib does not inhibit activation of the RAF/MEK/ ERK pathway in all human tumor cell lines [66] a finding that highlights the importance of in-vivo assays for the quantification of drug effects.

2.4.3 Discussion

In this section, we present an unbiased phosphoproteomic-based approach and an optimization formulation to construct cell-type specific pathways and to identify drug effects on those pathways. For the pathway construction, we track 13 key phopshorylation signals in 55 different conditions generated by the combinatorial treatment of stimuli and inhibitors. Using Integer Linear Programming (ILP) for pathway optimization we take a generic network of 74 proteins and 105 reactions and construct a cell-type specific network of 49 proteins and 44 reactions that spans between the 5 stimuli and the 13 measured phosphorylated proteins. In this network, we monitor 4 cases of drug-induced pathway alterations using a similar computational scheme.

In comparison to all other protein-based target identification approaches, our method is not based on measurements of drug affinities either by in vitro or in vivo assays. Instead, we use an "operative" signaling network and rely on key phosphorylation events and a-priori knowledge of possible connections to reveal the topology and monitor its alterations under the presence of the drug. Thus, our method is expandable to any type of intracellular perturbations such as ATP-based and allosteric inhibitors, RNAi, shRNA etc. Since no bait or MS is required, we have simple ELISA-type experimental procedure with minimal requirements of cell starting protein (30,000 cells per condition), without affinity immobilizations, protein fractionations, or carefully optimized wash conditions. With our current semi-automated procedures in our lab (robotic liquid handlers), we can achieve total experimental and computational time for a similar size experiment in less than a week. On the other side, our approach can only detect signaling

alterations in topologies bounded between the applied stimuli and the measured phosphorylated proteins and it misses off-target effects outside the constructed network. The expansion of the constructed network depends primarily on three factors: highly curated generic topology, multiplex assay availability for "key" phosphorylation measurements, and experimental cost. We believe that the explosive growth of multiplexed phosphoproteomic assays, the rapid reduction of the cost per datapoint, and the significant improvement in quality of several pathways databases will significantly increase the searching space for drug effects using our proposed methodology. However, our search space will always be significantly smaller compared to whole-genome based approaches because it requires (a) the input of a generic topology which is available only in well-studied pathways and (b) good quality antibodies for the detection scheme. By merging our phosphoproteomic method with genome-wide screening techniques, we might be able to combine the strengths of both approaches and increase the searching space for off-target drug effects.

When applied in HepG2s, our approach identifies both known and unanticipated results. As a positive control, it removes the TGFa branch upon EGRF drug treatments. Another easily understandable effect is Sorafenib's inhibition of the pathway downstream of p38 which can be explained by the drug's target affinity to p38a and p38b [65]. A surprising effect is the removal of the JNK $\rightarrow$ cJUN reaction under the influence 3 out of 4 cancer drugs Erlotinib, Gefitinib and Sorafenib. Interestingly, kinase profiles of those drugs shows no medium or high affinity for the directly upstream JNK1/2 kinases. Despite that, Gefitinib shows a significant reduction of the cJUN activity upon IL1a treatment. A possible explanation is that signal propagation can collectively be attenuated from the low or medium off-target inhibitions of several kinases upstream of JNK and cJUN. This also might explain the inhibition curve in Figure 2.15i, where Gefitinib inhibition of cJUN activation does not follow a typical dose-response curve. In this context, sensitivity analysis in ODE-based pathway models [67] have shown that slight changes of reaction constants can have significant attenu- ations on protein activities several steps downstream the network and thus inhibitory curves cannot be simulated by simplified dose- response models. Our findings also highlight a unique feature of our approach: we find effects of drug's promiscuity that cannot be identified by the direct binding of the drug to the upstream target but are the result of a collective effect of drug's interactions with several upstream molecules. Bait-based analysis cannot reveal those effects since there is no binding involved between the drug and the protein.

Understanding the interplay between cell function and drug action is a major endeavor in the pharmaceutical industry. Here, we provided a methodology to construct cell type specific maps and identify drug effects on those maps. Our ILP formulation was able to build the best possible topology from a set of a-priori determined reactions and choose those, where their presence is confirmed from high throughput phosphoprotein data. Since phosphorylation events are the ultimate reporters of protein/drug function the use of high-throughput phosphoproteomic datasets gave an advantage in data quality for modeling signaling network. We believe our approach complements standard biochemical drug profiling assays and sheds new light into the discovery of possible mechanisms for drug's efficacy and toxicity.

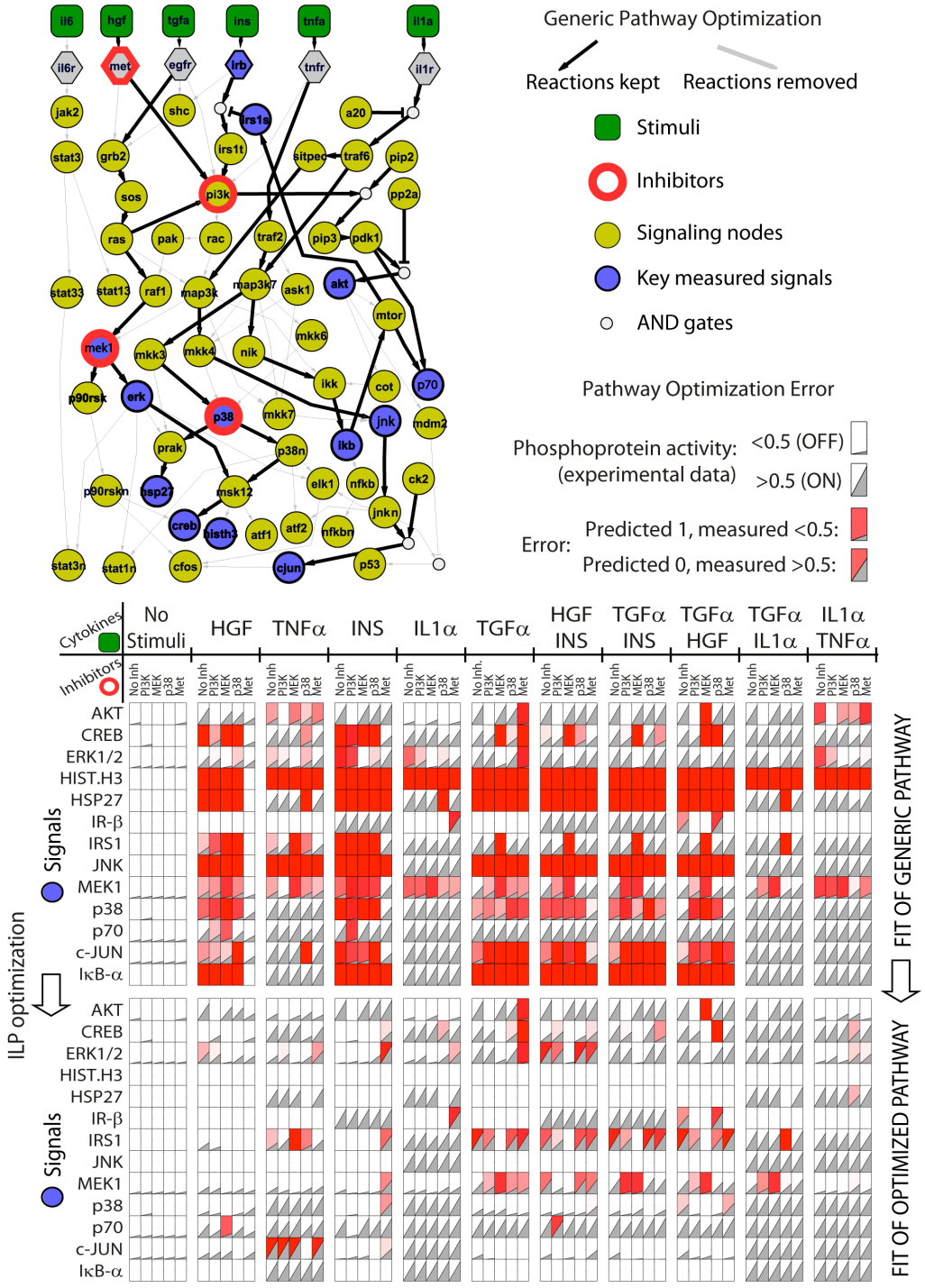


Figure 2.12: Cell type specific topology using Integer Linear Programming. The ILP algorithm is using a subset of postulated reactions denoted with black and gray arrows in a generic pathway to construct a HepG2 pathway map (black arrows in pathway diagram). Gray triangles show phosphoprotein activation level upon stimuli (columns in top and bottom panels) and inhibitors (subcolumns in top and bottom panels). Red background denotes an error between experimental and pathway-inferred responses. Generic topology can hardly represent the HepG2 signaling responses (red background in top panel) and pathway optimization is critical to obtain a pathway topology that captures HepG2 function (limited red background in bottom panel). Pathways are visualized using Cytoscape

.

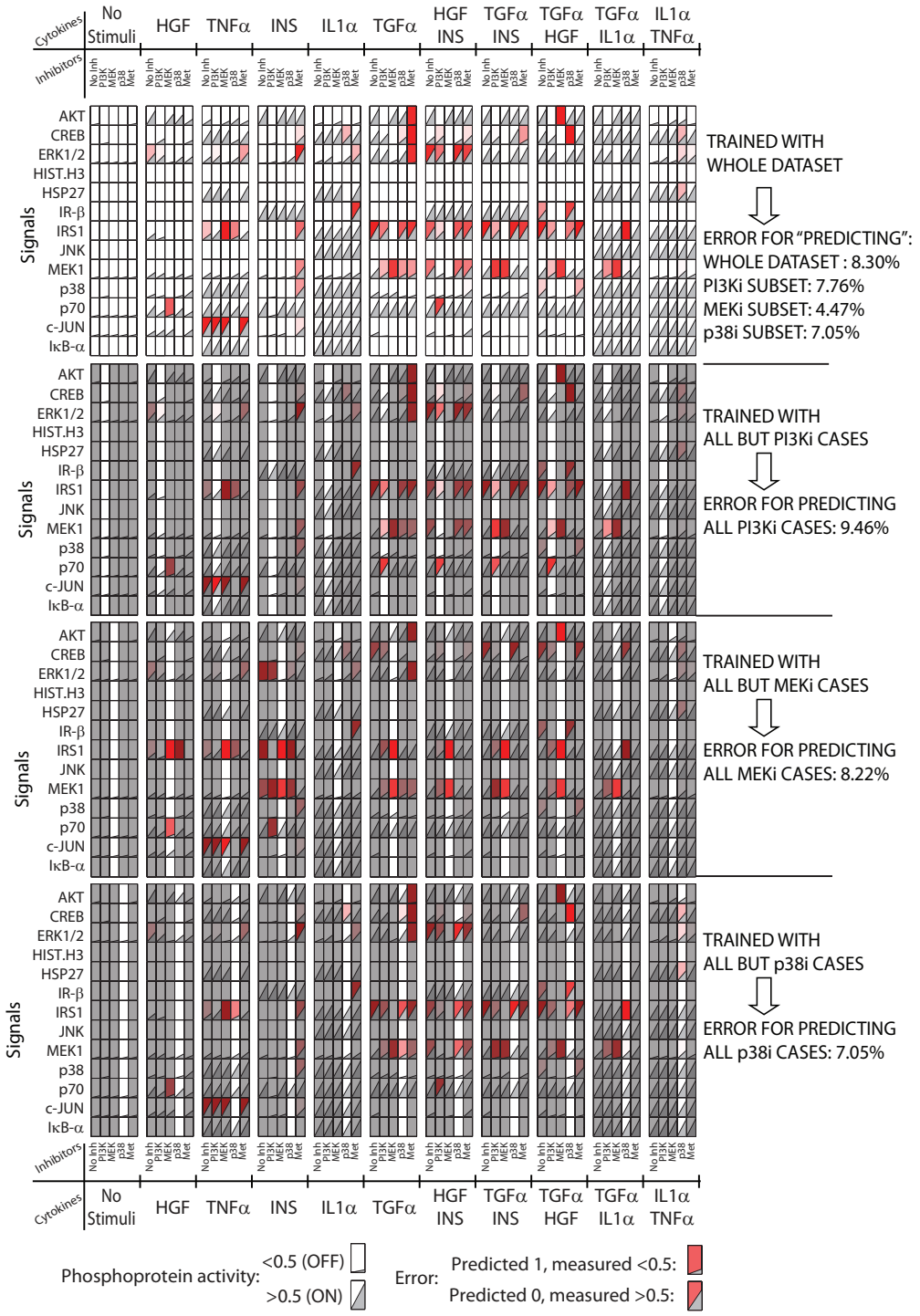


Figure 2.13: Model Validation. The first panel shows the optimization results when the full dataset (shown in Figure 2.12) has been used as training dataset. To validate our model, we created three subsets, in which 20% of our experimental cases are removed that correspond to the treatments with PI3K inhibitor (2nd panel), MEK inhibitor (3rd panel), and p38 inhibitor (bottom panel), and we trained our model against them. The data left out is then used as test dataset for prediction (see highlighted strips in each panel). The error of prediction of the test subsets (error = goodness of fit as described in section 2.2) is shown on the right of each panel.

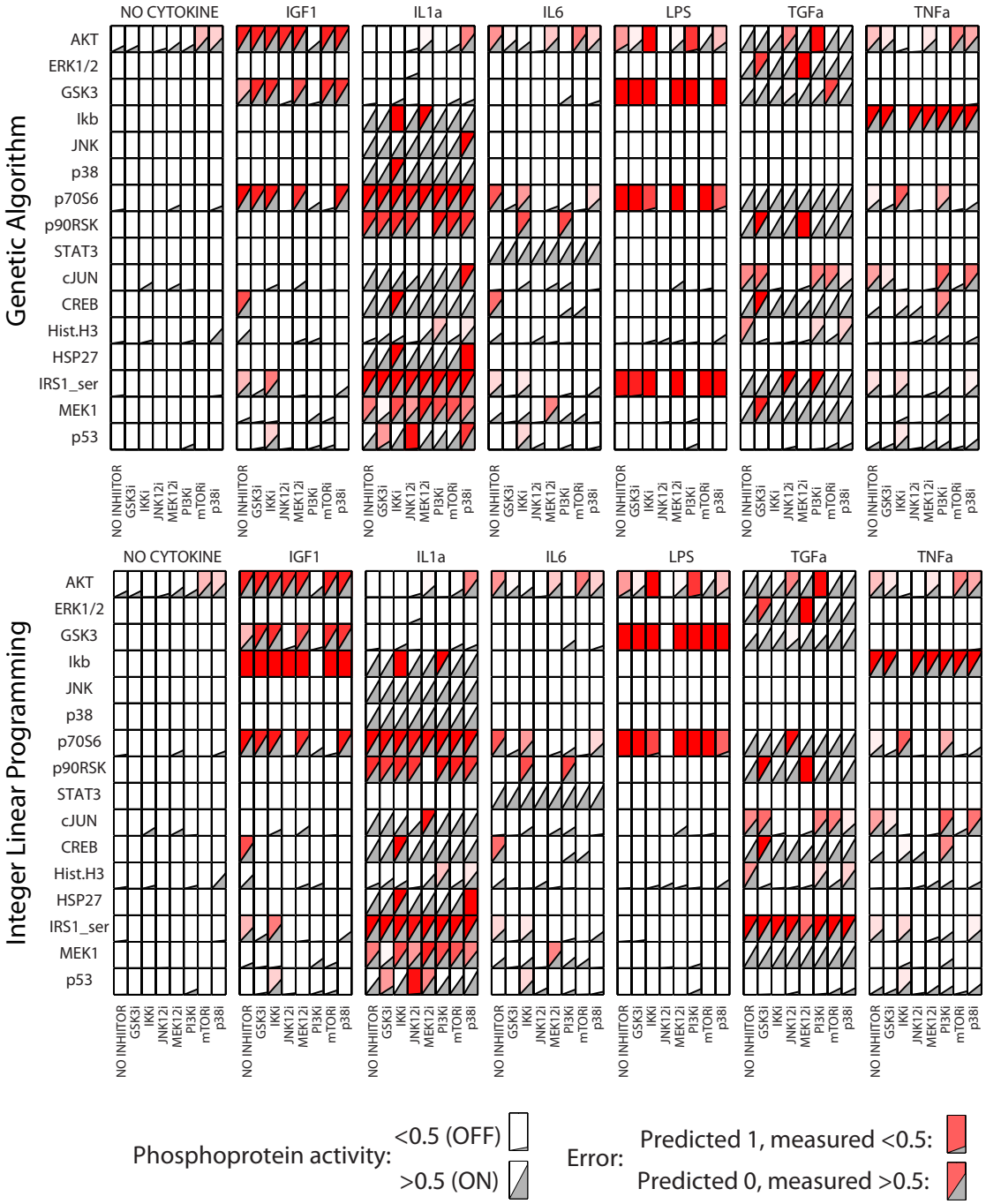


Figure 2.14: Comparison between genetic algorithm and ILP. Both algorithms performed well and achieved very similar solutions. Red background denotes inconsistency between predicted values and experimental data: ILP matched all but 98 out of 880 experimental data, as opposed to 110 mismatches in the topology furnished by the GA. The computational time for ILP was 14.3 sec as opposed to 1approximately one hour for GA.

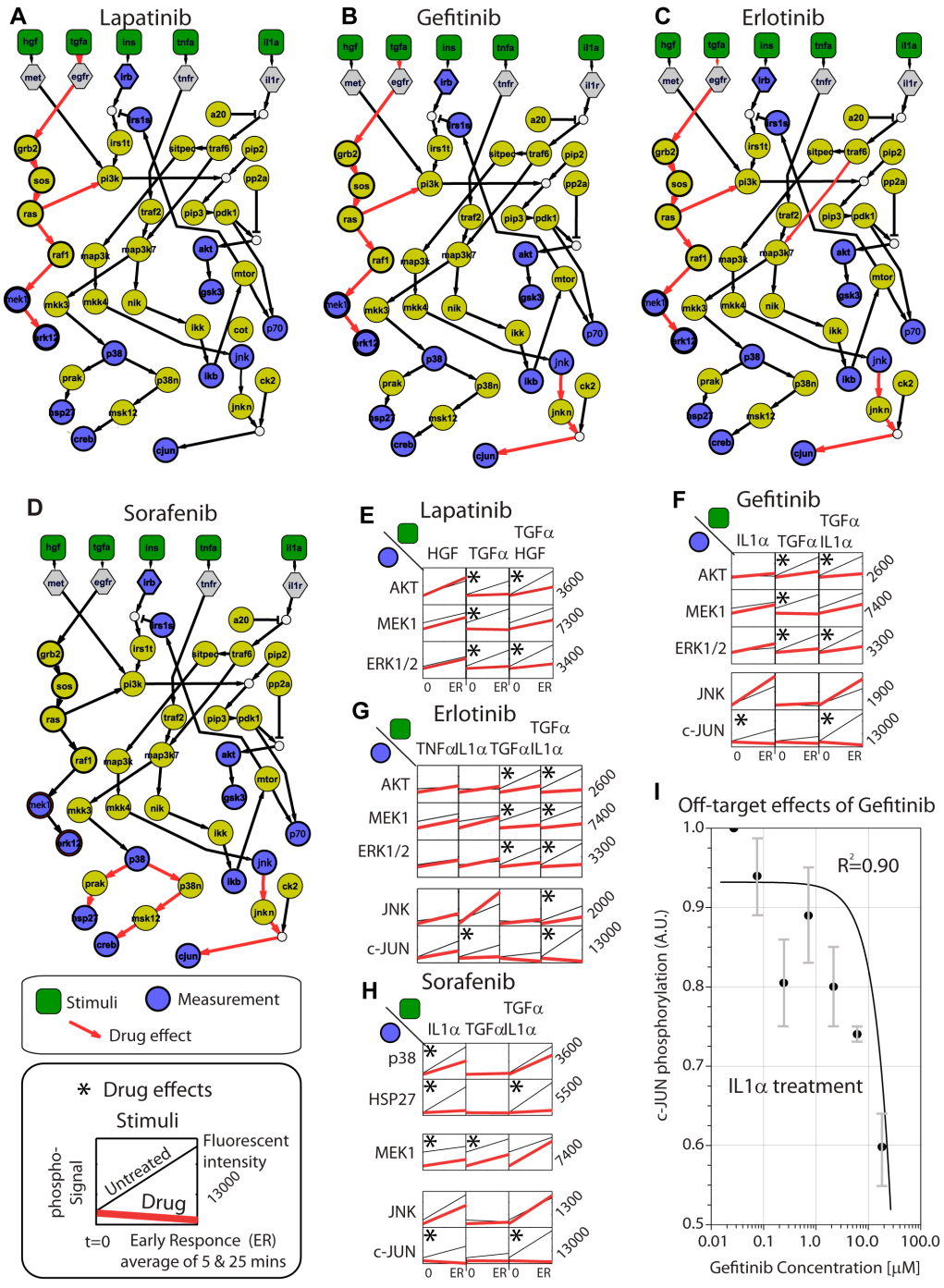


Figure 2.15: Drug-induced pathway alterations. (A–D) Red arrows denote drug effects, i.e., reactions that are removed from the HepG2 topology by the ILP algorithm in order to fit the drug-altered phosphoprotein dataset. (E–H) Raw data that correspond to drug effects. Lines indicates the signal between 0 minutes (untreated) and "early response" (average signal of 5 and 25 minutes post stimuli). (I) Off-target effect of Gefitinib. Dose response curve shows that the EGFR inhibitor reduces cJUN activation upon IL1a treatment. R^2 corresponds to linear fit.

2.5 Combining logical and data-driven models for linking signaling pathways to cellular response

In this work published in [51], a novel hybrid methodology is introduced to optimize signal transduction networks to high throughput proteomic data and link them to cellular response. This work was carried out in collaboration with Alexander Mitsos (at the time when this work was published an assistant professor in the Mechanical Engineering Department of MIT, Cambridge, MA, USA, currently a professor in RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany). The proposed methodology is based on an Integer Linear Programming (ILP) formulation that combines the Boolean modeling of signal transduction with regression driven analysis and links phosphoprotein signaling with cellular response.

Abstract

Background: signaling pathways are the cornerstone on understanding cell function and predicting cell behavior. Recently, logical models of canonical pathways have been optimized with high-throughput phosphoproteomic data to construct cell-type specific pathways. However, less is known on how signaling pathways can be linked to a cellular response such as cell growth, death, cytokine secretion, or transcriptional activity. Results: In this work, we measure the signaling activity (phosphorylation levels) and phenotypic behavior (cytokine secretion) of normal and cancer hepatocytes treated with a combination of cytokines and inhibitors. Using the two datasets, we construct "extended" pathways that integrate intracellular activity with cellular responses using a hybrid logical/data-driven computational approach. Boolean logic is used whenever a priori knowledge is accessible (i.e., construction of canonical pathways), whereas a data-driven approach is used for linking cellular behavior to signaling activity via non-canonical edges. The extended pathway is subsequently optimized to fit signaling and behavioral data using an Integer Linear Programming formulation. As a result, we are able to construct maps of primary and transformed hepatocytes downstream of 7 receptors that are capable of explaining the secretion of 22 cytokines. Conclusions: We developed a method for constructing extended pathways that start at the receptor level and via a complex intracellular signaling pathway identify those mechanisms that drive cellular behavior. Our results constitute a proof-of-principle for construction of "extended pathways" that are capable of linking pathway activity to diverse responses such as growth, death, differentiation, gene expression, or cytokine secretion.

2.5.1 Introduction on linking phosphoprotein signaling to cellular response

Construction of signaling pathways is a major endeavour in biology. signaling cascades, starting at the receptor level, orchestrate a variety of normal or pathological responses via a complex network of kinases, adaptor molecules, and other signaling proteins [3]. Several types of computational models have been proposed to elucidate the complex intracellular signaling network and are commonly classified as data- or topology- driven methods [68, 69]. Their main conceptual difference is their methodology for identifying intracellular connectivity: data-driven models are highly abstract and can identify molecular dependencies within experimental data based on regression analysis, i.e., principal component analysis-(PCA), Partial Least Square Regression (PLSR), Multi-Linear Regression (MLR), Bayesian or other probabilistic models [70, 71, 12]. On the other side, topology-driven models rely on a-priori knowledge of the signaling connectivity and depending on their signal-propagation assumption are classified as physicochemical, fuzzy, or logical.

Even though most experimental data conform on a Cue-Signal-Response (CSR) paradigm [72, 73] most of models -apart from limited cases [10, 74]- are capable of representing events from

either cue-to-signals or from signals-to-responses: topology-driven models are applicable on cue-to-signal datasets where a significant body of literature allows the construction of canonical maps, where data-driven models are applicable on signal-to-response datasets where the flow of information is not fully understandable. Thus, currently there is a lack of models that can answer how stimuli via their signaling mechanisms orchestrate diverse cellular responses such as gene expression, migration, growth, death, metabolic activity, or cytokine release.

In this section we present the construction of "extended" pathway models that aims to explain cellular responses based on pathway activity. The main idea behind the computational approach is a hybrid Boolean/ data-driven model where a logical model is used whenever a priori knowledge is accessible and a data-driven approach is used for adding non-canonical edges to reach out to cellular responses. A previously developed integer linear programming (ILP) framework [23] (see section 2.2) is modified to incorporate non-canonical edges with weights that correspond to regression coefficients and used to optimize the connectivity of the hybrid pathway. The resulting pathway is capable of linking signaling pathways to any type of quantifiable readout such as measurements of cell growth, necrosis, apoptosis, cytokine secretion, or transcriptional activity, as long as these data are available under the same experimental conditions as the phosphoproteomic dataset. As a case study, we construct extended pathways for studying hepatocellular carcinoma (HCC), a liver cancer disease that is the third leading cause of cancer death with inadequate therapeutic interventions [75]. As cellular response we choose the release of 22 cytokines and we ask what signaling activity downstream of 7 receptors, and 57 signaling molecules can explain the complex profiles of cytokine releases. Our computational approach is able to uncover well-known secretion pathways and identify significant differences between non- HCC and HCC cells. Our approach highlights the importance for construction of integrated CSR pathways that given a specific stimulus, can predict the intracellular activity that drives responses such as growth, death, differentiation, gene expression, or cytokine secretion.

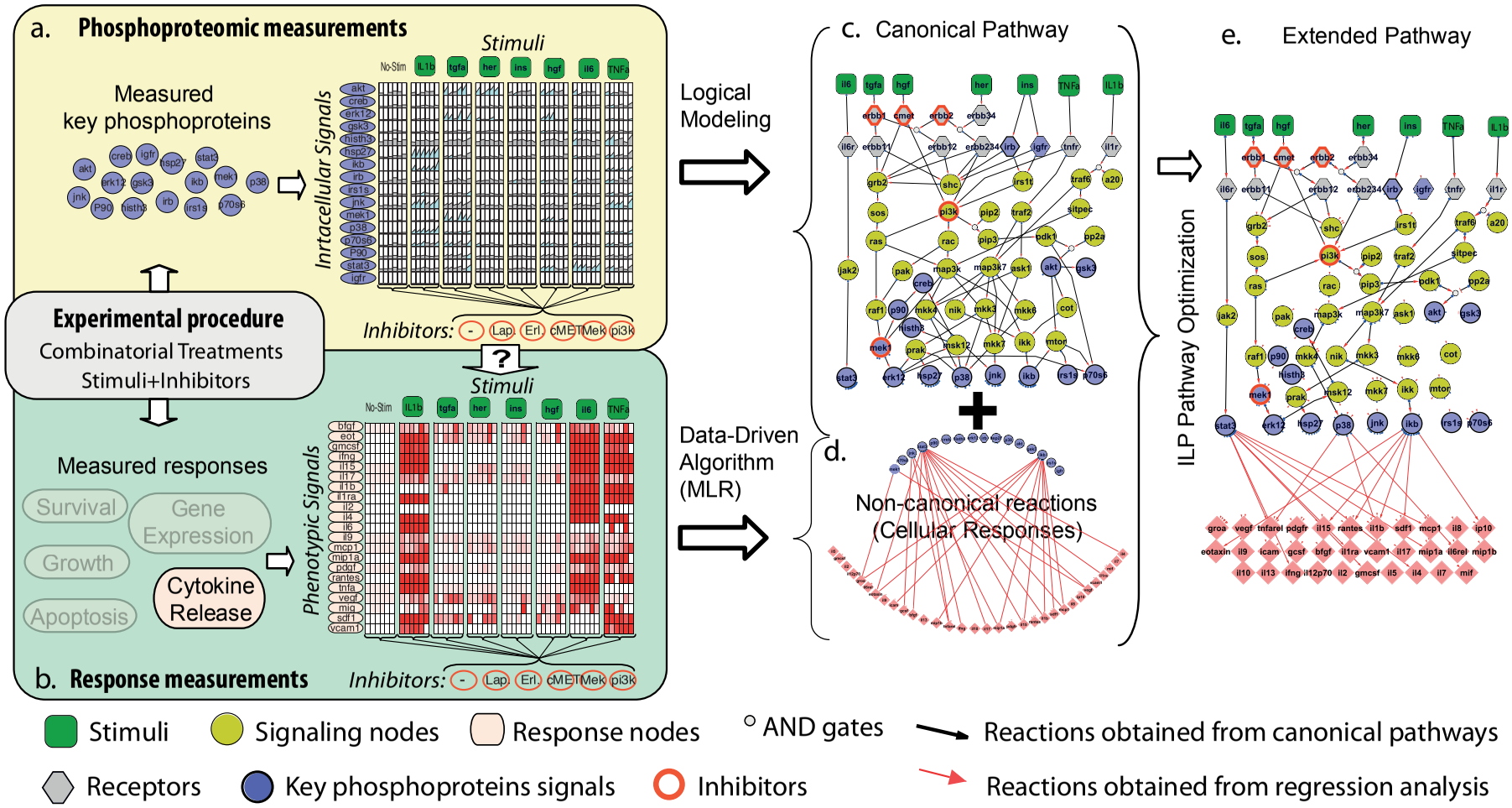


Figure 2.16: Experimental and Computational workflow: (a) The "signaling" dataset monitors the activity of 16 different key phosphoproteins (blue nodes) under the combinatorial treatment of stimuli (green nodes) and inhibitors (red circles). (b) The response dataset can include any quantifiable cellular response such as cytokine releases (22 pink nodes) that are monitored under similar treatments. (c) A canonical pathway that incorporates the stimuli and key phosphoproteins is constructed from the literature (d) Data-driven approach is used to connect the signaling nodes via "non-canonical" edges to cytokine releases. (e) Canonical and non-canonical edges that fit the experimental data are selected using an ILP optimization formulation and thus, the extended pathway topology is constructed.

2.5.2 Construction of extended signaling pathways

The construction of extended signaling pathways can be divided into three main steps: (a) the construction of canonical pathways, (b) the identification of new edges between signals and response from data-driven algorithms, and (c) the optimization of the extended pathway using an Integer Linear Programming (ILP) formulation.

The canonical pathway map (Figure 2.20a) is created around the 7 stimuli and the 16 key phosphoproteins using Ingenuity software (Redwood City, California) and manual curation based on literature search [23]. Non-canonical edges (Figure 2.20c) from key phosphoproteins to cytokine releases are then added to the generic topology and incorporated into the ILP objective function using stoichiometric representation with weights (in chemical reactions these are usually referred as "yields") that equal to the regression coefficients obtained from a multi linear regression (MLR) algorithm. This strategy allows us to enhance the canonical topology with response nodes using non-canonical edges from data-driven algorithms that have as dependent values (Y) the cellular response and as independent values (X) the key phosphoproteins nodes. With this strategy, any type of data-driven approach can be merged with canonical pathways. Herein, MLR was chosen because of its simplicity to connect signals to response in an intuitive way and without the need of intermediate nodes (e.g. nodes representing principal components if PLSR had been used).

Once the extended topology is created with canonical and non-canonical edges, an optimization formulation with binary variables and linear constraints is employed to identify a pool of pathway solutions that best describes the proteomic data. The main concept behind the ILP optimization is the minimization of an objective function that represents the deviation between the experimental measurements and the signaling and response values inferred from the network topology, penalized by a function of the map's size. Raw data were normalized to [0,1] as described previously [13] by taking into account the experimental noise, the saturation limits of the assay, and the basal level at time zero (see also the subsection entitled "Data Normalization"). There are three main terms in the ILP objective function:

$$\sum_{j,k} a_j^k \mid x_j^k - x_j^{k,m} \mid + \sum_k \sum_{j_{res}} a_{j_{res}}^k \mid x_{j_{res}}^{k,m} - \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res} j_{res}} \mid + \sum_i \beta_i \, y_i \tag{2.14}$$

The first term penalizes the measurement-prediction mismatch of the key phosphoproteins and removes all edges that contradict the "signaling" dataset. The second term penalizes the measurement-prediction mismatch of the response measurements and prunes non-canonical edges that contradict the response dataset. The third term removes all edges that have no effect on the measurement-prediction error and thus penalizes the map size.

In more detail,

- The set $j_{res} = \{1, ..., n_{s,res}\}$ represents the response species.
- The set $i_{res} = \{1, ..., n_{r,res}\}$ represents the edges linking signaling (j) with response (j_{res}) nodes.
- a_j^k, $a_{j_{res}}^k, \beta_i \geq 0$ are weights set by the user.
- x_j^k is the predicted value of species j in experiment k.
- $x_j^{k,m}$, $x_{j_{res}}^{k,m}$ is the measured value of species j in experiment k.
- $z_i^k \in \{0, 1\}$ denotes the activation or not of reaction i in experiment k.
- $y_i \in \{0, 1\}$ denotes whether reaction i is possible or not.

- $\sum_{i_{res}} z^k_{i_{res}} w_{i_{res}j_{res}}$ corresponds to $x^k_{j_{res}}$ and is the predicted value of response species j_{res} in experiment k. It equals to the sum of all reactions i_{res} leading to species j_{res} weighted by $w_{i_{res}j_{res}}$, i.e. the Multiple Linear Regression weights. In other words the summation is only over the reactions i that lead to response species j.

Therefore, the first term of the objective function $\sum_{j,k} a^k_j \mid x^k_j - x^{k,m}_j \mid$ corresponds to the measurement-prediction mismatch over all signaling species (j) and experimental conditions (k). Note that the summation is only taken over the species j that are measured in experiment k. The second term $\sum_k \sum_{j_{res}} a^k_{j_{res}} \mid x^{k,m}_{j_{res}} - \sum_{i_{res}} z^k_{i_{res}} w_{i_{res}j_{res}} \mid$ corresponds to the measurement-prediction mismatch over all response species (j_{res}) and experimental conditions (k). The middle summation is over the response species (j_{res}) that are measured in experiment k. The inner summation is over the reactions i that lead to response species j. The third term $\sum_i \beta_i y_i$ corresponds to the penalty imposed by the map size. For a complete reference to the original formulation see [23] or section 2.2. Here we will only discuss the extra constraints regarding the response species.

Concerning the term $\sum_k \sum_{j_{res}} a^k_{j_{res}} \mid x^{k,m}_{j_{res}} - \sum_{i_{res}} z^k_{i_{res}} w_{i_{res}j_{res}} \mid$, assuming $a^k_{j_{res}} \geq 0$, $\mid x^{k,m}_{j_{res}} - \sum_{i_{res}} z^k_{i_{res}} w_{i_{res}j_{res}} \mid \in [0,1]$ corresponds to the scaled measurement-prediction error. Let the minimal and maximal total yields (and thus expected measurements) of the species be given by

$$v^{min} = \sum_{i_{res}, w_{i_{res}j_{res}} < 0} z^k_{i_{res}} w_{i_{res}j_{res}} \tag{2.15}$$

$$v^{max} = \sum_{i_{res}, w_{i_{res}j_{res}} \geq 0} z^k_{i_{res}} w_{i_{res}j_{res}} \tag{2.16}$$

We want to minimize the weighted sum of the absolute differences $\hat{d}^k_{j_{res}} = \mid x^{k,m}_{j_{res}} - \sum_{i_{res}} z^k_{i_{res}} w_{i_{res}j_{res}} \mid$. Assuming that the measurement is consistent with the weights, we would have $x^{k,m}_{j_{res}} \in [v^{min}, v^{max}]$ which would give $\hat{d}^k_{j_{res}} \in [0,\, v^{max} - v^{min}]$. However, this cannot always be assumed and therefore we take the more general case that

$$\hat{d}^k_{j_{res}} \in [0,\, \hat{d}^{k,max}_{j_{res}}] \tag{2.17}$$

where,

$$\hat{d}^{k,max}_{j_{res}} = max(v^{max}, x^{k,m}_{j_{res}}) - min(v^{min}, x^{k,m}_{j_{res}}) \tag{2.18}$$

We can thus scale as

$$\hat{d}^k_{j_{res}}\, \hat{d}^{k,max}_{j_{res}} = \mid x^{k,m}_{j_{res}} - \sum_{i_{res}} z^k_{i_{res}} w_{i_{res}j_{res}} \mid \tag{2.19}$$

to

$$\hat{d}^k_{j_{res}}\, \hat{d}^{k,max}_{j_{res}} = x^{k,m}_{j_{res}} - \sum_{i_{res}} z^k_{i_{res}} w_{i_{res}j_{res}} \tag{2.20}$$

$$\hat{d}^k_{j_{res}}\, \hat{d}^{k,max}_{j_{res}} = -x^{k,m}_{j_{res}} + \sum_{i_{res}} z^k_{i_{res}} w_{i_{res}j_{res}} \tag{2.21}$$

and obtain the desired range i.e. $d^k_j \in [0,1]$. To ensure linearity we impose two inequality constraints, which are equivalent for $a^k_j \geq 0$,

$$-\hat{d}^k_{j_{res}}\, \hat{d}^{k,max}_{j_{res}} - \sum_{i_{res}} z^k_{i_{res}} w_{i_{res}j_{res}} \leq -x^{k,m}_{j_{res}} \tag{2.22}$$

$$-\hat{d}^k_{j_{res}}\, \hat{d}^{k,max}_{j_{res}} + \sum_{i_{res}} z^k_{i_{res}} w_{i_{res}j_{res}} \leq -x^{k,m}_{j_{res}} \tag{2.23}$$

The above constraints complete the formulation.

The ILP formulation is solved with the state-of-the-art commercial code CPLEX through GAMS. This solver guarantees minimal error between experimental data and the Boolean topology eliminating uncertainty associated with heuristic methods such as genetic algorithms. To overcome the existence of multiple near-optimal solutions, in the present work the ILP solver furnishes 100 distinct solutions within a 10% difference in the objective value. The resulting pathways are presented in Figure 2.20b where the width of each edge corresponds to its frequency in the pool of near-optimal solutions.

solution pool

As aforementioned, the objective function (2.14) consists of three major terms, namely $\sum_{j,k} a_j^k \mid x_j^k - x_j^{k,m} \mid$ and $\sum_k \sum_{j_{res}} a_{j_{res}}^k \mid x_{j_{res}}^{k,m} - \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res} j_{res}} \mid$ which are related to the goodness of fit, and $\sum_i \beta_i \, y_i$ which penalizes the size of the pathway. The need for the third term arises from the fact that there are many solutions fitting the measurements equally good. To reduce the number of optimal solutions the size of the pathway is also minimized. However, the biological significance underlying the minimization of the pathway's size is not evident. Thus, we introduce a tolerance of the global minimum and harvest 100 solutions lying within this tolerance. This modification allows us to consider a solution pool instead of a single solution. The frequency of each edge in the solution pool, expresses a level of confidence in the presence or absence of the respective edge in the optimal pathway. By taking into account suboptimal solutions we are sure to capture relations between the signaling cascades, and their probability of occurrence, that we might otherwise miss.

Data Processing and Linear Regression Analysis

Both signaling and response datasets were organized in data structures in the form of 5-D cubes using the DataRail software [50]; 4 of the dimensions of the cube correspond to the different experimental conditions (cell type, time point, stimuli treatment, inhibitor treatment) and the 5th to the measured readouts (response and signaling data). The raw data for both response and signaling datasets were then normalized using a hill function filter and scaled to the range [0,1] as described previously [13] (See the following paragraph for more details and figures 2.17 and 2.18 for an assessment of the proposed method's sensitivity to variables of the normalization procedure). The noise level of the assay has been estimated in [12] at the range of 160 fluorescent units, by considering the standard deviation of repeated measurements of unstimulated controls. The response matrix Y^{res} (an $m \times k$ matrix representing m response components under k conditions) was then regressed against the signaling matrix X^{sig} (an $m \times k$ matrix representing m intracellular signals under k conditions). The computed correlation matrix W is comprised by the correlation coefficients $w_{i,j}$, where i is the index of response components ($i = 1, ..., n_{s,res}$) and j the index of signals ($j = 1, ...n_{s,sig}$). The correlation coefficients $w_{i,j}$, were then used as the stoichiometric weights of the non-signalling reactions in the Boolean framework that originates from a signal j and ends to a response component i (see also section 2.2 and Additional information of this section for an estimation of the proposed method's sensitivity to $w_{i,j}$).

Data Normalization

In this section we apply a normalization of raw data from 0 to 1 as described previously [13] by considering 1) the percent change from basal to stimulated state, 2) the experimental noise, 3) the upper limit due to the saturation of the assay (30,000), and 4) the basal level at time zero. The most important parameter is the activation threshold where a signal is considered

"active" and should be mapped to a value greater than 0.5 in the 0/1 logic (a signal greater than 0.5 will favor a Boolean value of 1 in the optimization scheme in order to minimize the experimental/computational mismatch). To assess activity, the stimulated state (average of 10 and 30 mins for phosphoproteins, and 24 hours for cytokine release) is compared to its unstimulated state (basal levels at time zero). In previous work (Saez-Rodriguez et al, 2009) the default activation threshold value was set to 2.0, which implies that a two-fold increase compared to unstimulated state is considered an active signaling event. Here, in order to identify an optimal value for our particular dataset, we optimized the generic pathway for several different threshold values ranging from 1.1 fold increase (a 10% increase is considered significant) to 7 fold increase (a 700% increase is considered significant) and we look into the behavior of two parameters a) the number of edges conserved by the optimization algorithm (see Figure 2.17) and b) the optimization error (see Figure 2.18).

Number of edges conserved: For each value of fold increase we documented the number of edges that are conserved by the ILP algorithm when compared with the generic topology. Figure 2.17 shows a logarithmic decrease of conserved reactions as the fold change threshold increases. For low threshold values (1.10 $\rightarrow$ 1.40) almost all the original edges are conserved, since even the slightest increase of the signal is considered to be significant. The resulting pathway very much resembles the generic topology. Then a decrease of the edges number is observed as the threshold increases. For thresholds in the range of 1.50 $\rightarrow$ 2.5, the number of conserved edges lies in (60 $\rightarrow$ 80). As the threshold increases beyond 2.5, the number of reactions drops substantially until it reaches 0 at the threshold of 7.0

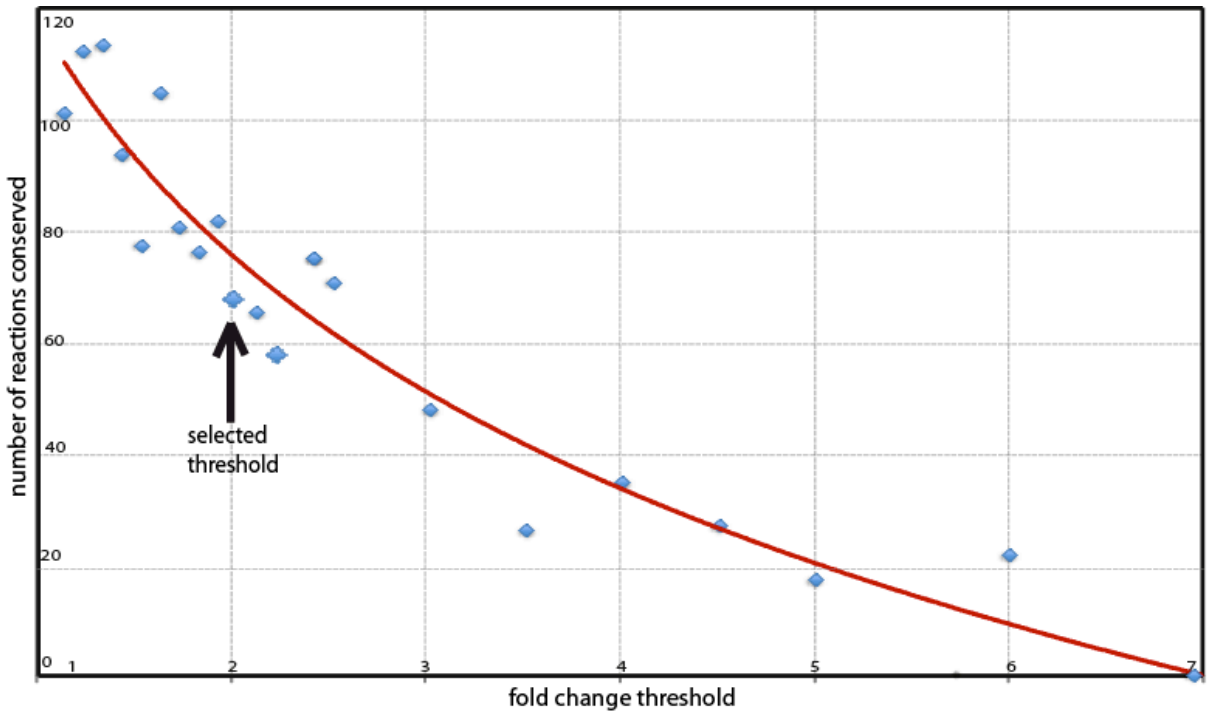


Figure 2.17: "Activity threshold" or "fold change threshold" is an important parameter for comparing Boolean pathways with experimental data. The number of edges conserved after optimization (y a-axis) depends strongly on the threshold set for assuming protein activity. Activity thresholds (x-axis) range from 1.1 (a 10% increase between basal level and stimulated level is considered active) and 7 (700% increase is considered active).

optimization Error: Figure 2.18 presents the total error between topology predictions and experimental data for "Activity thresholds" that range from 1.1 to 7. The error curve dictates that the optimization results are lower at threshold values between 1.8 and 2.0 where the minimum optimization error is 18%. Even though our point is not to find a threshold to the ILP formulation that does not cause problem, Figure 2.18 can help us understand how optimization algorithm performs and define ranges for optimal thresholds. For very low activity thresholds (left side of the curve) the normalized experimental data show significant activity and thus the dataset is too noisy for the algorithm to optimize the map, because of many active signals contradicting one another. As a result, the optimization algorithm cannot find a pathway that satisfies most of the signaling activities and the optimization error is high. For large activation threshold values (around 6.0 fold increase, right side of the curve in Figure 2.18) the normalization algorithm perceives many signals that are truly significant (a 5.0 fold increase is considered noise) as unstimulated signals. As a result, the respective edges are removed although they are functional and the total optimization error is increased (see threshold values above 2.0 in Figure 2.18).

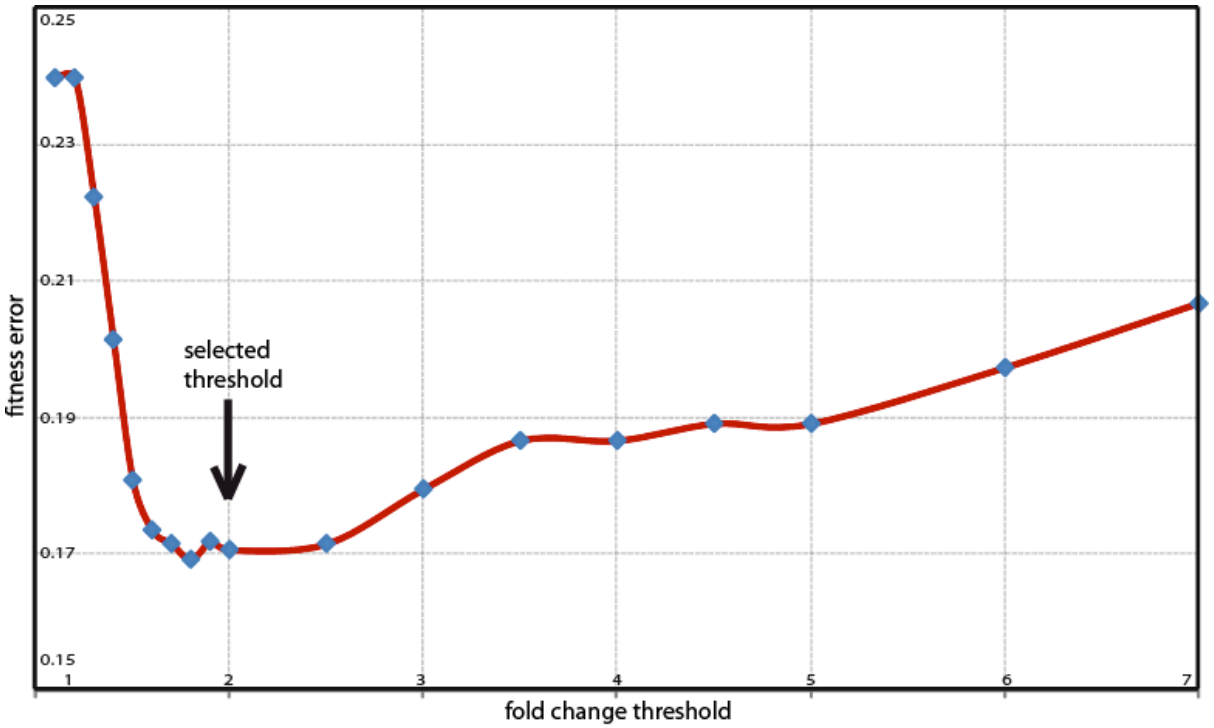


Figure 2.18: Total error after optimization (Y-axis) as a function of "Activity threshold" (X-axis).

Taking the above observations in consideration, a threshold between 1.8 and 2.0 should be used where the algorithm performs better.

2.5.3 Results

Construction of Cue-Signal-Response Datasets

For the construction of the extended pathways, a CSR dataset is created using the beads-based ELISA assays of xMAP technology (Luminex, Austin, TX) as shown in Figure 2.16. Our experimental data consists of the signaling subset (phosphoproteomic data) and the response subset

(cytokine releases) that were measured via multi-combinatorial treatments on two cell types: primary hepatocytes and a hepatocellular carcinoma cell type known as Huh7 [76]. Approximately 50 different perturbations are imposed to primary and HCC cells created by the combinatorial treatment of 7 diverse stimuli (+ no stimulus treatment) and 5 inhibitors (+no inhibitor treatment). As pro-growth stimuli, Tumor Growth Factor alpha (TGFa), Hepatocyte Growth Factor (HGF) and Heregulin (HER) have been chosen based on the response yielded on liver cells in previous experiments [12]. Interleukin 6 (IL6), IL1b and Tumor Necrosis Factor alpha (TNFa) have been chosen as inflammatory ligands. In addition, the Insulin (INS) pathway has been included because of its major role in liver homeostasis. To better constrain the optimization of pathways we impose additional perturbations using stimuli in combination of selective and potent inhibitors for MEK, PI3K, cMET, and EGFR/ERBB2 Lapatinib and Erlotinib [59, 60]. For each combination of stimulus and inhibitor, the phosphorylation state of 16 key intracellular proteins and the release of 33 cytokines were measured as shown in figure 2.19. Among the cytokines, 22 showed a significant activity in either primary or Huh7 hepatocytes. These are plotted in Figure 2.19 using the DataRail software [50].

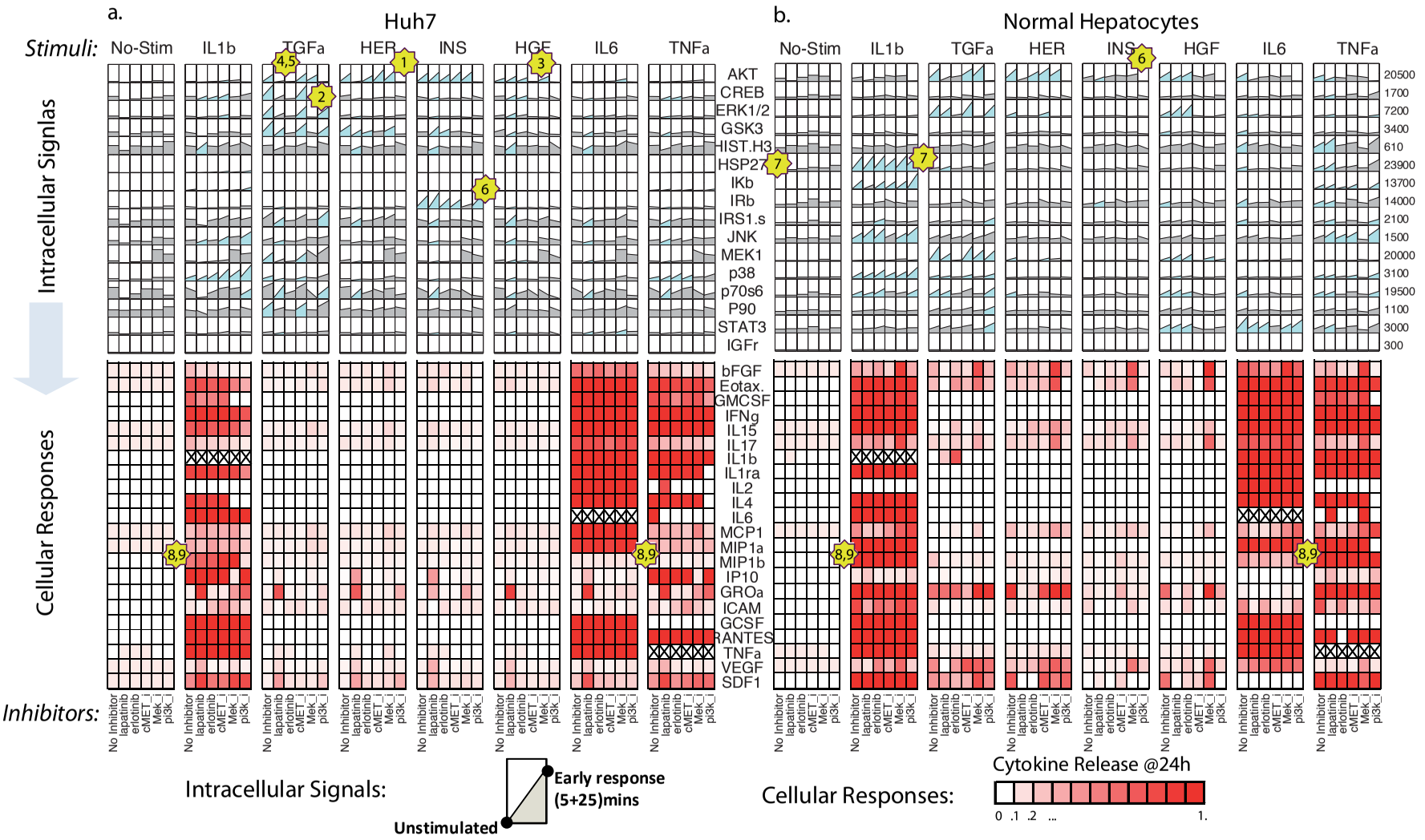


Figure 2.19: Experimental dataset in (a) Huh7 hepatocellular carcinoma cell type and (b) primary hepatocytes. Top panels correspond to the signaling dataset. Each small subplot consists of two datapoints: the zero "unstimulated" condition and the "early response" which is the average phosphorylation activity at 5 and 25 minutes post-stimuli treatment. Bottom panels correspond to the response dataset where 22 cytokines were measured 24 hours post-stimuli. The red color intensity is proportional to the percent increase of the cytokine release as compared to the basal (unstimulated) condition.

Several interesting signaling features can be observed simply by inspection of the data. As positive control observations, all inhibitors block their nominal downstream targets proving their potency and indicating an error-free execution of the multi-combinatorial pipetting procedure (see numbered stars in Figure 2.19; star1:PI3K inhibitor blocks AKT under any treatment, star2:MEK inhibitor blocks ERK under any treatment, star3:cMET inhibitor blocks AKT under HGF, star4:Erlotinib blocks AKT under TGFa, star5:Lapatinib blocks AKT under TGFa). In addition, significant differences can be observed between the two cell types: Huh7 cells respond stronger to insulin stimulus by activating the pro-growth signal AKT and their receptor IRb compared to primary cells that remain unaffected (Figure 2.19, star6). Furthermore, the basal and IL1b -induced phosphorylation activity of the pro-stress protein HSP27 is higher in hepatocytes (Figure 2.19 star7). With respect to cytokine data, primary cells appear to respond stronger under inflammatory stimuli by releasing the inflammatory cytokines MIP1a and MIP1b under TNFa and IL1b treatment, an observation that has been seen before as a mechanisms for HCC cells for evasion of immune surveillance (Figure 2.19 star8, star9). Even though significant differences can be observed simply by visual inspection of the data, the main question remains on how the cytokine release profile (bottom panels in Figure 2.19) can be explained by the pathway activity (upper panels in Figure 2.19). An answer to this question is the presented methodology for construction of extended pathways that incorporates the pathway activity as well as the cytokine release outcome.

Construction of signals-to-response pathways

The generic map includes a total of 7 receptors, 57 signaling molecules connected with 113 canonical edges, and 352 non-canonical edges that connect the 16 key phosphoproteins to the 22 cytokines. From the 352 non-canonical edges, a large percentage of those have correlation weights close to zero. To minimize the computational cost of the ILP solution, we choose to retain 60% of those weights as explained in the additional information of this section. Extended topologies were created for non-HCC and HCC (Huh7) hepatocytes. The mismatch between generic pathways and non-HCC or HCC datasets is 41.0% and 46.6% respectively. After optimization, a total of 47 canonical and non-canonical edges remained in Huh7 and 43 in non-HCC hepatocytes. The error of the cell-specific pathways drops to 18% in Huh7 and 17% in non-HCC hepatocytes. Several edges are removed due to conflict with the data. One example is the removal of TNFR $\rightarrow$ PI3K edge in both cell types in order to isolate the AKT and MEK activity from the TNFa stimuli (star1, Figure 2.20). In a similar manner the AKT $\rightarrow$ COT $\rightarrow$ IKK $\rightarrow$ IKB edges are removed because the measured AKT and IKB signals are not co-regulated as suggested by the Boolean logic (i.e., AKT = 1 then IKB = 1) (star2, Figure 2.20). Furthermore, the links for activating p70S6 on a PI3K independent manner remain only on the primary hepatocytes as suggested by the dataset (IL1b and TGFa activates p70S6 with or without the presence of a PI3K inhibitor, star3, Figure 2.20).

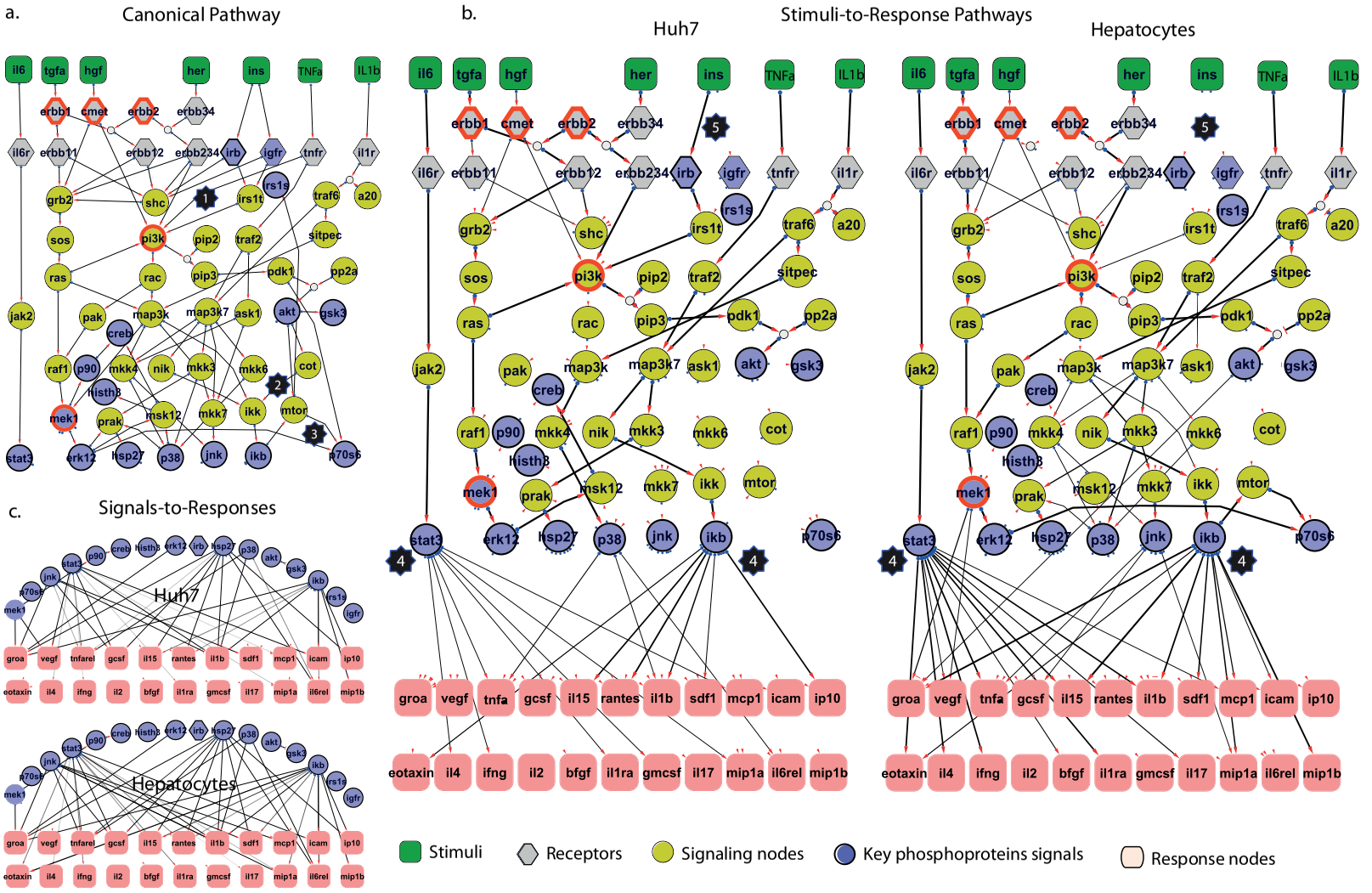


Figure 2.20: CSR pathways for primary non-HCC hepatocytes and HCC (Huh7) cell types. (a) Generic pathway comprised of canonical edges extracted from literature (b) non-canonical edges for Huh7 and primary hepatocytes extracted from a data-driven approach (multi linear regression) (c) extended pathways for Huh7 and primary hepatocytes constructed by fitting canonical and non-canonical edges to experimental data via an ILP formulation.

The presence of cellular response data significantly enhances the optimization of the signaling topology in two different ways. Firstly, non-canonical edges provide additional pathway information to the ILP formulation. In other words, the optimizer is forced to conserve edges that lead to cytokine nodes but do not affect measured phosphoproteomic signals (see "Impact of response measurements on pathway optimization" in the additional information of this section). Secondly, edges with marginal activations of intracellular signals that otherwise would be considered insignificant (either because of assay limitations or time-point selection) are retained in the topology if they correlate well with cellular response. An example of this case is the IL6 pathway: although IL6 activation of STAT3 in Huh7 is seemingly undetectable (see raw signaling data in Figure 2.19), the IL6 $\rightarrow$ ... $\rightarrow$ STAT3 pathway is conserved because even small chances in the STAT3 activation levels correlate well to the IL6-induced release of various cytokines (see star4 in Figure 2.20) (see "Impact of cellular response measurements on pathway optimization" in the additional information of this section). Taken together, when the ILP formulation uses both the phosphorylation and response data, it conserves pathways with barely detectable signaling activity as long as they correlate to cellular response.

The non-canonical edges in Figure 2.20 show that major pathways for the release of inflammatory cytokines are the IL6 $\rightarrow$ STAT3, IL1b $\rightarrow$ NFkB/p38, and TNFa $\rightarrow$ NFkB/p38 pathways [77]. Just three key phosphoprotein signals (STAT3, Ikb and to a lesser extend p38) are responsible for the release of most inflammatory cytokines including TNFa, GROa, RANTES, MCP1, ILb, and EOTAXIN (star4, Figure 2.20), an observation that is in accordance to a large body of literature. It is less known how many different pathways can lead to the release of a particular cytokine. A simple enumeration of paths that lead to cytokines for primary hepatocytes (Figure 2.20c), shows that more than 50% of the cytokines are induced by 2 or 3 edges that can be activated by up to 3 different stimuli following at most 3 different routes of activation. Since the constructed pathways are small subsets of the actual pathways, it is obvious that the mechanisms for a single cytokine secretion are numerous and complex. To tackle such complexity, graph theory analysis of the extended pathways (always limited by the lack of experimental approaches to decipher the whole signaling network) can identify central nodes or group of nodes for inhibiting cytokine secretion, and thus, increase the efficacy of pharmaceutical interventions. This is in particular applicable for multi-targeting of STAT3, NFkB, or p38 pathways to achieve anti-inflammatory effects, a major endeavor of pharmaceutical industry with significant investments on mono-targeted approaches for STAT3, NFkB, or p38 on several diverse diseases including p38 for rheumatoid arthritis [78], IkB for airway inflammation [79], or STAT3 and NFkB for HCC [80, 81].

2.5.4 Conclusions

In the present work, we developed a method for linking signaling data to cellular response. As a case study, we compare extended signaling topologies of primary hepatocytes and Huh7. The two pathway maps are significantly different. Huh7 are not as responsive as primary cells since only 17 non-canonical edges exists in Huh7 compared to 28 in primary hepatocytes (see also additional information of this section for a comparison of simulation runs for the two cell types). These findings are in agreement with recently published data that shows HCC cell types are less responsive to Toll Like Receptor (TLR) stimuli than primary hepatocytes [12], presumably to avoid detection and clearance by the innate immune system [82, 83, 84]. Major pathway differences related to a survival advantage for HCC can also be observed at the intracellular level: a closer look into the insulin pathway shows that INS $\rightarrow$ IRb and INS $\rightarrow$ IGFR edges are removed in hepatocytes but the INS $\rightarrow$ IRb is retained in Huh7 (star5, Figure 3). A closer look into the raw data (Figure 2.16) shows that insulin barely induces IRb and AKT

activation in primary cells. This is in accordance to recent findings that shows increased AKT activation correlates well with the formation of liver tumors [85]. However, in that study, the authors pinpoint the mechanisms of AKT over-activation to the reduced expression of p85a - a regulatory subunit of PI3K. Herein, we show that -at least for the Huh7 case- diminished Akt activation levels can be due to receptor's lower activation as shown from the phosphorylation of IRb.

Here we presented a method for constructing extended pathways that start at the receptor level and via a complex intracellular signaling pathway identify those mechanisms that drive cellular response. Because of the nature of response data - where detailed mechanisms are sparse and not easily searchable via text mining approaches- we used a data-driven approach to link intracellular activity to cellular responses via non-canonical edges. Those edges, together with well-defined intracellular pathways, were used for the construction of the "generic map" which is finally optimized to match high-throughput protein data. The resulting extended pathways revealed intracellular mechanisms that are responsible for the release of 22 cytokines and correlate well with a large body of literature that pinpoint at STATs and NFkB as major drivers of inflammatory stimuli. More importantly, comparison between cell types shows significant differences that lead to survival advantages of the HCC cells. Our results constitute a proof- of-principle for construction of "extended pathways" that are capable of linking pathway activity to diverse responses such as growth, apoptosis, differentiation, gene expression, or cytokine secretion.

2.5.5 Additional information/characterization of the proposed methodology

To evaluate the performance of our hybrid model, several in-silico tests were performed including comparison with a data-driven (regression) model (figure 2.21) and assessment of model sensitivity to i) optimization parameters, ii) experimental design, iii) data deterioration, iii) generic topology. Key points are highlighted in the following paragraphs.

Construction of a 2-step Multiple Linear Regression (MLR) model and comparison to the proposed approach

The performance of our hybrid ILP-MLR approach is compared against a data-driven 2-step MLR approach [12] that correlates i) stimuli and inhibitors to the measured phosphoproteins and ii) phosphoprotein activities to cytokine releases.

An $(n \times k)$ matrix X^{cue} is put together, where n equals to the number of stimuli and k equals to the number of experimental conditions. $X^{cue}(i,j) = 1$ implies stimulus i is included in experiment j, else $X^{cue}(i,j) = 0$. In similar fashion we introduce an $(m \times k)$ matrix X^{inh} where m equals to the number of inhibitors and $X^{inh}(i,j) = 1$ if and only if inhibitor i is included in experiment j. Furthermore, an $(s \times k)$ matrix Y^{sig} is put together, where $Y^{sig}(i,j)$ equals to the measured value of signal i in experiment j. The $(s \times n)$ matrix W^{cue} is defined such that, $Y^{sig} = W^{cue} \cdot X^{cue}$. W^{cue} is computed via Linear Regression using Matlab. The residual matrix, $RES = Y^{sig} \breve{} W^{cue} \cdot X^{cue}$ is then expressed as $RES = W^{inh} \cdot X^{inh} \cdot W^{inh}$ is computed via Linear Regression. Matrices W^{cue} and W^{inh} express the effects of each cue and inhibitor on the measured phosphoproteins. Subsequently, Y^{cyt} is introduced, consisting of the cytokine release measurements such that $Y^{cyt}(i,j)$ corresponds to the value of cytokine release i, in experiment j. Matrix W^{sig} is defined such that $Y^{cyt} = W^{sig} \cdot Y^{sig}$, where Y^{sig} was defined previously. Using the 2-step MLR approach, we performed the following tasks:

1. Correlate each stimuli and inhibitor with the intracellular protein activity (measured phosphoproteins)

2. Correlated phosphoprotein activities with the cytokine release

3. Construct an executable signal transduction model, able to predict cellular response upon external perturbation (in the form of stimuli and inhibitors).

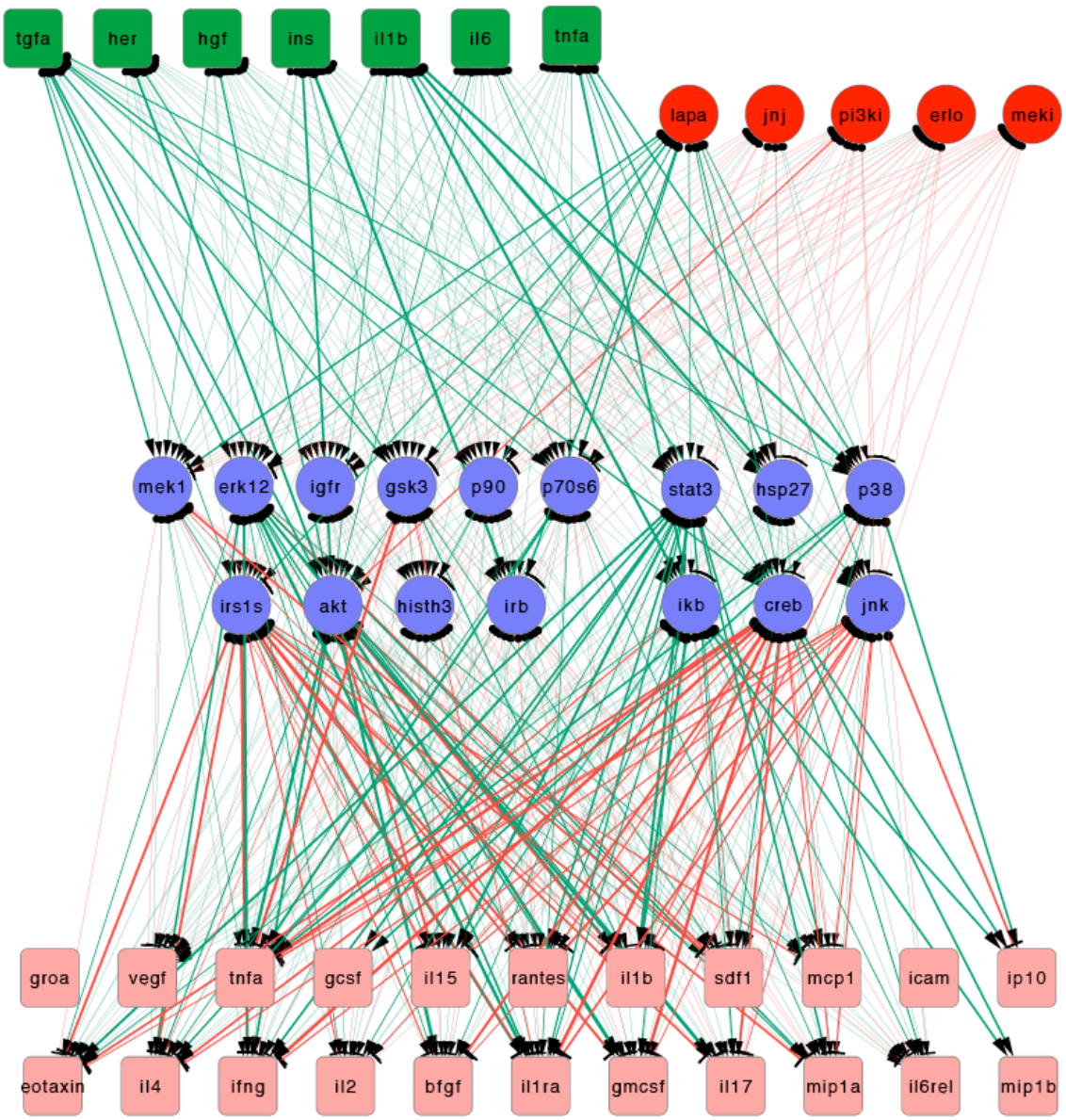


Figure 2.21: 2-step MLR model. Cytokines and inhibitors are linked to phosphoprotein signals via weights obtained by MLR. In a similar fashion, phosphoproteins are linked to cytokine releases. Edges thickness and opacity corresponds to the absolute value of the respective weights. Edges color corresponds to interaction type (green=activation, red=inhibition).

Figure 2.21 features the connectivity patterns underlying Huh7 data; the edge thickness corresponds to the absolute value of MLR weights and edge color corresponds to weights' sign. Simulation of the MLR model, upon external perturbation, can be performed by compiling the X^{cue} and X^{inh} matrices full of 0's and 1's in a way that reflects the experimental conditions, then simulation results on the phosphoprotein and cytokine release level are obtained: $Y^{sig} =$

$W^{cue} \cdot X^{cue} + W^{inh} \cdot X^{inh}$, $Y^{cyt} = W^{sig} \cdot Y^{sig}$. Simulation runs are rounded to 1 or 0, (denoting activation or not respectively) for easier comparison with our Boolean-based ILP approach and illustrated in Figure 2.22. An evaluation of the 2-step MLR approach can be obtained by computing the measurement – prediction mismatch via the following formula :

$$Error = \sum_{j,k} \frac{\mid x_j^k - x_j^{k,m} \mid}{x_j^{k,m}} = 12\%$$

Where,
x_j^k, is the predicted value of species j in the experiment k,
$x_j^{k,m}$, is the measured value (m) of species j in experiment k.

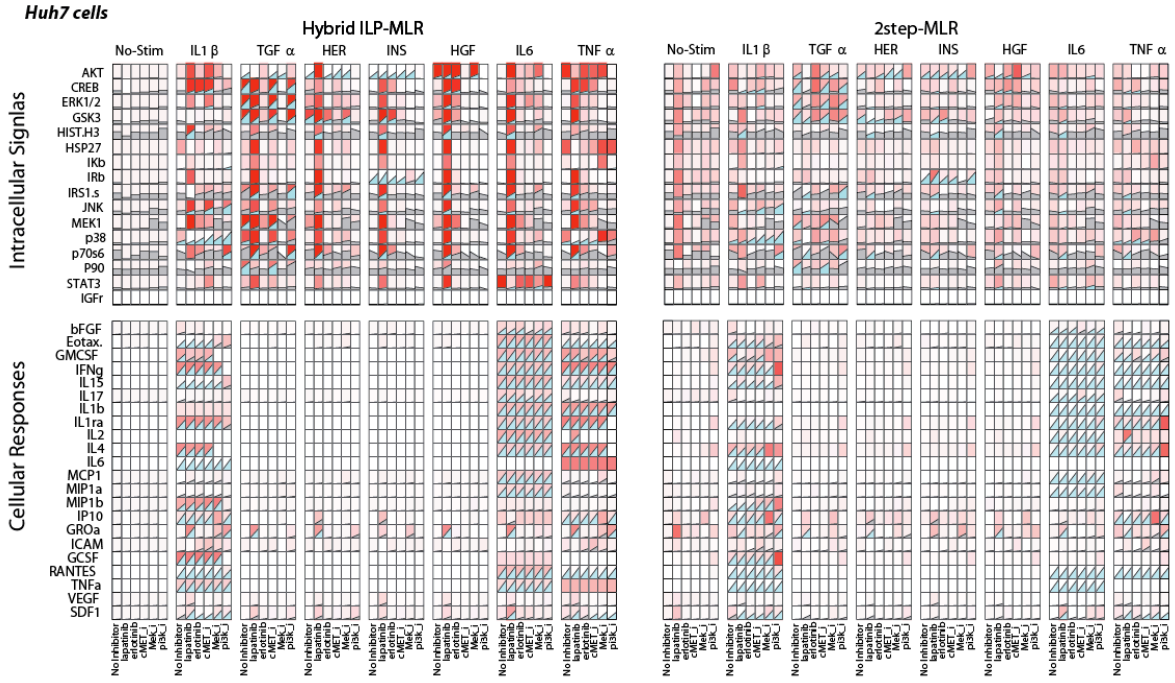


Figure 2.22: Simulation results of the hybrid ILP-MLR (left panel) and the 2-step MLR (right panel) models. The intensity of the red background corresponds to the fitness error.

The two methods are compared quantitatively for data fitting and qualitatively for capturing biological insight.

Regarding data fitting: the 2-step MLR is an unconstrained approach and as such, fits the experimental data better than the ILP approach as indicated by the measurement - prediction mismatch (12% for 2-step MLR, 18% for the hybrid ILP-MLR approach, see figure 2.22). It is expected for a data driven approach to perform better at data fitting and representation than a topology driven approach since no extra constraints are introduced from having to comply with a priori knowledge of the signaling network's connectivity. For instance, different activation patterns of MEK1 (or ERK12 and CREB) across various EGFR ligands such as TGFa, HER, HGF and even INS (although not an EGFR ligand) is only feasible in a data driven approach, since a topology driven approach will have to comply with a restrictive generic topology, such as the one used here, where TGFa, HER and HGF signal through overlapping pathways. Therefore if TGFa activates MEK1 so does HER and HGF. A topology driven approach, such as the ILP will either conserve the whole branch (RAS → RAF1 → MEK1 → ERK12) thus activating the

respective signals under all EGFR ligands, or remove the entire branch. The optimal solution is the one that least increments the measurement – prediction mismatch, however, both alternatives lead to an increase in the fitness error. Figure 2.22 features many examples like the one just described such as: different activation patterns for HSP27 and CREB under IL1b and TNFa, differential activation of GSK3 under TGFa, HER, HGF and INS.

Biological significance and relevancy of the results: Incorporation of a priori knowledge in the form of a generic pathway assists in obtaining biologically relevant and significant results, since the generated model need not only fit experimental data adequately, but also comply with literature. The proposed ILP-MLR hybrid approach by taking into account a priori knowledge of proteins connectivity guarantees biologically interpretable results, In contrast to the MLR approach that uncovers correlations that lack biological interpretation, such as Lapatinib induced IRB, MEK1, HSP27 and P70S6 activation (see Figure 2.21). Moreover, the activation state of latent nodes can be inferred given the activation state of others, while 2-step MLR ignores any intermediate nodes. Further comparison of data driven versus topology driven methods goes beyond the purpose of this section, the reader may find an elaborate characterization of these classes of methods in [8].

Model Assessment: Sensitivity of the proposed approach to experimental design, data, generic topology, linking weights and a_j^k constants

For better assessment of the proposed methodology, its sensitivity to changes in

1. experimental design
2. measured data
3. generic topology
4. linking weights
5. a_j^k constants

is assessed, in terms of i) remaining fitness error and ii) topology alterations.

The "remaining fitness error" corresponds to the measurement-prediction mismatch in the pathway after the optimization and it is evaluated by the following formula:

$$Error = \sum_{j,k} \frac{\mid x_j^k - x_j^{k,m} \mid}{x_j^{k,m}} \tag{2.24}$$

where,
x_j^k, is the predicted value of species j in the experiment k,
$x_j^{k,m}$, is the measured value (m) of species j in experiment k.

"Topology alterations" aim to identify the differences between i) the pathway optimized with a reduced/altered dataset and ii) the pathway optimized with the full dataset. The differences are captured by comparing simulation runs, following the formula:

$$ErrorTopol = \sum_{j,k} \frac{\mid x_j^k - x_j^{k,r} \mid}{x_j^{k,r}} \tag{2.25}$$

where,
x_j^k, is the predicted value of species j in the experiment k, full-dataset optimized pathway
$x_j^{k,r}$, is the predicted value of species j in experiment k, reduced-dataset (r) optimized pathway

Sensitivity to changes in the experimental design

It is apparent that the optimized topology is based on the experiments performed, the number of signals that were measured, and the number of perturbations imposed in the network. More specifically, a single stimulus experiment with only one measured signal can provide information for a very small subset of the generic topology, and as such the optimized map will be very small. On the other hand, an extensive experiment with all different combinations of stimuli and inhibitors is not possible due to time and cost limitations. In this study, we created a dataset that is experimentally feasible and includes all possible combinations of single stimulus with single inhibitor. Removal of 50% of these treatments randomly shows a significant deterioration of the constructed pathway and 35% increase of the fitness error (see Figure 2.23). Finding an optimal experimental design for maximally constraining a generic topology is a very important aspect in the field of pathway optimization that can include pathway controllability, pathway observability, experimental limitations, and definitely several other experimental constraints imposed by how the assays are performed.

In more detail, the phosphoproteomic dataset in Figure 2.19 consists of 48 experimental conditions (8 stimuli including the no-stimulus treatment, times 6 inhibitors, including the no-inhibitor treatment) and 16 signals; the response dataset consists of the same experimental conditions and 22 signals. In this part of model assessment, we exclude random subsets of the 48 experimental conditions and monitor how the ILP algorithm performs. 10 in-silico experiments are carried out, leaving out 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45% and 50% of the experiments.

It becomes apparent that the greater the subset of experimental conditions left out of the optimization procedure is, the poorer the fit we obtain becomes, in what seems to be a linear fashion (R2=0.88995) (Figure 2.23). Finally, having excluded 50% of the experimental conditions, the remaining fitness error reaches 35% (almost two times the fitness error of the full-dataset optimized pathway).

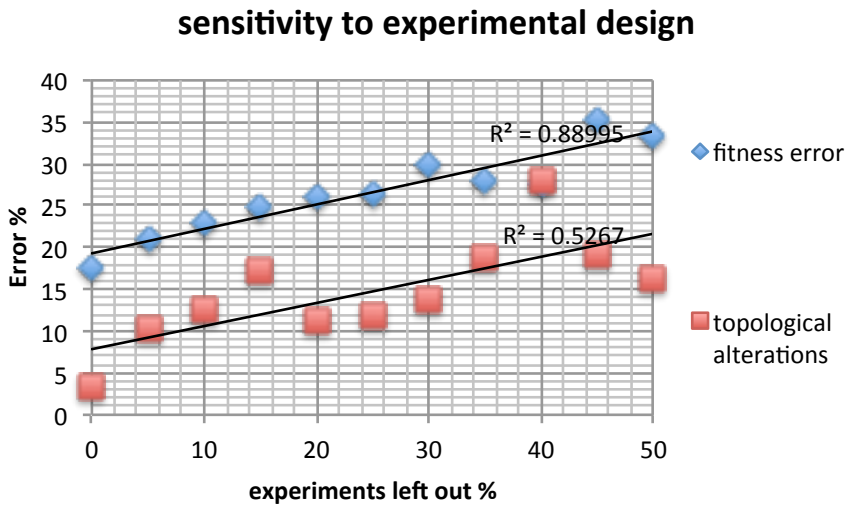


Figure 2.23: Sensitivity of proposed methodology to experimental design. First curve (blue points) corresponds to remaining fitness error upon excluding subsets of the experimental conditions. Second curve (red points) corresponds to topological alterations between the full-dataset optimized map and reduced-dataset optimized map.

Sensitivity to data deterioration

In this part of model assessment, we scramble random subsets of the original phosphoproteomic and response datasets. 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45% and 50% of the total datapoints are substituted with random numbers in (0,1). Like before, the remaining fitness error and topology alterations are computed via formulae (2.24), (2.25) and the results are plotted in Figure 2.24. Scrambled data are characterized by internal conflicts (e.g., activated values in no-stimuli experiments), the ILP algorithm cannot emulate this behavior, resulting in increasing fitness error with limited alterations in the topology of the optimized pathway.

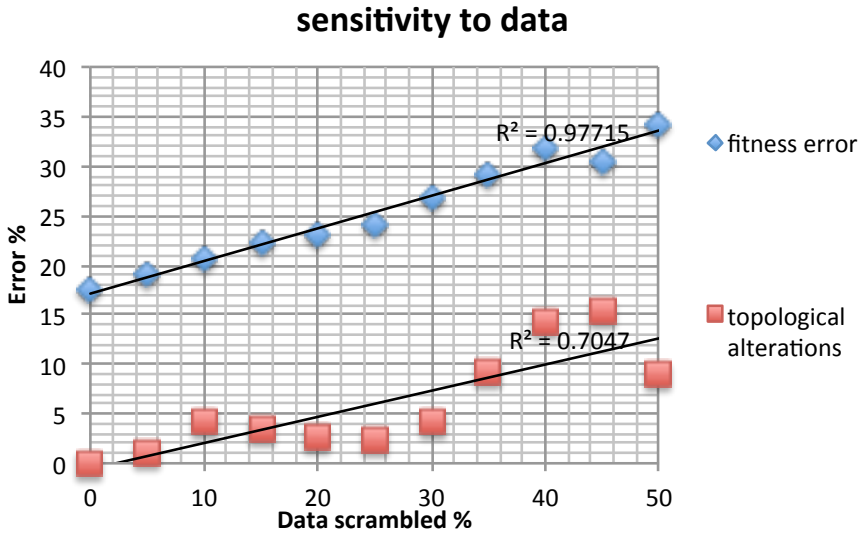


Figure 2.24: Sensitivity of proposed methodology to data scramble. First curve (blue points) corresponds to remaining fitness error upon substitution of datapoints with random numbers in (0,1). Second curve (red points) corresponds to topological alterations between the full-dataset optimized map and scrambled-dataset optimized map.

Sensitivity to changes in the generic topology

Despite the wealth of information found for pathway construction mainly from pathway databases, conflicting reports in pathway connectivity makes the construction of the generic topology a non-trivial task with significant manual curation and with no guaranteed for the "right" generic topology. In order to assess the sensitivity of our hybrid model to changes in the generic topology, we substitute up to 10% of our generic reactions with random reactions.

In more detail, subsets of reactions (2%, 4%, 6%, 8%, 10% of the generic pathway) are substituted with random connections. Scrambled pathways are optimized using phosphoproteomic and cytokine release datasets and the remaining fitness error and topology alterations are plotted in Figure 2.25. As expected the generated models are very sensitive to changes in the generic topology, even a single reaction removed can cause drastic changes in the model behavior. For instance, removal of EGF $\rightarrow$ EGFR prevents the ILP from successfully fitting all experiments where EGF is introduced. Deviation of simulation runs also increases with increasing number of scrambled reactions, but to a smaller extent, implying that removed reactions have disrupted the signal transduction and no alternative paths exist.

A possible way to reduce the sensitivity to the generic topology is to allow the addition of less known or conflicting reactions with weight based on literature findings. However, such a

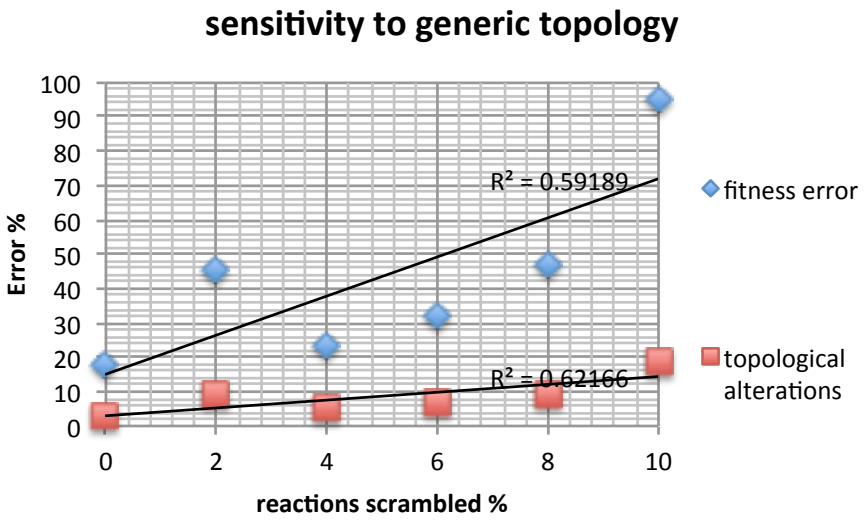


Figure 2.25: Sensitivity of proposed methodology to changes in the generic topology. First curve (blue points) corresponds to remaining fitness error upon substitution of reactions with random connections between the species. Second curve (red points) corresponds to topological alterations between the map optimized using the original generic topology and map optimized using the scrambled generic topology.

method should be coupled to a text mining approach and a pathway database which is beyond the scope of this study.

Sensitivity to changes in modeling decisions: w_{cr}

Weights obtained via MLR are used to link intracellular signal transduction pathways with cellular response (herein cytokine releases), in a consistent, integrative model. MLR generates connections from each of the measured phosphoprotein to every cytokine released, resulting in a total of 352 reactions. However, most of the respective weights are very close to zero, suggesting very little (if any) effect on the released cytokines. To increase computational efficiency, only reactions with weights of absolute value greater than w_{cr} are considered in the optimization procedure. Herein, we test model's sensitivity to this arbitrary threshold by running the optimization procedure for a range of values (0, 0.1, 0.2, 0.3, ..., 2.0). The remaining fitness error for the whole model as well as the non-signaling part alone is plotted in Figure 2.26. The arbitrary threshold (w_{cr}) has little effect on the signaling part of the model. Concerning the non-signaling part, increasing threshold values leads to drastic increase in the fitness error. For small values (0.0, 0.1, 0.2) fitness error is relatively stable (0.082 $\rightarrow$ 0.10), however, for values greater than 0.3 fitness error increases significantly.

Sensitivity in modeling decisions: $a^k_{j_{res}}$

In this part of model assessment we examine the model sensitivity to changes in the user defined constants $a^k_{j_{res}}$. The ratio $a^k_j/a^k_{j_{res}}$, determines how the ILP prioritizes signaling over cytokine release measurements. Throughout the analysis presented in this section we have selected $a^k_j = a^k_{j_{res}} = 100.0$. Herein we test a range of values (10, 20, 40, 76.1905, 100, 120, 150, 200, 250, 300, 350, 400, 450, 500) for $a^k_{j_{res}}$, and monitor changes in the remaining fitness error for the

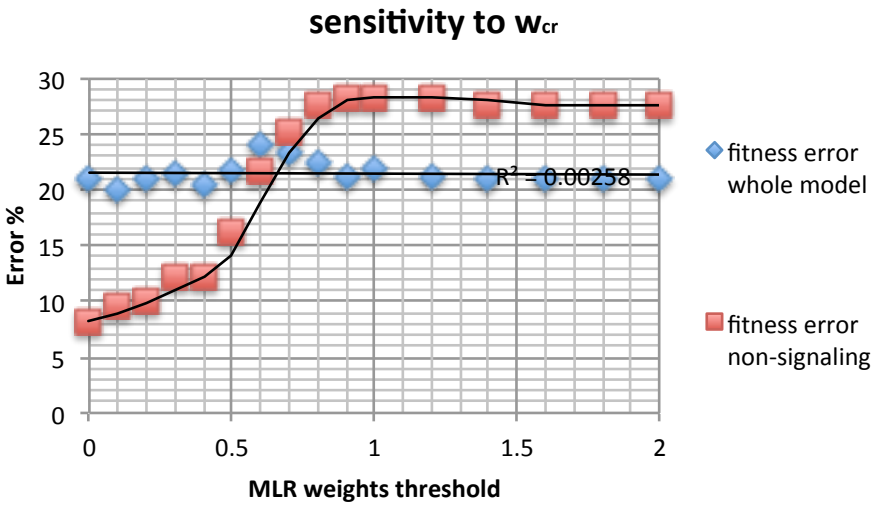


Figure 2.26: Sensitivity of proposed methodology to changes in w_{cr}. First curve (blue points) corresponds to remaining fitness error for the whole model as a function of the arbitrary threshold w_{cr} . Second curve (red points) corresponds to remaining fitness error for the non-signaling part of the model as a function of the arbitrary threshold w_{cr} .

whole model, as well as the non-signaling part alone (Figure 2.27). $a^k_{j_{res}}$ has little effect on the behavior of the model as a whole, since no conflicts exist between the phosphoprotein and cytokine release datasets. Concerning the non-signaling part of the model, small values of $a^k_{j_{res}}$ ($a^k_{j_{res}} = 10.0$) imply the cytokine release measurements have little weight on the objective function, and the respective reactions are excluded from the solution leading to an increase in fitness error of the response dataset. For greater values, error drops significantly and stabilizes to 10%. The value $a^k_{j_{res}} = 76.1905$ implies term $\sum_{j,k} a^k_j \mid x^k_j - x^{k,m}_j \mid$ equals to $\sum_k \sum_{j_res} a^k_{j_{res}} \mid x^k_{j_{res}} - \sum_{i_{res}} z^k_{i_{res}} w_{i_{res}} \mid$, thus, cytokine release data and signaling data are treated equally.

Impact of response measurements on pathway optimization

To assess the effects of response measurements on pathway optimization we optimized canonical pathways with the Huh7 dataset on three different ways: a) using the signaling data (left panel of Figure 2.28) as described previously [23], b) using response data (cytokine secretion, middle panel of Figure 2.28) and c) using both signaling and response data. Pathway results are presented in Figure 2.28.

When optimized with response data only, we find that pathways which are not connected with any cytokine releases (ex. INS, HGF) are removed during the optimization process. On the other side, because of HuH7 cells release cytokines only upon IL6, TNFa and IL1b stimulation, the ILP algorithm conserves those paths and connects them to the cytokine releases via the nodes that are highly correlated to the release such as IKB, STAT3, and P38. An interesting observation is the IL6 → STAT3 → Cytokine release. Despite the fact that the IL6 → STAT3 pathway had been removed when only signaling data were used (because STAT3 activation is below threshold as shown in IL6-induced STAT3 data in Figure 2.19) the same pathway is conserved when response data are used. The reason of that seemingly contradictory observation is because the small changes of STAT3 activation correlate well with a large number of cytokine releases and thus, the MLR algorithm attributed large correlation weight on the corresponding non-canonical edges.

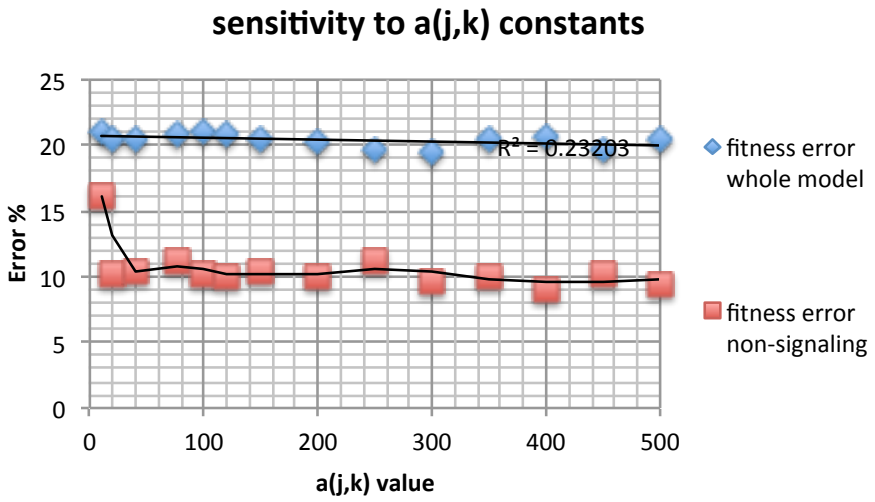


Figure 2.27: Sensitivity of proposed methodology to changes in $a^k_{j_{res}}$. First curve (blue points) corresponds to remaining fitness error for the whole model as a function of $a^k_{j_{res}}$. Second curve (red points) corresponds to remaining fitness error for the non-signaling part of the model as a function of $a^k_{j_{res}}$.

Subsequently, the ILP formulation conserves those non-canonical edges by connecting them to the respective stimulus. Thus, response data further constrain the optimization formulation and take advantage of the power of statistical analysis that identifies connections between signals and cellular response.

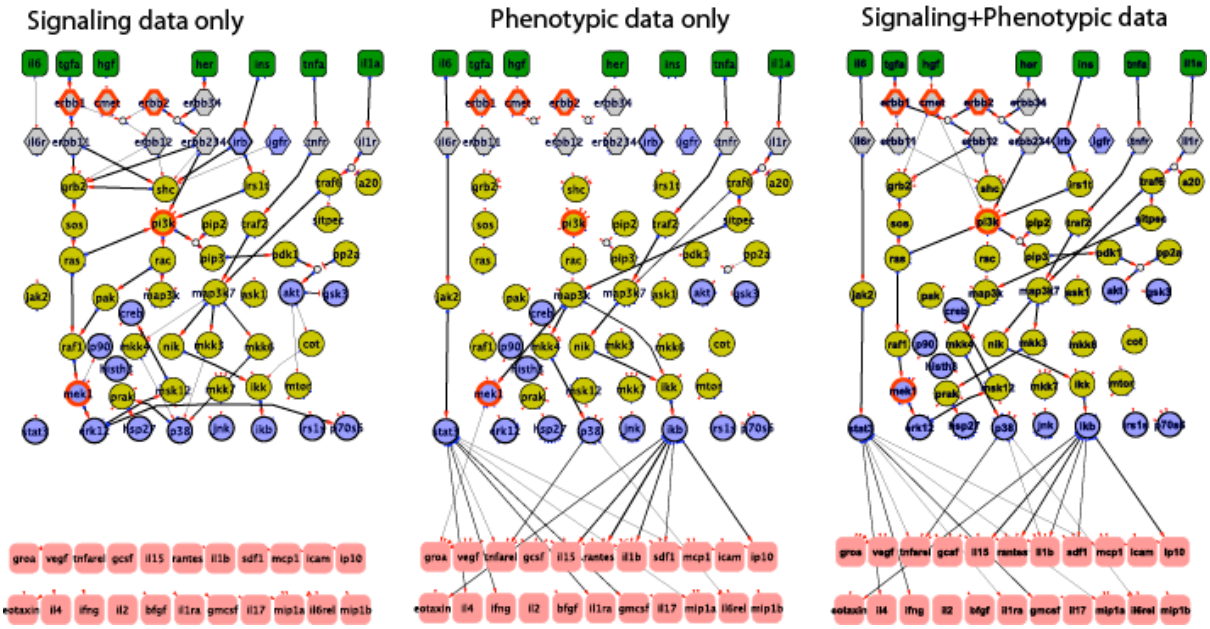


Figure 2.28: optimization of the canonical pathway using either (a) signaling data, (b)response data, and (c) both signaling and response data

Comparison of simulation runs for Huh7 and Normal cells

Differences regarding the signal transduction mechanisms of Primary and Huh7 cells are illustrated in the "Results" subsection with emphasis given on connectivity patterns of the optimized topologies. Herein we demonstrate how these differences are reflected on simulation runs of the generated models. Figure 2.29, features the signed matrix: $Y_{sim}^{DIFF} = Y_{sim}^{NORMAL} - Y_{sim}^{HUH7}$

The phosphoproteins part of Figure 2.29 reveals differential activation patterns for JNK, p70s6 and CREB signals, in addition to activation of the INS pathway present only in Huh7 cells. Concerning the prediction of cytokine releases, Primary cells release bFGF, Eotaxin, IL15, IL17, IL1b, MIP1a, MIP1b, GROa, and VEGF in contrast to Huh7 cells that release GMCSF, IL6, MCP1, IP10 and RANTES. Simulation runs presented here are in accordance with the connectivity patterns described in the results subsection.

Calibrating the weights of the three objective terms

The two prediction mismatch terms were given equal weights(= 1). In contrast, for the map-size reduction term a significantly smaller weight was selected (= 1/ 20). This weight was chosen based on the longest chain of consecutive reactions, namely 12, with the purpose to force the optimizer not to remove edges if they are essential for satisfying experimental results. For example, consider a chain reaction R1 $\rightarrow$ R2 $\rightarrow$ ... $\rightarrow$ R12 that should be kept because experimentally we found the relation "R1 = 1 implies R12 = 1". The reward for the optimizer to satisfy this chain reaction should be more than the penalty that it has to pay for keeping all 12 reactions. Therefore, if by keeping all reactions the map size reduction term increases the objective function by 12 units, then the reward for satisfying a chain of 12 reactions (mismatch term) should be higher than 12. The maximum chain in our pathway is 12 reactions but we choose 20 in case that further refinements in the generic topology increase the maximum chain.

Independent experimental validation of the model

In order to evaluate the predictive power of our hybrid model, we asked how well the Huh7 model shown in Figure 2.20b captures the correlation of cellular response to phosphoprotein activity. To achieve that, we choose the pathways IL1b/TNFa to P38/IKB that play major role in cytokine secretion, we block them with potent and selective IKB and P38 inhibitors, and we ask how well our model can predict the IP10 and RANTES, two major players for cytokine release. Figure 2.30 shows the experimental results and the mismatches with the hybrid model. Our hybrid model was able to recapitulate the IP10 release upon introducing IL1b, TNFa or both in an IKK dependent but p38/HSP27 independent manner. On the other hand, the hybrid model did not fit the induction of RANTES upon IL1b or TNFa stimulation probably not because there was no induction (an almost two fold trend can be seen in the IL1b induced RANTES) but because the induction does not pass the 0.5 threshold so the logic model to consider it an "ON" event. This issue highlights the importance of data normalization: currently data are normalized to the maximum cytokine value among all treatments. In the follow up experiments, one treatment is the combination of IL1b and TNFa where Huh7 cells show a super-induction of RANTES and makes all other RANTES values to be considered low. Logic models cannot handle such non-linear behavior and lead to predictive errors. When Huh7 treated with the combination of IL1b and TNFa then the hybrid model was able to perfectly recapitulate the RANTES release in an IKK dependent and p38/ HSP27 dependent manner.

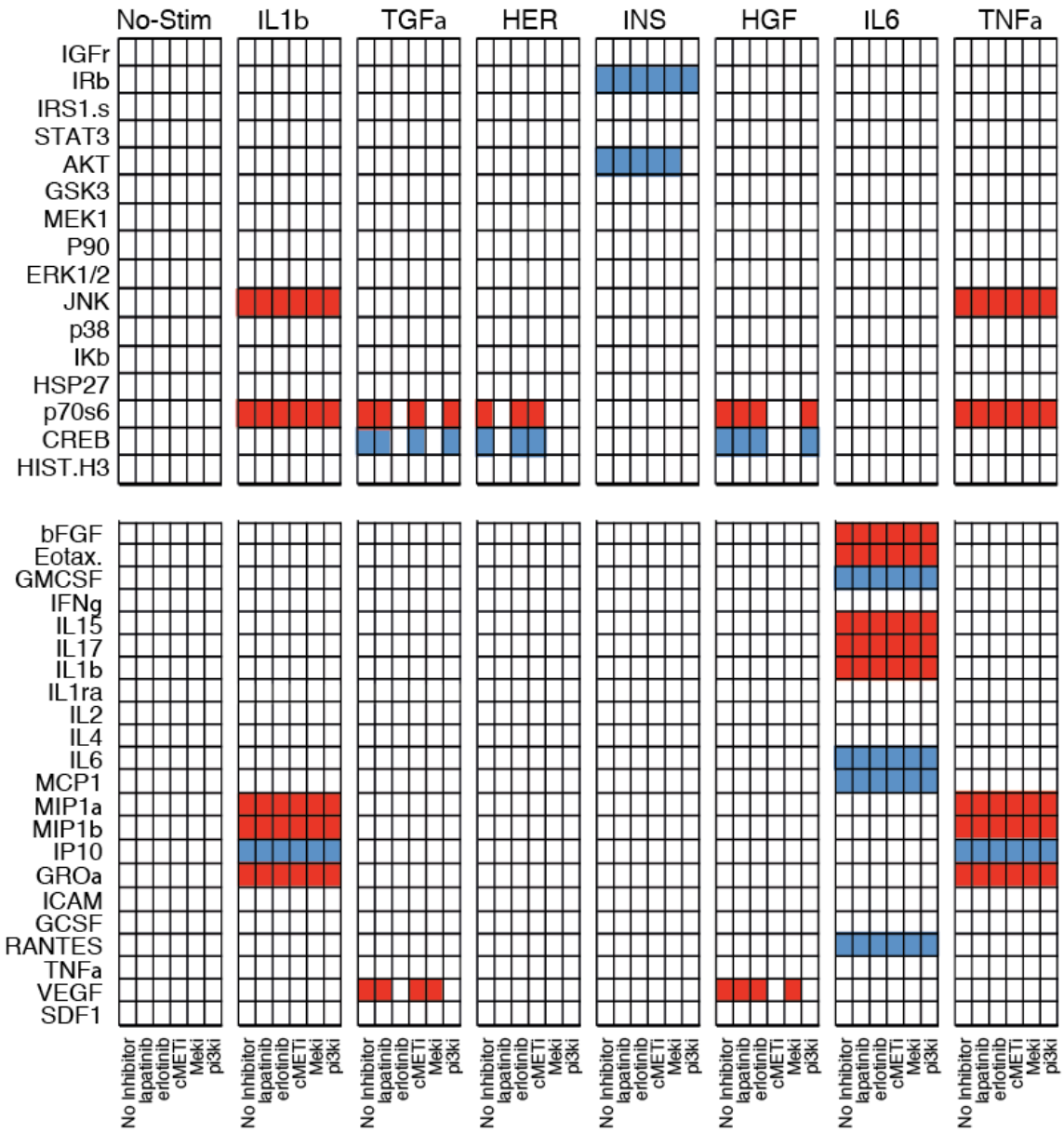


Figure 2.29: Comparison of simulation runs for Huh7 and Normal cells. Negative values (YNORMAL < YHUH7) are plotted in blue, positive values ($Y^{NORMAL} > ^{YHUH7}$) are plotted in red.

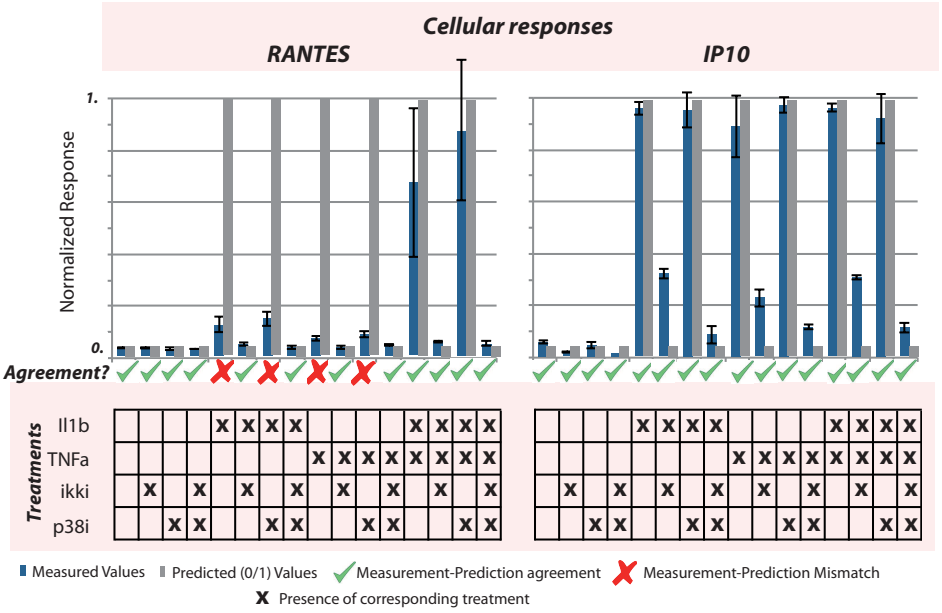


Figure 2.30: Validation of the hybrid model predictive power and evaluation NFkB and p38/HSP27 pathways in the release of RANTES and IP10 cytokines in Huh7 cells. Cytokines were measured upon combinatorial treatments of IL1b, TNFa stimuli and inhibitors for ikk (ikki) and p38/HSP27 (p38i). Agreement between hybrid model and experimental results is denoted with YES/NO symbols.

Chapter 3

A Non Linear Programming (NLP) Formulation for Quantitative Modeling of Protein Signal Transduction Pathways

In this work published in [86], An NLP formulation was introduced that models signal transduction using constrained fuzzy logic and by optimizing the model parameters, fits high throughput phosphoproteomic data. This work was carried out in collaboration with Alexander Mitsos (at the time when this work was published an assistant professor in the Mechanical Engineering Department of MIT, Cambridge, MA, USA, currently a professor in RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany) and Douglas A. Lauffenburger (head of the Biological Engineering department of MIT, Cambridge, MA, USA). In contrast to previous approaches the use of an NLP formulation speeds up the run time significantly, allowing the interrogation of large scale topologies and datasets.

3.1 Introduction on Constrained fuzzy logic for modeling signal transduction networks

Cross-referencing experimental data with signaling topologies, is a cornerstone in the attempt to gain biologically relevant insight for the interrogated signaling network. Traditionally, a mathematical formalism is adopted to model how signal propagates from one node to the next, and an optimization algorithm trains this model to measured data. Boolean modeling is amongst the most widely used approaches to model intracellular signal transduction [41, 42, 43, 44, 45]. In Boolean modeling, protein nodes assume only binary values $\{0, 1\}$, denoting the activation (or not) of the corresponding signaling molecule, and signal is propagated from the receptor level to downstream nodes using a combination of OR and AND gates. Even though Boolean modeling successfully captures the qualitative nature of signal transduction, it cannot model frequently occurring instances where protein activation assumes intermediate values.

To this effect, constrained fuzzy logic (cFL) adopts a quantitative yet static view of the signaling network (in contrast to ODEs that adopt a dynamic view of the signaling network). Proteins assume real values and a transfer function (TF) is introduced to propagate the signal from one protein to the next [31, 10]. A set of parameters in the TF defines its behavior and allows the calibration of the model to signaling data, in similar fashion to the pruning of the pathway in Boolean modeling. In [10] a two-step method was proposed, wherein first a GA

was used to remove all reactions that appear not to be functional based on the data at hand and estimate a rough approximation of transfer function parameters and in a subsequent step, a gradient based/ greedy algorithm was used to give a better estimate of the parameters. The cFL approach performed significantly better than Boolean modeling in terms of fitting the data but resulted in more parameters, raising concern about model over-parameterization and causing the training process to be computationally more expensive.

Computational efficiency and availability of data are amongst the main limiting factors in modeling via cFL. In the present work we introduce two new approaches for more efficient optimization of signaling pathways in a fuzzy logic framework. Firstly, we formulate the signaling activities as a regular optimization problem (i.e., a nonlinear program (NLP)), solved through IPOPT [87] under GAMS. Secondly, we introduce an aggressive compartmentalization scheme similar to the equivalent classes concept published in [32], to simplify the model at hand so it can be constrained with small datasets. In contrast to previous compression methods, the new compartmentalization procedure is capable of addressing complex connectivity patterns and feedback loops, decreasing in a more efficient manner network size, CPU time, and over-parameterization/non-identifiability caused by the lack of data [88]. As a result, the proposed NLP formulation allows for fast optimization of medium-scale topologies, and can also address the quantitative modeling of large scale signaling pathways. As a case study, we tackle the construction of cell type specific pathways in normal and transformed hepatocytes, to prove that our approach works for pathways as large as 15 receptors wide, numbering around 120 nodes and 230 reactions.

3.2 NLP formulation

Our approach is based on the utilization of a transfer function (TF) to model how signal propagates between nodes of the signaling network. Briefly, we implemented and tested the following transfer functions: (i) Unity function $f(x) = x$ (ii) linear function $f(x) = a\,x$ and (iii) normalized Hill function $f(x) = a(p^n + 1)\dfrac{x^n}{x^n + p^n}$. The normalized Hill function was chosen for being continuous, differentiable, monotonic, and fitting the expected qualitative trends of signaling reactions (sigmoid curve). The normalized Hill function was used in modeling signal transduction in [31, 10]. Reactions with multiple inputs are supported via AND and OR gates. In the case of an AND gate, all of the upstream nodes must be activated for the signal to propagate downstream, while in the case of an OR gate, one of the upstream nodes is enough to activate the downstream node. Normalized Hill function, AND and OR gates are shown in figure 3.1. In this work, we implement an NLP formulation to optimize the value of reaction parameters (a, p and n for every reaction), minimizing the difference between model predictions and measured data, resulting in a cell-type specific model of the signaling pathway. We then investigate if all reactions were necessary to fit the data by examining the parameters of the reactions and testing to determine if their removal significantly affect model fit.

The proposed NLP formulation is built based on a pre-existing ILP (Integer Linear Programming) formulation first published in [23] and thus uses the same nomenclature, repeated here for consistency.

Definitions

A pathway is defined as a set of reactions $i = 1, ..., n_r$; and species $j = 1, ..., n_s$. Each reaction has three corresponding index sets. Namely the index set of signaling molecules (or reactants) R_i , inhibitors I_i , and products P_i. These sets are all subsets of the species index set ($R_i, P_i, I_i \subset \{1, ..., n_s\}$); Typically, these subsets have very small cardinality (few species), e.g., $\mid R_i \mid = 0, 1, 2$; $\mid I_i \mid = 0, 1$; $\mid P_i \mid = 1, 2$; $\mid R_i \mid + \mid I_i \mid = 1, 2$. A set of in-silico experiments is performed mimicking the conditions of each actual experiment. The experiments are indexed by the superscript $k = 1, ..., n_e$. In each experiment a subset of species is introduced to the system and another subset is excluded from the system, in similar fashion to the "actual" experiments where a combination of stimuli and inhibitors are introduced to the cells. The predicted activation value of species j in experiment k is represented by the constant $x_j^k \in [0, 1]$. If available, the corresponding measured value is represented by $x_j^{k,m} \in [0, 1]$. The last group of variables introduced, $z_i^{k,m} \in [0, 1]$, represent the activity of reaction i in experiment k.

Objective Function

The objective function to be minimized is

$$\sum_{j,k} a_j^k \mid x_j^k - x_j^{k,m} \mid \tag{3.1}$$

and represents the weighted measurement-prediction mismatch; $a_j^k \in [0, 1]$ are user-set weights that may favor the fit of specific nodes in the pathway. In the present study, all nodes are considered equally important (have equal weights a_j^k).

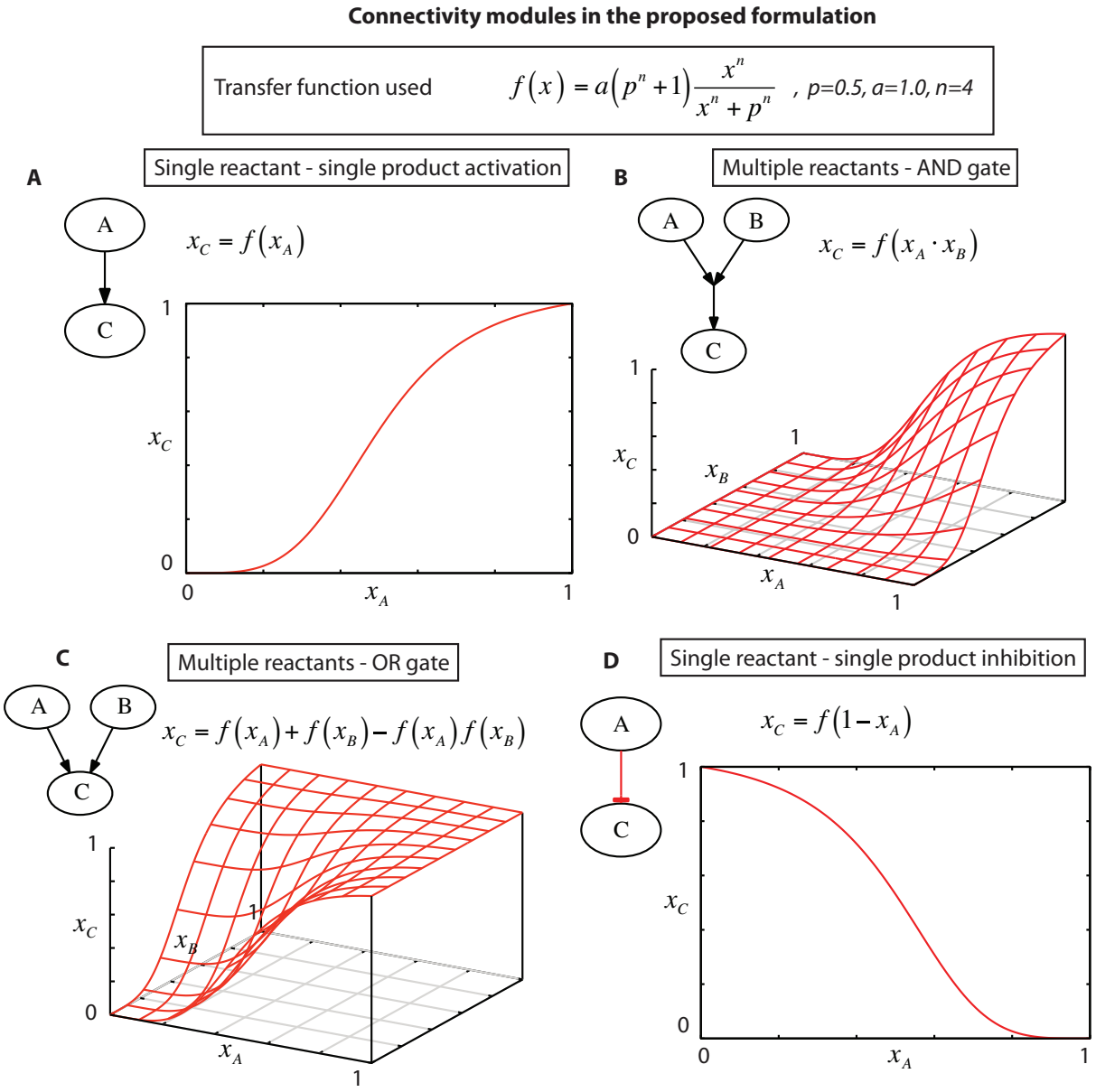


Figure 3.1: Connectivity modules of signaling pathways in the proposed constrained fuzzy logic formulation. The transfer functions supported by the proposed constrained fuzzy logic (cFL) formulation are illustrated. (A) "single reactant – single product" activation. (B) AND gate with two reacting species. (C) OR gate with two signaling species., (D) "single reactant – single product" inhibition. In all instances, function $f(x)$ refers to the normalized hill function, with $p = 0.5$, $a = 1.0$ and $n = 4$.

Single reactant – single product reactions

Reactions with a single reactant and a single product are modeled using the following transfer function (TF):

$$f(x) = a(p^n + 1)\frac{x^n}{x^n + p^n} \tag{3.2}$$

Equation (3.2) represents a normalized Hill function. Parameter p defines the midpoint of the curve (i.e. the value of x for which $f(x)$ equals to 0.5), n is the Hill coefficient and defines the steepness of the curve, whereas a is a scaling factor. The activity of reaction i in experiment k equals to: $z_i^k = f(x_j^k)$, where $j \in R_i$. The activation value of the downstream node equals to: $x_j^k = z_i^k$, where $j \in P_i$. In case species j is inhibitory we use: $z_i^k = f(1 - x_j^k)$, where $j \in R_i$.

Multiple reactants – single product reactions (AND gates)

In case more than one reactants are needed to propagate the signal to the downstream species, the activity of reaction i is modeled as a function of the bilinear product of the reacting species:

$$z_i^k = f(\prod_{j \in R_i} x_j^k \times \prod_{j \in I_i} (1 - x_j^k)) \tag{3.3}$$

The activation value of the downstream node equals to: $x_j^k = z_i^k$, $j \in P_i$. The bilinear product is chosen for satisfying key properties, such as continuity, differentiability and for reproducing the Boolean AND gate for 0 and 1 values of the reacting species.

Multiple reactions leading to same product (OR gates)

In case more than one reactions lead to the same product, the activation value of the downstream species is given by the following formulation:

$$x_j^k = b_{|T_j|}^k \tag{3.4}$$

where,

$$T_j = \{i \in \{1, ..., n_r\} : j \in P_i\} \tag{3.5}$$

T_j is the set of all reactions that have species j as their product. Let $i_1, i_2, ..., i_{|T_j|}$ denote the elements of T_j. Then, b_m^k is calculated recursively as:

$$b_m^k = b_{m-1}^k + z_{i_m}^k - b_{m-1}^k z_{i_m}^k; \quad 2 < m \leq \mid T_j \mid \tag{3.6}$$

$$b_2^k = z_{i_1}^k + z_{i_2}^k - z_{i_1}^k z_{i_2}^k \tag{3.7}$$

Implementation

The goal of the NLP formulation, described above, is the identification of optimal values for a, p and n parameters of each reaction to minimize the difference between model predictions and experimental data, as captured by the objective function in (3.1). The NLP was solved through IPOPT under GAMS. Additionally, an interface was developed in BASH scripting language to preprocess the PKN and generate the input files for the NLP algorithm. The DataRail toolbox was employed in MATLAB to handle and plot the dataset [50]. The optimization was run on Dual Quad Core IntelH XeonH Processors E5530 2.4 GHz, 12 GB, DDR3 RDIMM Memory, 1066 MHz. All results presented in this section were computed using a single core.

Definition of the search space

A systematic definition of the search space is vital for obtaining the best possible solutions within reasonable CPU time. A wider search space accounts for a bigger number of feasible solutions, possibly including some that minimize the objective function, but often increases the CPU time.

The model parameters to be estimated are: a, p and n; a serves as a scaling factor to limit protein activity in case the reaction appears not to be functional based on the data at hand,

and is defined in $[0, 1]$; p defines the midpoint of the curve (i.e. when x_j^k equals to 0.5) and can be any real number; n can be any positive integer, but here is fixed to 4, since the remaining parameters suffice to fit the data. In the toy model p was arbitrarily defined in $[0.3, 0.7]$. For the medium and large-scale topologies, we test a number of different upper-lower bound pairs, ranging from 0.1 to 2.0, to determine the one for which the algorithm performs best, in terms of goodness of fit, as well as decrease the required CPU time, facilitating the generation of a family of solutions. Goodness of fit is quantified by the mean absolute error (MAE) as calculated by the following formula

$$MAE = \frac{\sum_{j,k} \mid x_j^k - x_j^{k,m} \mid}{n_e \, n_s^m} \tag{3.8}$$

Results for the medium-scale topology are shown in Figure 3.2. The x-axis ($0.1 \rightarrow 2.0$) corresponds to the lower bound of p range; y-axis ($0.1 \rightarrow 2.0$) corresponds to the upper bound; while the z-axis corresponds to the MAE of the solution. Figure 3.2A shows that the quality of the solution mostly depends on the lower bound and less on the upper bound of p. In Figure 3.2B the corresponding CPU time is shown. As expected widening the range of p drastically increases the CPU time, since the search space becomes bigger. Based on these graphs the bounds of choice for p is $0.1 \rightarrow 0.4375$, since they provide both an excellent fit and low CPU time

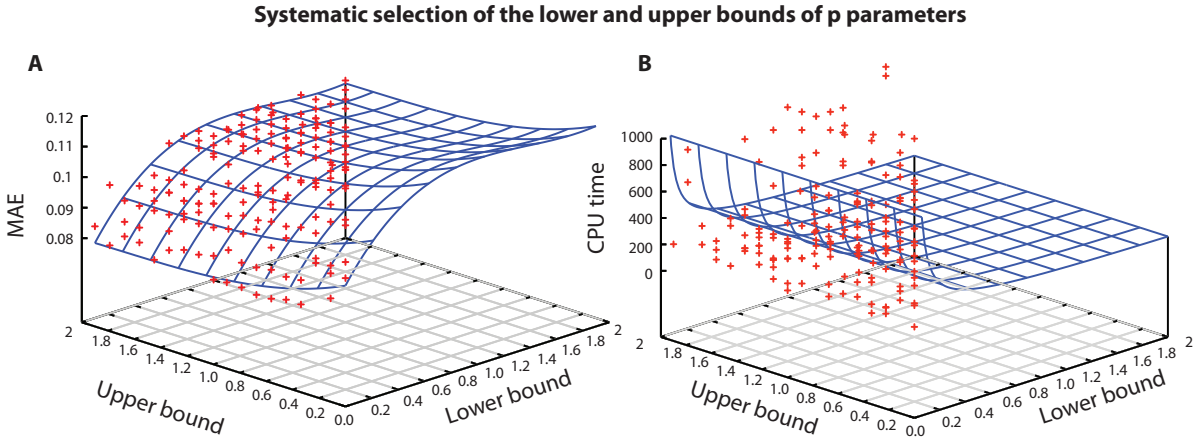


Figure 3.2: Systematic selection of the lower and upper bounds of p parameters. (A) Mean Absolute Error (MAE) as a function of the lower and upper bounds of p parameters of each reaction. The x-axis ($0.1 \rightarrow 2.0$) corresponds to the lower bound of p range; y-axis ($0.1 \rightarrow 2.0$) corresponds to the upper bound; while the z-axis corresponds to the MAE of the solution. The figure shows that MAE is mostly affected by the lower bound of p, smaller values of the lower bound lead to a better fit of the signaling data. (B) CPU time as a function of the lower and upper bounds of p parameters. CPU time is mostly affected by the lower bound of p. smaller values of the lower bound lead to increased CPU time.

Generation of a family of solutions

Instead of collecting a single solution that minimizes the objective function in (3.1), we collect a family of 500 near optimal solutions to account for slightly suboptimal pathways that may bare

strong biological significance, and avoid as much as possible terminating with a significantly suboptimal local minimum.

The proposed NLP approach optimizes the values of a and p to minimize the measurement – prediction mismatch as shown in equation (3.1). However, as long as the optimizer used is local, there is no guarantee that the obtained solution is a global minimum of (3.1). Moreover, there might be more than one solution (with different values for a and p), scoring the same (optimal) goodness of fit, which should be taken in consideration when biological insight about the interrogated system is to be extracted. Therefore, a large number of runs is performed each one starting from different (random) initial guesses, to obtain a family of near optimal solutions. Figure 3.3A shows the MAE of 500 solutions, obtained from equal runs of the proposed NLP approach each one starting from a different initial guess for the parameters a and p. Most of the runs resulted in solutions with very similar MAEs. This indicates that although the IPOPT optimizer, used herein, is not global, it furnishes near-optimal solution points independently on the initial guess. In Figure 3.7C, an "average" pathway for these 500 runs is illustrated. The opacity of each of these edges corresponds to the average activity of the respective reactions over the 500 runs.

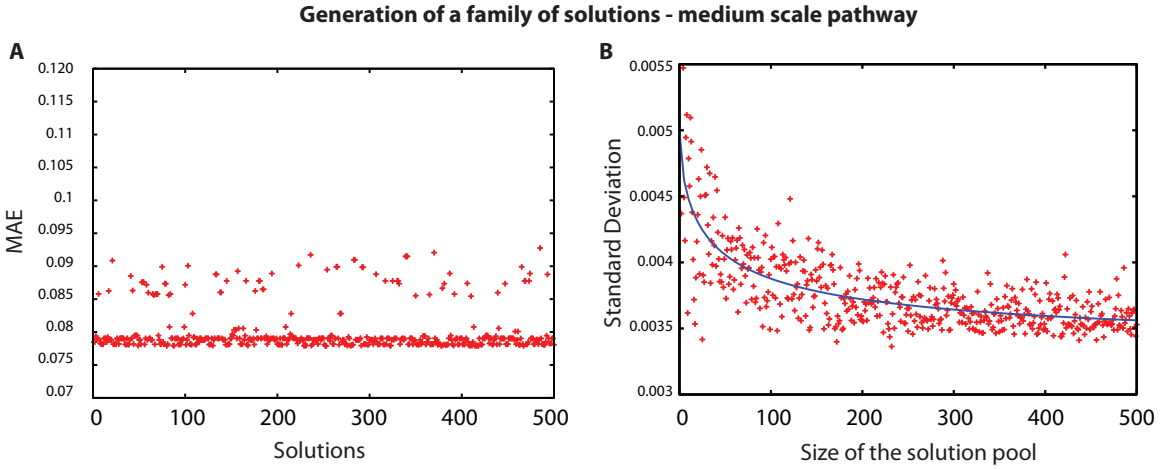


Figure 3.3: Generation of a family of solutions – medium-scale pathway. (A) The MAEs of a family of 500 near optimal solutions. The x-axis corresponds to the different runs; the y-axis corresponds to the MAE of the solution. (B) Standard deviation of the MAEs in a family of solutions as a function of the family's size. The x-axis represents the size of the family of solutions; y-axis represents the standard deviation of the solutions. The bigger the size of the family of solutions the smaller the standard deviation of the solutions becomes, indicating decreased sample variability. Optimum size would be around 150-200 solutions where the standard deviation has dropped close to its final value.

Removing conflicting and redundant reactions from the PKN

Optimization of the PKN to the data at hand results in a set of values for the model parameters (a and p) that minimize the measurement prediction mismatch, as defined in equation (3.1). Subsequently, we iteratively remove reactions from the PKN (every time a reaction is removed we re-optimize the PKN) while monitoring the fitness error to identify all reactions that are not vital in fitting the signaling dataset, either because they directly contradict the data, or

because they are non-identifiable. Non- identifiable reactions are those whose presence in the model cannot be validated nor disproven based on the data at hand. This may occur when signal transduction from a cytokine to a measured protein can be achieved by a number of different pathways, and there is no definite way to identify which one is really functional. Consequently, removing a non-identifiable reaction from the PKN has no effect on the fitness error.

In an attempt to remove conflicting reactions and tackle over- parameterization, we gradually remove reactions from the PKN until the fitness error starts increasing (i.e., the algorithm can no longer fit the dataset at hand). At that point there are no more conflicting or non-identifiable reactions left in the model, but all of the remaining ones are vital for fitting the data. At every iteration, the reaction with the lowest activity is removed (variable z_i^k in the formulation). The activity of each reaction mostly depends on the parameter a (gain) of the reaction and directly correlates to the "amount of signal" propagating downstream. In this manner, the least significant reaction is removed at every iteration. Even though the sequence reactions are removed by will affect the obtained solution (i.e., the solution is not unique), it is guaranteed to be optimal since only conflicting/non-identifiable reactions are removed and key property of these reactions is that their removal does not affect the fitness error of the solution.

Results are illustrated in Figure 3.4. Figure 3.4 shows how the algorithm performs when reactions of the PKN are removed in order of increasing significance. The x-axis corresponds to the number of removed reactions, while the y-axis corresponds to the MAE of the solution. As illustrated in Figure 3.4, up to 10 reactions can be removed (20% of the initial topology) without affecting the goodness of fit of the solution. More than that, vital reactions are missing and the MAE increases significantly. Small fluctuations in the figure are attributed to variations of the fitness error of the solutions (63%). Figure 3.7C, 3.7D shows the solution after removing conflicting and non-identifiable/redundant reactions.The above-mentioned procedure results in the identification of one of possibly many optimal and identifiable solutions, the superposition of which is the family of solutions.

Compartmentalization of the large-scale topology

Before optimizing large-scale networks, the PKN is compartmentalized by grouping together nodes that share identical response under all experimental conditions, to reduce the parameter space.

In similar fashion to the medium-scale model in Figure 3.7, the large-scale pathway in Figure 3.11 also includes a number of non-identifiable reactions, in the sense that signal transduction from a cytokine to a measured protein can be achieved by a number of different pathways and there is no definite way to identify which one is truly functional. In pathways of this size, however, is not efficient to exhaustively remove reactions until the optimizer can no longer fit the data at hand. Instead we propose an alternative method for reducing the parameter space. We propose a compartmentalization scheme, based on the "equivalent classes" concept introduced in [32], for "grouping" nodes that share identical responses under all experimental conditions; thus resulting in an equivalent (compartmentalized) model where nodes have been replaced with their respective compartments, and reactions between nodes are now reactions between compartments. In more detail, we define a compartment (C) as every set of non-measured species ($j \in C$), such that $x_1^k = x_2^k = ... = x_{|C|}^k$ for every $k = 1, ..., n_e$. Where $k = 1, ..., n_e$, is the set of experiments. x_j^k is the predicted value of species j in experiment k.

In this case study, we simulate the pathway running the NLP formulation under all experimental conditions present in the signaling dataset with nominal values for all parameters; subsequently, we format the simulation results in a 2d matrix, rows corresponding to the nodes in the pathway and columns corresponding to the different experimental conditions; we identify

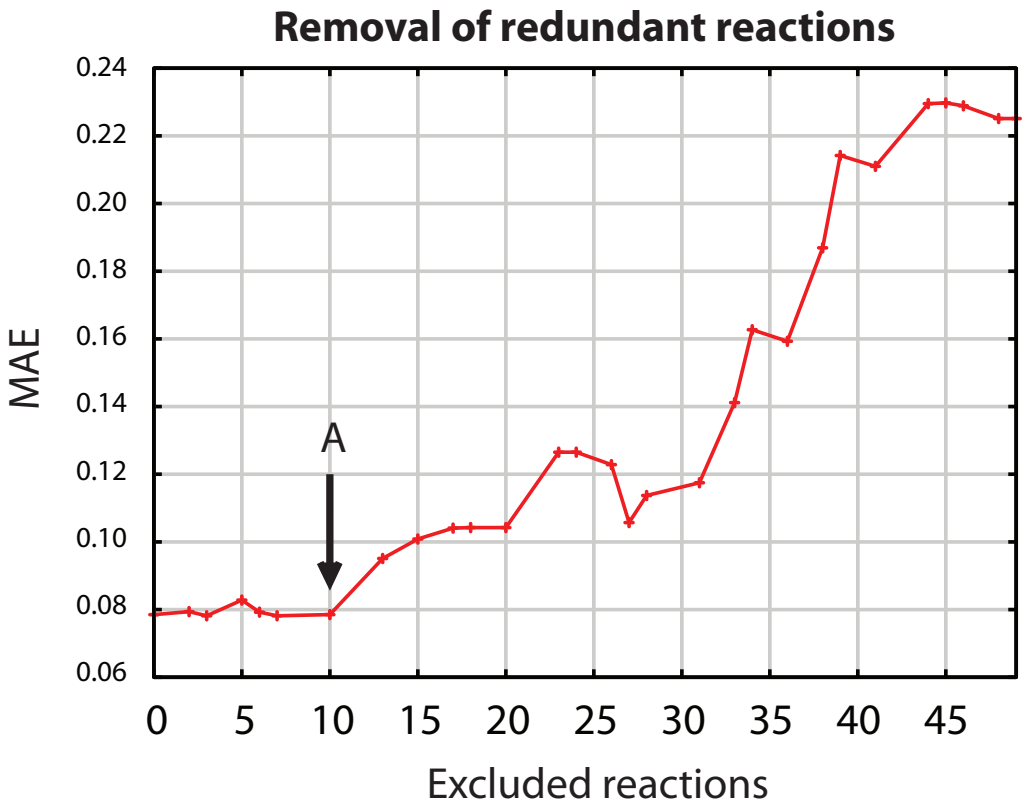


Figure 3.4: Addressing over-parameterization (medium scale pathway). Reactions are exhaustively removed from the PKN in order of increasing activity, and the fitness error is monitored. The x-axis shows the number of reactions excluded; the y-axis shows the Mean Absolute Error of the solution. The figure shows the dependency of the MAE from the subset of excluded reactions. Up to 10 reactions can be removed from the PKN without affecting the MAE of the solution (arrow A), implying these 10 reactions are not vital in fitting the signaling data (redundant reactions). Beyond this point vital reactions are removed, the optimization algorithm can no longer fit the data at hand and the fitness error increases drastically.This is where the final (optimal and identifiable) solution is obtained. Small fluctuations in the figure are attributed to variations of the fitness error of the solutions ($\pm 3\%$).

the nodes that share the same response under all conditions (i.e., identify replicate lines) and group them together in compartments; we replace every node in the PKN with its corresponding compartment and remove replicate reactions. This procedure is implemented using BASH. Since the nodes in a compartment share identical responses under all experimental conditions, their connectivity inside the compartment cannot, in principle, be interrogated based on the data at hand. Thus, it is purposeful to group these nodes together and update the PKN replacing nodes with the compartments they belong into. By doing so, we drastically decrease the parameters space.

Application of the compartmentalization scheme to an illustrative example

To better illustrate how the proposed compartmentalization scheme works to simplify the interrogated model, we construct the example model of Figure 3.5A. Node "A" serves as input to the pathway (stimuli), and activates nodes B1, B2; these interact with each other and finally activate node "C" that serves as a readout (signal). The proposed scheme groups B1-B2 into "Cmp" and simplifies the model as illustrated in Figure 3.5B. If data dictates: $A = 1$; $C = 1$, then reactions A $\rightarrow$ Cmp and Cmp $\rightarrow$ C are conserved. Else if $A = 1$; $C = 0$, then at least one of the above mentioned reactions have to be removed.

Figure 3.5C, 3.5D demonstrate how the compartmentalization scheme can be too restrictive and may decrease the quality of the solution. In Figure 3.5C input nodes A1, A2 are connected to latent nodes B1 and B2; B1 activates C1 and B2 activates C2. After the compartmentalization procedure, B1 and B2 are replaced with compartment "Cmp" that activates C1 and C2 (Figure 3.5D). In the case where C1 is activated by A1, and C2 by A2; then either C1, or C2 will be misfitted in the compartmentalized model, since differential activation of C1 and C2 is possible only if either CMP $\rightarrow$ C1, or CMP $\rightarrow$ C2 are removed from the pathway. However, if either one of the two reactions are removed, then the respective signal (C1 or C2) will remain inactive under all conditions, thus misfitting the data. If no compartmentalization is performed, then the pathway can be optimized by removing (or decreasing the activity) of A1 $\rightarrow$ B2 and A2 $\rightarrow$ B1. This increase in fitness error caused by the compartmentalization procedure implies that grouping nodes B1 and B2 in the compartment Cmp should not have taken place if data were to fit perfectly. Cases like this may arise when limited experimental conditions are available, since it is more likely for nodes to be grouped together. E.g. If only one condition is available, then all nodes will be grouped in a single compartment. In such cases compartmentalization of the PKN is not recommended. In all cases the solution should be manually inspected to ensure that the remaining fitness error is not caused by the aggression of the compartmentalization scheme.

3.2.1 Results

Optimization of a toy model

To illustrate how the proposed formulation fits parameters a, p and n to signaling data, we used the 10-node toy model shown in Figure 3.6A consisting of two stimuli (green nodes); two inhibitors (red nodes); 5 measured signals (gray nodes); 4 OR gates (e.g.,TNFa OR PI3K $\rightarrow$ JNK); 4 AND gates (e.g.,TGFa AND NOT MEK1/2i $\rightarrow$ MEK1/2); and 4 NOT gates (total number of parameters = 20). In-silico data are shown in Figure 3.6B and consist of 3 stimuli (green nodes); the activation levels of 5 signals (gray nodes); and 2 inhibitors (red nodes) (total number of data points = 45). The red background color in the data (Figure 3.6B and D) represents the initial and after-optimization measurement- prediction mismatch of the model. For example, MEK1/2 signal under TNFa, without any inhibitor being present, was initially misfitted by the PKN. i.e. The data showed no activation, while in the PKN, MEK1/2 was clearly activated by TNFa. After the optimization procedure the red background was removed, implying that, in the optimized model, TNFa did not activate MEK1/2.

The goal of the NLP formulation is to minimize the fitness error by searching for optimum values of the parameters a, p and n within predefined bounds. For the toy problem the bounds were: $p = 0.3 \rightarrow 0.7$ while the exponent was held constant $n = 4$. The upper and lower bounds for p were defined in such a manner that $p = 0.3$ corresponded to an over-responsive transfer function and $p = 0.7$ corresponded to an under-responsive transfer function, while $p = 0.5$ was the initial guess for the p parameter. Parameter a acts as a scaling factor and serves to limit the

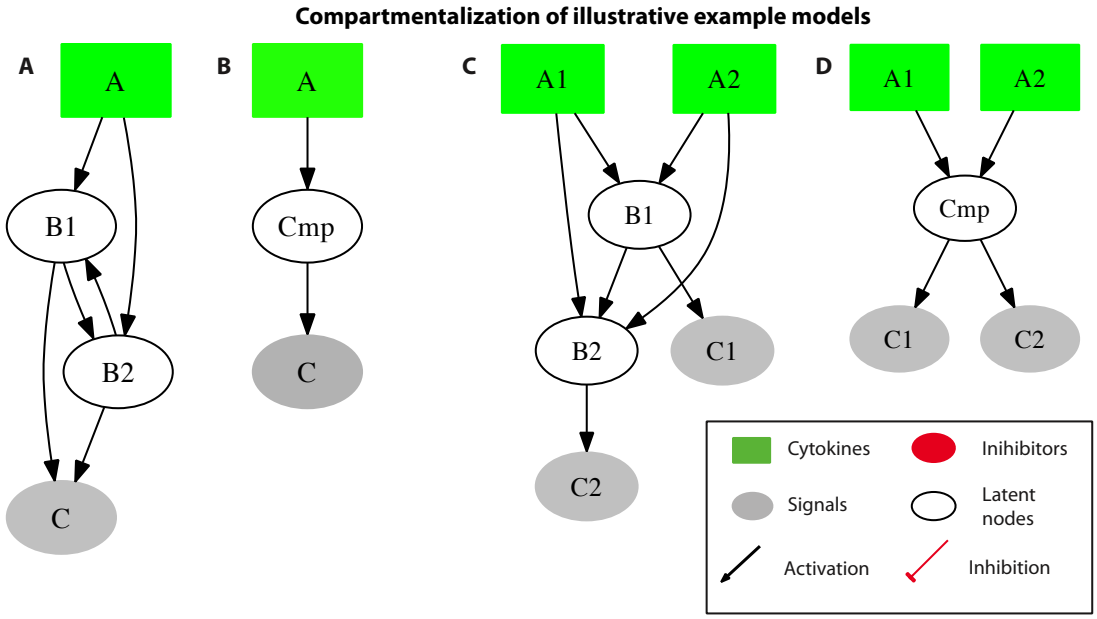


Figure 3.5: Compartmentalization of illustrative example models. The compartmentalization of two example models is featured. (A) example model with a single input (green node) single output (grey node) and 2 latent nodes (white nodes). (B) Compartmentalized version of the example model in (A). The two latent nodes are grouped in compartment Cmp. (C) example model with two inputs, two outputs and two latent nodes. (D) Compartmentalized version of the example model in (C). The proposed compartmentalization scheme is over-aggressive decreasing the quality of the solution in case the two measured proteins have different response under A1 and A2.

activity of those reactions that appear not to be functional based on the data at hand. Although the initial selection of upper and lower bounds for the p parameters together with the value of n was done arbitrarily, in case of high remaining fitness error these values can be updated and the algorithm rerun to guarantee the best possible solution.

Figures 3.6C and 3.6D present the optimization results of the toy model. In Figure 3.6C, the activity of each reaction is visualized using arrows in gray scale; reactions with larger a parameters effectively transmit more signal downstream (are more active) and have a more solid color. The transfer functions themselves are illustrated in Figure 3.6E. The efficiency of our approach is validated by the eradication of most of the fitness error as shown in Figure 3.6D (red background). The optimization eliminated the PI3K to JNK, PI3K to P38, and PI3K AND NOT MEK1/2i to MEK1/2 reactions (bottom right panels in Figure 3.6E). Manual inspection of the data and the initial topology can confirm this decision: JNK and P38 were activated upon TNFa stimulation alone; therefore reactions from the TGFa pathway to JNK and P38 were not active. On the other hand, TNFa stimulation induced AKT activation but did not affect MEK1/2 or ERK1/2, implying that the PI3K to MEK1/2 reaction was not active. To validate that reactions i) PI3K to MEK1/2, ii) PI3K to JNK and iii) PI3K to P38 were not active in the optimized model; we manually removed them from the initial model and run the NLP algorithm once again. No significant differences were observed between the two optimized models, indicating that these three reactions were not vital to fit the data (data not shown).

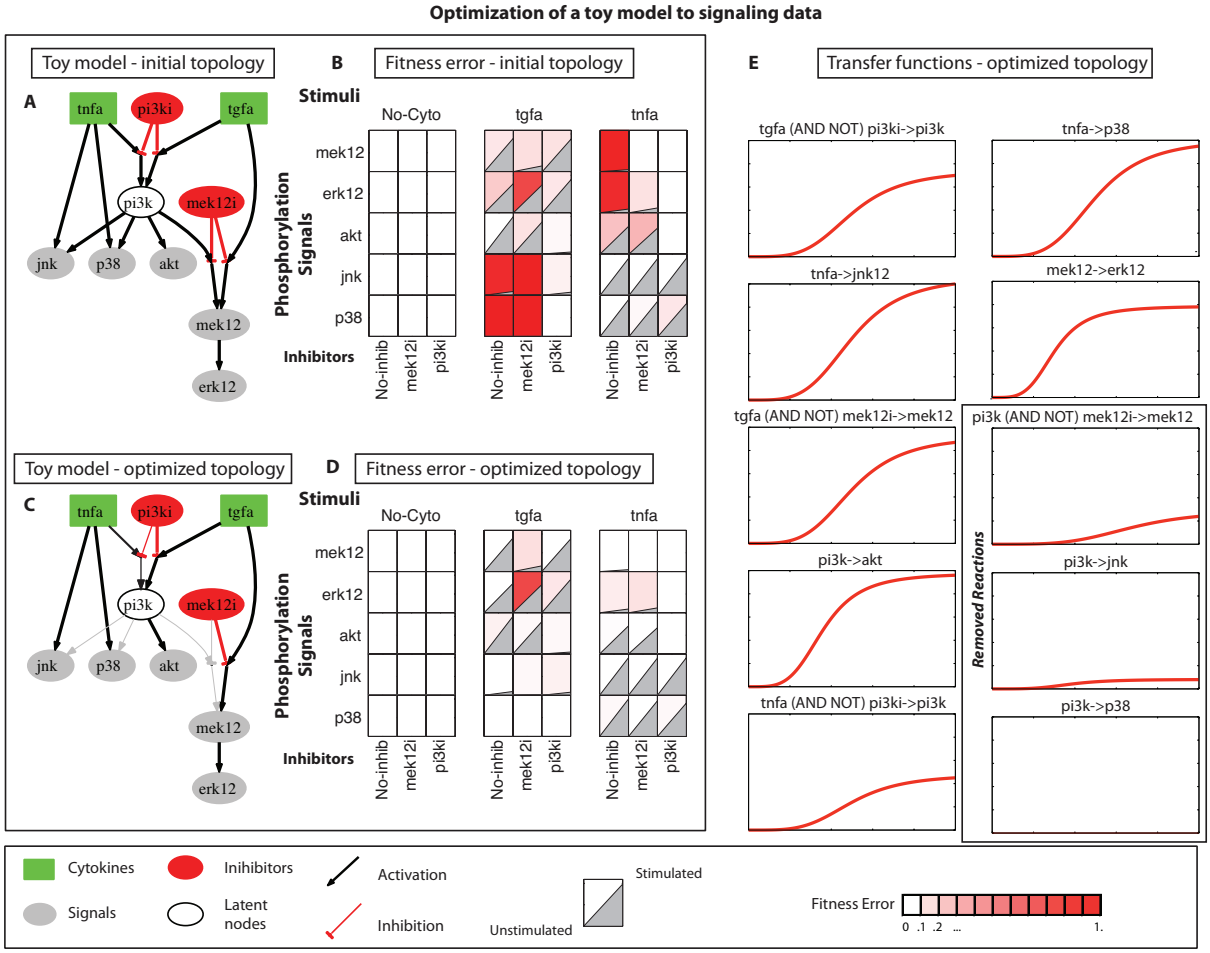


Figure 3.6: Optimization of a toy model to signaling data. (A) Generic pathway is represented as a signed directed graph, also refers as PKN. Green nodes refer to different cytokines (ligands) where the signaling process initiates; Red nodes refer to inhibitors present in the in-silico dataset; Grey nodes refer to measured proteins; White nodes refer to latent species, i.e. proteins whose activation state is not measured. (B) In-silico signaling data under combinatorial treatment with stimuli (TGFa, TNFa, no-treatment) and inhibitors (mek12i, pi3ki, no-inhibitor). Each subplot shows the average activation level within 30 minutes upon stimulation [12]. Red background refers to model-prediction mismatches (C) Optimized pathway, grey arrows refer to reactions with limited activity (z_i^k) (caused by a parameters being close to 0). The opacity of each edge corresponds to the activity (z_i^k) of the corresponding reaction. (D) In silico signaling dataset and fitness error after the optimization procedure. Decrease in the red background color shows the optimized model is in accordance to the signaling dataset. (E) Optimized transfer functions presented in C.

Optimization of a medium-scale signal transduction pathway

Background. Next, we tested the proposed NLP approach to the medium-scale signaling pathway used in [10], which numbers a total of 52 reactions and 37 species (total number of model parameters = 104). The training dataset was constructed using the xMAP technology on transformed human hepatocytes (HEPG2 cells) [12] and numbers a total of 728 datapoints. The initial topology and the experimental dataset are illustrated in Figure 3.7A and B. The pro-

posed approach was implemented in 3 steps: (i) definition of search space for the p parameters of each reaction, (ii) generation of a family of solutions and (iii) exhaustive removal of reactions from the PKN to address over-parameterization (see previous subsections).

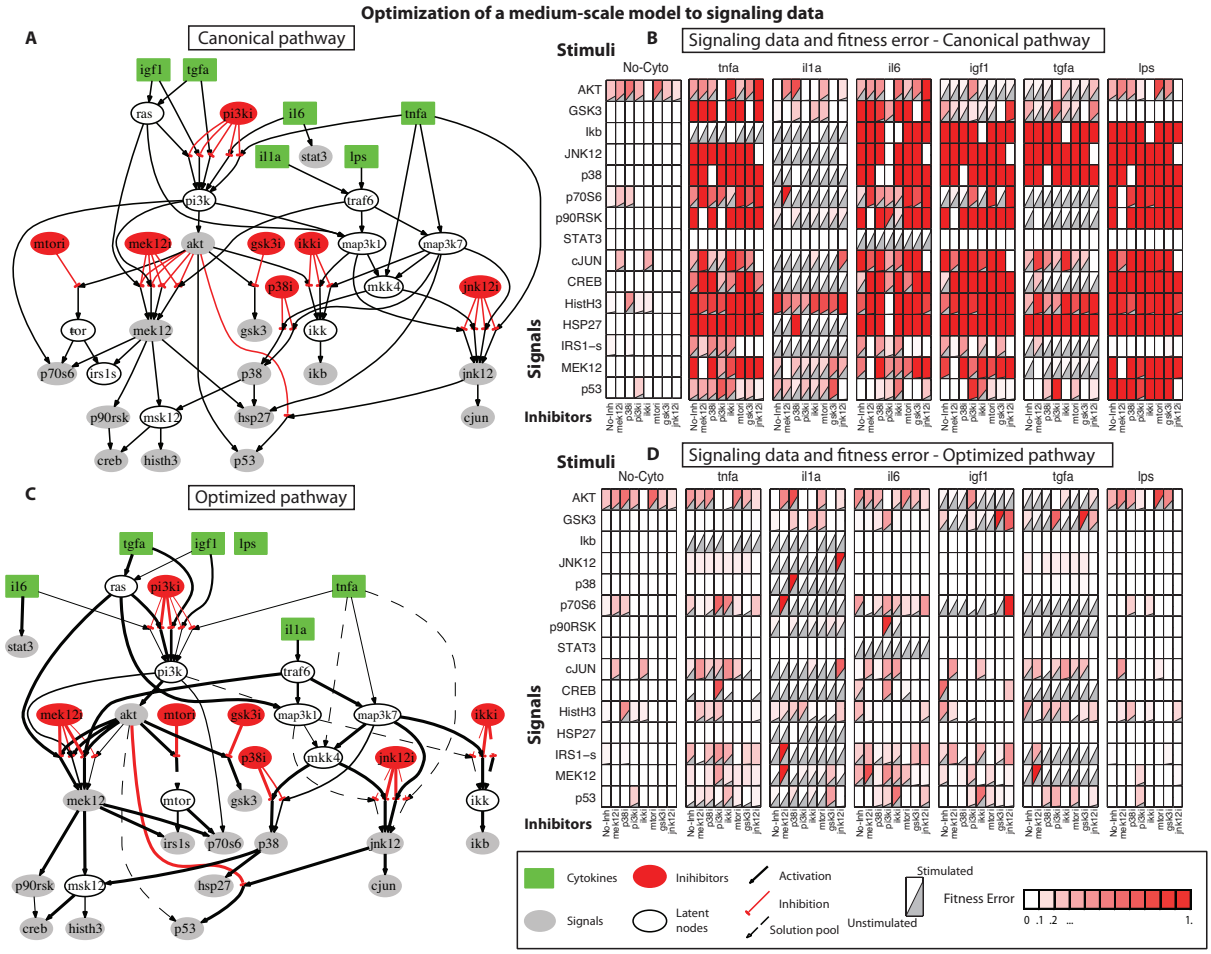


Figure 3.7: Optimization of a medium-scale model to signaling data. (A) Initial topology as presented in [10]. (B) Signaling data under combinatorial treatments of 6 stimuli (green nodes) and 7 inhibitors (red nodes) reporting 15 signals (grey nodes). The red background represents the measurement – prediction mismatch of the initial topology (46%) (mean fitness error). To generate model predictions, the initial guesses of all model parameters were used ($a = 1.0$, $p = 0.5$). (C) Optimized pathway. Bold lines refer to the optimized pathway after removing redundant/conflicting reactions. Dashed lines refer to reactions present in the family of solutions that although being redundant are reported since they may bare biological significance. The opacity of each edge corresponds to the activity (z_i^k) of the corresponding reaction. (D) Signaling dataset and remaining fitness error (8%) (mean fitness error). The red background refers to the fitness error of the solution. Decrease in the red background compared to (B) implies the optimized model successfully fits the signaling dataset (mean fitness error went from 46% to 8%). A and C were generated using graphviz package (http://www.graphviz.org/). B and D were generated using Datarail toolbox [50].

Optimization results. Figure 3.7C contains an "average" pathway for 500 solutions. The solid lines are the minimum set of reactions needed to fit the experimental dataset and the opacity of each of these edges corresponds to the maximum activity (z_i^k) of the respective reactions.

The dashed lines are reactions that were present in the family of models but could be removed for being redundant based on the analysis in a previous subsection. Figure 3.7D presents the signaling dataset together with the measurement prediction mismatch for the optimized model (red background). The average CPU time of each run was 10 minutes.

Several interesting features can be uncovered from the proteomic-driven optimization of the generic pathway: LPS pathway was deactivated altogether since it only partially affected the AKT signal. IGF1 and TGFa signaled through PI3K and activated AKT, GSK3 and P70. Moreover, TGFa activated MEK1/2, P90, CREB, IRS1S and HISTH3 via RAS. TNFa and IL1a also had partially overlapping pathways signaling through the MAP3Ks. IL1a signaled through TRAF6 to MAP3K7 and then to JNK, CJUN, P38, HSP27 and IKB. IL1a also activated MEK1/2 via TRAF6 and then P70S6, P90RSK, CREB, IRS1S and HISTH3. TNFa, on the other hand, signaled through MAP3K7 but had clear effects only on IKB, while partially activated a number of signals such as CJUN and P53. Moreover, TNFa partially activated P70S6, CREB, IRS1S and MEK via PI3K.

As shown in Figure 3.7D, most of the measurement-prediction mismatch has essentially been removed by the optimization procedure. The remaining fitness error is below 8% (mean fitness error). Residual errors appear either in areas of the pathway where the a priori knowledge was poor, or where erroneous measurements in the experimental dataset conflicted each other. The latter is shown in the JNK signal under IL1a and JNKi. Even though JNKi was supposed to have inhibited JNK activation upon IL1a stimulation, the data shows that JNK remained active. In such cases the NLP algorithm is not able to reproduce the respective datapoint. Similar case consisted the misfitting of i) CJUN under IL1a and JNKi, ii) MEK1/2 under IL1a, IL6, TGFa and MEK1/ 2i, iii) P38 under IL1a and P38i, iv) GSK3 under IGF1, TGFa and GSK3i, and so forth. Those residual errors appeared in almost all optimization procedures [13, 10]. In conclusion, despite the residual error, the optimized model successfully captured the patterns underlying the signaling dataset.

Cross-validation. For the optimization of the PKN, the signaling dataset in its entirety is used. Herein, however, to better evaluate the performance of the proposed formulation, we performed a cross validation study where random portions of the dataset, of increasing size, were left out of the training process, model predictions corresponding to this data were computed and then compared to the measured data evaluating the measurement prediction mismatch. Figure 3.8 illustrates the fitness error corresponding to all measured data (total fitness error), in blue, together with the error corresponding to the excluded data (in red). Interestingly, up to 40% of the dataset could be removed before the fitness error started increasing significantly, implying the proposed formulation is robust against missing data. Moreover, the algorithm performed relatively well even with 80% of the dataset missing. After this point, a steep increase in the overall error was observed, since key pathways were removed and the fitness error quickly reached that of the null solution.

Optimization of a large-scale signal transduction pathway

Background. In order to evaluate the performance of our optimization procedure, we asked whether we could apply the procedure to larger pathways. Here, we focused on pathways that are experimentally identifiable using ELISA type of assays and thus are limited in well-known signal transduction mechanisms. The resultant PKN accounts for dozens of stimuli and their downstream nodes [46]. The pathway contains 228 reactions and 117 species (total number of model parameters = 456). The corresponding data were measured using the xMAP technology on primary human hepatocytes and consist of a total of 120 multi-combinatorial experiments. Cells were perturbed with combinations of 15 stimuli and 3 inhibitors (including the No-inhibitor

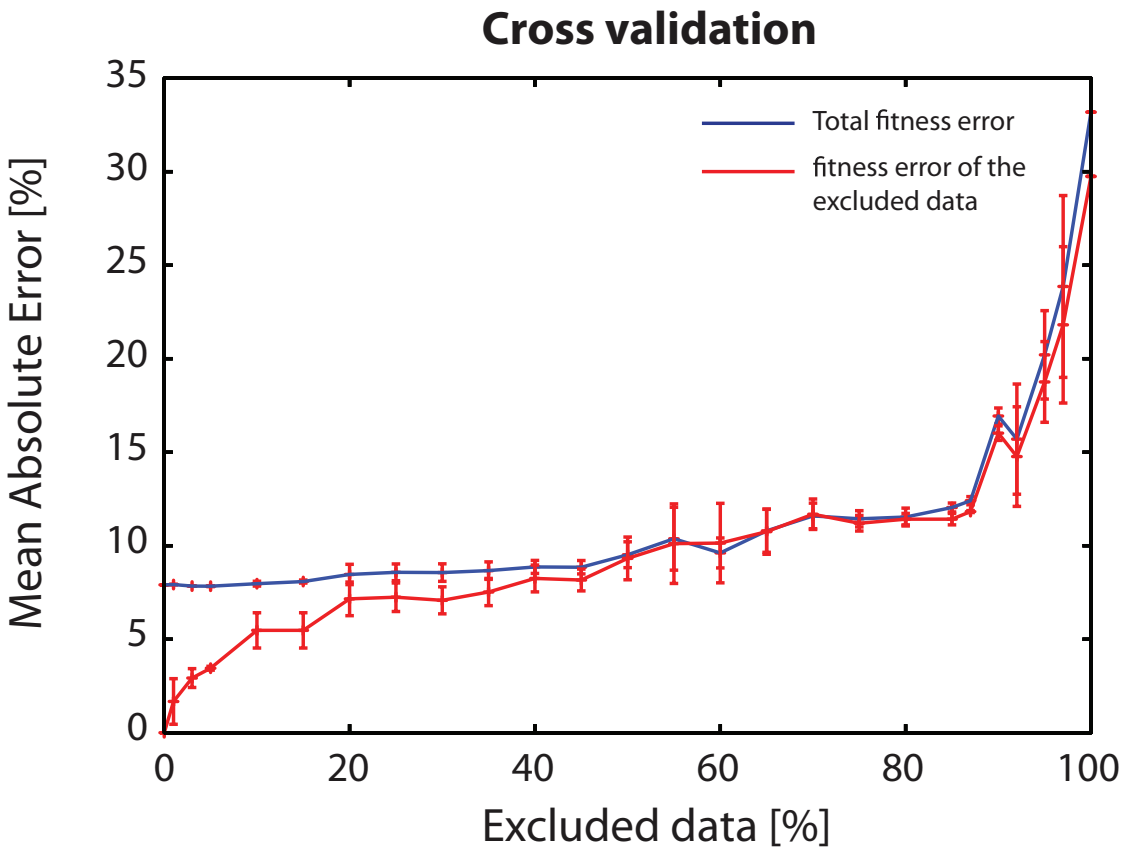


Figure 3.8: Cross validation of the NLP algorithm (medium-scale pathway). Blue line represents the fitness error corresponding to all measured data (total fitness error); red line corresponds to the fitness error of the predicted (excluded) data. Total fitness error initiates at ,8% (mean fitness error) and stays relatively stable for excluded portions of the dataset smaller than 40% of the total. Implying that the proposed approach handles efficiently missing data. Even when no data is excluded (0% point in the plot) the total fitness error is at 8% (mean fitness error) because of conflicts in the data or poor prior knowledge of protein connectivity in the PKN. The fitness error corresponding to the excluded data (red line) initiates at 0% since the removal of random portions of the dataset may leave out of the training process datapoints that are easily inferred from the remaining data. E.g. measurement of MEK1/2 under TGFa and IKKi is easily inferred from TGFa and no-inhib experiment. As increasing portions of the data are left out of the training process (excluded data 40%) the fitness error increases significantly. For excluded portions greater than 80% the fitness error quickly reaches that of the null solution.

treatment), while 14 key phosphoproteins were measured (total number of data points = 1680). Before the optimization procedure, the pathway was compartmentalized to reduce the parameters space (the compartmentalized pathway numbers 44 species and 69 reactions, total number of model parameters = 138), while a family of solutions was obtained to guarantee that the algorithm is not trapped in a local minimum.

Optimization Results. In Figure 3.10 the optimized, compartmentalized version of the large-scale pathway is shown, together with the measurement-prediction mismatch. To demonstrate

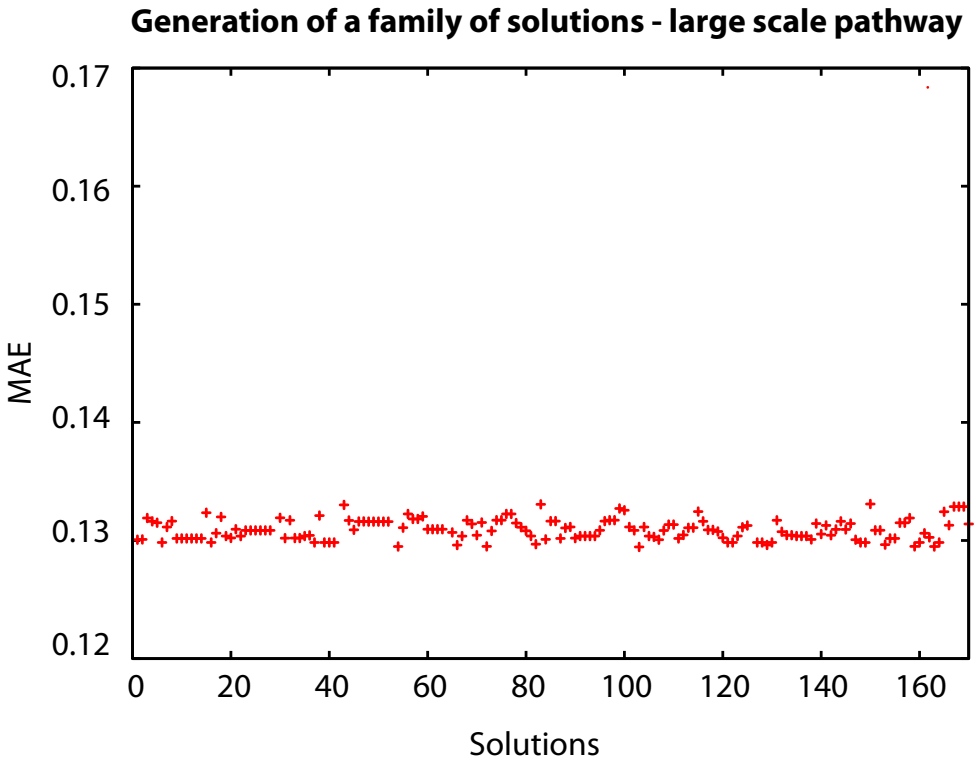


Figure 3.9: Generation of a family of solutions – large-scale pathway. The MAEs of a family of 170 near optimal solutions are illustrated. The x-axis corresponds to the different runs; the y-axis corresponds to the MAE of the solution. Most of the solutions share the same ,optimal, goodness of fit ensuring the algorithm is not trapped in local minima.

how the compartmentalization scheme works, we first examined the pathways downstream of EGF, TGFa, BTC, NRG1 and IL6. ERBB3 was placed in a group alone (C11) since it was the only node activated by NRG1; ERBB4 was also placed alone (C12) for having been activated by BTC and NRG1. ERBB2 and SHC were grouped together since they were both activated by EGF, TGFa, BTC and NRG1. Moving further downstream, INPP5D, JAK1, JAK2, INPPL1, GRB2, GAB2, GAB1, SOS, RAS, CRK, CRKL, DOCK1, BRAF, RAC1 and the MAP3Ks were grouped into C2 since all of them were activated by EGF, TGFa, BTC, NRG1 and IL6. This example demonstrates how the proposed compartmentalization scheme is based on the experimental treatments present in the dataset. If for example, another ligand was introduced activating via a different pathway RAC1, then the extensive compartment C2 would be broken into 2 smaller ones. First, INPP5D, JAK1, JAK2, INPPL1, GRB2, GAB1, GAB2, SOS, RAS, CRK, CRKL and BRAF, activated by EGF, TGFa, BTC, NRG1 and IL6; and second, RAC1 and the MAP3Ks (MAP3K2,3,4,6,9,10,11,12,13,15), activated by EGF, TGFa, BTC, NRG1, IL6 and the new ligand. With the proposed compartmentalization scheme, the interrogated pathway is never larger than what can be constrained by the data at hand.

In Figure 3.11, the optimized pathway of Figure 3.10 was mapped back to the PKN. Reactions within the same compartment were plotted in blue and were not involved in the optimization procedure. The rest of the reactions were plotted in black and their thickness corresponds to

the maximum activity of each reaction in the optimized model. The resulting pathway reveals well known characteristics of signaling cascades (See [46]): EGF, TGFa, BTC and NRG1, all signaled through the EGFR and then through the cluster of SHC, GRB2, GAB1, SOS, RAS to either activate MAP2K1, ERK, RPS6KA1, GSK3 and STAT3, or go through PI3K to AKT and subsequently to RPS6KB1 and IRS1S. On the other hand IL1b, FLAGELLIN and IL1a signaled through TRAF6 and mainly activated IKB, JNK, MAPK14 and HSP27. CD40LG and TNF activated the same signals but went through TRAF5, TRAF2 and MAP3K7.

The solution obtained herein, when compared to the Boolean solution in [46] was able to decrease the remaining fitness error up to 75% (mean fitness error). The algorithm completed within 20 minutes. Even though the two solutions share the same basic connectivity patterns, the constrained fuzzy logic approach handles conflicts in the data more efficiently, since it allows partial activation of the signaling species. For instance, GSK3 was removed from the Boolean solution for having been activated in an inconsistent manner (it was activated under very few combinatorial treatments and remained unaffected by either PI3Ki or MEKi). Under the constrained fuzzy logic approach, however, GSK3 was activated by RPS6KA1. By fitting the p and a parameters of this and the upstream reactions the model predictions for GSK3 matched the data and the fitness error was reduced. Similarly, IRS1S and RPS6KB1 were activated under constrained fuzzy logic, in contrast to the Boolean approach.

3.2.2 Discussion

n this section we introduced a Non Linear Programming (NLP) formulation for the quantitative modeling of signal transduction pathways, based on signaling data. We employed a fuzzy logic approach to model signal transduction mechanisms and coupled it to an NLP optimization formulation. The proposed method allowed for fast optimization of signaling pathways to high throughput signaling data in a quantitative framework. As case study, three pathways of different scale were interrogated, a small, medium and a large-scale one. For the latter two, i) the systematic definition of the search space, ii) the generation of a family of solutions, iii) and the identifiability/over-parameterization of the pathway were addressed to ensure the best possible performance of the proposed formulation. The systematic definition of search space guaranteed that a representative set of solutions was obtained while at the same time minimized the required CPU time. The collection of a family of near optimal solutions decreases the probability of biologically relevant solutions remaining unreported. The proper size for the family of solutions was also addressed. By addressing over-parameterization either by exhaustively removing reactions from the PKN, or via the proposed compartmentalization scheme, we decreased CPU time and guaranteed that only reactions vital for fitting the data were included in the solution. Finally, results on both the medium and the large scale signaling pathways were compared with the ones obtained by alternative approaches [10, 46].

Our NLP formulation presents several advantages and limitations in pathway optimization. On the negative side, it is clear that verification of the presence or absence of each reaction in the generic topology, or unique identification of its parameters is not possible given the relatively small dataset at hand. Figure 3.11 shows the un-compartmentalized version of the initial pathway where 116 out of 228 reactions (,50%) could not be identified if they are present or not given the data at hand (blue lines in Figure 3.11). This implies that the optimization problem incorporates more parameters than what it is possible to constrain. However, the exhaustive removal of reactions from the PKN, in the case of the medium scale topology, and the adoption of the equivalent classes concept (introduced in [32]) as a compartmentalization scheme, in the case of the large-scale topology, limited the number of redundant/non-identifiable reactions left in the model. Another inherent limitation of the proposed approach is our restriction to connectivity

present in the PKN. The formulation we use, by optimizing the values of model parameters (a and p), minimizes measurement prediction mismatch. Essentially reactions can be removed by setting the gain parameter of the respective reactions to zero, however, there is no support for adding new connections. Thus, the connectivity of proteins in the solution is a subset of the connectivity in the PKN. If the data dictates connectivity that is not supported by the PKN, there will be remaining fitness error in the solution. Even though methods have been developed to address this [89] based on the inference of physical interactions of proteins from the signaling data, adding new connectivity in the PKN can lead to poorly confined solutions and further research is needed to tackle this issue. Another limitation is the single time point measurement of the signaling activity. All the incorporated signaling data from HepG2 cells were obtained from the same time-point (30 min). Consequently, any activity that takes place earlier or later on will not be accounted for. To alleviate this limitation an average "early" time point was employed in the phosphoprotein activity of primary hepatocytes that incorporates the average activity of 5 and 25 minutes [46].The single time point measurements also prevent us from capturing the dynamics of the signaling reactions. Even though a dynamic representation is closer to reality, and can be potentially handled within a logic framework [90], both the experimental cost and the number of parameters required, make it difficult to model large topologies. On the positive side, our approach is a significant advancement of the Boolean Logic that successfully addresses both the protein connectivity and the activity/intensity of reactions in large signaling pathways that – as shown- number 120 species and 230 reactions.

When compared to Boolean modeling, the proposed approach provides a quantitative view of the signaling pathway, supporting continuous values for the activation of the included species. Moreover, each reaction is modeled via a sigmoid curve (normalized hill function) that more closely replicates its actual mechanics. As a result, the proposed approach gives lower fitness error than the Boolean counterpart. When compared to other fuzzy models, the proposed algorithm performed equally good to previous approaches [10] interrogating the optimization of the medium scale pathway to signaling data. Even though the two procedures follow different workflows, the topology of the solutions is very similar and the goodness of fit is of the same level, whereas CPU times favors the NLP approach (60 minutes per run for CellNOpt-cFLagainst 15 mins for NLP).

The computational efficiency of the NLP approach allowed the interrogation of large-scale pathways, namely the one introduced in [46]. It performed significantly better than the Boolean approach in terms of goodness of fit, decreasing the fitness error up to 75% (mean fitness error). Although the CPU time was increased, the solution remained computationally feasible.

Overall, the proposed approach addressed successfully the optimization of medium and large-scale signal transduction networks. It allowed the fast optimization of signaling topologies by combining the versatile nature of logic modeling with state of the art optimization algorithms.

3.2.3 Additional information

An alternative Mixed Integer Non Linear Programming formulation

Apart from the Non Linear Programming (NLP) formulation described in the main text of this section, we derived a Mixed Integer Non Linear Programming (MINLP) formulation to address the optimization of the Prior Knowledge Network (PKN) to signaling data. The MINLP formulation not only solves for the reaction parameters (a, p and n) but also interrogates the presence or absence of each reaction by introducing a set of binary variables $y_i \in \{0, 1\}$, where $i = 1, ..., n_r$ is the set of reactions, $y_i = 0$ implies reaction i is absent, $y_i = 1$ implies reaction i is present (see also [23]). Even though the MINLP is capable of optimizing the connectivity of the proteins in the signaling network together with the mechanics of each reaction, it introduces an

additional parameter for each reaction, increasing the complexity of the optimization problem and subsequently the CPU time.

Apart from the constraints (3.3)-(3.7) the following inequality must be incorporated.

$$z_i^k \leq y_i; i = 1, ..., n_r; \quad k = 1, ..., n_e \tag{3.9}$$

Constraint (3.9) implies that reaction i can be active only if it is present.Thus, constraint (3.3), in the case of AND gates, becomes:

$$z_i^k = y_i \cdot f(\prod_{j \in R_i} x_j^k \times \prod_{j \in I_i} (1 - x_j^k)) \tag{3.10}$$

In the case of OR gates the activation value of the downstream species is given by:

$$x_j^k = b_{|T_j|}^k \tag{3.11}$$

where,

$$T_j = \{i \in \{1, ..., n_r\} : j \in P_i\} \tag{3.12}$$

T_j is the set of all reactions that have species j as their product. Let $i_1, i_2, ..., i_{|T_j|}$ denote the elements of T_j. Then, b_m^k is calculated recursively as:

$$b_m^k = b_{m-1}^k + z_{i_m}^k - b_{m-1}^k z_{i_m}^k; \quad 2 < m \leq \mid T_j \mid \tag{3.13}$$

$$b_2^k = z_{i_1}^k + z_{i_2}^k - z_{i_1}^k z_{i_2}^k \tag{3.14}$$

where, $z_i^k = y_i \cdot f(x_{j \in R_i}^k)$.

The y_i variables allow for the explicit removal of reactions that appear to contradict the data at hand, in contrast to the proposed NLP formulation where reactions are removed implicitly by setting the a parameters to 0. Although the MINLP approach successfully optimized small and medium scale topologies, the proposed NLP approach performed significantly better in terms of goodness of fit to the data and CPU time, thus it is the method of choice for the analysis presented here. Both the NLP and the MINLP variants are susceptible to local minima of the objective function, thus a family of solutions must be obtained to guarantee that all biologically significant solutions have been accounted for.

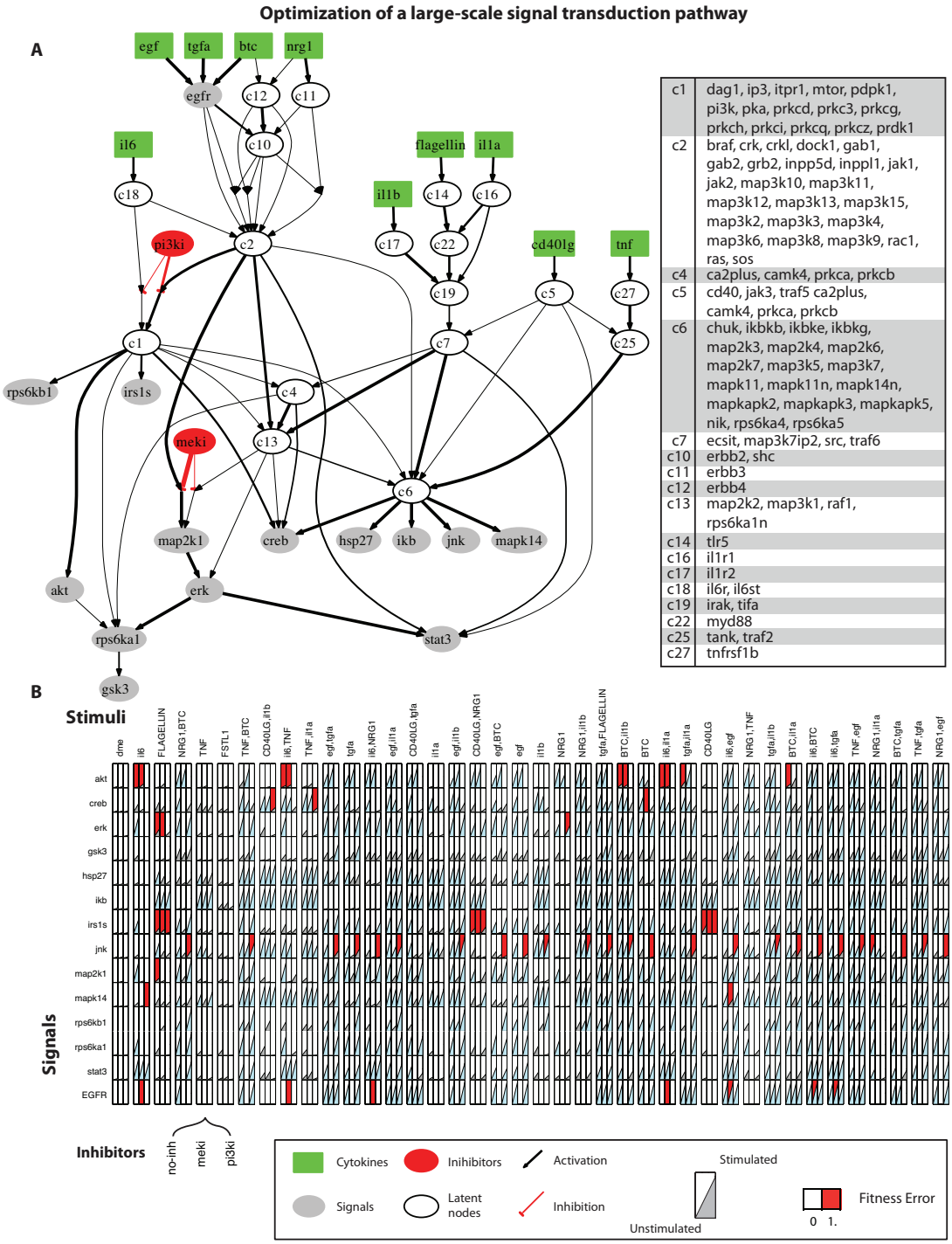


Figure 3.10: Optimization of a large-scale signal transduction pathway. (A) optimized pathway upon compartmentalization based on the equivalent classes concept (right panel). The proposed compartmentalization scheme groups together nodes that share identical in-silico responses under all experimental conditions, thus decreasing the parameters space. (B) Signaling dataset, consisting of 15 cytokines in combinations of two, and 3 inhibitors (including the no-inhibitor treatment), total of 120 experimental treatments (see [46]). The red background color corresponds to the measurement prediction mismatch of the solution.To generate model predictions the optimized values of all model parameters were used (i.e., parameter values obtained from the optimization procedure)

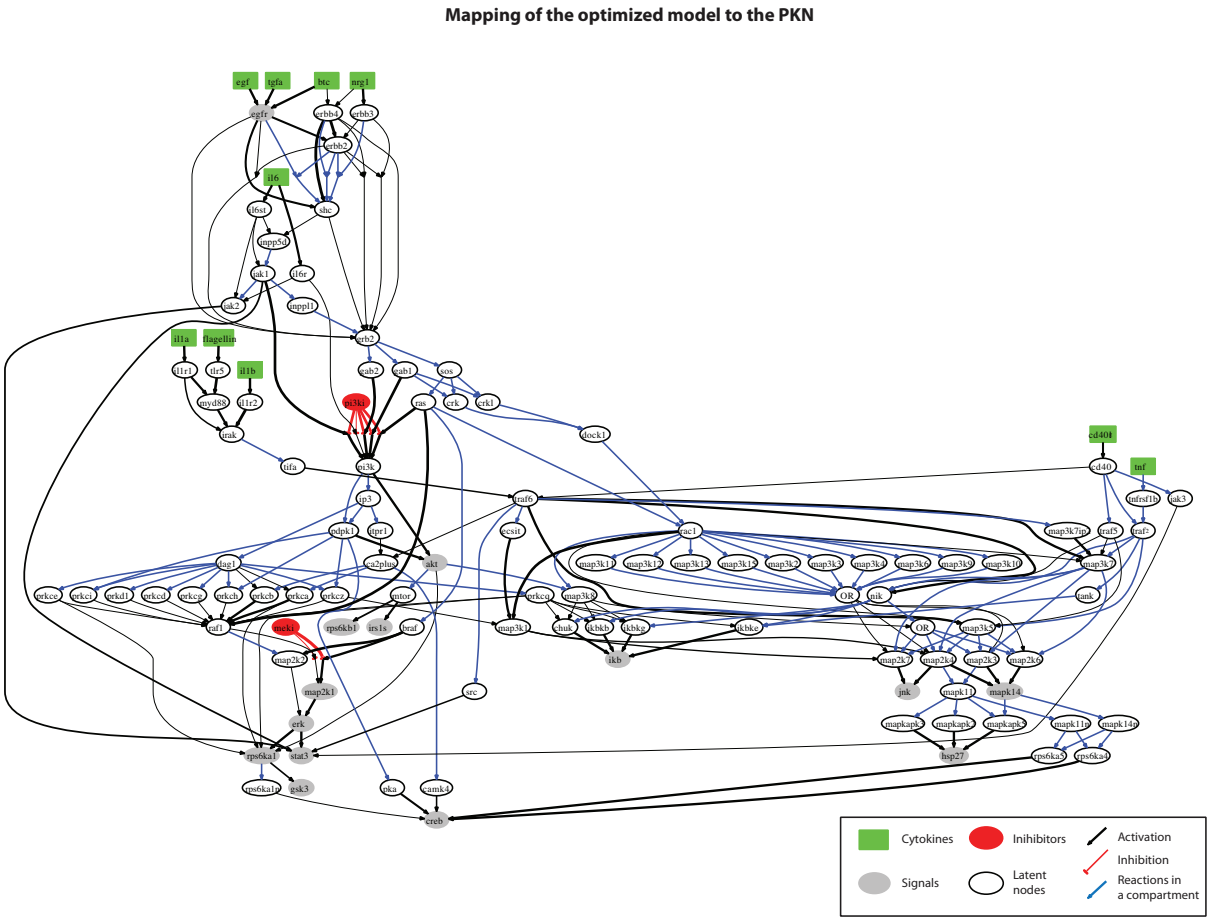


Figure 3.11: Mapping of the optimized model to the PKN. Mapping of the optimization results to the PKN by removing the compartmentalized components. Reactions within the same compartment are plotted in blue and were not included in the optimization procedure. Reactions in black are the ones whose parameters were interrogated. Their opacity corresponds to their activity in the optimized model, with reactions that propagate more signal downstream being more opaque than the rest.

3.3 Modeling of signaling pathways in chondrocytes via a Non Linear Programming (NLP) formulation on phosphoproteomic and cytokine release data

In this work published in [91], the NLP formulation in [86] was used to model signal transduction in primary human chondrocytes. A signaling pathway was constructed downstream 78 receptors of interest and was optimized to high throughput phosphoproteomic data obtained via Luminex technology. The optimized model best captures the signaling patterns of human chondrocytes.

Abstract

Protein signaling is widely identified as a key component in the etiology of osteoarthritis (OA). Traditionaly, OA is attributed to overactivation of NF-κB in chondrocytes under prototypical inflammatoy players such as IL1α and TNFα. However, as high throughput proteomics evolve and being employed in the study of complex disease, such as the disease of OA, the community moves away from the reductionist approaches that focus on these few well known players and aims towards a systems level understanding of the disease. In this section, we interrogate the signal transduction pathways in chondrocytes downstream 78 receptors of interest, on both the phosphoproteomic and cytokine release level. On the phosphoproteomic level, 17 key phosphoproteins are measured upon stimulation with single treatments of 78 ligands. On the cytokine release level, 55 cytokines are measured in the supernatant upon stimulation with the same treatments. Using a Non Linear Programming (NLP) formulation, the proteomic data is combined with a priori knowledge of proteins' connectivity to construct a mechanistic model, predictive of signal transduction in chondrocytes.

3.3.1 Introduction on modeling of signaling pathways in chondrocytes

Osteoarthritis, a debilitating joint disease, is characterized by the imbalance of anabolic and catabolic and inflammatory processes in articular cartilage. Traditionally, the disease is attributed to the overactivation of NF-κB pathway in chondrocytes, leading to the release of MMPs, resulting in degradation of cartilage tissue and ultimately loss of its structural integrity [19, 20, 21, 22]. Major players connected to the etiology of the disease include inflammatory mediators like IL1α,β, and TNFα, together with the WNT [92] and BMPs [93] pathway. OA is traditionally tackled by interrogating these few major players, without taking into account other less known pathways.

Recently, the emerging field of high throughput proteomics and systems biology have created a paradigm shift in the study of OA, from studying individual proteins into studying cartilage degeneration on a systems level. The community comes to realize that complex disease, such as OA, cannot be tackled by reductionist approaches, but a systems level understanding of the disease has to be established for effective treatments to be discovered [94, 95, 96, 97]. On this front, high throughput proteomics have found the following applications in the study of OA: *(i)* direct analysis of cartilage protein content [98, 99, 100, 101], *(ii)* analysis of cartilage related biological fluids [102, 103] (Synovial fluid and plasma) and *(iii)* study of chondrocytes secretion upon treatment with pro-catabolic mediators [104, 105]. Proteomic analysis of cartilage explants and chondrocytes have lead to the identification of several hundreds of proteins in articular cartilage, as well as characterization of their expression patterns in normal and OA patients. Findings of this analysis may lead to better understanding of the etiology underlying the disease and potential drug targets. Study of the chondrocytes secretion, leads to deeper understanding of the cells inflammatory response and extracellular signaling, while proteomic analysis of the synovial fluid and plasma has identified proteins differentially regulated in OA and normal

patients, and aims mostly at biomarker discovery.

Eventhough the first steps towards a systems level understanding of the disease have been made, collecting all this data alone is not enough to explain the disease's complex phenotype and intricate mechanisms of progression [1]. What is missing is *(i)* interrogating, in rigorous fashion, the chondrocytes' signaling pathways and *(ii)* integrating this data in functional models, predictive of cells' biology. The study of chondrocytes' signaling mechanisms on both the phosphoproteomic and the cytokine release levels and construction of functional models may provide valuable insight on the cells' function and the etiology of the disease.

In this section, we interrogate the signal transduction mechanisms of primary chondrocytes on both the phosphoproteomic and the cytokine release level, downstream 78 receptors of interest. Moreover, we implement a Non Linear Programming formulation to integrate the two types of data and construct a mechanistic model of their signal transduction pathways [23, 51]. On the experimental front, the xMAP technology is used to measure the activation level of 17 key phosphoproteins and the release of 55 cytokines in the supernatant, upon stimulation with single treatments of the 78 ligands. Eventhough the xMAP technology does not provide for signal multiplexability as high as other proteomic technologies (e.g. the various Mass Spectrometry methods), fast turnaround times using the Luminex equipment and low requirements in protein content allows the design of the experiment on 96 well plates, leading to high sample throughput [11, 12]. On the computational front, a Non Linear Programming (NLP) formulation is used to fit a prior knowledge network to the proteomic data, resulting in a mechanisitic model, predictive of the function and response of human chondrocytes [23]. The proposed apporach is based on the utilization of a transfer function (TF) to model signal trasduction from one node to the next [10]. By adopting this modeling approach and using a Prior Knowledge Network (PKN) -obtained from literature citations of signaling reactions- as a scaffold, we construct an initial model of the signal transduction network. Subsequently, the NLP formulation is used [86] to train this initial model to proteomic data. The optimized model best captures the signaling patterns of human chondrocytes.

3.3.2 Results

Phosphoproteomic data

Chondrocytes were stimulated with single treatments of 78 ligands while measuring the activation level of 17 key phosphoproteins via xMAP technology. The phosphoproteomic dataset is plotted in Figure 3.12 via Datarail toolbox [50]. Figure 3.12 is a collection of subplots representing the time course of the 17 signals from the unstimulated state to 20 minutes for each of the imposed ligands. The filling color in each subplot corresponds to the normalized value of the respective datapoint. Activated signals are plotted in blue.

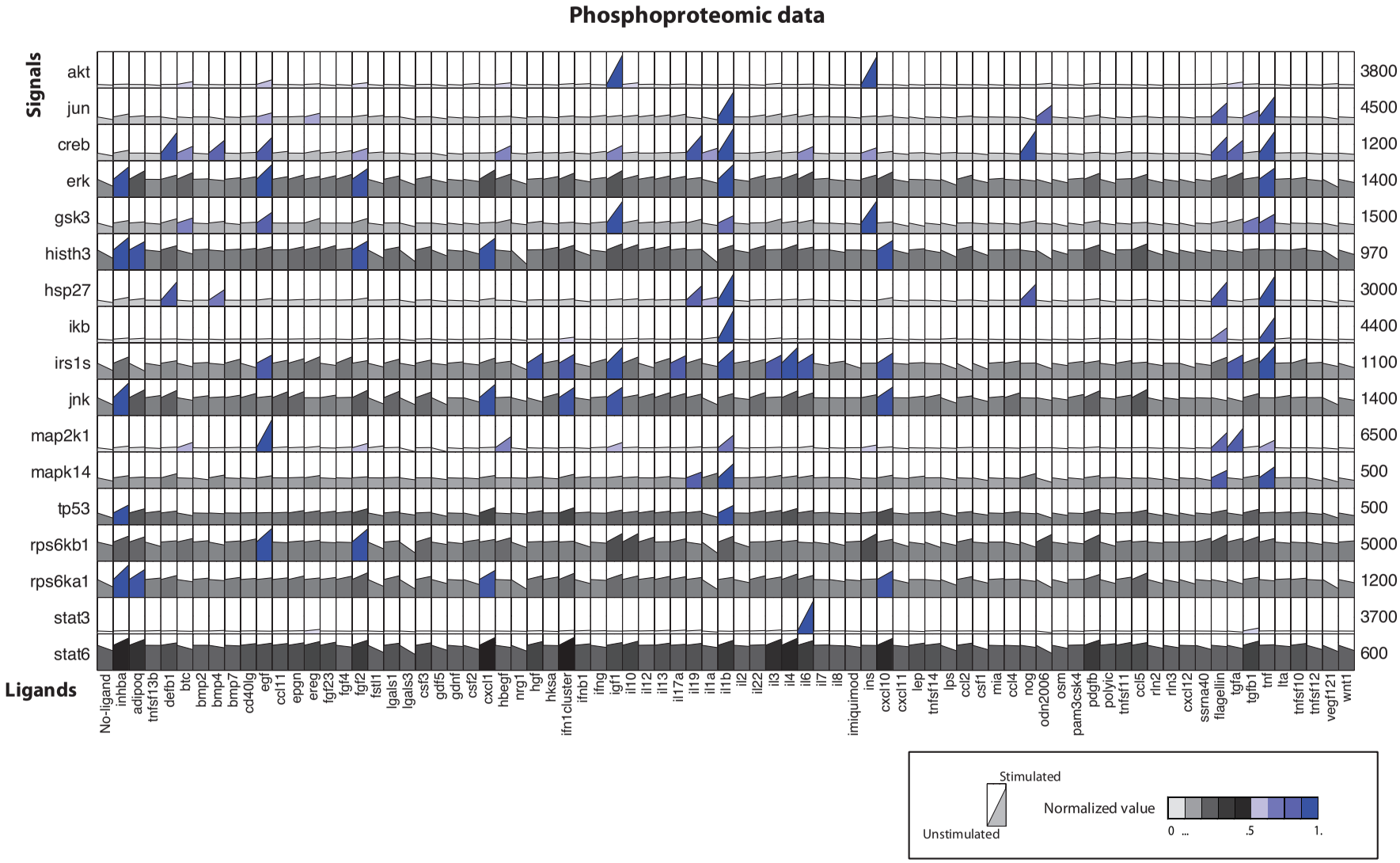


Figure 3.12: **Phosphoproteomic data**: The time course of the phosphoprotein signals from the unstimulated state to the average early response is illustrated. The rows correspond to the 17 phosphoproteins measured and the columns to the 79 ligand treatments (including the No-ligand treatment). In each subplot, the first point shows the unstimulated activity of the respective signal (zero time point); the second point shows the raw measurement of the signal (in fluorescent units) 5+25 minutes after stimulation; while the color code corresponds to the normalized value (between 0 and 1) of the signal.

1. **Major pathways**

   (i) Regarding prototypical inflammatory players such as IL1α, IL1β and TNF: **IL1**β activated most signals related to inflammation e.g. JUN, CREB, IκB, MAPK14, HSP27 and TP53. Additionally activated ERK, IRS1S, MAP2K1 and GSK3. **TNF**, in similar fashion to IL1β, activated the inflammation related signals apart from TP53, together with ERK, GSK3, IRS1S and MAP2K1. **IL1**α, on the other hand, activated only CREB while decreased the activation level of HISTH3, JNK, TP53, RPS6KB1 and RPS6KA1. *(ii)* Regarding growth factor signaling: **EGF** and **TGF**α appear to have similar effects on the measured phosphoproteins, activating most growth related proteins such as MAP2K1, ERK, GSK3 and partially AKT, leaving unaffected inflammation related signals such as IκB, MAPK14 and HSP27. Similar response yielded **INS** and **IGF1**, activating mostly MAP2K1, GSK3 and AKT. *(iii)* **IL6** activated STAT3, IRS1S and CREB. *(iv)* **BMP4** activated HSP27 and CREB while the rest of the BMPs (**BMP2** and **BMP7**) raised no significant response. *(v)* **TGF**β**1** activated STAT3, JUN and GSK3.

2. **Under-reported pathways**

   Apart from major players, significant activity was measured upon stimulation with several under-reported ligands, such as

   - **INHBA** (Inhibin, Beta a) activated RPS6KA1, TP53, JNK, HISTH3 and ERK.
   - **DEFB1** (Defensin, beta 1), a member of the Defensin family, activated HSP27 and CREB.
   - **BTC** (Betacellulin), an EGFR ligand, activated signals related to growth factor signaling such as GSK3, MAP2K1 and AKT; also activated CREB.
   - **FGF2** (Basic Fibroblast Growth Factor) activated mostly growth related proteins e.g. RPS6KB1, MAP2K1, HISHT3, ERK and partially AKT; also activated CREB.
   - **CXCL1** (GROa), mostly expressed by macrophages, activated RPS6KA1, JNK and HISTH3.
   - **HBEGF** (Heparin-binding EGF-like growth factor), a member of the EGF family of proteins, activated growth related signals such as MAP2K1 and AKT; also activated CREB.
   - **IL19**, a protein that belongs to the IL10 subfamily and is mostly expressed in monocytes, was found to activate inflammation related proteins such as MAPK14 and HSP27. Also activated CREB.
   - **CXCL10**, also known as Interferon gamma-induced protein 10 (IP-10), activated RPS6KA1, JNK, IRS1S and HISTH3.
   - **NOG** (Noggin), a protein known to bind to TGFβ family ligands, activated HSP27 and CREB.
   - **ODN2006**, a Toll-like Receptor 9 ligand, activated mostly JUN.
   - **FLAGELLIN**, a bacterial component and Toll-like Receptor 5 ligand, has significant effects on chondrocytes' signaling by activating MAPK14, MAP2K1, IκB, HSP27, CREB and JUN.

Cytokine release data

Chondrocytes were stimulated with single treatments of 78 ligands while measuring the release of 55 cytokines in the supernatant. The cytokine release data is shown in Figure 3.13. Figure 3.13

is a collection of subplots representing the time course of the 55 signals from the unstimulated state to 24 hours, for each of the imposed ligands. The filling color in each subplot corresponds to the normalized value of the respective datapoint. Activated signals are plotted in blue.

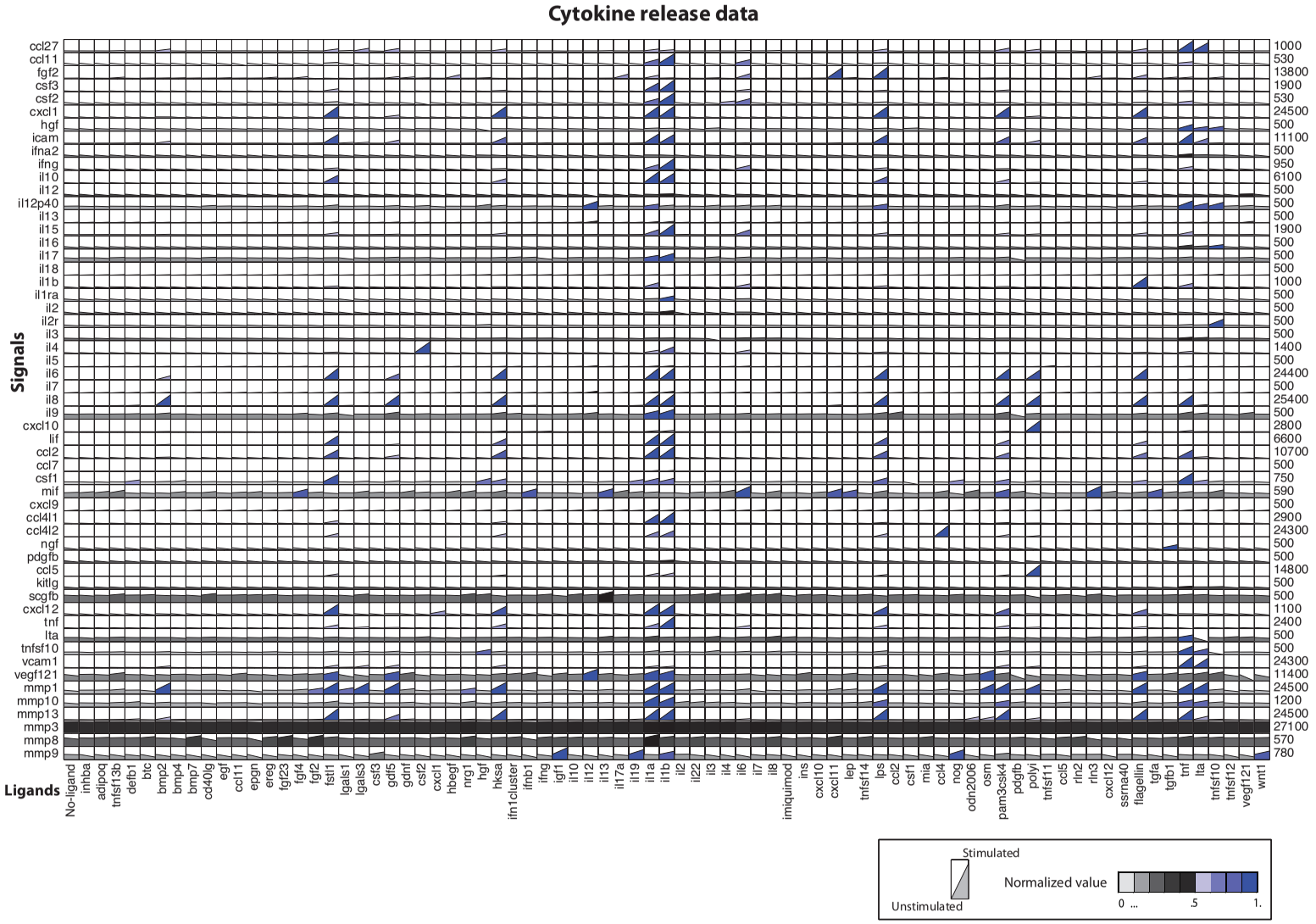


Figure 3.13: **Cytokine release data**: The time course of the cytokine releases from the unstimulated state to 24 hours is illustrated. The rows correspond to the 55 cytokine releases measured in the supernatant and the columns to the 79 ligand treatments (including the No-ligand treatment). In each subplot, the first point shows the unstimulated concentration of the respective cytokine in the supernatant (zero time point); the second point shows the raw measurement of the signal (in fluorescent units) 24 hours after stimulation; while the color code corresponds to the normalized value (between 0 and 1) of the signal.

1. **Major pathways**

   (i) Regarding prototypical inflammatory players such as IL1α, IL1β and TNF: **IL1**α and **IL1**β raised very similar responses leading to the release of most major inflammatory mediators such as IL6, IL17, TNF, IFNg, LIF, ICAM, VCAM1, IL8, CCL2, CCL4L1, CCL4L2, CXCL1, CCL11, CCL5 and CXCL12. Additionally several MMPs were released such as MMP1, MMP9, MMP10 and MMP13, known to degrade cartilage tissue. Anti-inflammatory cytokine levels also increased significantly e.g. IL4, IL10 and IL1RA. **TNF** lead to extensive inflammatory response as well, inducing the release of IL1β, IL6, IL8, IL12p40, IL15, IFNG, CCL2, CCL27, CCL11, ICAM and VCAM1, together with the release of MMP1, MMP10 and MMP13. FGF2 (Basic FGF) and HGF growth factors together with CSF1 (Colony stimulating factor) were also released upon TNF stimulation. *(ii)* **IL6** lead to the release of IL1β, IL15, TNF, IFNG, FGF2, CSF1, CSF2, CCL11 and MMP13 suggesting a pro-inflammatory action, together with the release of anti-inflammatory IL4. *(iii)* Regarding the BMPs family, **BMP2** lead to the release of MMP1, IL8, IL6, ICAM and CCL27, while the rest of the BMPs (**BMP4**, **BMP7**) raised no significant response.

   (iv) Regarding Toll-like Receptors (TLR) signaling, **PAM3CSK4** (TLR1/2 agonist), **HKSA** (TLR2 agonist), **POLYIC** (TLR3 agonist), **LPS** (TLR4 agonist), **FLAGELLIN** (TLR5 agonist) and **FSTL1** (TLR6/2 agonist) were the most potent inducers of inflammatory signals, leading to the release of VCAM1, TNF, CXCL12, CCL4L2, CCL2, LIF, IL8, IL6, IL15, IL10, ICAM, CXCL1, CCL27 and MMP1,10,13.

2. **Under-reported pathways**

   Apart from the major players, significant response was raised by several under-reported ligands, such as

   - **LGALS1, LGALS3** (Galectin 1,3) lead to the release of MMP1.
   - **GDF5** (Growth/differentiation factor 5) lead to the release of MMP1, MMP13, VCAM1, CCL2, IL8, IL6, ICAM, CXCL1 and CCL27.
   - **CSF2** (GMCSF) lead to the release of IL4.
   - **CXCL11** induced MIF and FGF2 release.
   - **OSM** (Oncostatin) released MMP1.
   - **TGF**β**1** released NGF.
   - **LTA**, member of the TNF superfamily, lead to extensive inflammatory response, inducing the release of MMP13, MMP1, VCAM1, CCL2, IL8, ICAM and CCL27.

Pathway construction and optimization

Prior Knowledge Network: A PKN was constructed downstream the 78 receptors of interest, based on literature citations of signaling reactions [46]. Several online databases were querried (Reactome [27], PathwayCommons [28], KEGG [26]), but most of the reactions were obtained from Ingenuity (`http://www.ingenuity.com/`). The PKN was constructed in such a manner that it includes all interrogated receptors and measured phosphoproteins.

Upon its construction, the PKN was preprocessed to remove non-observable and non-controllable parts of it according to the analysis in [106]. Non-observable, are nodes in the pathway whose activation state cannot be inferred based on the measured phosphoproteins (i.e. nodes, downstream of which there are no measured signals). Non-controllable are nodes whose activation state cannot be controlled by the imposed perturbations (i.e. nodes with no upstream stimuli).

Removing non-observable, non-controllable parts of the pathway facilitates the optimization process by reducing the size of the pathway. The PKN is plotted in Figure 3.14, it includes a total of 207 species and 426 reactions.

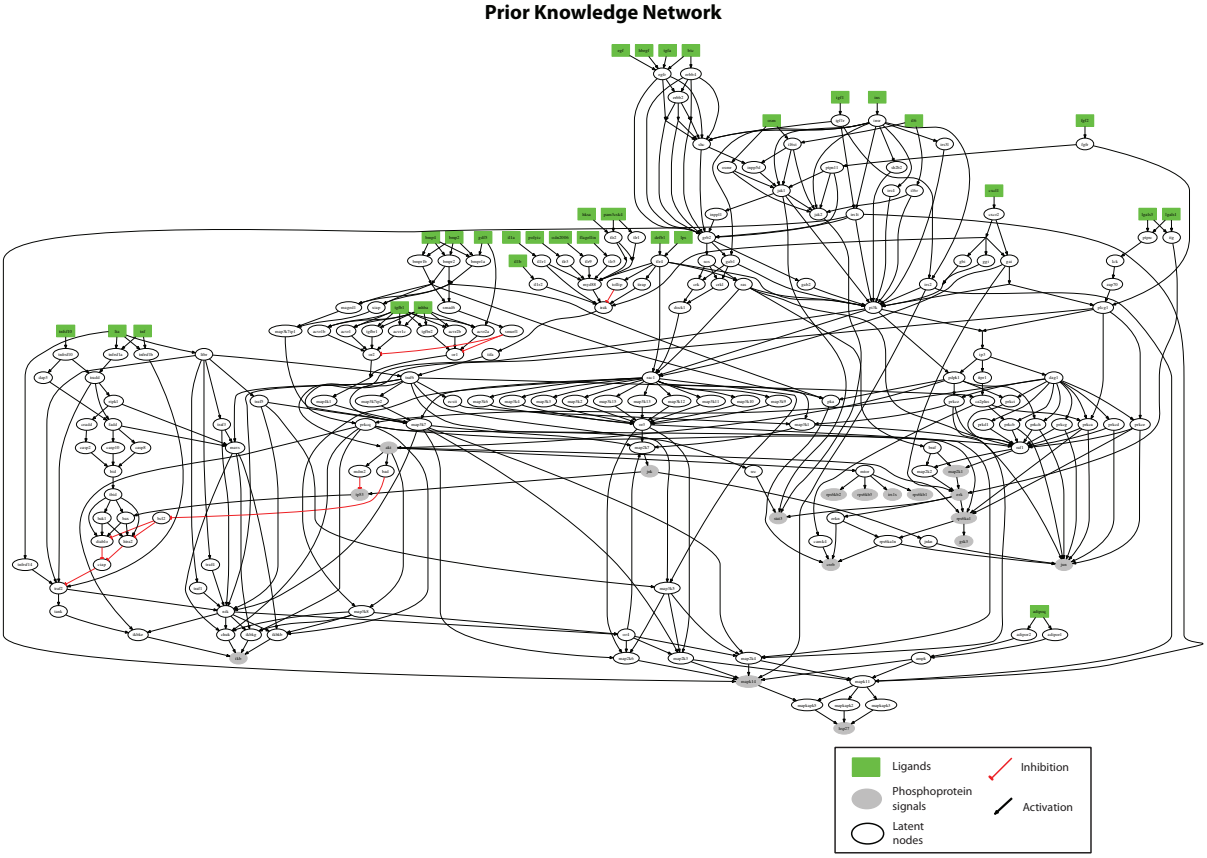


Figure 3.14: Prior Knowledge Network: The signal transduction network that served as a PKN for the analysis presented in this section is illustrated, upon removing non-observable and non-controllable parts of it [13]. Green nodes correspond to the imposed ligands, grey elliptic nodes to the measured phosphoproteins, grey rectangle nodes to the measured cytokine releases and clear (white) nodes to latent signaling proteins in the network. Non-observable, are nodes in the pathway whose activation state cannot be inferred based on the measured phosphoproteins (i.e. nodes, downstream of which there are no measured phosphoproteins). Non-controllable are nodes whose activation state cannot be controlled by the imposed perturbations (i.e. nodes with no upstream ligands). Removing non-observable, non-controllable parts of the pathway facilitates the optimization process by reducing the size of the pathway. The observable-controllable part of PKN includes a total of 207 species and 426 reactions, downstream of 30 ligands.

Pathway compartmentalization: Before implementation of the NLP formulation and optimization of the PKN to proteomic data, we compartmentalized the signaling topology to address over-parameterization [107]. For details of the compartmentalization scheme see [86], where it was first introduced. Herein we describe its main principles for consistency purpose.

The implemented compartmentalization scheme aims at fixing the parameters of certain reactions to nominal values, to decrease the parameters space and counter over-parameterization of the model. Over-parameterization is caused by overlapping pathways and lack of experimental data to interrogate all of them. As a result, the presence or absence of certain reactions in

the pathway cannot be validated (nor disproven) based on the experimental data at hand. The compartmentalization scheme groups together nodes that share identical in-silico response under all experimental conditions [32], resulting in an equivalent (compartmentalized) pathway, where nodes have been replaced with their respective compartments and reactions between proteins have been replaced with reactions between compartments. Grouping nodes that share identical response, simplified the PKN and decreased the parameters' space significantly (together with CPU time of the optimization algorithm). Upon compartmentalization, its size decreased to 30% its initial value with respect to the included nodes (a total of 63 compartments are constructed) and to 50% with respect to the included reactions (a total of 238 reactions). The compartmentalized pathway is show in Figure 3.15.

Data normalization: Both phosphoprotein and cytokine release data, obtained via xMAP technology, is measured in fluorescent units and is dependent on the antibody pair used for detection. E.g. MAP2K1 ranged from 280 units (untreated condition) to 6500 units (under EGF), while GSK3 ranged from 500 units (untreated condition) to 1500 units. Variations such as these do not necessarily reflect that MAP2K1 is more activated than GSK3, but may be attributed to protein abundance or assay calibration issues. Consequently two challenges emerge, firstly, identifying wether a signal is activated or not, and secondly, normalizing the raw data in a way that the optimization algorithm is not biased in favor of the highest values [106].

Herein, we implemented the following normalization procedure [46]: *(i)* A bimodal distribution was assumed for the activation values of each measured signal (both phosphoprotein and cytokine release), consisting of an activated and a deactivated mode. The bimodal distribution was calculated via Matlab's statistic toolbox (`gmdistribution.fit()` function). *(ii)* For each datapoint, the frequencies respective to the two modes were evaluated (`pdf()` function), *(iii)* their ratio was subsequently passed through a Hill function filter and normalized in the range between 0 and 1. The normalized data was imported in the NLP algorithm for the optimization process.

Pathway optimization. The NLP formulation introduced in [86] is utilized to train the PKN to the proteomic data (both phosphoroteomic and cytokine releases) resulting in an integrative model, predictive of the signal transduction mechanisms of human chondrocytes. The optimized -compartmentalized- model is shown in Figure 3.15. Of the 78 ligands used to stimulate the cells, 49 of them did not raise any significant response, thus, the corresponding pathways were excluded from the solution. On the other hand, 29 ligands activated a significant portion of the measured signals, either on the phosphoproteomic or the cytokine release level, and their pathways were conserved in the solution and illustrated in Figure 3.15. The edges' opacity represents the activity of the corresponding reactions. In Figure 3.17, the solution was mapped back to the original topology by substituting the compartments of Figure 3.16 with the corresponding proteins. Reactions within the same compartment are plotted in blue and have not been interrogated by the optimization algorithm.

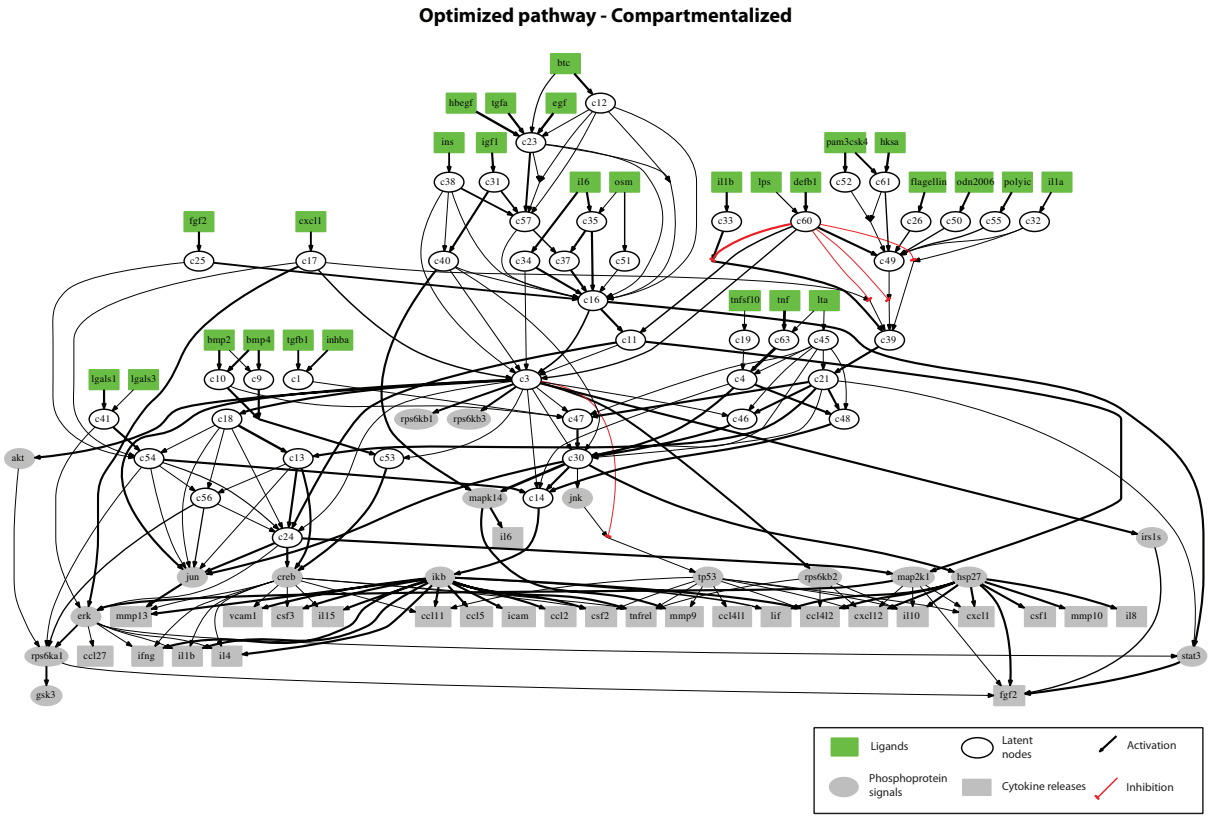


Figure 3.15: Optimized pathway, Compartmentalized: The compartmentalized version of the signaling network upon optimization via the NLP formulation is illustrated. The implemented compartmentalization scheme groups together nodes that share the same in-silico response under all experimental conditions, reducing the parameters space (thus, addressing the over-parameterization of the pathway) and decreasing its size and complexity together with the CPU time of the optimization algorithm. Green nodes correspond to the imposed ligands, grey elliptic nodes to the measured phosphoproteins, grey rectangle nodes to the measured cytokine releases and clear (white) nodes to compartments of signaling proteins. The proteins included in each compartment are listed in Figure 3.16. The optimization procedure evaluated the parameter values of each signaling reaction to minimize the difference between model predictions and experimental measurements. The opacity of the edges corresponds to the activity of the respective reactions.

Compartments of the PKN

c1	acvr1, acvr1b, acvr1c, acvr2b, map4k1, tgfbr1, tgfbr2	c30	hsp27, jnk, jnkn, map2k3, map2k4, map2k6, map2k7, mapk11, mapk14, mapkapk2, mapkapk3, mapkapk5, nik
c2	acvr2a		
c3	akt, bad, irs1s, map3k10, map3k11, map3k12, map3k13, map3k15, map3k2, map3k3, map3k4, map3k6, map3k8, map3k9, mdm2, mtor, pdpk1, pi3k, prkch, prkci, prkcz, rac1, rps6kb1, rps6kb2, rps6kb3	c31	igf1, igf1r
		c32	il1a, il1r1
		c33	il1b, il1r2
		c34	il6, il6r
c4	bak1, bid, casp10, casp2, casp8, cradd, fadd, ripk1, tank, tbid, tradd, traf2	c35	il6st
		c36	inhba
c5	bax, diablo, htra2, tp53	c37	inpp5d
c6	bcl2, ciap	c38	ins, insr, irs3l, irs4, sh2b2
c7	bmp2	c39	irak, tifa
c8	bmp4	c40	irs1t, irs2
c9	bmpr1a	c41	itg, lck, ptprc, zap70
c10	bmpr1b, bmpr2, maged1, map3k7ip1, smad6, smurf1, xiap	c42	lgals1
c11	braf, ras	c43	lgals3
c12	btc, erbb4	c44	lps
c13	ca2plus, camk4, prkcb	c45	lta, ltbr, tnfrsf14, traf1, traf3, traf4, traf5
c14	chuk, ikb, ikbkb, ikbke, ikbkg, jun	c46	map3k1, map3k5
c15	creb	c47	map3k7
c16	crk, crkl, dock1, gab1, gab2, grb2, inppl1, jak1, jak2, sos	c48	mavs
c17	cxcl1, cxcr2, gai, gbi, ggi	c49	myd88
c18	dag1, ip3, itpr1, prkcd, prkcg, prkd1	c50	odn2006, tlr9
c19	dap3, tnfrsf10, tnfsf10	c51	osm, osmr
c20	defb1	c52	pam3csk4, tlr1
c21	ecsit, map3k7ip2, src, traf6	c53	pka
c22	egf	c54	plcg1, prkce, prkcq
c23	egfr, erbb2	c55	polyic, tlr3
c24	erk, erkn, gsk3, map2k1, map2k2, raf1, rps6ka1, rps6ka1n, stat3	c56	prkca
		c57	shc
c25	fgf2, fgfr, ptpn11	c58	tgfa
c26	flagellin, tlr5	c59	tgfb1
c27	gdf5	c60	tirap, tlr4, tollip
c28	hbegf	c61	tlr2
c29	hksa	c62	tnf
		c63	tnfrsf1a, tnfrsf1b

Figure 3.16: Compartments of the PKN: The latent nodes included in each compartment are listed. The PKN includes a total of 63 compartments consisting of 207 nodes. The implemented compartmentalization scheme has decreased the size of the pathway to 30% that of the PKN, respective to the nodes and 55% respective to the reactions.

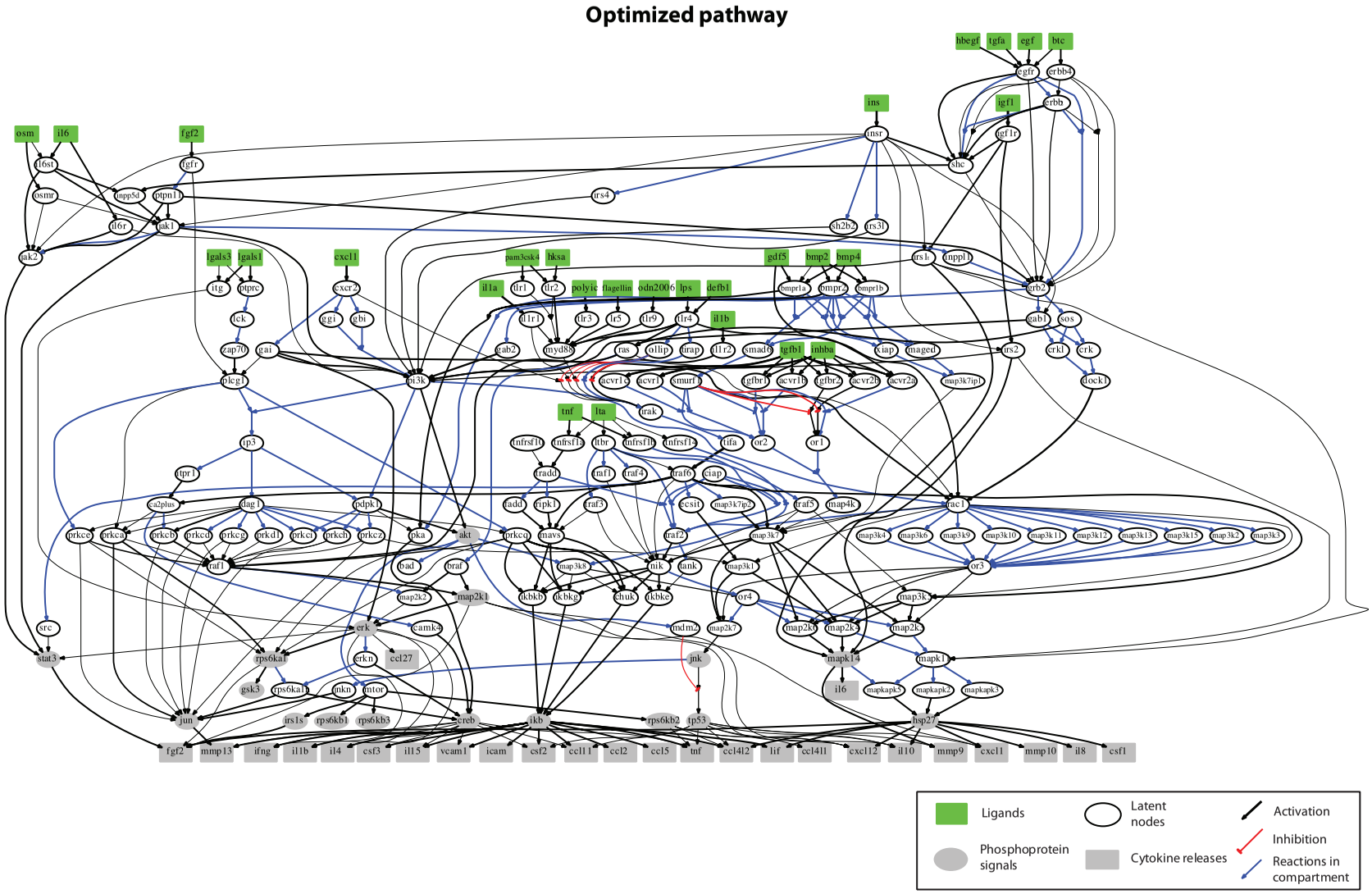


Figure 3.17: **Optimized pathway**: The signal transduction network upon optimization via the NLP formulation. Green nodes correspond to the imposed ligands, grey elliptic nodes to the measured phosphoproteins, grey rectangle nodes to the measured cytokine releases and clear (white) nodes to latent signaling proteins in the network. The optimization procedure evaluated the parameter values of each signaling reaction to minimize the difference between model predictions and experimental measurements. The opacity of the edges corresponds to the activity of the respective reactions. Reactions plotted in solid blue color are included in a compartment and thus were not interrogated by the NLP algorithm.

1. **Major pathways**

   (i) Regarding the major inflammatory players IL1α, IL1β and TNF: **IL1**α, **IL1**β signaled through overlapping pathways activating MYD88, IRAK and TIFA (C49, C39); subsequently went through TRAF6 (C21) to MAP3K1,3,7 (C46, C47, C48) and then either to the IκB cluster (CHUK, IκBKB, IκBKE, IκBKG) C14, or to MAPK14, JNK, HSP27 (C30) and JUN. Another branch leads from TRAF6 (C21), to PRKCB (C13) and activated CREB. In similar fashion, **TNF** activated MAPK14, JNK, HSP27, JUN and IκB, but signaled through TRAF2 (C4) instead of TRAF6. Upon activation of IκB and HSP27 most of the measured outputs were secreted (CCL11, FGF2, CSF3, CSF2, CXCL1, ICAM, IFNG, IL10, IL15, IL4, IL6, IL8, LIF, CCL2, CSF1, CCL4L2, CCL5, CXCL12, TNF, VCAM1, MMP10 and MMP13).

   (ii) On the other hand, major growth related ligands such as **EGF** and **TGF**α, signaled through the EGFR (C23) to SHC (C57) and then to GRB2 (C16). From GRB2 they either activated AKT, IRS1S, RPS6KB1,2,3, RPS6KA1 and GSK3 via PI3K (C3), or MAP2K1, CREB and JUN via SOS, RAS, BRAF and RAF1 (C11, C24). No significant cytokine releases were induced apart from the release of FGF2. **INS** and **IGF1** signaled through pathways partially overlapping with EGF and TGFα, activating MAP2K1, AKT, GSK3, ERK, CREB, RPS6KA1 and RPS6KB1, through GRB2 (C16), PI3K (C3), and SOS, RAS, BRAF, RAF1 (C11, C24). IGF1 additionally activated MAPK14 through IRS1T and IRS2 (C40). MAPK14 activation under IGF1 induced the release of MMP9, IL6 and FGF2. INS raised no significant extracellular activity.

   (iii) **IL6** signaled mostly through JAK2 (C16) to activate STAT3 and subsequently induced the release of FGF2, or through PI3K (C3) to activate MAP2K1, CREB, JUN, AKT and GSK3 following the same pathways as EGF and TGFα. *(iv)* The BMPs connectivity in the network is limited, with **BMP4** activating mostly CREB through PKA (C53) while **BMP2** slightly activated MAPK14, IκB and HSP27 through MAP3K7 (C47) and subsequently through NIK and MAP2K3,4,6,7 (C30). IκB and HSP27 activation lead to the release of several cytokines in similar fashion to IL1α, IL1β and TNF treatments.

2. **Under-reported pathways**

   - **BTC, HBEGF** (EGFR ligands) signaled through the same pathway as EGF and TGFα and activated most growth related proteins, such as AKT, MAP2K1, ERK, GSK3, RPS6KA1 and RPS6KB1,2,3. They additionally induced the release of FGF2.
   - **TNFSF10, LTA**, members of the TNF superfamily, had no significant effects on any of the measured phosphoproteins, however, they induced the release of several cytokines (e.g. MMP13, VCAM1, CCL2, IL8 and ICAM) and were subsequently conserved in the solution.
   - **HKSA, LPS, POLYIC, PAM3CSK4, FLAGELLIN** (TLRs ligands) signaled through pathways overlapping with other major inflammatory players such as IL1α and IL1β. Upon binding at their respective receptors, they signaled through MYD88 (C49), IRAK and TIFA (C39) to TRAF6 (C21) and then via MAP3K1,3,7 (C46,C47,C48) to the IκB cluster (C14), or to MAPK14, JNK, HSP27 (C30) and JUN. Upon activation of IκB and HSP27, several cytokines were released (CCL11, FGF2, CSF3, CSF2, CXCL1, ICAM, IFNG, IL10, IL15, IL4, IL6, IL8, LIF, CCL2, CSF1, CCL4L2, CCL5, CXCL12, TNF, VCAM1, MMP10 and MMP13), in similar fashion to the IL1α, IL1β and TNF treatments. **ODN2006** signaled through the same pathway, however, only raised average response on the cytokine release level inducing the release of MMP13.

- **DEFB1** (Defensin, beta 1) signaled mostly through RAS (C11), RAF (C24) to activate MAP2K1, ERK and GSK3, or through PI3K (C3) to activate AKT, JUN and HSP27. JUN activation lead to the release of MMP13, while borderline activation of IκB and HSP27 induced the release of ICAM and IL8.
- **CXCL1** (GROa) signaled through CXCR2 to GAI, GGI and GBI (C17) and then either to ERK, or to PI3K (C3) to activate AKT, GSK3, RPS6KA1, JNK and IRS1S.
- **FGF2** (basic FGF) signaled through FGFR (C25) to GRB2, GAB1, SOS (C16) and activated MAP2K1, RPS6KB1,2,3, AKT, CREB, IRS1S and GSK3.
- **LGALS1, LGALS3** (Galectin 1,3), even though raised no significant response on the phosphoproteomic level, induced the release of VCAM1 and CCL27. Thus, the pathway leading from ITG and PTPRC (C41) to ERK and JUN was conserved.
- **OSM** (Oncostatin) had no significant effects on any of the measured phosphoproteins, however, lead to the release of MMP13. Thus, the pathway leading from OSMR (C51) to JUN was conserved.
- **TGFβ1** signaled through TGFBR1, TGFBR2, ACVR1B, ACVR2A, ACVR2B (C1) and activated STAT3, GSK3 and JUN via MAP3K7 (C47) and C30.
- **INHBA** (Inhibin, Beta A) signaled through the same pathway as TGFβ1 and activated RPS6KA1, TP53, JNK, HISTH3 and ERK.

Performance of the optimization procedure

The NLP formulation aims at optimizing the values of a and p parameters of each reaction to better fit the proteomic data at hand. The goodness of fit was quantified by the absolute difference of experimental measurements and model predictions. To better evaluate the performance of the optimization procedure, we plotted the measurement-prediction mismatch for the responsive subset of ligand treatments, for both the initial (Figure 3.19) and the optimized (Figure 3.18) model, on phosphoproteomic and cytokine release levels.

Measurement-prediction mismatch on the phosphoproteomic level is illustrated in Figure 3.19A (initial model) and Figure 3.18A (optimized model). Measurement-prediction mismatch is indicated by the red background color. As seen in Figure 3.19A, in most cases initial model prediction misfitted the experimental data (mismatch = 49.88%). Upon the optimization procedure the measurement-prediction mismatch was significantly decreased (from 49.88% to 14.62%, Figure 3.18A). Remaining fitness error is observed mostly in cases where the data conflict each other, or equivalently in areas of the PKN where there is poor prior knowledge of the proteins connectivity. E.g. measurement of TP53, RPS6KB1, MAP2K1 and IRS1S signals under IL1α and IL1β treatments. Since, IL1α and IL1β signaled through the same pathways, if connectivity of IL1r to the above mentioned signals was conserved, then TP53, RPS6KB1, MAP2K1 and IRS1S would share identical (or very similar) in silico response under IL1α and IL1β. This contradicts the experimental data where the two ligands had different signaling patterns. In cases such as this, the NLP algorithm fits the ligand that minimizes the objective function, while misfitting the other (i.e. fits IL1α and misfits IL1β).

Measurement-prediction mismatch on the cytokine release level is illustrated in Figure 3.19B (initial model) and Figure 3.18A (optimized model). Only the responsive subset of the 55 cytokine releases measured originally is plotted here. As seen in Figure 3.19B, the initial model failed to capture the cytokine release mechanisms of primary chondrocytes (fitness error = 58.98%). Upon the optimization procedure the fitness error was decreased significantly as seen in Figure 3.18B (fitness error = 11.97%). In similar fashion to the phosphoproteomic part of the network, most of the remaining fitness error is attributed to conflicting data and poor prior

knowledge of the proteins connectivity. E.g. measurement of MMP13, VCAM1, TNF, CCL2, LIF, IL8, IL6, ICAM, CXCL1, FGF2 and CCL27 release under GDF5. Since the release of the above mentioned cytokines was induced upon IκB, ERK, or HSP27 activation, GDF5 would have to activate at least one of these phosphoproteins to lead to the release of any cytokines. This contradicts the phosphorylation data that shows no significant response upon GDF5 stimulation. The NLP algorithm, in an attempt to minimize the objective function, fitted the phosphoprotein data and misfitted the cytokine releases under GDF5.

3.3.3 Discussion

In this section, we have presented an unbiased approach for studying signal transduction in chondrocytes on a systems level. We interrogated their signaling mechanisms, downstream 78 receptors of interest, on both the phosphoproteomic and the cytokine release level and integrated this data in a consistent model, predictive of their function and response. On the experimental front, we adopted the xMAP technology to measure 17 key phosphoproteins and the release of 55 cytokines in the supernatant, upon stimulation with single treatments of the 78 ligands. On the computational front, we implemented a previously published optimization formulation, based on the modeling of signal transduction via constrained fuzzy logic, to train a PKN to proteomic data. The optimized model successfully captures the signaling mechanisms of primary human chondrocytes.

The analysis presented herein, was able to validate previous findings regarding major players of cartilage homeostasis (e.g. IL1α, IL1β, TNF, EGF, TGFα), as well as identify novel players that may have an important role in chondrocytes' signaling (e.g. INHBA, DEFB1, FGF2, CXCL1, IL19, NOG, GDF5, LTA). In more detail, novel (or under reported) pathways include: *(i)* the TLRs pathways, that initiate at the various TLRs and propagate downstream through MYD88 to IRAK and TIFA, then merge with the IL1α, IL1β pathways and ultimately activate major inflammation related signals such as IκB, HSP27, MAPK14 and JUN. Activation of IκB, HSP27 and JUN leads to the release of several cytokines and MMPs as widely reported in literature [20]. *(ii)* Growth factors such as HBEGF, BTC (EGFR ligand) and FGF2, that signal through pathways partially overlapping with EGF and TGFα and activate growth related signals such as MAP2K1, ERK and AKT. Interestingly, FGF2 induces the release of MMP13 (see also [108]), via JUN activation. *(iii)* DEFB1 (Defensin, Beta 1), that signals through RAS, RAF to activate growth related phosphoproteins, or through PI3K to activate AKT, JUN and HSP27. JUN activation induces the release of MMP13.

The integration of phosphoproteomic and cytokine release data into a consistent model adds to its robustness against experimental noise and assay calibration issues, since small changes in the phosphorylation value of certain proteins that would otherwise go unnoticed, may induce the release of cytokines, implying these changes are in fact significant. Moreover, this approach is easily extensible to other kinds of response data such as gene expression, cell growth, GAG release, PG loss, or even mechanical properties of cartilage tissue, setting the example of how extensive models spanning on many different levels may be constructed to model cartilage degeneration on a systems scale.

Regarding major limitations of the proposed approach, these include *(i)* the static view of the signaling pathway together with *(ii)* the limited number of measured phosphoproteins. The static view of the signaling pathway results from obtaining phosphoproteomic measurements in a single time point, representing the average early response (average activity of 5 and 25 minutes). Even though protein connectivity can be interrogated properly in this manner, more complex structures (e.g. feedback loops) cannot. Concerning the limited number of measured phosphoproteins, as proteomic technologies evolve, more assays will become available allowing

us a broader perspective on the signaling mechanisms of chondrocytes.

Overall, the proposed approach successfully addressed the construction of a functional, predictive model of the signaling mechanisms in human chondrocytes. We interrogated signal transduction downstream 78 receptors of interest, on both the phosphoproteomic and the cytokine release level, validating previous findings on major players of cartilage homeostasis and identifying novel players that may have a significant effect on chondrocytes signaling. Extension of this approach to include other types of response data such as gene expression, cell growth, or mechanical properties of cartilage tissue, may lead to the construction of integrated models spanning on many different levels facilitating the systems scale study of OA.

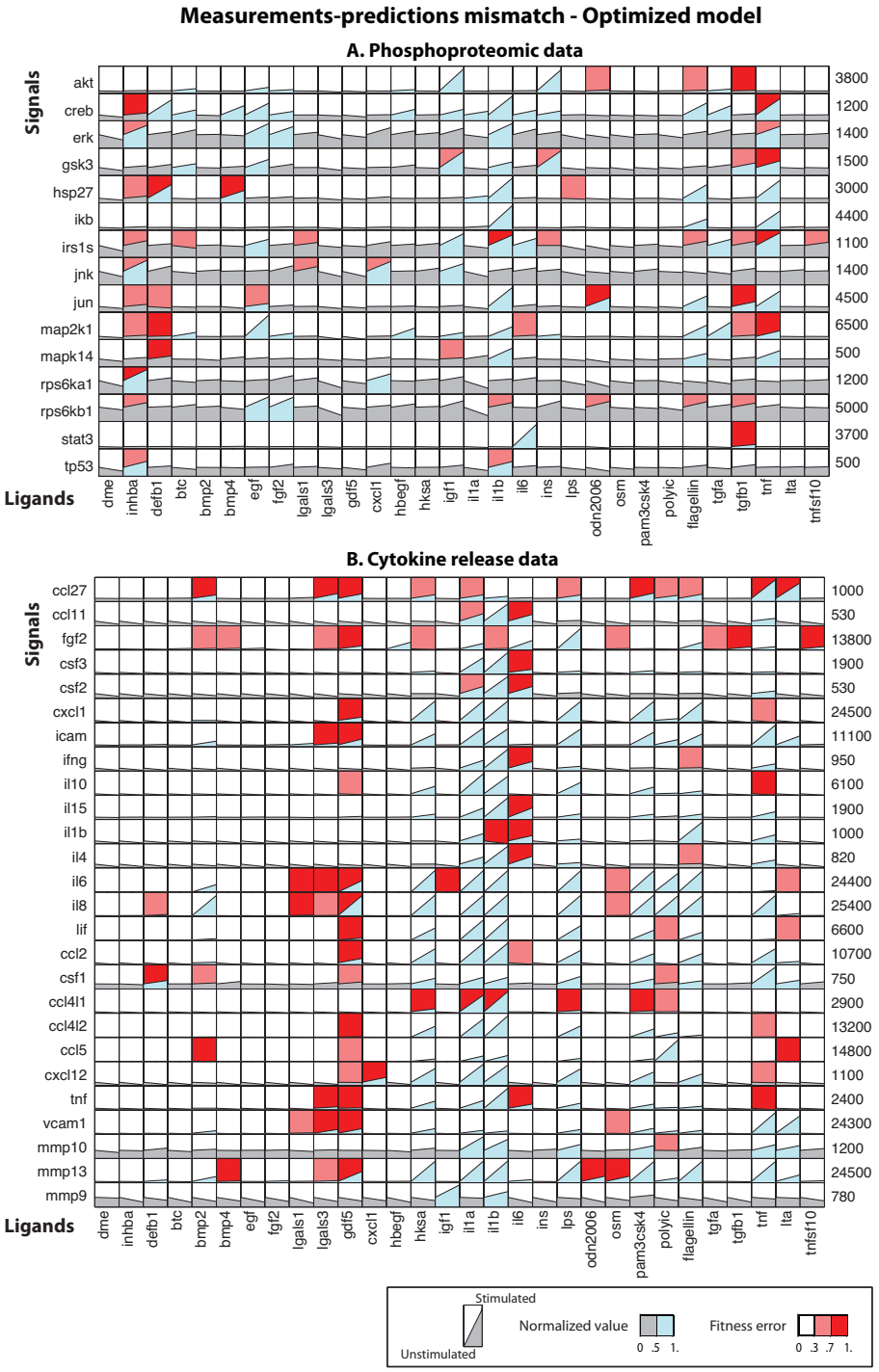


Figure 3.18: **Measurements-predictions mismatch for the optimized model**: The difference between model predictions and experimental data is illustrated (red background color in A and B). (A) Measurements-predictions mismatch respective to the normalized phosphoproteomic data. The average mismatch (or fitness error) is 14.62%, decreased from 49.88% for the initial model. (B) Measurements-predictions mismatch respective to the normalized cytokine release data. The average mismatch (or fitness error) is 11.97%. Remaining fitness error is mostly attributed either to conflicting measurements in the phosphoproteomic or the cytokine release datasets, or to poor prior knowledge of the proteins' connectivity in the signaling network.

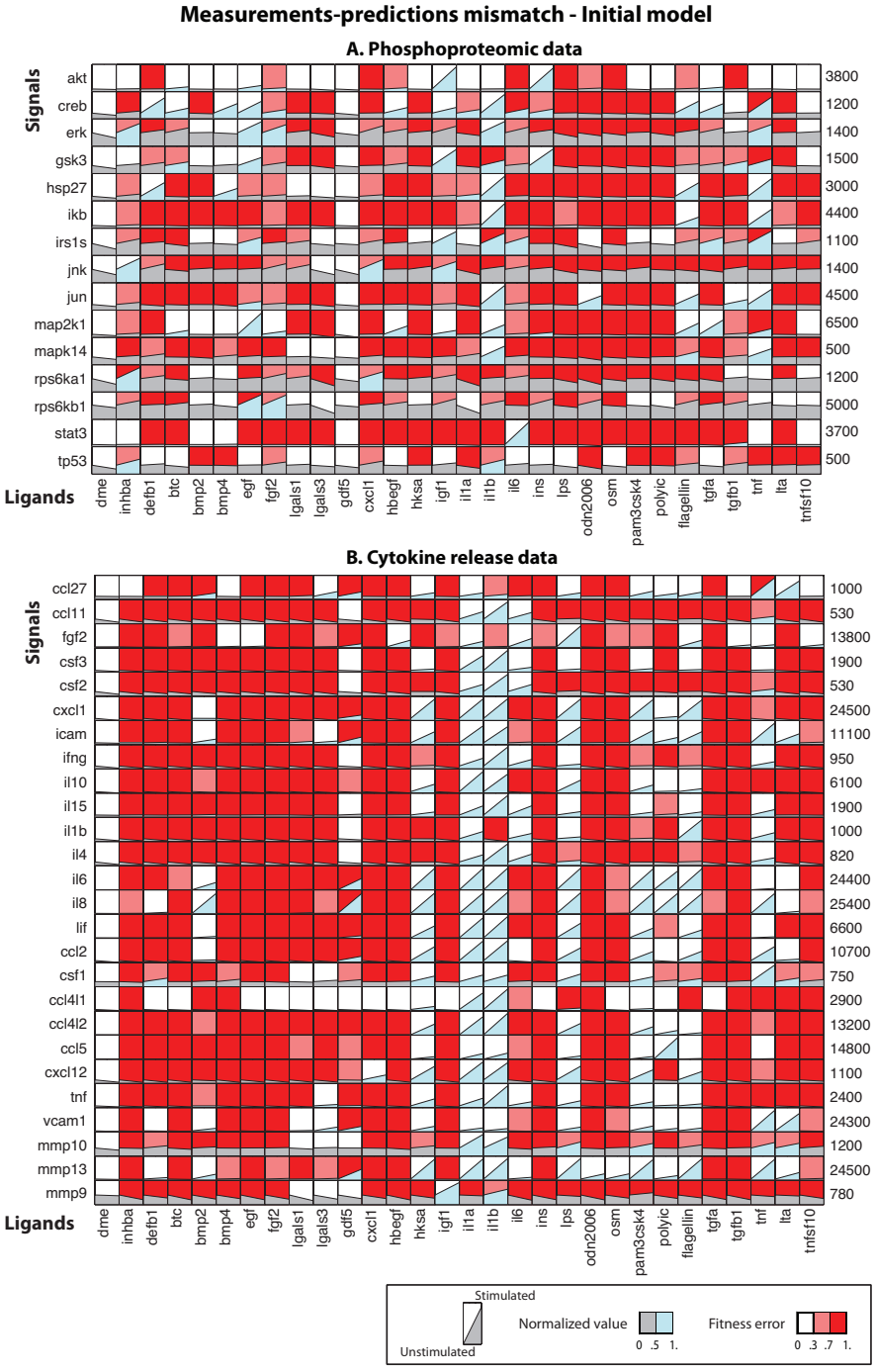


Figure 3.19: Measurements-predictions mismatch for the initial model: The difference between model preditions and experimental data for the initial model is illustrated (red background color in A and B). (A) Measurements-predictions mismatch respective to the normalized phosphoproteomic data. (B) Measurements-predictions mismatch respective to the normalized cytokine release data. As shown by the average mismatch / fitness error (49.88%) the initial model fails to capture the signaling mechanisms of human chondrocytes. The large amount of fitness error is attributed to the generic nature of signaling reactions in on-line databases, and since the PKN is used as a scaffold for the subsequent modeling and analysis, the model predictions will only be as accurate as the imported prior knowledge. After the parameter fitting by the NLP formulation the fitness error is decreased significantly to 14.68% and 11.97% (phosphoprotein and cytokine release data respectively).

Chapter 4

Detecting and Removing Inconsistencies Between Experimental Data and Signaling Network Topologies using Integer Linear Programming on Interaction Graphs

In this work submitted for publication in January 2013, An ILP formulation was introduced for detecting and removing inconsistencies between experimental data and signaling network topologies. This work was carried out in collaboration with Steffen Klamt (group leader in the Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany). The proposed ILP provides a novel, flexible, framework for optimizing signaling pathways to measured data. The computational efficiency of the proposed formulation and the fact that a global solution is obtained makes it an appealing approach for interrogating a variety of interaction graphs and accompanying signaling datasets.

Abstract

Cross-referencing experimental data with our current knowledge of signaling network topologies is one central goal of mathematical modeling of cellular signal transduction networks. We present a new methodology for data-driven interrogation, correction and inference of signaling networks. While most published methods for signaling network inference operate on Bayesian, Boolean, or ODE models, our approach uses integer linear programming (ILP) on interaction graphs to encode constraints on the qualitative behavior of the nodes. These constraints are posed by the network topology and their formulation as ILP allows us to predict the possible qualitative changes of the activation levels of the nodes for a given stimulus. We provide four basic operations to detect and remove inconsistencies between measurements and predicted behavior: (i) find a topology-consistent explanation for responses of signaling nodes measured in a stimulus-response experiment (if none exists, find the closest explanation); (ii) determine a minimal set of nodes that need to be corrected to make an inconsistent scenario consistent; (iii) determine the optimal subgraph of the given network topology which can best reflect measurements from a set of experimental scenarios; (iv) find possibly missing edges that would improve the consistency of the graph with respect to a set of experimental scenarios at most.

We demonstrate the applicability of the proposed approach by interrogating a manually curated

interaction graph model of EGFR/ErbB signaling against a library of high-throughput phosphoproteomic data measured in primary hepatocytes. Our methods detect interactions that are likely to be inactive in hepatocytes and provides suggestions for new interactions that, if included, would significantly improve the goodness of fit.

Our framework is highly flexible and the underlying model requires only easily accessible biological knowledge. All related algorithms were implemented in a freely available toolbox *SigNetTrainer* making it an appealing approach for various applications.

4.1 Introduction on optimization of interaction graphs

Recent advancements in high-throughput phosphoproteomic technologies have led to the generation of large datasets, capturing the cell's response to factors of its biochemical microenvironment [11, 12]. However, interpreting the increasing amounts of available data in such a manner that biologically relevant insights can be drawn for the interrogated system is far from trivial. To this end, signaling data are often examined in conjunction with network models that represent our current knowledge of the causality of cellular signal flows (as stored, for example, in online pathway databases [27, 28, 26]). Finding, in a rigorous fashion, causal explanations for experimental data in the context of a given network topology is one of the key challenges for systems biology of cellular signaling.

Significant work has been published on this front attempting to identify inconsistencies between measured data and signaling topologies [23, 13, 10, 39, 109, 110, 111, 112, 113]. Some methods also facilitate an optimization of the network structure to identify the wiring diagram that can best fit the data at hand [13, 23, 114]. However, before such an analysis can be conducted one has to choose an appropriate modeling formalism. Common approaches used for modeling signal transduction networks are based on graphs [111, 112, 115, 32], Bayesian networks [114], some form of logical modeling including Boolean or constrained fuzzy logic [115, 9, 116], hybrid intelligent systems [31, 9, 32, 33, 34], or Ordinary Differential Equations (ODEs) [35, 36, 37].

Deciding on the mathematical formalism to be used for representing and modeling signal transduction networks is often not trivial and depends on many factors such as the amount and type of available data, the quality of prior knowledge, whether transient or steady-state behavior needs to be addressed, the biological questions that are to be answered, the computational efforts and so forth. For example, ODE modeling or constrained fuzzy logic are closer to the actual mechanics of signal transduction than Boolean logic as they support continuous values for the activation states of signaling species, but at the cost of numerous free parameters. These parameters must be known (in addition to the actual (initial) network structure) or estimated from experimental data. A great number of parameters in the model often gives rise to identifiability problems requiring a more elaborate training dataset.

Graph models are probably the simplest models of signaling networks one can think of. In particular, *signed directed graphs* (also called interaction graphs, dependency graphs, or influence graphs), where each edge indicates either a positive or a negative effect of one node upon another, have frequently been used to investigate basic functional properties of biological networks with signal or information flows. Despite their simplicity, interaction graphs (IGs) capture the most important biological information and are useful to uncover fundamental network properties such as feedback and feedforward loops or global interdependencies between the involved players. The fact that each Boolean and each ODE model has an underlying IG renders the analysis of IGs directly relevant also for other modeling formalisms. A famous example is the fact that a system (in an ODE or Boolean model representation) exhibiting bistability must contain a positive feedback loop in its underlying network structure [117, 118]. Properties that are uniquely identifiable from a given IG immediately hold for all ODE and Boolean models that have this

IG as underlying wiring diagram, whereas the opposite direction does not hold. For example, in Figure 4.1A we see that there is (exactly) one path in the IG leading from node A to node G and that this path is negative. We can therefore uniquely conclude from the IG that, in any Boolean or ODE model derived from it, a perturbation in A cannot lead to an increase in the (activation) level of G. In contrast, there is a positive and a negative path from A to F, hence, nothing can be concluded from the graph alone when perturbing A. In fact, it will depend on the kinetics and parameters in an ODE model (and the logical functions in a logical model) whether the level of B will increase, decrease, or, in the extreme case, remain constant.

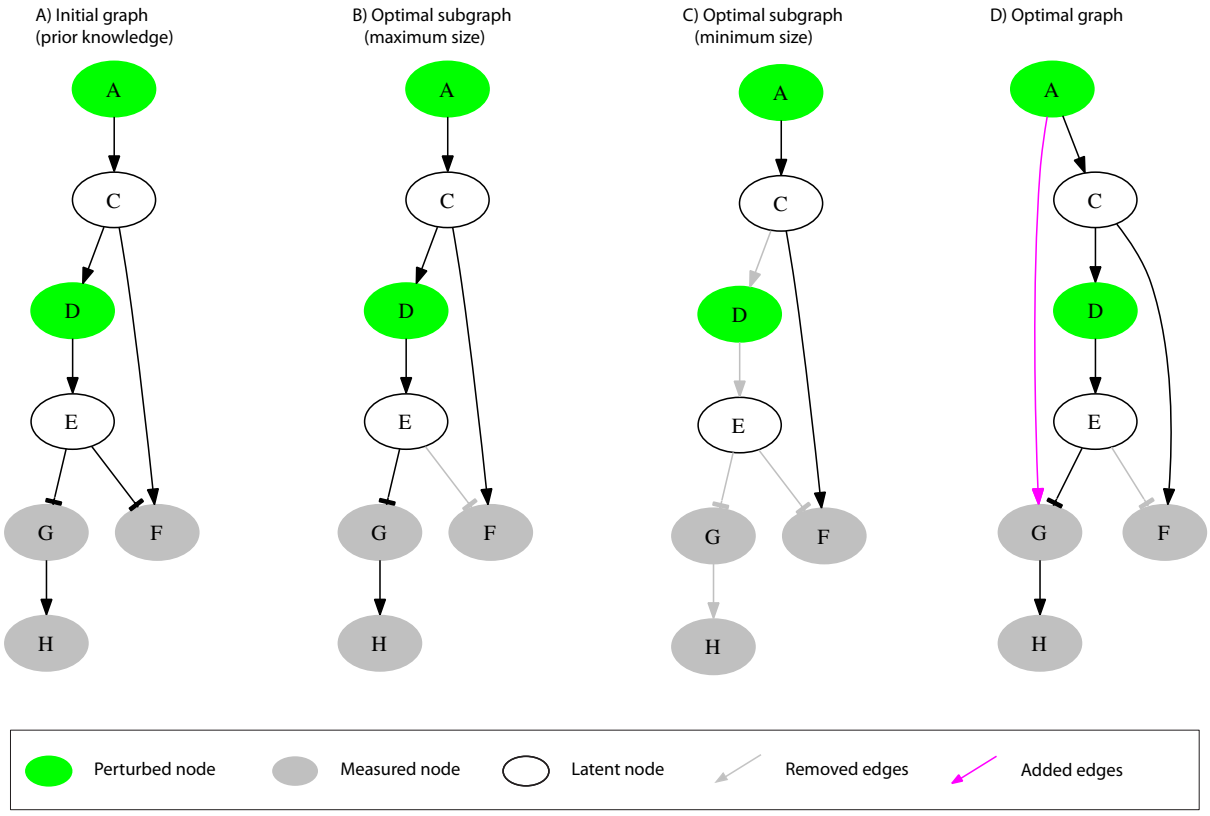


Figure 4.1: **A simple example network used for illustration purposes**. The interaction graph consists of 7 nodes and 7 edges. The green nodes A and D can be perturbed externally; the grey nodes F, G and H are the readouts of the network whose activation state is measured in the experiments; the white nodes C and E are latent nodes which are neither perturbed nor measured. (A) The initial topology of the interaction graph representing the prior knowledge. (B) The optimal sub-graph with maximum number of edges retained from the original graph minimizing the fitting error on the experimental scenarios (see Table 4.1). Edge $E \to F$ is removed, as F remains inactive in scenario 3 upon up-regulation of D (Table 4.1), suggesting that $E \to F$ is not active. (C) The optimal sub-graph with the minimum number of remaining edges (yielding the same fitting errors as solution (B). (D) The optimal interaction graph obtained by combining insertion of new and removal of existing edges. By adding edge $A \to G$ and removing $E \dashv F$ the fitting error is eradicated completely (see also Table 4.1).

The previous example shows that IGs can be used to make predictions (without needing any further parameters) on the qualitative behavior of signaling and regulatory networks. These predictions can easily be compared with (qualitative trends of) experimental data, typically from

Table 4.1: **Example scenarios and optimizations for the example network in Figure 4.1.**

	Perturbations		Measurements			Initial fitting error (Fig. 1A)			MCoS	Remaining fitting error Fig. 1B/Fig. 1C/Fig. 1D		
	A	D	F	G	H	F	G	H		F	G	H
sc1	1	-1	1	1	1	0	0	0		0/0/0	0/1/0	0/1/0
sc2		1	0	-1	-1	1	0	0	$F = -1$	0/0/0	0/0/0	0/0/0
sc3	1		1	1	1	0	2	2	$E = -1$/ $G = 1$	0/0/0	2/1/0	2/1/0

The rows correspond to scenarios 1 to 3. The "Perturbations" column shows the externally imposed state of the nodes A and D which can be -1 (down-regulation), 0 (state of the node did not change), or 1 (activation level is increased). No value is given if the node was not perturbed. The "Measurements" column shows the measured change of the activation level of F, G and H in the respective scenarios. The "Initial fitting error" column shows the total mismatch of predictions and measurements with respect to the initial topology (shown in Figure 4.1A). The "MCoS" (Minimal Correction Sets) column shows virtual corrections in the activation level of some nodes which would lead to a perfect fit of the data (resulting fitting error is 0). The "Remaining fitting error" columns show the remaining mismatches for the optimal subgraphs depicted in Figures 4.1B and 4.1C and for the optimal graph displayed in Figure 4.1D. The original netwok in Figure 4.1A has a total fitting error of 5; it is 4 for the optimal subgraphs in Figures 4.1B and 4.1C and it becomes 0 in the optimal graph in Figure 4.1(D).

stimulus-response experiments. The concept of the dependency matrix introduced in [115] is consequently based on the idea used above, namely to check—for each (ordered) pair (A, B) of nodes A and B—the existence of positive and negative paths (and negative feedback loops) to make predictions on the effect of perturbations in A. This concept has been applied, for instance, in [32] to experimental data of the epidermal growth factor (EGF) receptor signaling network. The comparisons of the predictions from the dependency matrix with the measured behavior from several combinatorial stimulations showed several inconsistencies from which some (cell-type specific) conclusions on missing or probably inactive interactions could be made. However, these conclusions were drawn by inspection and not in a systematic way. It is therefore one goal of this study to develop methods that find, in an automatic way, corrections in the network structure improving the consistency. The dependency matrix is useful to get an overview on how a node can potentially influence the other nodes in the network; however, it may become limiting if multiple node values are measured in one experiment. Given the IG topology, state changes measured for certain nodes are, in general, not independent and therefore require stronger constraints. For example, assume there would be another node Z in Figure 4.1A that is activated by F (edge $F \to Z$). From the IG topology we know that F and Z can both decrease or increase their levels if A is perturbed (as correctly predicted by the dependency matrix); however, it is not possible that their new steady state levels change in different directions.

A related class of methods for detecting discrepancies between IG topology and experimental data relies on the *sign consistency rule* [110, 111, 112]. The key idea is that, in a steady-state shift experiment, the direction of change of the state of a node must be explainable by the direction of change of at least one of its predecessor nodes (except for the directly perturbed

node(s)). For example, in Figure 4.1A, after a perturbation in node A , the steady-state level of node F may have become larger only if node E decreased its activation level (as E inhibits F) or if node C increased its level (as C activates F). The sign consistency rule gives rise to constraints on the possible patterns of "ups and downs" of the nodes' activation level in a given IG. These constraints can be encoded, for example, by Answer Set Programming [112]. Confronting these constraints with experimental data may then lead to the detection of topological inconsistencies, namely if no sign pattern complying with the given measurements and perturbations can be found [110, 111, 112].

The novel methods we will present herein are based on a similar sign consistency rule; however, they differ in a number of aspects. First, we will encode the sign constraints as an Integer Linear Programming (ILP) problem which has not been described before. This formulation gives us the opportunity to utilize the large corpus of efficient algorithms developed for ILP problems. Furthermore, for the situation that multiple stimulus-response experiments are available, we will address aspects that go beyond the detection of inconsistencies from single experiments, namely to correct a given network structure such that the number of mismatches is minimized. For the structure optimization process we will consider edge removals as well as edge additions.

As starting point, we assume that we are given (i) an initial IG topology, for example, a "master topology" of a signaling pathway subsuming all reported (potential) interactions and (ii) a set of stimulus-response experiments (scenarios) in each of which some nodes were perturbed and the resulting up- or downregulation of some readout nodes was measured. The IG is a signed directed graph $G = (V, E, \sigma)$, where V is the set of nodes (species), E is the set of edges (interactions) and σ is the set of signs corresponding to edges in E ($\sigma_e \in \{-1, 1\}$, $e \in E$). Figure 4.1A and the three experimental scenarios in Table 4.1 (defined by the columns "Perturbations" and "Measurements") provide an illustrative example. Here, A and D are nodes that can be perturbed; for F, G and H we get measurements, and C and E are the latent nodes which are neither perturbed nor measured.

Our goal is now to analyze and improve the consistency of an IG topology with respect to a given set of experimental data. Central to all algorithms presented herein is the following definition of sign consistency:

Definition1 (Sign Consistency). We are given an IG and a node labeling (sign pattern) $\mathbf{s}$ which stores for each node X a sign $s_X \in \{-1, 0, 1\}$. We say that $\mathbf{s}$ is *sign-consistent with respect to the IG* if the following conditions hold for each node X:
a) If $s_X = -1$: either s_X was fixed (marked as perturbed node), or there is a predecessor node Y and an edge $e : Y \to X$ with $\sigma_e \cdot s_Y = -1$.
b) If $s_X = 1$: either s_X was fixed (marked as perturbed node), or there is a predecessor node Y and an edge $e : Y \to X$ with $\sigma_e \cdot s_Y = 1$.
c) If $s_X = 0$: either (i) s_X was fixed (marked as perturbed node), or (ii) X has no predecessor, or (iii) for all edges $Y \to X$ we have $s_Y = 0$, or (iv) there is an edge $e : Y \to X$ with $\sigma_e \cdot s_Y = -1$ and another edge $h : Z \to X$ with $\sigma_h \cdot s_Z = 1$.

In our setting, the signs of the external perturbations as well as the measured signs of the readout nodes can be described by a specific node labeling (which we call the *associated* labeling of the scenario). In realistic applications, one usually has latent nodes which are neither perturbed nor measured, hence, the associated node labeling of an experimental scenario may contain unknown values which we denote by NaN. We call incomplete sign patterns *partial labelings*. A partial labeling $\tilde{\mathbf{s}}$ is sign-consistent if there exists a complete sign-consistent labeling $\mathbf{s}$ for which we have $\tilde{s}_X = s_X$ wherever $\tilde{s}_X \neq NaN$. In this sense, we say that an experimental scenario is sign-consistent if its associated (partial) labeling is sign-consistent. Finally, if we have a *collection* of scenarios we say that this collection is sign-consistent with the IG if all the

(partial) labelings associated with the scenarios are sign-consistent.

We can now consider four fundamental problems on the consistency of experimental scenarios with respect to a given IG:

(1) SCEN_FIT

Given a single experimental scenario, we fix the states of the perturbed nodes (according to the experimental interventions) and search then for an optimal sign-consistent node labeling which has a minimal mismatch with the given measurements. In the ideal case, namely if the associated labeling of the experimental scenario is sign-consistent, the *fitting error* will be 0. The latter is defined as the absolute difference between measurements $\mathbf{m}$ and the optimal sign pattern $\mathbf{s}$: $\sum_{X:m_X \neq NaN} |m_X - s_X|$.

In Figure 4.1A/ Table 4.1, we see that scenario 1 is sign-consistent: A was externally increased and D decreased, and with $s_A = s_C = s_G = s_F = s_H = 1$ and $s_D = s_E = -1$, we obtain a sign-consistent labeling giving us a possible explanation for the measurements. In contrast, scenario 2 is not consistent with the IG topology: if D is increased externally (no perturbation in A), then we expect to see a decrease in F, G and H which is not seen in F (unchanged). The minimal resulting fitting error for an optimal sign pattern is thus 1. Generally, an error of 1 or -1 occurs if a change was expected/ not expected but was not seen/ was seen in the experiments. For scenario 3, the predictions are even worse: increase in A (no perturbation in D which thus depends on C) should lead to down-regulation of G and H, but an increase is measured for both. We thus get an absolute error of 2 for each of the two predictions. The fitting error of a sign-consistent node labeling closest to scenario3 can thus not be smaller than 4.

It may happen that several solutions exist explaining a given scenario equally well. For example, assume again that there was another node Z in Figure 4.1A that is activated by F through an edge $F \rightarrow Z$. If we now measured $G = H = F = 1$ and $Z = -1$ after positively perturbing A ($A = 1$), then the best scenario fit would result in an error value of 2 since F and Z must have the same value. However, there are three optimal solutions regarding F and Z, namely $F = Z = 0$, $F = Z = 1$, or $F = Z = -1$, all leading to the same minimal fitting error of 2. For some applications it will be helpful to know all these optimal solutions and we will therefore also address their enumeration.

(2) Minimal Correction Sets (MCoS)

Another optimization problem for a single scenario directly follows if a given scenario is not sign-consistent, i.e., if no sign-consistent labeling can be found that results in a fitting error of 0. We can then try to identify a minimal set of nodes that, if corrected by fixing their states to a certain value, would lead to a consistent scenario. We call these sets *Minimal Correction Sets* (MCoS), the minimality property demanding that no subset of a MCoS would lead to a consistent labeling. For example, regarding scenario 3 in Table 4.1, there are two MCoS suggesting that there was either a down-regulation of E ($E = -1$) or an upregulation of G ($G = 1$), both of unknown cause. Thus, MCoS show possible places in the network that have a high probability to cause the observed inconsistencies. With the MCoS problem we identify the enumeration of MCoS of minimal size for a given scenario (a simple extension not considered herein is to enumerate all MCoS irrespective of their size).

(3) OPT_SUBGRAPH

The first two problems focus on a single scenario; now we intend to optimize the network structure in such a way that the total fitting error over all scenarios is minimized. Initially, we allow only the removal of edges in the network, that is, we search for an optimal subgraph. As there

might be several solutions to this optimization problem, we consider the following sub-problems: computation of any/ of the sparsest/ of the largest sub-network of the initial IG minimizing the mismatches. In addition, we may also be interested in an enumeration of all sub-networks minimizing the number of inconsistencies between IG topology and data. As an example, Figure 4.1B shows the maximum and Figure 4.1C the minimum solution (maximum and minimum with respect to the number of remaining edges) of optimal subgraphs of the IG given in Figure 4.1A considering all three scenarios in Table 4.1. Both solutions reduce the total fitting error from 5 to 4 (and there are no solutions that could reduce it further).

(4) OPT_GRAPH

The removal of certain edges may significantly improve the agreement between measurements and network topology, but some fitting errors can often only disappear if we have additionally the opportunity to add new interactions. This fourth optimization problem therefore intends to minimize the fitting error by allowing edge removals *and* insertions in parallel. Obviously, the fit cannot be worse than one obtained by problem (3). For smaller networks, a full enumeration of all optimal solutions might be possible. However, as the insertion of new interactions increases the solution space dramatically in large networks, we may consider a *greedy* strategy which determines, in each iteration, the optimal edge whose inclusion (in combination with the pruning step (3)) decreases the fitting error at most. One may then add this edge permanently and repeat the algorithm described above until no further significant improvement can be obtained by inserting a new edge.

Figure 4.1D shows the result of this optimization step in our example: the edge $A \rightarrow G$ is identified as missing edge which, in combination with a pruning step, completely eradicates the original fitting errors in all scenarios. The resulting network is thus fully consistent with the entire set of experimental data.

4.2 ILP formulation

Basic definitions and ILP formulation of sign consistency

As described in the Introduction section, we assume that we are given an interaction graph (signed digraph) $G = (V, E, \sigma)$ capturing our prior knowledge on the signaling topology and, additionally, a set of experimental scenarios each consisting of a specific set of perturbed nodes and a set of taken measurements. The edges (also called interactions) are indexed by $i = 1, \ldots, n_E$ ($n_E = |E|$), the nodes by $j = 1, \ldots, n_V$ ($n_V = |V|$), and the scenarios by $k = 1, \ldots, n_S$. The experimental scenarios are specified by two matrices: (i) the $n_V \times n_S$ perturbation matrix $\mathbf{p}$ with $p_{j,k} \in \{-1, 0, 1\}$ storing the (enforced) state of node j in scenario k through external perturbation, and (ii) the $n_V \times n_S$ measurement matrix $\mathbf{m}$ with $m_{j,k} \in \{-1, 0, 1\}$ storing the measured change of the (steady) state level of node j in scenario k. Perturbation and measurement values thus indicate enforced/measured upregulation (1), downregulation (-1), or unchanged state (0). Usually, only a small subset of nodes is perturbed, and only a subset of nodes can be measured; unperturbed and non-measured states are therefore marked by NaN in the matrices $\mathbf{p}$ and $\mathbf{m}$, respectively.

In what follows we translate sign-consistency of a node labeling (according to Definition 1) into equality and inequality constraints of an Integer Linear Programming (ILP) problem. In this formulation, the predicted state of a node j in experiment k will be represented by an integer variable $x_{j,k} \in \{-1, 0, 1\}$. Again, $x_{j,k} = 1$ encodes upregulation and $x_{j,k} = -1$ downregulation of node j in scenario k, whereas $x_{j,k} = 0$ indicates that the activation level of j remained unchanged.

The i-th signaling edge is defined as $S_i \to P_i$, where $S_i \in V$ is the start node and $P_i \in V$ the end node of edge i. Furthermore, for each edge i we encode its sign by two binary variables σ_i^+ and σ_i^- : if i is an activation ($\sigma_i = 1$), then $\sigma_i^+ = 1$ and $\sigma_i^- = 0$; else if i is an inhibition ($\sigma_i = -1$), then $\sigma_i^- = 1$ and $\sigma_i^+ = 0$.

In addition, the binary variables $u_{i,k}^+$ and $u_{i,k}^-$ are introduced to represent the potential of edge i to up- or downregulate its end node P_i in experiment k. Edge i with start node $j = S_i$ has the potential of upregulating its target node P_i in experiment k (i.e., $u_{i,k}^+ = 1$) if and only if $\sigma_i^+ = 1$ and $x_{j,k} = 1$; or $\sigma_i^- = 1$ and $x_{j,k} = -1$. In any other case, $u_{i,k}^+ = 0$. Accordingly, edge i with start node $j = S_i$ has the potential of downregulating its target P_i in experiment k (i.e. $u_{i,k}^- = 1$) if and only if $\sigma_i^+ = 1$ and $x_{j,k} = -1$; or $\sigma_i^- = 1$ and $x_{j,k} = 1$. In any other case, $u_{i,k}^- = 0$. Thus, with $j = S_i$,

$$u_{i,k}^+ = \max(0,\, x_{j,k} + \sigma_i^+ - 1,\, -x_{j,k} + \sigma_i^- - 1) \tag{4.1}$$

$$u_{i,k}^- = \max(0,\, -x_{j,k} + \sigma_i^+ - 1,\, x_{j,k} + \sigma_i^- - 1) \tag{4.2}$$

Introducing the binary variables $d1_{i,k}, d2_{i,k}, \ldots, d6_{i,k}$, constraints (4.1) and (4.2) can be reformulated as an ILP in the following way:

$$x_{j,k} + \sigma_i^+ - 1 = aux1_{i,k} \tag{4.3}$$

$$-x_{j,k} + \sigma_i^- - 1 = aux2_{i,k} \tag{4.4}$$

$$-x_{j,k} + \sigma_i^+ - 1 = aux3_{i,k} \tag{4.5}$$

$$x_{j,k} + \sigma_i^- - 1 = aux4_{i,k} \tag{4.6}$$

$$u_{i,k}^+ \geq aux1_{i,k} \tag{4.7}$$

$$u_{i,k}^+ \geq aux2_{i,k} \tag{4.8}$$

$$u^-_{i,k} \geq aux3_{i,k} \tag{4.9}$$

$$u^-_{i,k} \geq aux4_{i,k} \tag{4.10}$$

$$u^+_{i,k} \leq 1 - d1_{i,k} \tag{4.11}$$

$$u^+_{i,k} \leq aux1_{i,k} + 3 - 3\,d2_{i,k} \tag{4.12}$$

$$u^+_{i,k} \leq aux2_{i,k} + 3 - 3\,d3_{i,k} \tag{4.13}$$

$$u^-_{i,k} \leq 1 - d4_{i,k} \tag{4.14}$$

$$u^-_{i,k} \leq aux3_{i,k} + 3 - 3\,d5_{i,k} \tag{4.15}$$

$$u^-_{i,k} \leq aux4_{i,k} + 3 - 3\,d6_{i,k} \tag{4.16}$$

$$d1_{i,k} + d2_{i,k} + d3_{i,k} = 1 \tag{4.17}$$

$$d4_{i,k} + d5_{i,k} + d6_{i,k} = 1 \tag{4.18}$$

Finally, two more binary variables, $x^+_{j,k}$ and $x^-_{j,k}$ are introduced to represent the potential for node j of being up- or downregulated depending on the activity of its upstream edges. Node j has the potential of being upregulated ($x^+_{j,k} = 1$) if and only if an edge i exists such that $j = P_i$ and $u^+_{i,k} = 1$, and node j has the potential of being downregulated ($x^-_{j,k} = 1$) if and only if i exists such that $j = P_i$ and $u^-_{i,k} = 1$. Thus,

$$x^+_{j,k} \geq u^+_{i,k}, \quad \forall i \text{ with } P_i = j \tag{4.19}$$

$$x^-_{j,k} \geq u^-_{i,k}, \quad \forall i \text{ with } P_i = j \tag{4.20}$$

$$x^+_{j,k} \leq \sum_{i:j=P_i} u^+_{i,k}, \quad \forall i \text{ with } P_i = j \tag{4.21}$$

$$x^-_{j,k} \leq \sum_{i:j=P_i} u^-_{i,k}, \quad \forall i \text{ with } P_i = j \tag{4.22}$$

The state $x_{j,k}$ of node j in scenario k is constrained by the values of $x^+_{j,k}$ and $x^-_{j,k}$ according to the definition of sign-consistency (see Definition 1): (i) Node j may be be upregulated ($x_{j,k} = 1$) if it has the potential of being upregulated ($x^+_{j,k} = 1$). (ii) Node j may be downregulated ($x_{j,k} = -1$) if it has the potential of being downregulated ($x^-_{j,k} = 1$). (iii) Node j may stay unchanged ($x_{j,k} = 0$) if it has the potential of being both up- and downregulated ($x^-_{j,k} = x^+_{j,k} = 1$) or neither of the above ($x^-_{j,k} = x^+_{j,k} = 0$). These rules are encoded in inequalities as follows:

$$x_{j,k} \leq x^+_{j,k} \tag{4.23}$$

$$x_{j,k} \geq -x^-_{j,k} \tag{4.24}$$

$$x_{j,k} \leq 2\,x^+_{j,k} - x^-_{j,k} \tag{4.25}$$

$$x_{j,k} \geq -2\,x^-_{j,k} + x^+_{j,k} \tag{4.26}$$

The equations and inequalities derived in this subsection describe sign-consistent node labelings and provide the frame within which we can now address the four basic optimization problems posed in the Introduction section.

SCEN_FIT

The goal of SCEN_FIT is to identify, for a given scenario k, a vertex labeling that is closest to the measurements of this scenario. We first have to constrain the values of the perturbed nodes in scenario k:

$$x_{j,k} = p_{j,k} \quad \forall j \text{ with } p_{j,k} \neq NaN \tag{4.27}$$

Realistic perturbations typically affect either input nodes (e.g., ligands) or internal nodes in the case where a specific inhibitor was added or where a constitutive activation or a knock-in/knock-out is introduced. The state of the perturbed nodes are thus fixed to the enforced value and the constraints (4.23) – (4.26) are omitted for these nodes to preserve the consistency of the formulation.

We now search for a sign-consistent labeling (fulfilling thus all constraints of the previous subsection) that minimizes the measurement-prediction-mismatch. The following objective function expresses this goal:

$$\sum_{j:m_{j,k}\neq NaN} a_{j,k}|m_{j,k} - x_{j,k}| = \min! \tag{4.28}$$

The $a_{j,k}$ are user-defined constants set to 1 by default (in principle, they could be used for a weighted sum of measurement errors). By introducing $abs_{j,k} = |m_{j,k} - x_{j,k}|$, $abs_{j,k} \in \{0, 1, 2\}$, the absolute value above is reformulated as follows:

$$abs_{j,k} \geq m_{j,k} - x_{j,k} \tag{4.29}$$

$$abs_{j,k} \geq x_{j,k} - m_{j,k} \tag{4.30}$$

The resulting states $x_{j,k}$ for scenario k represent an optimal solution as desired for SCEN_FIT.

As discussed in the Introduction section, we also consider the enumeration of *all* optimal SCEN_FIT solutions for a given scenario. To this end, we solve the ILP repeatedly, and, after each run, we exclude previously found solutions by adding the following constraints for each previous solution s:

$$\sum_j |x_{j,k} - x_{j,k,s}| \geq 1, \tag{4.31}$$

where $x_{j,k,s}$ represent the value of $x_{j,k}$ in solution s. We set $|x_{j,k} - x_{j,k,s}| = dx_{j,k,s}$. Constraint(4.31) is reformulated as follows:

$$\sum_{j,k} dx_{j,k,s} \geq 1 \tag{4.32}$$

$$-x_{j,k} + dx_{j,k,s} - 4\,dx2_{j,k,s} \leq x_{j,k,s} \tag{4.33}$$

$$x_{j,k} + dx_{j,k,s} - 4\,dx1_{j,k,s} \leq -x_{j,k,s} \tag{4.34}$$

$$dx1_{j,k,s} + dx2_{j,k,s} = 1 \tag{4.35}$$

where $dx1_{j,k,s}$ and $dx2_{j,k,s}$ are binary variables. We may then compute a new sign-consistent labeling of the nodes by optimizing again objective function (4.28). To ensure that only solutions with minimum fitting error are found, we replace, after the first iteration, the objective function in (4.28)by forcing instead the algorithm to find solutions with the same (minimum) fitting error as in the first run:

$$\sum_{j:m_{j,k}\neq NaN} a_{j,k}|m_{j,k} - x_{j,k}| = objval \tag{4.36}$$

Here, *objval* is the value of the objective function after the first run of the algorithm. The resulting problem becomes thus a simple search for a feasible solution and is repeated until no further solution can be found.

Minimal Correction Sets - Computing a single Minimal Correction Set

Next we address the identification of a Minimal Correction Set (MCoS) for a sign-inconsistent scenario k. An MCoS indicates possible causes of discrepancies between measured data and assumed IG topology. As described in the Introduction section, MCoS correspond to artificial perturbations of certain nodes which render the measurements from a given scenario consistent with the given network. Let a new set of binary variables $B^+_{j,k}$ and $B^-_{j,k}$ denote these artificial perturbations. The state $x_{j,k}$ of node j can be enforced to 1 by adding a positive input, $B^+_{j,k} = 1$. Accordingly, $x_{j,k}$ can be enforced to -1 by adding a negative input, $B^-_{j,k} = 1$. To enforce the state of $x_{j,k}$ to 0, either a positive ($B^+_{j,k} = 1$) or a negative ($B^-_{j,k} = 1$) input might be required. To account for these artificial perturbations, we modify the constraints (4.23)– (4.26) in the following manner:

$$x_{j,k} \leq x^+_{j,k} + B^+_{j,k} \tag{4.37}$$

$$x_{j,k} \geq -x^-_{j,k} - B^-_{j,k} \tag{4.38}$$

$$x_{j,k} \leq 2\, x^+_{j,k} - x^-_{j,k} + 2\, B^+_{j,k} \tag{4.39}$$

$$x_{j,k} \geq -2\, x^-_{j,k} + x^+_{j,k} - 2\, B^-_{j,k} \tag{4.40}$$

Having introduced the correction terms $B^+_{j,k}$ and $B^-_{j,k}$, we set as an extra constraint the perfect fit of the data (which is now always feasible):

$$\sum_{j:m_{j,k}\neq NaN} |m_{j,k} - x_{j,k}| = 0 \tag{4.41}$$

The absolute value is reformulated as described in section "SCEN_FIT".

As we are interested in MCoS with a *minimum* number of corrections, the following objective function is minimized:

$$\sum_j B^+_{j,k} + \sum_j B^-_{j,k} = \text{min!} \tag{4.42}$$

Enumeration of Minimal Correction Sets

In the previous subsection, we formulated an ILP problem for the identification of a single minimum MCoS for a given inconsistent scenario. Since potentially many MCoS might in general exist, we address the enumeration of *all* minimum MCoS. To this end, we solve the ILP repeatedly, and after each run, we exclude previously found solutions by adding the following constraints for each previous solution s:

$$\sum_{j,k} (|B^+_{j,k} - B^+_{j,k,s}| + |B^-_{j,k} - B^-_{j,k,s}|) \geq 1, \tag{4.43}$$

where $B^+_{j,k,s}$ and $B^-_{j,k,s}$ represent the value of $B^+_{j,k}$ and $B^-_{j,k}$ in solution s. We set $|B^+_{j,k} - B^+_{j,k,s}| = dB^+_{j,k,s}$ and $|B^-_{j,k} - B^-_{j,k,s}| = dB^-_{j,k,s}$. Constraint(4.43) is reformulated as follows:

$$\sum_{j,k} dB^+_{j,k,s} + dB^-_{j,k,s} \geq 1 \tag{4.44}$$

$$-B^+_{j,k} + dB^+_{j,k,s} - 2\, dB2^+_{j,k,s} \leq B^+_{j,k,s} \tag{4.45}$$

$$B^+_{j,k} + dB^+_{j,k,s} - 2\, dB1^+_{j,k,s} \leq -B^+_{j,k,s} \tag{4.46}$$

$$-B^-_{j,k} + dB^-_{j,k,s} - 2\, dB2^-_{j,k,s} \leq B^-_{j,k,s} \tag{4.47}$$

$$B^-_{j,k} + dB^-_{j,k,s} - 2\,dB1^-_{j,k,s} \leq -B^-_{j,k,s} \tag{4.48}$$

$$dB1^+_{j,k,s} + dB2^+_{j,k,s} = 1 \tag{4.49}$$

$$dB1^-_{j,k,s} + dB2^-_{j,k,s} = 1, \tag{4.50}$$

where $dB1^+_{j,k,s}$, $dB1^-_{j,k,s}$, $dB2^+_{j,k,s}$ and $dB2^-_{j,k,s}$ are binary variables. We may then compute a new MCoS by optimizing again objective function(4.42). To focus only on MCoS with the minimum number of corrections we replace, after the first iteration, the objective function(4.42) by forcing the algorithm to find a solution with the (same) minimum number of corrections:

$$\sum_{j,k} B^+_{j,k} + \sum_{j,k} B^-_{j,k} = objval. \tag{4.51}$$

Here, *objval* is the value of the objective function found in the first run of the algorithm. The resulting problem becomes thus a simple search for a feasible solution and is repeated until no further solution can be found.

OPT_SUBGRAPH - Computing a single optimal subgraph

As introduced in the Introduction section, OPT_SUBGRAPH searches for an optimal subgraph of the original topology (i.e., for a set of suitable edge removals) minimizing the total fitting error *over all* scenarios. In this subsection we describe how we can identify one particular solution to this problem before turning to the enumeration of optimal subgraphs.

The removal of edges is implemented using binary variables y^+_i and y^-_i. The algorithm will set $y^+_i = 1/\ y^-_i = 1$ if the positive/negative edge i is removed by the optimization procedure to improve the fit of the data (otherwise $y^+_i = 0/\ y^-_i = 0$). We use again the same constraints for sign-consistency as in section "Basic definitions and ILP formulation of sign consistency". The actual pruning is implemented by modifying constraints (4.1) and (4.2) as follows:

$$u^+_{i,k} = \max(0,\, x_{j,k} + \sigma^+_i - y^+_i - 1,\, -x_{j,k} + \sigma^-_i - y^-_i - 1) \tag{4.52}$$

$$u^-_{i,k} = \max(0,\, -x_{j,k} + \sigma^+_i - y^+_i - 1,\, x_{j,k} + \sigma^-_i - y^-_i - 1) \tag{4.53}$$

The max operator is reformulated in a similar fashion to section "Basic definitions and ILP formulation of sign consistency". The updated constraints are listed below:

$$x_{j,k} + \sigma^+_i - y^+_i - 1 = aux1_{i,k} \tag{4.54}$$

$$-x_{j,k} + \sigma^-_i - y^-_i - 1 = aux2_{i,k} \tag{4.55}$$

$$-x_{j,k} + \sigma^+_i - y^+_i - 1 = aux3_{i,k} \tag{4.56}$$

$$x_{j,k} + \sigma^-_i - y^-_i - 1 = aux4_{i,k} \tag{4.57}$$

$$u^+_{i,k} \geq aux1_{i,k} \tag{4.58}$$

$$u^+_{i,k} \geq aux2_{i,k} \tag{4.59}$$

$$u^-_{i,k} \geq aux3_{i,k} \tag{4.60}$$

$$u^-_{i,k} \geq aux4_{i,k} \tag{4.61}$$

$$u^+_{i,k} \leq 1 - d1_{i,k} \tag{4.62}$$

$$u^+_{i,k} \leq aux1_{i,k} + 4 - 4\,d2_{i,k} \tag{4.63}$$

$$u^+_{i,k} \leq aux2_{i,k} + 4 - 4\,d3_{i,k} \tag{4.64}$$

$$u^-_{i,k} \leq 1 - d4_{i,k} \tag{4.65}$$

$$u^-_{i,k} \leq aux3_{i,k} + 4 - 4\,d5_{i,k} \tag{4.66}$$

$$u^-_{i,k} \leq aux4_{i,k} + 4 - 4\,d6_{i,k} \tag{4.67}$$

$$d1_{i,k} + d2_{i,k} + d3_{i,k} = 1 \tag{4.68}$$

$$d4_{i,k} + d5_{i,k} + d6_{i,k} = 1 \tag{4.69}$$

The following constraints were also added, stating that edge i can only be removed if it is present:

$$y^+_i \leq \sigma^+_i \tag{4.70}$$

$$y^-_i \leq \sigma^-_i \tag{4.71}$$

We then reuse objective function(4.28), but now optimize *over all* scenarios:

$$\sum_{j,k:m_{j,k}\neq NaN} a_{j,k}|m_{j,k} - x_{j,k}| = \min! \tag{4.72}$$

This optimization will deliver an optimal sub-network of the original IG which can best explain the data. Usually, many optimal solutions may exist yielding the sameresidual fitting error in eq.(3.100)). One might then be interested to focus on particular solutions, for example, on those containing the minimal/maximal number of edges in the remaining subgraph (see Figure 4.1B and 4.1C). For this purpose, we may replace (3.100) by

$$\sum_{j,k:m_{j,k}\neq NaN} a_{j,k}|m_{j,k} - x_{j,k}| + \sum_i b_i y^+_i + \sum_i b_i y^-_i = \min! \tag{4.73}$$

(the absolute value in the equation is reformulated as in section "Basic definitions and ILP formulation of sign consistency"). The constants $a_{j,k}$ and b_i can be defined by the user, we choose 1 as default value for $a_{j,k}$. Regarding b_i, in order to arrive at a solution with minimal error between predicted and measured values, the absolute value $|b_i|$ needs to be less than or equal to $1/n_E$. Furthermore, constants b_i assume negative values ($-1/n_E \leq b_i < 0$) for obtaining a maximum subgraph and positive values ($0 < b_i \leq 1/n_E$) for obtaining a minimum subgraph.

Another way to deal with non-unique solutions is to enumerate all of them which we address next.

Enumeration of optimal subgraphs

To identify all optimal subgraphs minimizing the inconsistencies between IG topology and measurements of all scenarios, we solve the ILP repeatedly and, after each run, we exclude previous solutions s by adding the following constraints:

$$\sum_i (|y^+_i - y^+_{i,s}| + |y^-_i - y^-_{i,s}|) \geq 1, \tag{4.74}$$

where $y^+_{i,s}$ and $y^-_{i,s}$ represent the value of y^+_i and y^-_i in solution s. We set $|y^+_i - y^+_{i,s}| = dy^+_{i,s}$ and $|y^-_i - y^-_{i,s}| = dy^-_{i,s}$. Constraint(4.74) is reformulated as follows:

$$\sum_i (dy^+_{i,s} + dy^-_{i,s}) \geq 1 \tag{4.75}$$

$$-y_i^+ + dy_{i,s}^+ - 2\,dy2_{i,s}^+ \leq y_{i,s}^+ \tag{4.76}$$

$$y_i^+ + dy_{i,s}^+ - 2\,dy1_{i,s}^+ \leq -y_{i,s}^+ \tag{4.77}$$

$$-y_i^- + dy_{i,s}^- - 2\,dy2_{i,s}^- \leq y_{i,s}^- \tag{4.78}$$

$$y_i^- + dy_{i,s}^- - 2\,dy1_{i,s}^- \leq -y_{i,s}^- \tag{4.79}$$

$$dy1_{i,s}^+ + dy2_{i,s}^+ = 1 \tag{4.80}$$

$$dy1_{i,s}^- + dy2_{i,s}^- = 1, \tag{4.81}$$

where $dy1_{i,s}^+$, $dy1_{i,s}^-$, $dy2_{i,s}^+$ and $dy2_{i,s}^-$ are binary variables. Moreover, after the first run we replace the objective function in (3.100) by enforcing the algorithm to obtain the same, optimal, goodness of fit as in the first run:

$$\sum_{j,k:x_{j,k,m}\neq NaN} a_{j,k}|x_{j,k,m} - x_{j,k}| = objval, \tag{4.82}$$

where *objval* is the value of the objective function(3.100) after the first run of the algorithm. In the same way we may also consider the enumeration of minimum and maximum subgraphs where we then have to fix (4.73) to its optimal value instead of (3.100).

OPT_GRAPH: a greedy algorithm for identifying missing edges

As motivated in the Introduction section, optimizing the IG topology by edge removals may eliminate some but often not all mismatches. One reason could be that some real effects cannot be transduced in the model due to missing edges. We therefore propose an algorithm suggesting de-novo interactions whose addition would minimize the fitting error. As the possibility to insert new interactions increases the solution space dramatically in large networks, we consider the following greedy strategy: for each interaction not contained yet in the IG, we temporarily insert this edge and determine the resulting optimal solution for the fitting error by applying the OPT_SUBGRAPH algorithm introduced above. The single interaction that reduces the fitting error at most is picked by the greedy algorithm and permanently inserted in the IG. This process is repeated until no further edge exists that could improve the goodness of fit to the data significantly (significance can be quantified by a certain threshold). Importantly, at the beginning of each iteration, a list of eligible edges is computed consisting only of those edges that do not form a positive cycle (see below).

Positive cycles and steady-state assumption

(Feedback) cycles often hamper the analysis of causality and many network inference techniques therefore exclude cycles from the network or assume that no cycles exist (see e.g., [13, 114]). In contrast to many other approaches, our method can readily deal with negative cycles without any problems. However, positive cycles may become problematic as they can provide explanations for state changes without any external perturbation. A simple example for such "self-explaining" state changes is the following network: $A \rightarrow B \rightarrow C \rightarrow B$ (all edges are positive). Node A would normally serve as an input. However, assuming that A has not changed, a measured up-regulation of B would be explainable by the sign-consistent labeling (0,1,1), that is, B activates C which then activates B again. Although such a shift without external perturbations could indeed happen in realistic systems (due to fluctuations in bistable systems), we recommend that the initial IG should not contain a positive feedback (otherwise, many observations might become sign-consistent just through the existence of positive cycles). This is also the reason why

a new candidate edge can only be added to the network if it does not give rise to a new positive cycle (see previous subsection). In many applications, this requirement is not a real limitation, in particular when describing early events in signaling networks.

We also restate another assumption for the analysis followed herein, namely that the system moves from one steady state to another upon imposing the perturbations (see also [110]; similar assumptions are also required in other studies, e.g., [13, 46]). However, this does not necessarily mean that we have to wait until the system has reached its new steady state completely; instead, we can take the measurements if we can assume that the *signs* of the state variations will not change anymore. It will therefore be important to determine a suitable time point where all relevant state changes induced by the perturbation have become visible in the measurements. For example, if measurements are taken too early, a signal has possibly not yet been propagated to all downstream nodes at the bottom of the network resulting in inconsistencies with the predictions made from the IG.

Model compression

In the previous sections we presented several ILP formulations related to detecting and resolving inconsistencies between IGs and experimental data. As long as one searches for a single (optimal) solution it is likely that a solution will be found even in very large networks due to an evolved library of efficient ILP algorithms. However, the related enumeration approaches may quickly become intractable, at least if one aims at an exhaustive enumeration. In those cases one may stop the calculation if no new solution is found within a given time interval. Another useful strategy is to use (loss-free) network compression techniques by which (compressed) solutions can be calculated from a smaller network and then subsequently decompressed to solutions of the full network. Other advantages of network compression are that differences between the original and the compressed network structure may indicate non-identifiabilities in the original network and that found optimal solutions can be represented in a condensed manner (not explicitly displaying all combinatorial solutions existing due to non-uniqueness). We use four simple compression rules (illustrated in Figure 4.2) in an iterative manner which, as shown in the EGF scenario below, may reduce the network size dramatically so that enumeration of solutions in large networks become possible (some but not all rules are identical to those used in [13]). Compressing the network is particularly useful for enumerating solutions for OPT_GRAPH and OPT_SUBGRAPH.

Rule 1 (removal of non-controllable and non-observable nodes): *Non-controllable* nodes (which cannot be affected by any of the perturbed nodes in any scenario) and *non-observable* nodes (which do not influence any measured (readout) node in any scenario) define non-identifiable parts of the network and can therefore be removed, including all edges they are connected to. Non-observable and non-controllable nodes can easily be identified by shortest path algorithms (cf. [13]).

Rule 2 (removal of parallel edges): If there are two parallel edges of the same sign, we may safely remove one of them (Figure 4.2A).

Rule 3 (absorbing a node with one input edge): If a latent node (neither measured nor perturbed in any of the experimental scenarios) has only one single incoming edge, then we can remove this node (together with the incoming edge) and reconnect all the outgoing edges of this node to its only predecessor node (under consideration of edge signs; see example in Figure 4.2B).

Rule 4 (absorbing a node with one output edge): If a latent node has only one single outgoing edge, then we can remove this node (together with the outgoing edge) and reconnect all its incoming edges to its only successor node (under consideration of edge signs; see example

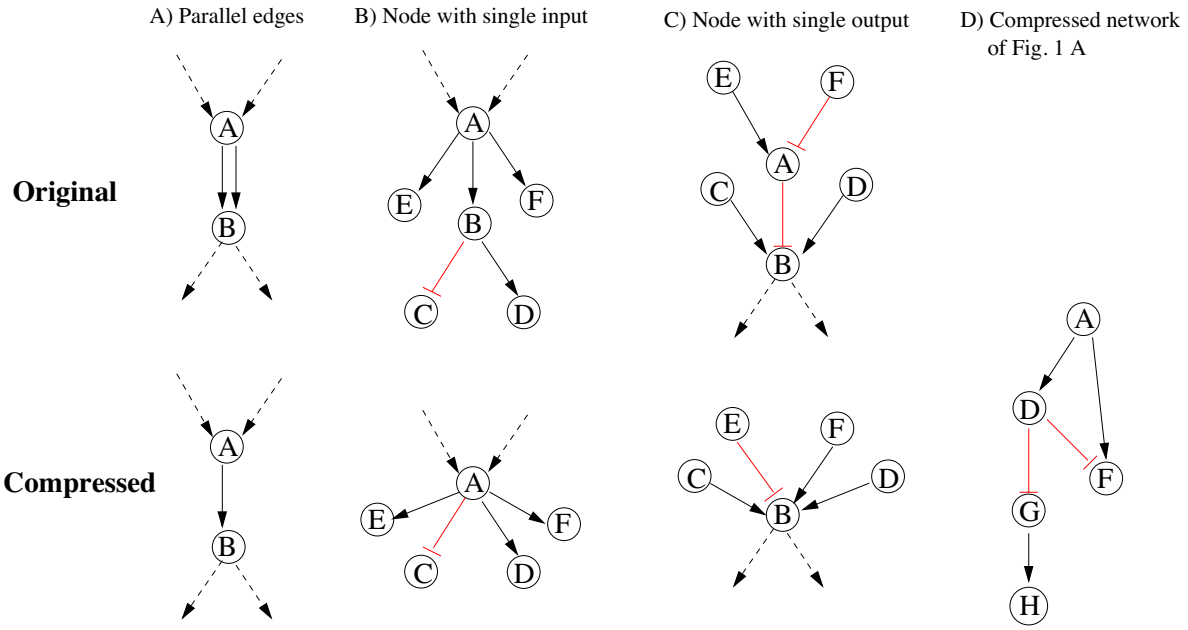


Figure 4.2: **Basic network compression rules.** For further explanations see main text.

in Figure 4.2C).

Rule 1 is performed once at the beginning, whereas rules 2–4 are iteratively used until no further rule can be applied (note that new parallel edges may arise after applying rules 3 or 4). The compressed version of the example network in Figure 4.1A is shown in Figure 4.2D).

By keeping track of the made compression steps it is, in principle, possible to decompress solutions found by the described optimization algorithms in the compressed network. However, as mentioned above, it is often useful to discuss instead the found solutions directly for the compressed network, thereby avoiding the interpretation of a typically much larger number of decompressed solutions arising due to non-uniqueness. For example, instead of listing all possible (parallel) pathway combinations connecting node A with B, one might conclude that "at least one pathway between A and B must exist" which can easier be represented in a compressed network.

Implementation

The ILP formulations presented in the previous sections were implemented in the new software *SigNetTrainer*. This toolbox is written in C and uses routines from GUROBI (`http://www.gurobi.com`) to benefit from state-of-the-art-solvers for ILP problems. *SigNetTrainer* is easy to use; the user has to provide three files to define network training problems: (i)the network topology in .sif format (also used by Cytoscape `http://www.cytoscape.org`), (ii)an ASCII file describing the experimental scenarios (i.e., the imposed state changes), and (iii)an ASCII file containing the experimentally measured state changes for each scenario. The user may then call different functions implementing the optimization routines as described herein. The source code and binaries of *SigNetTrainer* together with a manual are available for testing on the following website: `http://www.mpi-magdeburg.mpg.de/project/cna/etcdownloads.html`. Preprocessing routines, in particular the network compression algorithm, were implemented as MATLAB functions which are also part of the package.

4.3 Results

In order to demonstrate the performance of the proposed approach in a realistic situation, we apply it to a recently published network topology of EGFR/ErbB signaling [32]. This network is based on the logical modeling framework introduced in [119] and describes signal transduction downstream of the members of the EGF receptor family, ErbB1–4. In [32], qualitative predictions derived both from the logical model and its underlying interaction graph were compared with a dataset (a subset of the phosphoproteomic data published in [12]) consisting of combinatorial treatments of primary human hepatocytes with/without TGFα and seven specific molecular inhibitors (including the no-inhibitor treatment; see FigureS1). Note that the measurements were taken at an optimal time point such that the perturbation-induced changes in the phosphorylation level of the proteins are well-reflected by the measurements [12]. The interaction graph-based data analysis in [32] made use of the dependency matrix of the network (see Introduction section): for pairs of experiments (e.g., Exp.1: stimuli A, inhibitor B, Exp.2: stimuli A, no inhibitor) it was checked whether the ratio of the measured responses (e.g., Exp.1/Exp.2, showing the effect of inhibitor B) is consistent with the causal dependencies in the network topology (e.g., if B has a positive/negative/no influence on a readout C, inhibiting B should lead to decreased/increased/unchanged C). Resulting from this analysis, changes in the network structure were proposed that would improve the agreement between experimental data and model predictions. These changes were derived solely by inspection; the ILP approach presented herein can be seen as a step forward as it adapts the model structure to the experimental data in an automatic way and searches systematically for all possible solutions resolving discrepancies between model and data.

Preprocessing of signaling data

Before applying the ILP formulation to the ErbB network and phosphoproteomic data at hand, both the data and signaling topology have to be preprocessed. The phosphoproteomic data, obtained via xMAP technology, is measured in fluorescent units and is dependent on the antibody pair used for detection [12]. For example, JNK ranges from 100 units to 500 units, while MEK1/2 ranges up to 25000 units. Variations such as these do not necessarily reflect that JNK is less activated than MEK1/2, but may be attributed to protein abundance or assay calibration issues. Furthermore, the proposed formulation requires a qualitative view of signal transduction, supporting only three discrete states indicating the variation of the activation state of signaling nodes when changing external inputs or adding inhibitors ("-1", i.e., downregulated, "0", i.e., unchanged, and "1", i.e., upregulated). Thus, the raw data is preprocessed and discretized before being imported in the ILP. To this end, the methodology introduced by Samaga et al. in [32] is adopted: the ratios of all experiments that differ only by a single perturbation (ligand or inhibitor treatment) are evaluated and the respective measurement is considered to be (i)upregulated if the fold-increase of the signal (with versus without perturbation) is above 1.5, (ii)downregulated if the fold decrease of the signal (with versus without perturbation) is below 0.66 and (iii)unchanged otherwise. The dataset analyzed in [32] contains measurements with JNK inhibitor showing an effect of the inhibitor on many of the measured signals. As these inhibitions are likely to be off-target effects [12], we decided to exclude the JNK inhibitor data for our analysis. The complete set of discretized data can be seen in Figure 4.4. The original interaction graph used by Samaga et al.[32] was preprocessed to remove non-observable and non-controllable nodes (see [13] and Rule1 of the model compression described in the Methods section; the full compression will be applied in a later step). The resulting graph is shown in Figure 4.5A.

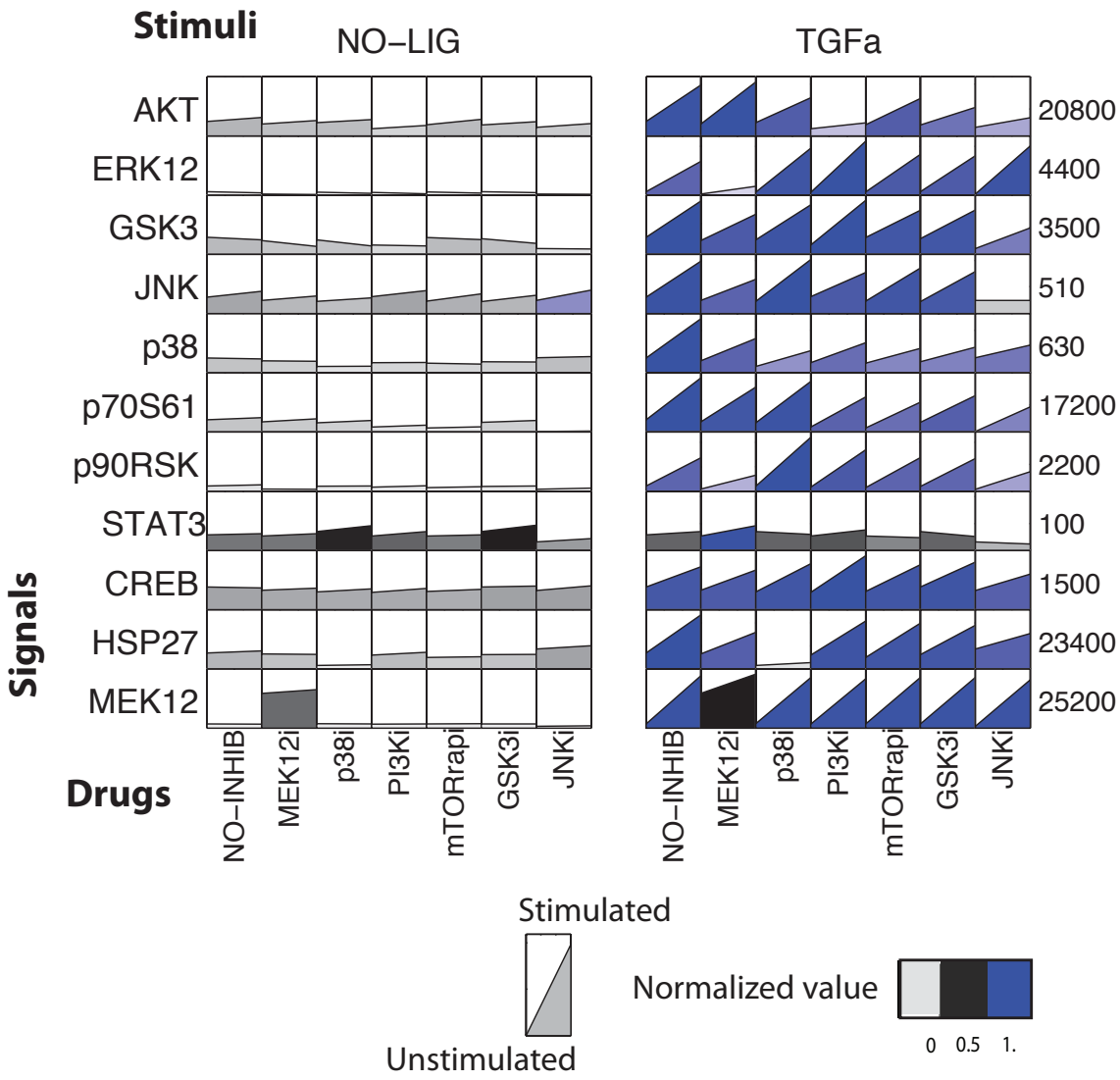


Figure 4.3: Figure S1. Raw training data. A subset of the phosphoprotein data published in [12], capturing the signaling response of primary human hepatocytes to TGFα in combination with six specific molecular inhibitors (including the no-inhibitor treatment) (MEK12-i, p38-i, PI3K-i, mTORrap-i, GSK3-i, no-inhib). Each subplot shows the phosphorylation state of the respective protein in fluorescent units (obtained via xMAP technology), measured 0 minutes (left border) and 25 minutes (right border) after stimulation.

SCEN_FIT and Minimal Correction Sets

In Figure 4.4, the discretized measurements and, for each scenario, the corresponding SCEN_FIT derived from the EGFR network topology given in Figure 4.5 is shown. Recall that the SCEN_FIT algorithm determines for a given scenario a sign-consistent node labeling which is closest to the made measurements. Deviations between this determined sign pattern and the measured state changes (as indicated in Figure 4.4) reflect the inconsistencies between network structure and observed behavior. For example, scenario1 reflects the influence of the ligand TGFα, that is, TGFα is the perturbed node and its state is fixed to 1. As depicted in Figure 4.4, the SCEN_FIT for

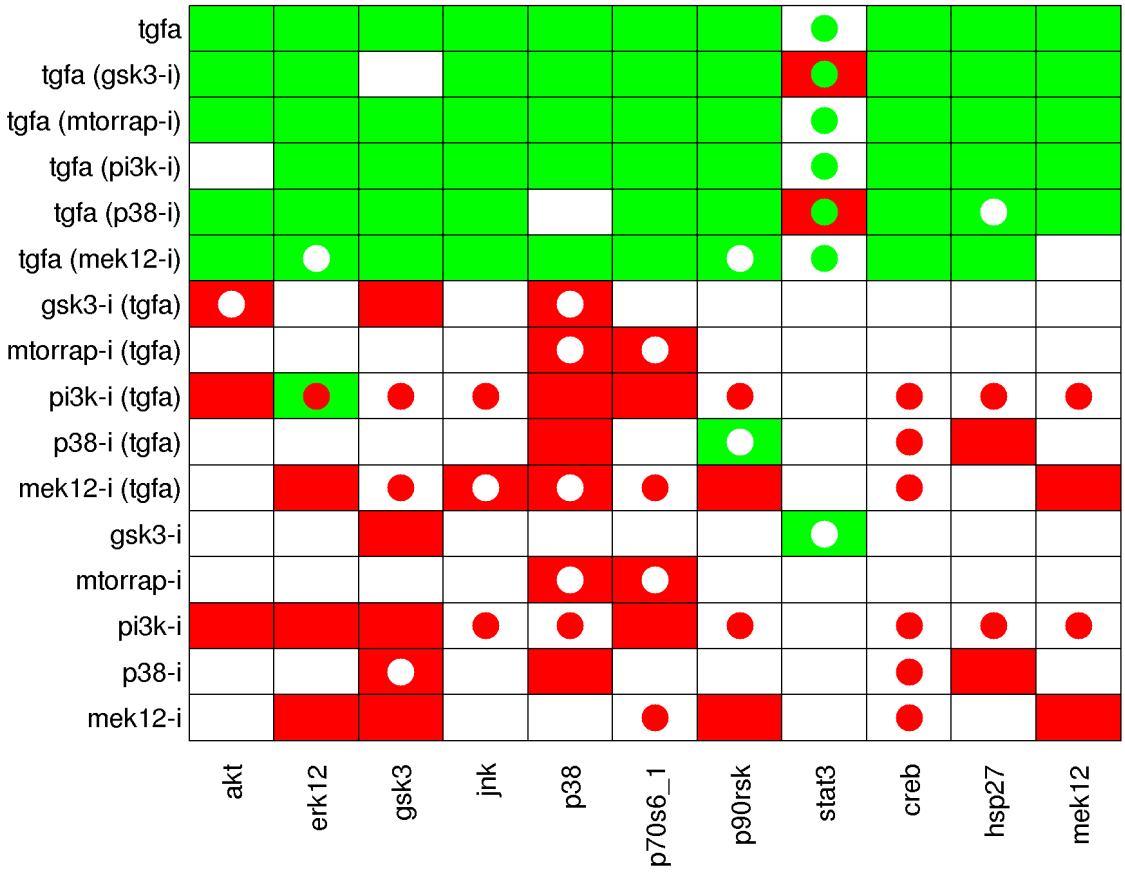


Figure 4.4: **Discretized measurements and SCEN_FIT derived from the EGFR/ErbB network.** Each row corresponds to one experimental scenario, each column contains the measured species. The discretized measurements are mapped to the fill color of the respective fields: if a node is upregulated in the respective scenario, the corresponding field is filled green, if it is downregulated, the field is filled red, and if it shows no significant change, it is filled white. Accordingly, the color of the added circles shows the sign of the node in the node labeling as derived by SCEN_FIT: green circles correspond to sign 1, red circles to sign -1 and white circles to sign 0. Note that as a consequence, circles only appear if the measurement is not in accordance with the respective state in the sign-consistent labeling.

this scenario shows a fitting error of 1: in the optimal sign-consistent node labeling, all measured nodes have sign 1 as they are connected to TGFα by positive paths only; this is in accordance with the measured state of all nodes but STAT3, which shows no significant change in response to TGFα inducing thus a fitting error. Scenarios2–6 reflect the influence of TGFα in presence of different inhibitors. We assume that an inhibitor completely blocks the signal flow through the inhibited species and thus define these scenarios by two perturbed nodes: TGFα fixed to 1 and the respective inhibitor fixed to 0. The remaining scenarios reflect the influence of the inhibitors in presence (scenarios7–11) and absence (scenarios12–16) of TGFα. In each of these scenarios the perturbed node is the respective inhibitor and its state is fixed to -1. Importantly, by using the enumeration algorithm for SCEN_FIT we could also prove that, for each scenario, the found solution for the optimal fit is unique, hence, no other optimal solutions need to be considered.

Figure 4.4 shows that there are a number of inconsistencies between experimental data and

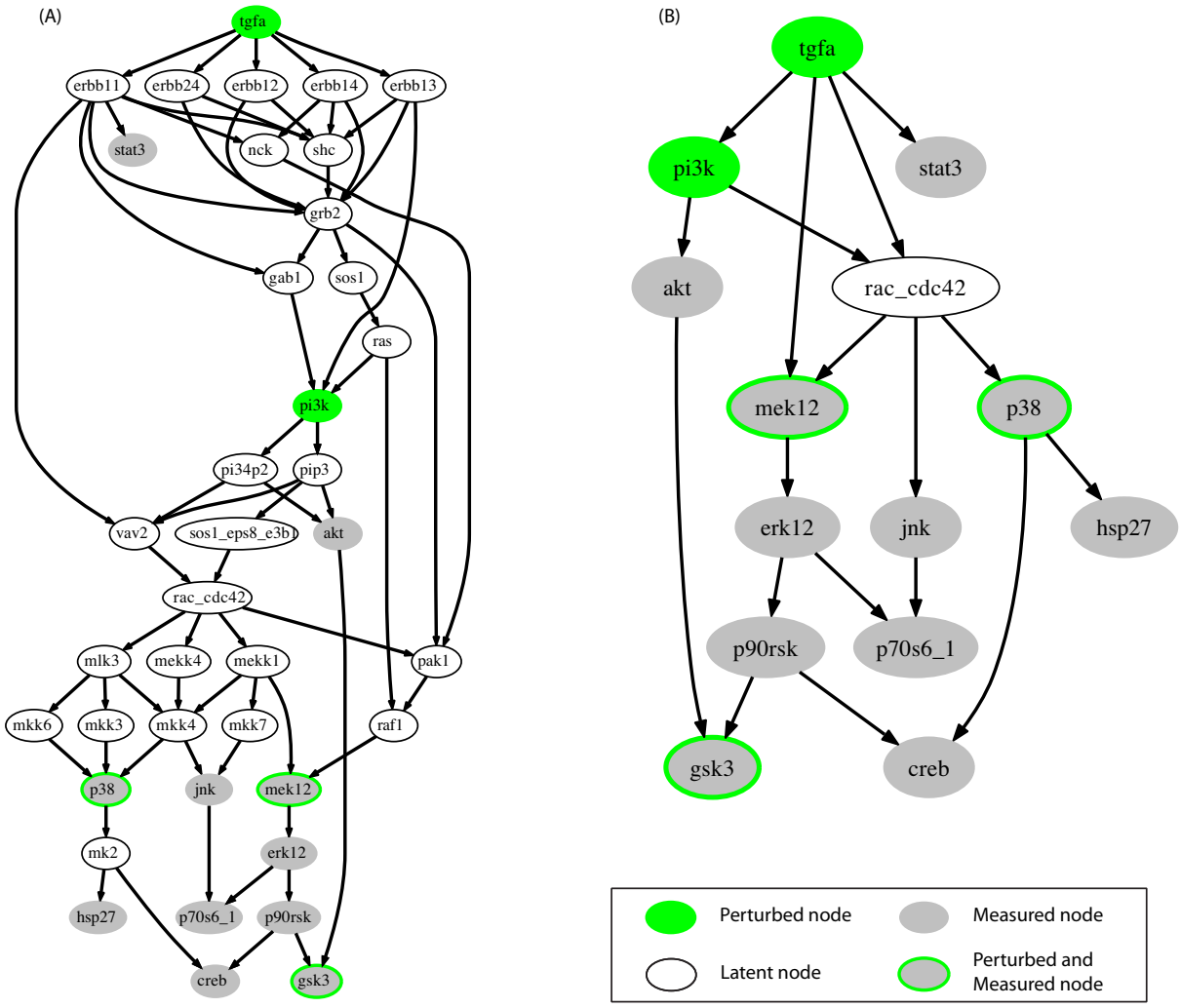


Figure 4.5: **Interaction graph for the EGFR/ErbB model adapted from [32]**. Non-observable and non-controllable nodes have been removed. Measured nodes are depicted in grey, perturbed nodes in green. Nodes that are perturbed and measured are plotted with grey filling and green bound. All edges are activating edges (having positive signs).

the SCEN_FITS derived from the initial network topology. In order to understand possible causes of some of these inconsistencies in single scenarios, we first address the identification of minimal correction sets (MCoS). We recall that MCoS are minimum sets of (artificially) enforced changes of node states (e.g., from up- to downregulated) which will lead to a perfect fit of the data. As described in the Methods section, the identification of all MCoS with minimum size is applied for a single scenario. Exemplarily, we focus herein on scenario14 of Figure 4.4 (PI3K-i is introduced without the presence of TGFα) which produces a total error value of 6.

As shown in Table 4.2, five MCoS are identified, each containing three corrections (virtual perturbations) rendering the experimental scenario14 sign-consistent. Common trend in all MCoS is to remove the downregulating effect of PI3K on signals downstream of Rac_Cdc42 by setting Rac_Cdc42 to unchanged (0) or one of the nodes SOS1_Eps8_E3b1, Vav2, PI(3,4)P2 or PIP3 to upregulated (1). Introducing this change, the states of p38, JNK, MEK1/2, Hsp27, CREB and p90RSK are now in accordance with the measurements (i.e., they show now response

upon adding PI3K inhibitor); however, the states of ERK1/2 and p70S6_1 have changed from the measured "downregulated" to "unchanged". This is corrected in all MCoS by setting ERK1/2 to -1. Again, this correction has an undesired effect, namely changing p90RSK from 0 to -1, which is countered by assigning p90RSK the value 0 in all MCoS. Clearly, three required corrections indicate that the observed behavior for this scenario is not well-reflected by the network topology. It would therefore be useful to consider all scenarios at the same time to detect common points of errors produced in all or many scenarios.

Table 4.2: **MCoS for scenario 14 in Figure 4.4.**

	MCoS 1			**MCoS 2**			**MCoS 3**			**MCoS 4**			**MCoS 5**		
Node id	B_i^+	B_i^-	**Val**	B_i^+	B_i^-	**Val**	B_i^+	B_i^-	**Val**	B_i^+	B_i^-	**Val**	B_i^+	B_i^-	**Val**
rac_cdc42	1		**0**												
p90rsk	1		**0**	1		**0**	1		**0**	1		**0**	1		**0**
erk12		1	**-1**		1	**-1**		1	**-1**		1	**-1**		1	**-1**
sos1_eps8_e3b1				1		**1**									
vav2							1		**1**						
pi34p2										1		**1**			
pip3													1		**1**

Five MCoS are identified for the EGFR network model (Figure 4.5) with respect to scenario 14 in Figure 4.4. Each MCoS would lead to a perfect fit for this scenario and all five MCoS contain three nodes to be enforced to a certain value. Nodes p90rsk and erk12 are common in all MCoS. Nodes rac_cdc42, sos1_eps8_e3b1, vav2, pi34p2 and pip3 are perturbed respectively in MCoS 1–5. In columns MCoS 1–5, three sub-columns are shown: sub-column "Val" shows the corrected state of the node (the actual MCoS), the entry 1 in sub-column "B_i^+" indicates that a positive input edge is added to the node in order to alter its state, and the entry '1' in sub-column "B_i^-" indicates that a negative input edge is added to the node (see Methods section).

OPT_SUBGRAPH

As a first step we use the OPT_SUBGRAPH algorithm to find—by appropriate edge removals—an optimal subgraph of the EGFR network structure which minimizes the fitting errors over all experimental scenarios.

To be able to make valid conclusions, we need to find all optimal solutions. However, enumerating all solutions for OPT_SUBGRAPH in the full model structure becomes quickly intractable as the highly branched network structure (e.g., various feedforward routes running over different combinations of erbb dimers and adapter proteins connect TGFα with PI3K) leads to an immense number of different optimal solutions. Therefore, we compress the model structure as described in section "Model compression" before searching for optimal subgraphs. As can be seen in Figure 4.5B, the model structure can be compressed substantially from 39 nodes and 67 edges to 14 nodes and 18 edges. Strikingly, Rac_Cdc42 remains as the only latent node in the compressed structure. The compressed IG shows the essential dependencies in the original network structure that can be addressed by the given set of perturbed/measured nodes. For example, parallel signaling paths leading from a perturbed node to a measured node without passing any other measured/perturbed node cannot be distinguished in the analysis performed herein and are therefore condensed to one single edge in the compressed graph.

The computation of all optimal subgraphs of the compressed network resulted in six solutions

having the same minimal fitting error of 26 which has thus reduced much in comparison to 45 in the original model. Figure 4.6 shows a combined view of the six optimal solutions. In more detail, a positive influence of TGFα on STAT3 is not reflected in the measurements (see Figure 4.4); consequently, the edge TGFα $\rightarrow$ STAT3 is removed in all optimal solutions. Another edge that is removed in all solutions is PI3K $\rightarrow$ Rac_Cdc42, as a number of signals downstream of Rac_Cdc42 did not show the expected downregulated response to the PI3K inhibitor in the measurements. Finally, by removing the edge ERK1/2 $\rightarrow$ p70S6_1 in all solutions, the missing influence of MEK inhibitor on p70S6_1 is accommodated. The edges TGFα $\rightarrow$ MEK1/2 and Rac_Cdc42 $\rightarrow$ MEK1/2 are only removed in some of the solutions. This is an example for two parallel routes that cannot be distinguished: the model structures containing both routes or either route give rise to the same sign-consistent labeling. In contrast, removing either of the edges p90RSK $\rightarrow$ CREB and p38 $\rightarrow$ CREB results in different sign-consistent labelings, both showing the same number of discrepancies to the measurements: the phosphorylation state of CREB is neither affected by MEK inhibitor nor by p38 inhibitor. However, removing both edges at the same time would interrupt all routes from TGFα to CREB what is contradictory to the observed positive effect of TGFα in scenarios1–6. Thus, in this case, allowing only the removal of edges is not sufficient to fully explain the observed measurements. This can be seen in Figure 4.7, where the two possible optimal sign-consistent labelings that SCEN_FIT would find for the six pruned model structures are shown in comparison to the discretized measurements: in each solution, there are three different remaining errors in the CREB column. The errors for STAT3 as well as the errors in response to PI3K inhibitor (scenarios9 and 14) could be significantly reduced by removing the respective edges.

OPT_GRAPH

Next, we use the OPT_GRAPH procedure to identify edges that may be missing from the EGFR network and whose addition would therefore improve the goodness of fit to the data. Table 4.3 displays the edges that lead to the highest improvement as determined by OPT_GRAPH. All these edges have in common that they give rise to an additional route from TGFα to CREB not running over p38 or MEK1/2. By adding any of these edges to the model structure before doing the OPT_GRAPH, we can further reduce the fitting error to 23 (compared to 26 if only edge removals are allowed).

As an example, we show the optimized model structures when adding the edge TGFα $\rightarrow$ CREB. A combined view of the three optimal solutions (that can be found by OPT_GRAPH after adding this edge) is shown in Figure 4.8. As it was the case for the optimization in the original network, the edges TGFα $\rightarrow$ STAT3, PI3K $\rightarrow$ Rac_CDC42 and ERK1/2 $\rightarrow$ p70S6_1 are removed in all solutions, while the edges TGFα $\rightarrow$ MEK1/2 and Rac_Cdc42 $\rightarrow$ MEK1/2 are each removed in one of the solutions. With the added edge TGFα $\rightarrow$ CREB the model structure comprises an activation route from TGFα to CREB that is independent of p38 and p90RSK, and removing both the p90RSK $\rightarrow$ CREB and p38 $\rightarrow$ CREB edge in all solutions is now optimal.

All three solutions induce the same optimal sign-consistent node labeling. Figure 4.9 shows the mismatches to the experimental data in comparison to the mismatches derived from the initial model structure. The measurements for CREB are now in full accordance with the model structure and the errors for STAT3 could be significantly reduced. Furthermore, a number of errors in scenarios9 and 14 showing the influence of PI3K inhibitor could be eliminated, although at the same time new mismatches for some nodes have been introduced. Finally, the influence of MEK inhibitor on p70S6_1 is now predicted correctly. Here, we just considered the addition of a single edge to improve the fit to data. In principle, one could remove all remaining discrepancies

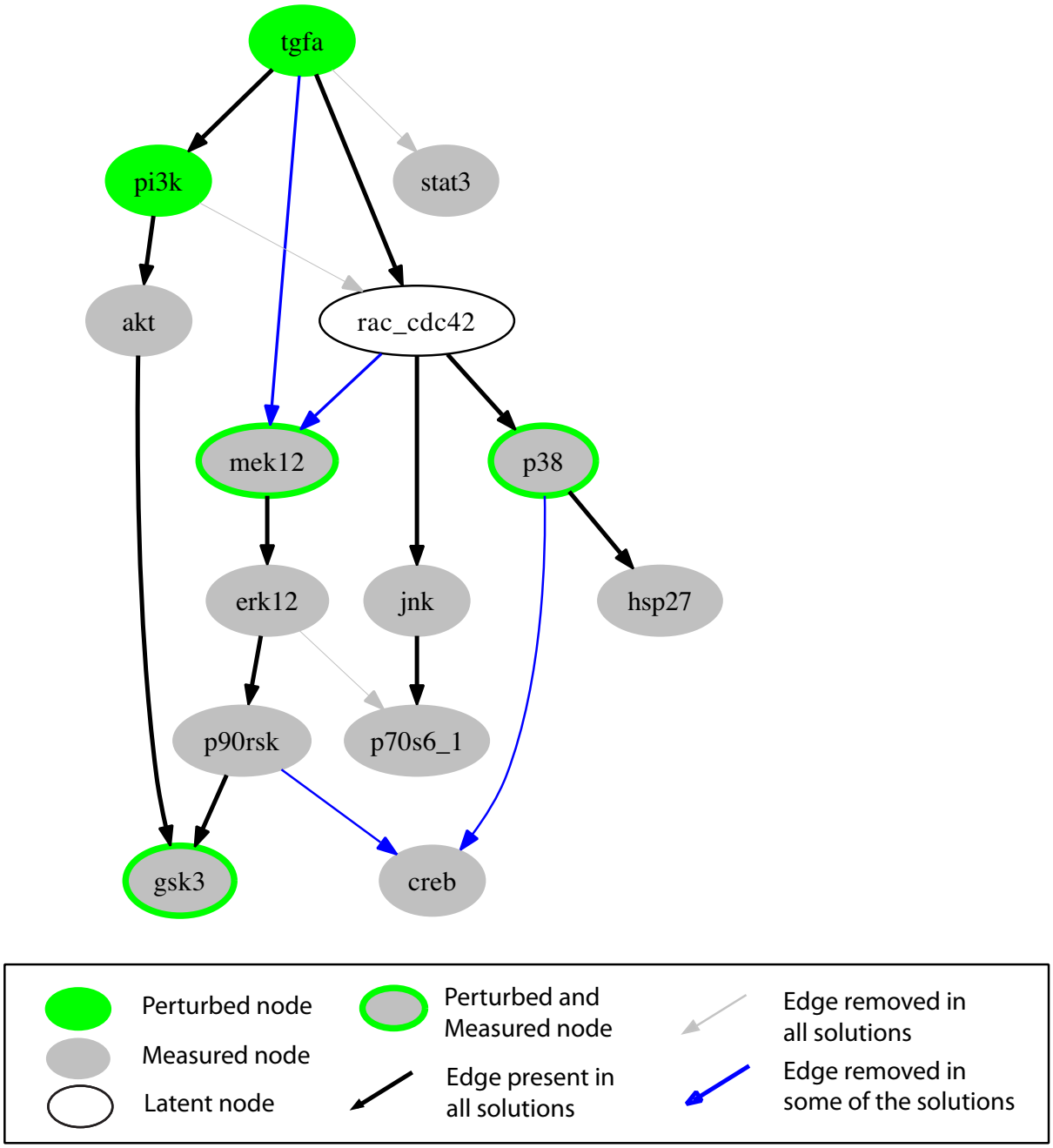


Figure 4.6: **Combined view of all optimal model structures derived from the compressed EGFR/ErbB model by applying the OPT_SUBGRAPH procedure with enumeration.** Measured species are depicted in grey, perturbed species in green, and species that are measured and perturbed are depicted in grey filling with green bound. Black edges are contained in all solutions, grey edges are removed in all solutions and blue edges are removed in some, but not in all solutions.

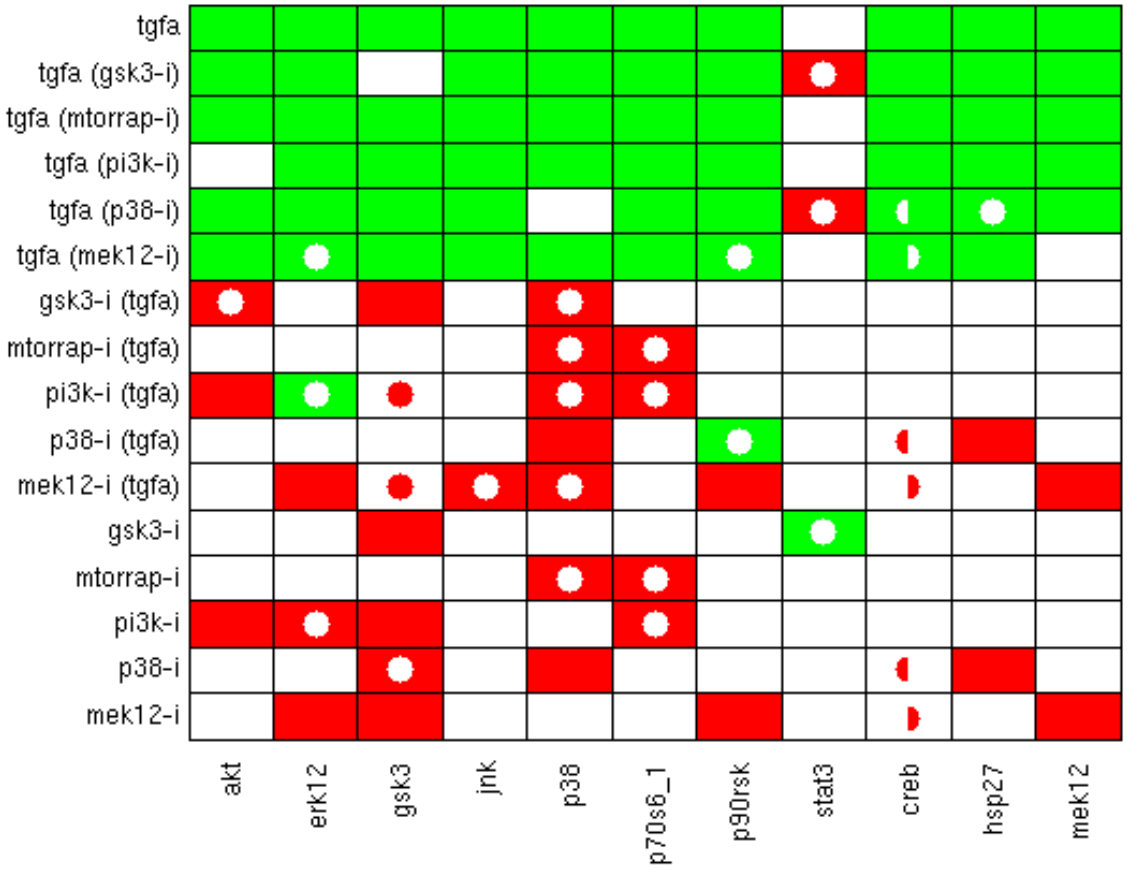


Figure 4.7: **Discretized data and the (two) SCEN_FIT solutions derived from the optimized model structures given in Figure 4.6**. The color coding is the same as in Figure 4.4. All six optimal solutions give rise to the same SCEN_FIT, except for the CREB column. Here, three solutions show a mismatch in scenarios 5, 10, and 15 (indicated by the left semicycles), while the other three solutions show a mismatch in scenarios 6, 11, and 16 (indicated by the right semicycles).

by adding further edges. However, in particular if the measurements show inconsistencies (e.g., the different effect of PI3K inhibitor on ERK1/2 with/without TGFα), some errors can only be removed by introducing a positive and a negative edge between a pair of nodes. Furthermore, adding an edge might only lead to a minor improvement of the fitting error such that it unlikely represents a real effect. In any case, additional experiments should be conducted to confirm suggested interactions.

To summarize, essential findings of the network structure optimization in the EGFR/ErbB network—which may indicate important specifics of this signaling pathway in hepatocytes—are: (1)STAT3 is not activated by TGFα; (2)Phosphorylation of the autocatalytic domain of p70S6 (termed p70S6_1 in the model) is independent of ERK1/2; (3)The activation of CREB in response to TGFα is likely to be caused by a p38 and MEK1/2 independent species; and (4)The activation of Rac/Cdc42 is independent of PI3K. These results, generated in an automated way, confirm several of the conjectures formulated in [32] that were derived by inspection only. In addition, by identifying parallel activation routes that cannot be distinguished with the experimental data at hand, the presented approach contributes to a better understanding of

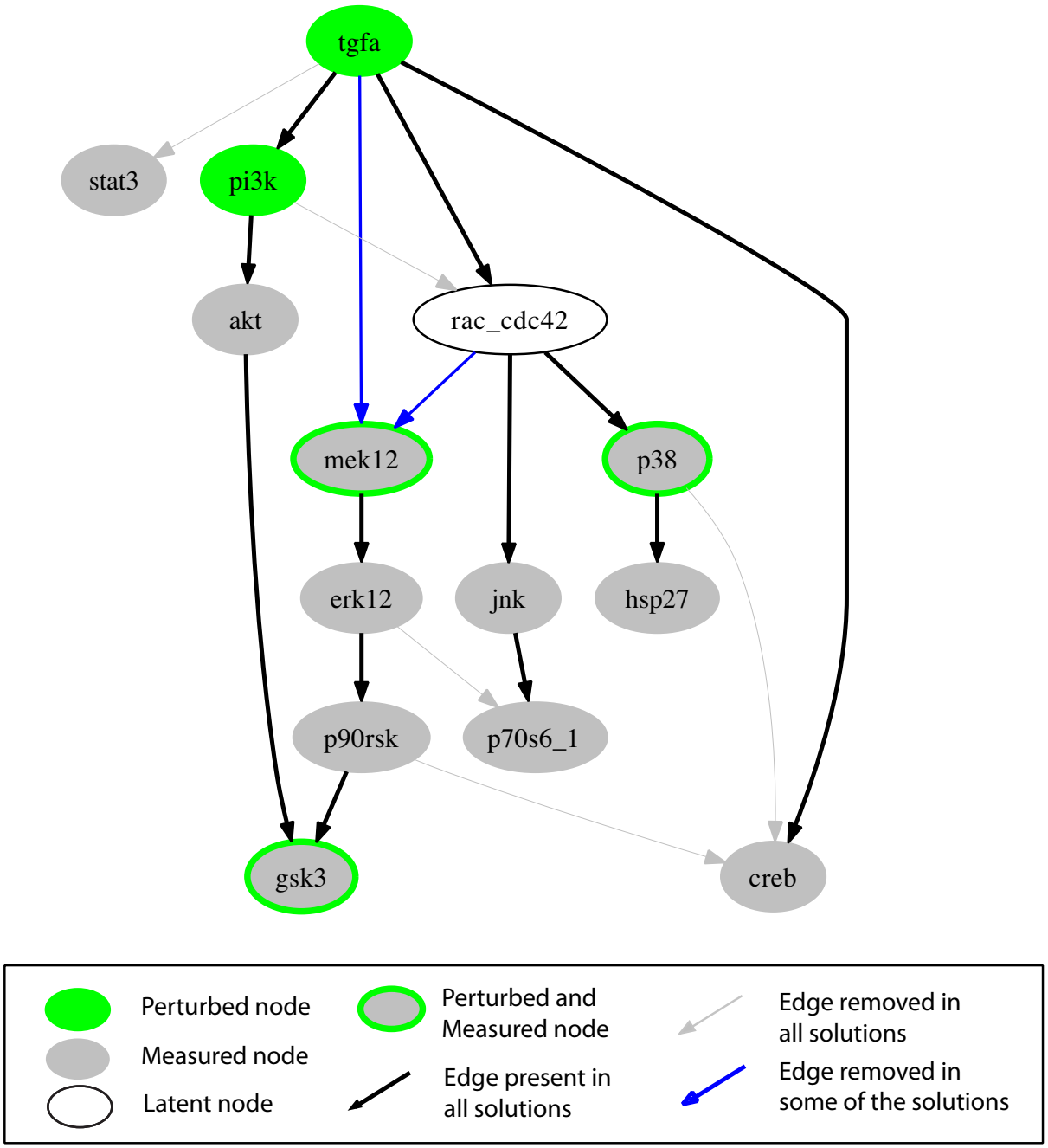


Figure 4.8: **Combined view of all optimal subgraphs resulting when adding TGFα to CREB to the initial model structure.** The color code is the same as in Figure 4.4. In all three solutions, the edges erk12→p70s6_1, tgfa→stat3, p90rsk→creb and p38→creb are removed. The edges tgfa→mek12 and rac_cdc42→mek12 are removed in one of the solutions.

the network topology and helps to suggest further experiments for uncovering the true wiring diagram of this important signaling pathway in the given cell type.

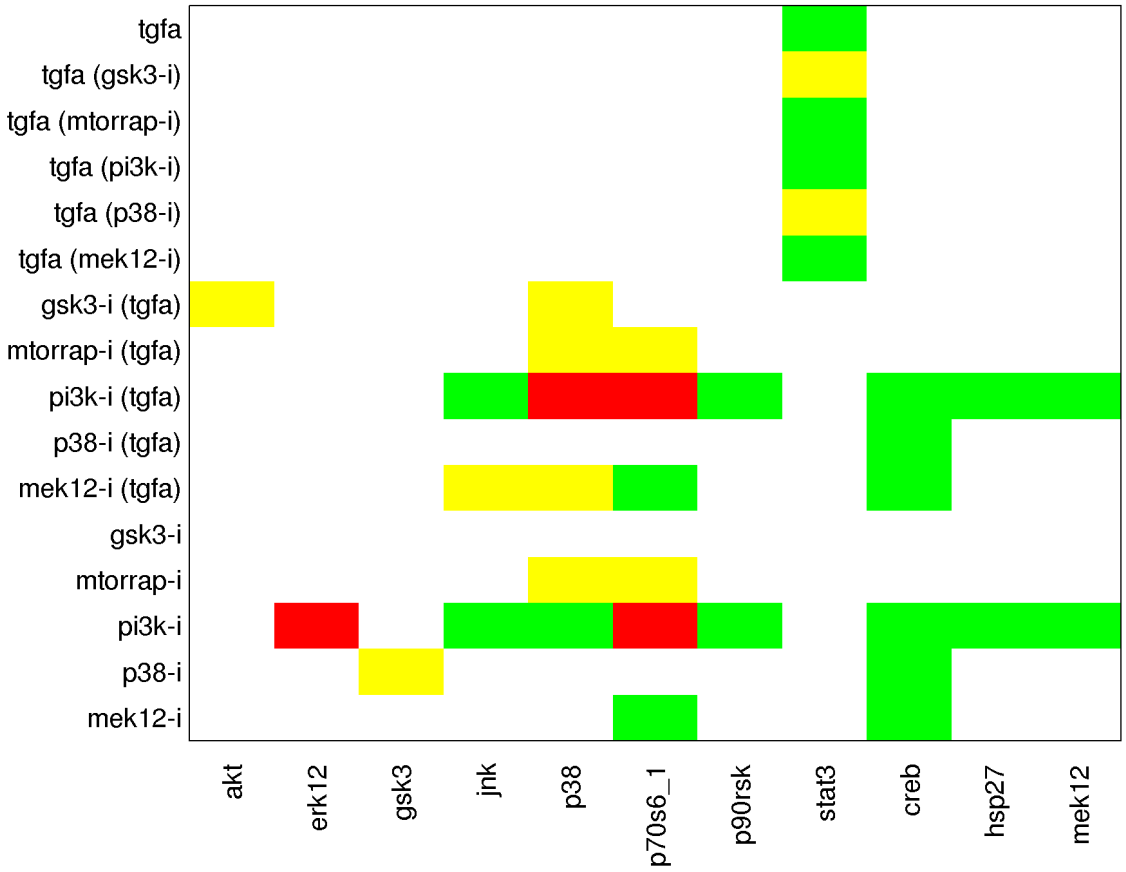


Figure 4.9: **Comparison of the fitting errors of the initial model structure (Figure 4.4) and of the optimal interaction graph shown in Figure 4.8.** Green fields indicate an error that has been present in the original model structure, but could be removed by optimizing the model structure. Yellow fields refer to errors that could not be resolved, and red fields refer to errors that have not been present in the original model structure, but were introduced by the optimization.

Table 4.3: **Suggestions for new edges as computed by OPT_GRAPH.**

tgfa → creb
jnk → creb
p70s61 → creb
rac_cdc42 → creb
tgfa → erk12
jnk → erk12
rac_cdc42 → erk12

Adding any of these edges to the model structure leads to a decrease of the fitting error from 26 to 23.

4.4 Discussion

We presented a new framework for interrogating and training signaling networks based on measurements from stimulus-response experiments. The proposed approach represents signaling networks as interaction graphs and can thus immediately be applied to network topologies stored in many databases without the need to convert these graphs into other modeling formalisms . Interaction graphs capture merely the positive and negative edges between the components in the network; however, this information already sets constraints on the possible qualitative behaviors of the nodes when stimulating or perturbing the network. Our approach uses Integer Linear Programming to encode these constraints and to predict the possible changes (down, neutral, up) of the activation levels of the involved players for a given experiment. Based on this ILP formulation we presented four basic optimization routines useful to detect and remove inconsistencies between measurements and predicted behaviors:

(1) SCEN_FIT: Determination of a causal explanation for the measured activation changes of readout nodes under a given perturbation scenario. If the measurements are inconsistent with the network topology, identification of the closest feasible scenario.
(2) Minimal Correction Sets: In case of an inconsistent scenario, determination of a minimal set of nodes whose state needs to be corrected to make a single inconsistent scenario consistent.
(3) OPT_SUBGRAPH: Determination of an optimal subgraph of a given network topology that can reflect the measurements for a set of scenarios at best.
(4) OPT_GRAPH: Identification of edge candidate(s) whose insertion would improve the consistency of the graph with respect to a set of experimental scenarios at most.

The first two optimization problems seek to match the network topology with measurements from a *single* stimulus-response experiment. In contrast, (3) and (4) operate on a *set* of scenarios and seek to optimize (train) the network structure over all scenarios, either by removing or by adding edges. For the first three problems we also provided enumeration algorithms to find multiple or all solutions that solve the optimization problem equally well (e.g., for problem (3), all optimal subgraphs that minimize the number of inconsistencies between measurements and predictions). The enumeration of all solutions is necessary to be able to draw general conclusions, for example, that a certain edge is removed in all (not only in some) optimal solutions. However, the enumeration of optimal solutions may quickly become prohibitive in larger networks. We therefore employ effective and partially new compression techniques to deal with the combinatorial complexity arising in large-scale networks. In fact, this even allowed us to also address the enumeration of multiple optimal solutions in the EGFR/ErbB case study.

In contrast to the globally optimal solutions that will be delivered for problems (1)–(3), the identification of (a set of) missing edges reducing the fitting error at most (problem (4)) is based on a greedy algorithm which may deliver local instead of globally optimal solutions. However, given the huge search space of potentially missing (sets of) edges, the employed greedy algorithm appears to be a suitable and useful heuristics to suggest missing interactions in the IG model. If only one candidate edge is to be added (instead of a set), it even delivers the globally optimal solution, also in large networks.

To the best of our knowledge, our presented approach is the first that uses Integer Linear Programming directly on *interaction graphs* to systematically interrogate and train the wiring diagrams of signaling networks. Our framework shares some similarities with the approach of Saez-Rodriguez et al.[13] for which recently also an ILP formulation was conceived[46]. This method also starts with an IG representing the prior knowledge; however, the IG is then translated to a superstructure of Boolean networks within which the optimal (sub)model fitting the data at best is identified. Although a correctly reconstructed Boolean network can potentially provide a more specific view on the network structure than an IG, the search space is consider-

ably larger since usually a vast number of possible Boolean networks can be constructed from a given IG. This may lead to highly underdetermined problems and enumeration strategies as discussed herein can become intractable. Furthermore, Boolean networks require a strict binarization of the nodes' states whereas in the IG formulation we consider "influences". This may lead to different results. For example, the Boolean function for a node Z may read $Z = A$ OR B. Assume that we consider the influence of (external) activation of node B with respect to a given "standard (resting) state" where node A is active (1) and B inactive (0); hence, where Z is activated by A. The Boolean model will tell us that Z remains in state 1, hence, the influence of B seems to be not relevant. However, Boolean functions are discrete approximations of the true mechanisms and what one could probably see in the measurements is that the level of Z goes from "high" to "very high". In the IG, we can still account for this effect stating that an elevated level of B induces a positive effect on Z. So discretized node states need not to be considered in the IG model; however, similar as for the Boolean model, some kind of discretization of the data will be required as well when classifying a change of an activation level to be significant or not.

The approach that is arguably closest to ours is the method introduced in [110, 111, 112]. This framework is also based on IG and uses a similar consistency rule as we did herein. However, there are a number of key differences. First, we explicitly allow a "0" change to mark non-affected states of nodes. This extension seems to be essential, for example, when perturbation of a node A cannot affect another node B simply because (in the true topology) a path from A to B does not exist. Second, the four basic problem formulations presented herein go beyond the techniques introduced in [110, 111, 112]. In particular, the training of the topology, that is, the identification of inactive or missing interactions based on a library of stimulus-response experiments, was not considered in these works. A third key difference is that we formulated the constraints resulting from the consistency rules as an ILP problem, whereas [112] uses Answer Set Programming (ASP). Both ILP and ASP deliver globally optimal solutions and highly optimized solvers exist. Using ILP or ASP solvers is not straightforward for non-experts and with *SigNetTrainer* we provide an easy-to-use toolbox. However, it would be an interesting aspect for future work to compare ASP and ILP formulations of the training and enumeration problems formulated herein.

We demonstrated the power of our proposed approach by interrogating and (re-)training a manually curated IG model of EGFR/ErbB signaling against a library of high-throughput phosphoproteomic data measured in primary human hepatocytes. Our algorithms could systematically uncover all inconsistencies between measurements and network topology and gave possible explanations for them. Novel biological insights could be revealed by listing interactions that are likely to be inactive in hepatocytes and by giving suggestions for possibly missing interactions that, if included, would significantly improve the goodness of fit. Clearly, these predictions await experimental validation.

This study gave a proof of principle for our framework, showing its flexibility and that it can be applied to a wide range of problems arising when confronting signaling network topologies with experimental datasets. Given that only fairly accessible biological knowledge is required and that all related algorithms were implemented in a freely available toolbox make it an appealing approach for various applications.

Chapter 5

Identification of signaling pathways related to drug efficacy in HCC via integration of phosphoproteomic, genomic and clinical data

In this work submitted for publication in February 2013, a novel approach is proposed for the identification of signaling pathways related to drug efficacy in HCC via integration of phosphoproteomic, genomic and clinical data. This work was carried out in collaboration with Douglas A. Lauffenburger (head of the Biological Engineering department of MIT, Cambridge, MA, USA). The proposed approach uncovers signaling motifs that can be predictive of drug efficacy in HCC and constitute potential drug targets.

Abstract

Hepatocellular Carcinoma (HCC) is one of the leading causes of death worldwide, with only a handful of treatments effective in unresectable HCC. This is amongst the first studies where the mode of action of some of the compounds that are extensively used in drug trials is interrogated on the phosphoproteomic level, in an attempt to identify signaling pathways related to clinical drug efficacy. Signaling data is combined with previously published gene expression and clinical data within a consistent, mechanistic framework that identifies drug effects on the phosphoproteomic level and translates them to the gene expression level where they are correlated with genes differentially expressed in normal versus tumor tissue, and genes predictive of patient survival. Despite the limited number of clinical trials results, our approach uncovers signaling motifs that can be predictive of drug efficacy in HCC.

5.1 Introduction on the prediction of drug efficacy for Hepatocellular carcinoma

HCC is one of the leading causes of death worldwide [15]. Traditionally, the etiology of the disease is attributed to genetic alterations that accumulate during chronic inflammation of the liver. Mutations are found in several important genes including p73, p53, Rb, APC, DLC-1 (deleted in liver cancer), p16 , PTEN, IGF-2, BRCA2, SOCS-1, Smad2 and Smad4, β-catenin, c-myc, and cyclin D1 [16, 17, 14]. Moreover, as in other cancers, HCC is characterized by an imbalance in growth promoting signals and the MAPK cascade [14]. Approved treatments so far

for unresectable HCC include sorafenib and erlotinib [120, 60] that target the VEGFR, PDGFR and RAF kinase, and the EGFR respectively. However, with the average survival benefit of these treatments at about 3 months, it is evident that identification of new targets for HCC is of the utmost importance.

As high throughput technologies evolve, a paradigm shift occurs in drug discovery. Fields like systems biology attempt to take advantage of the vast datasets generated by the new -omic technologies to identify suitable genes/proteins whose biological activity can be directly linked to pathological processes. To this end an increasing number of studies tackle the complete characterization of tumors' gene expression profiles and protein content [121, 122, 123, 124, 125, 126, 127, 128]. These approaches have succeeded in identifying several hundreds if not thousands of genes and proteins that are differentially expressed in tumor vs normal tissue on the same patient, or genes that are differentially expressed across different patients and are predictive of cancer metastasis, or patient survival. However, applying this knowledge in drug discovery and the identification of drug targets is not a straightforward procedure. Data must also be incorporated that capture the way cells function and respond to factors of its microenvironment (i.e. signaling data). Signaling data can provide the causality / directionality much needed in gene expression networks and uncover the genes that truly regulate/govern the disease phenotype [129].

The importance of intracellular signaling in HCC has been well established and interrogated [18], while a number of new drugs target certain kinases or receptors that are differentially expressed in disease. However, with most of these drugs (especially the approved ones) being highly promiscuous and their effects on the cell's signaling pathways not yet studied in a systematic manner [23], we have yet to discover the key features that are predictive of drugs efficacy, reflecting also the fact that there are key aspects of this disease that elude us. Recently, light was shed into the mode of action of a number of the most widely used compounds with the construction of the *"Genomics of Drug Sensitivity in Cancer"* database (`http://www.cancerrxgene.org/`) where the effects of several hundreds compounds on some of the major cancer genes were interrogated. Such a large scale compound screening has yet to be undertaken on the phosphoproteomic level, where data is much sparser. Nevertheless, studying the effects of these drugs on the phosphoproteomic level where they act is critical for identifying the key features that govern drug efficacy and also for gaining a deeper understanding of the disease itself.

Herein, we propose a consistent framework for the integration of signaling, gene expression and clinical data, aiming at the identification of signaling pathways related to drug efficacy in HCC. We have put together a signaling dataset consisting of the phosphoproteomic response of 3 HCC cell lines, presence of 8 drugs for unresectable HCC, most of which of known clinical efficacy, and attempted using recursive feature extraction to identify the phosphoproteomic signatures that are most predictive of drug efficacy. The following drugs were interrogated: Lapatinib, Gefitinib, Sorafenib, Erlotinib, Vandetanib, Sunitinib, Dasatinib and Bortezomib. Of these Lapatinib, Gefitinib, Vandetanib, Sunitinib failed in clinical trials [59, 130, 131, 132], Sorafenib and Erlotinib passed [120, 60, 133], while Dasatinib and Bortezomib are still under investigation. In more detail,

- Lapatinib is a dual inhibitor of epidermal growth factor receptor (EGFR) tyrosine kinase 1 and 2 (Her2/Neu). Even though Lapatinib in combination to trastuzumab is approved for HER2-positive early breast cancer [134] and well-tolerated in HCC patients, therapy with Lapatinib did not meet the predefined efficacy rate for HCC.

- Gefitinib, also an EGFR inhibitor, successfully prevented hepatocellular carcinoma development in rat liver with cirrhosis [61], but as a single agent is not active in advanced HCC.

- Sorafenib is a small molecular inhibitor of several Tyrosine protein kinases (VEGFR and PDGFR) and Raf kinases (C-Raf than B-Raf), it has been approved for the treatment of primary kidney cancer [62] and advanced primary liver cancer.
- Erlotinib, another EGFR kinase inhibitor, apart from HCC is also approved for the treatment of lung cancer and pancreatic cancer (in combination to Gemcitabine) [135, 136].
- Vandetanib is an antagonist of VEGFR and EGFR, it is approved for the treatment of inoperable medullary thyroid cancer [137], while failed in phase III for non-small-cell lung cancer [138] and HCC because of limited clinical activity.
- Sunitinib, is a multi-targeted receptor kinase inhibitor, blocking PDGFR and VEGFR, approved for treatment of renal cell carcinoma [139], and advanced gastrointestinal stromal tumor [140], however, failed in phase III clinical study for HCC.
- Dasatinib, a multi- BCR/ABL and Src family kinase inhibitor, is approved for treatment of chronic myelogenous leukemia [141]. Its effects on HCC are still under investigation, the study presented in [142] showed that a subgroup of liver cancers may be more likely to benefit from treatment with dasatinib than others.
- Bortezomib is a proteasome inhibitor approved for multiple myeloma [143], also, arrested the proliferation of hepatocellular carcinoma cells HepG2 and JHH6 by differentially affecting E2F1, p21 and p27 levels, making it attractive for future experimentation in animal models of HCC [133]

By using signaling data as the basis for our analysis, we studied the effects of the above-mentioned 8 drugs directly on the phosphoproteomic level where they act. We subsequently translated our findings to the gene expression level, where we inferred regulatory networks between the identified phosphoprotein features and gene sets known to be correlated/implicated to HCC (either differentially expressed between tumor and normal tissue on the same patient, or differentially expressed across different patients and predictive of metastasis, or survival) (see Figure 5.1). This offered both a validation of our findings as well as lead to the identification of a subset of differentially expressed genes that could possibly govern/regulate patient survival and/or drug efficacy. The analysis presented herein could serve both for the identification of drug targets, as well as a new framework for the integration of signaling, gene expression and clinical data, aiming towards the holistic study of mechanisms implicated in drug efficacy.

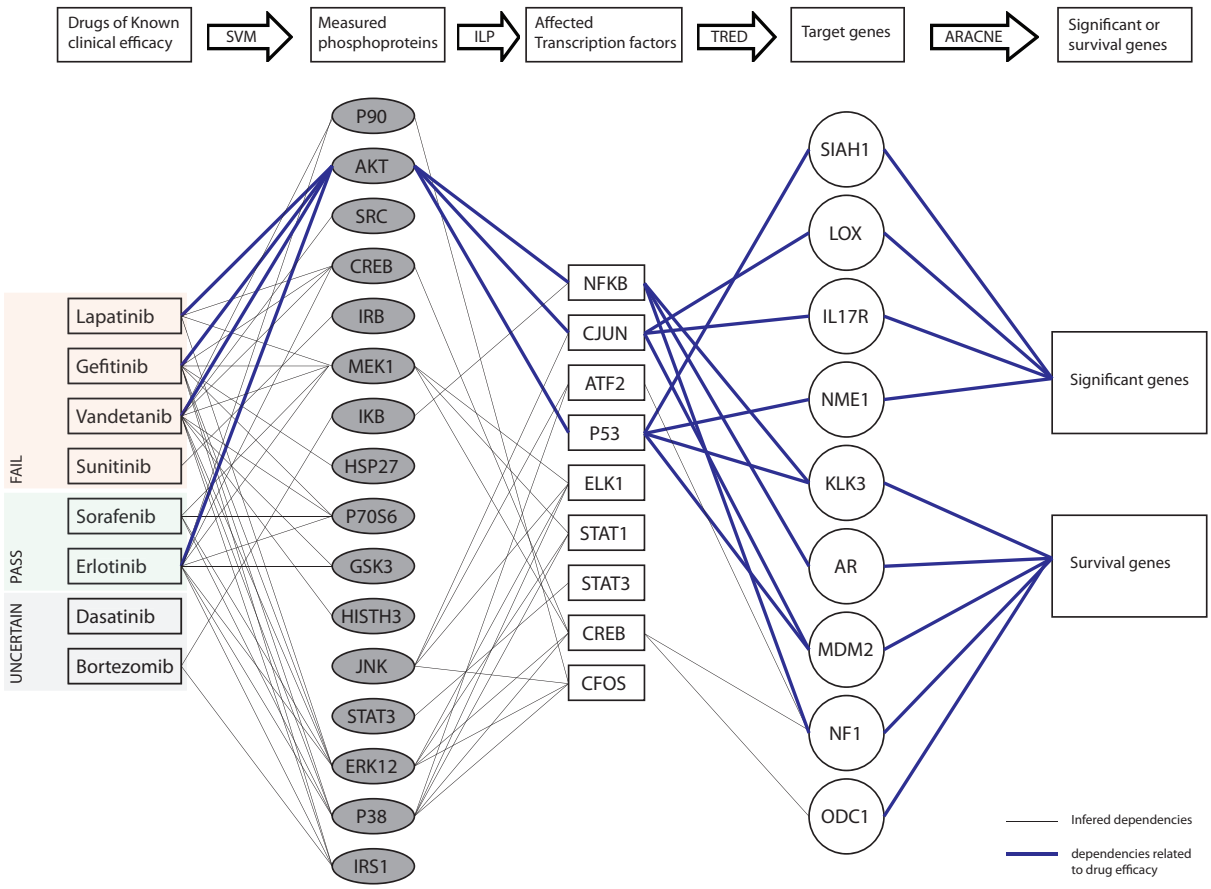


Figure 5.1: **Integrating phosphoproteomic, gene expression and clinical data - flowchart** Dependencies between phosphoproteomic, gene expression and clinical data are shown. In the first step, 8 drugs of known clinical efficacy are screened based on their effects on 16 phosphoprotein signals. In the second step the phosphoprotein signals are connected to transcription factors via the optimized topology of the signaling pathway. In the third step, transcription factors are connected to their target genes using the TRED database. In the fourth step, target genes are connected to significant genes (genes differentially expressed in cancer versus normal tissue in HCC) and survival genes by constructing a mutual information network using ARACNE through R. The dependencies implicated in drug efficacy are plotted in blue thick lines.

5.2 Results

Data collection and normalization

3 HCC cell lines were interrogated (huh7, hep3b, hepg2), by measuring the activation level of 16 key phosphoproteins (P90RSK, AKT, SRC, CREB, IRβ, MEK1, IKβ, HSP27, P70S6, GSK3, HISTH3, JNK, STAT3, ERK12, P38, IRS1), under 6 stimuli (IL1β, TGFα, Heregulin (HER), Insulin (INS), IL6 and TNFα) and presence of 8 drugs for unresectable HCC (Lapatinib, Gefitinib, Sorafenib, Erlotinib, Vandetanib, Sunitinib, Dasatinib, Bortezomib) (plus the JNJ, MEKPD32, PI103 inhibitors and the no-drug treatment). The 16 signals were chosen based on assay availability and quality controls performed at early stages of the experimental setup. The 6 stimuli were chosen to perturb most of the pathways observed by the 16 signals, and of the 11 drugs, the 8 were chosen based on the availability of clinical trial data and their target proteins, while the remaining 3 (JNJ, MEKPD32 and PI103 inhibitors) were chosen to better constrain the optimization of the Prior Knowledge Network (PKN) to signaling data according to the study by Mitsos et al. [23].

The 3 HCC cell lines were averaged in an "average cancer cell type" to better simulate the signaling response of tumor tissue [144]. Subsequently, the raw data was normalized using a linear regression model, that modeled the measured value of each signal as a linear function of the stimuli and inhibitor introduced and analyte measured. The normalized data was then scaled between 0 and 1 by evaluating the fold change of the signal before and after stimulation and dividing it by the maximum fold change for that signal. Normalized data for the average cancer cell type are shown in figure 5.2. Raw data for the individual cell types are shown in Figures 5.3-5.5.

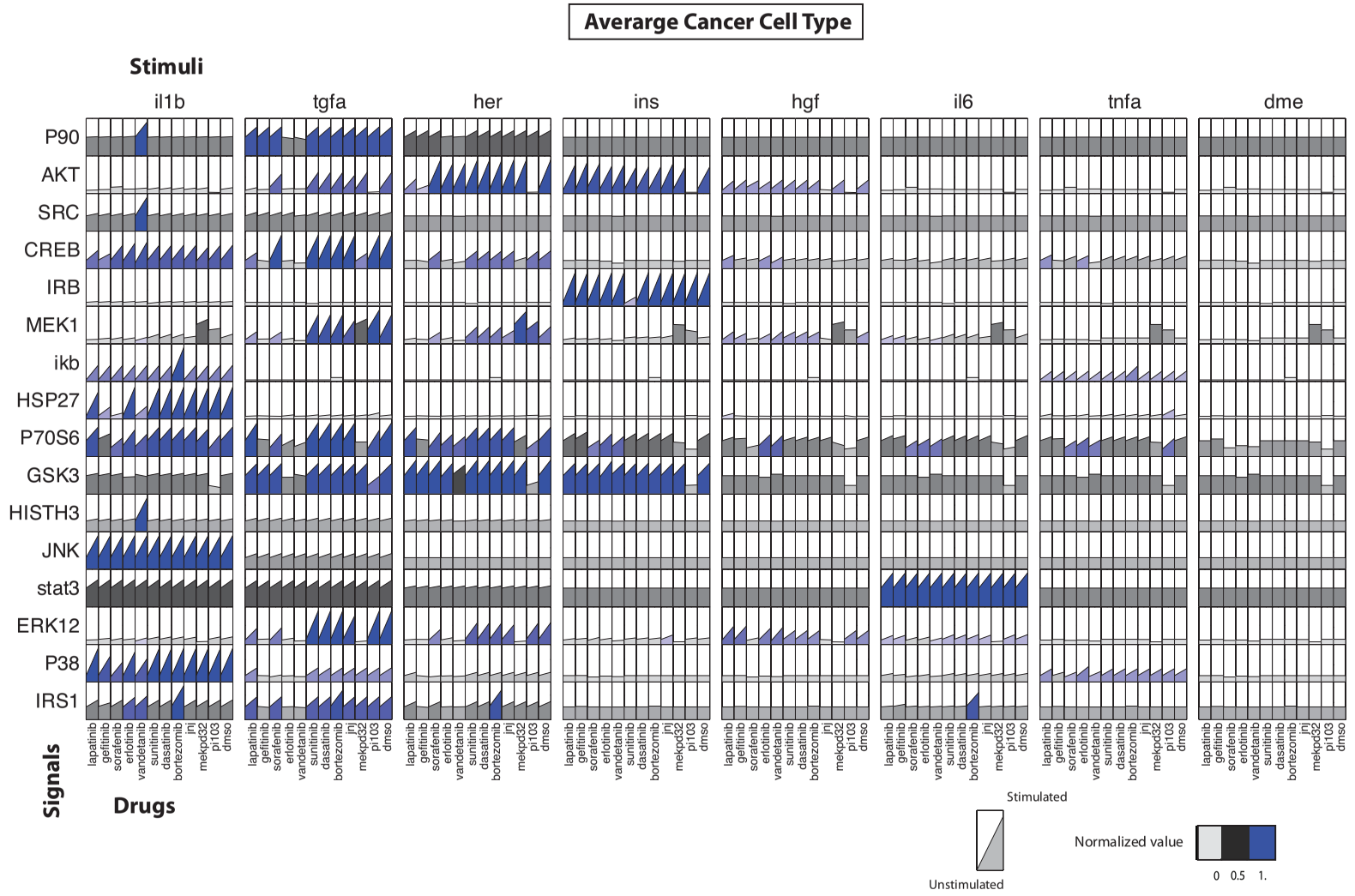


Figure 5.2: **Phosphoproteomic data** Phosphoproteomic data of the average cancer cell type under 8 stimuli (including the no-stimuli treatment) and 12 molecular inhibitors (including the no-inhibitor treatment). The time course of the 16 phosphoprotein signals from the unstimulated state to the average early response is illustrated. The rows correspond to the 16 signals, the main columns to the 8 stimuli treatments and the 12 subcolumns to the inhibitors. In each subplot, the first point shows the unstimulated activity of the respective signal (zero time point) and the second point shows the normalized value of the signal 5+25 minutes after stimulation.

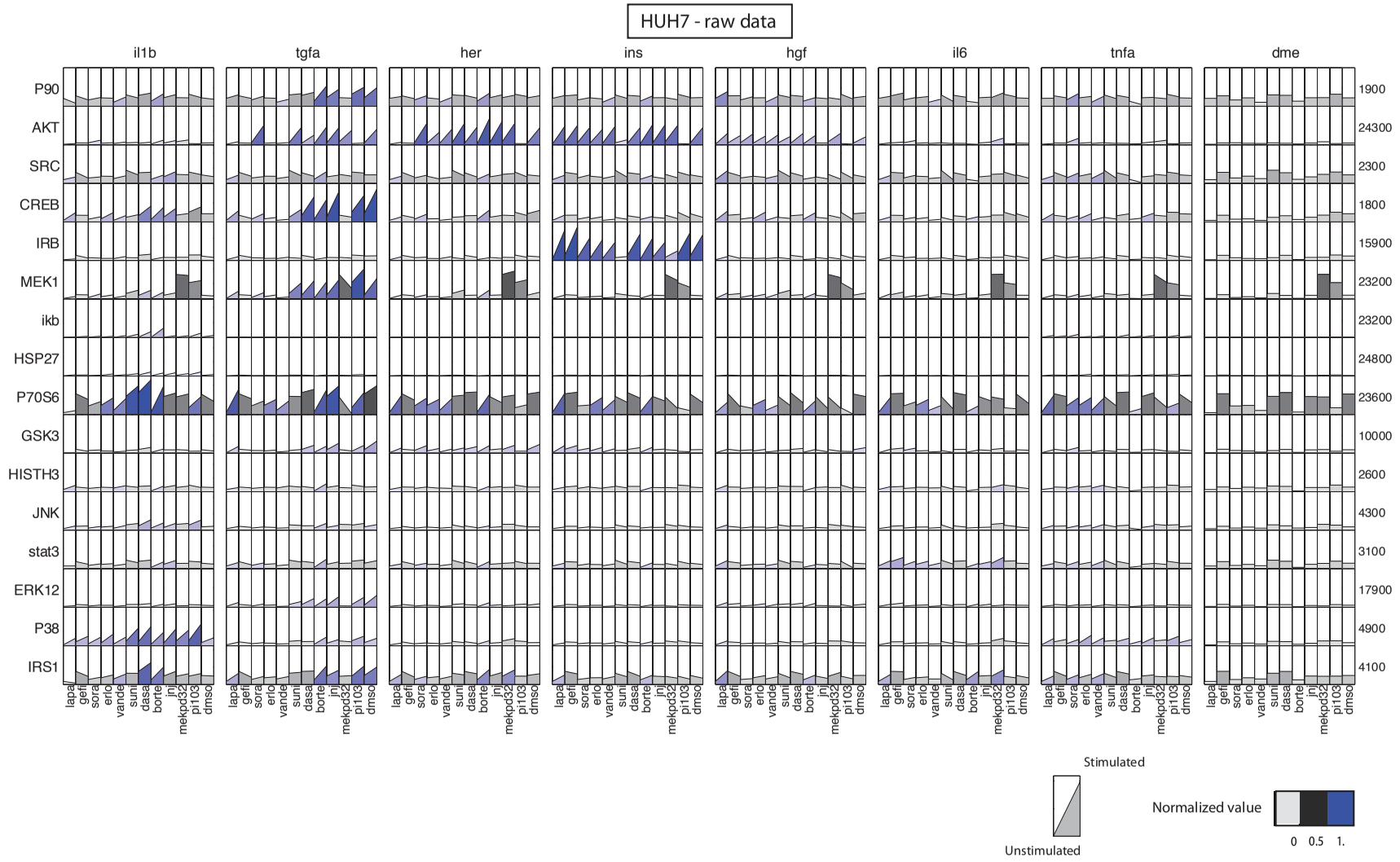


Figure 5.3: Raw phosphoproteomic data for the HUH7 cell line Phosphoproteomic data of the HUH7 cell type under 8 stimuli (including the no-stimuli treatment) and 12 molecular inhibitors (including the no-inhibitor treatment). The time course of the 16 phosphoprotein signals from the unstimulated state to the average early response is illustrated. The rows correspond to the 16 signals, the main columns to the 8 stimuli treatments and the 12 subcolumns to the inhibitors. In each subplot, the first point shows the unstimulated activity of the respective signal (zero time point) and the second point shows the raw value (in fluorescent units) of the signal 5+25 minutes after stimulation.

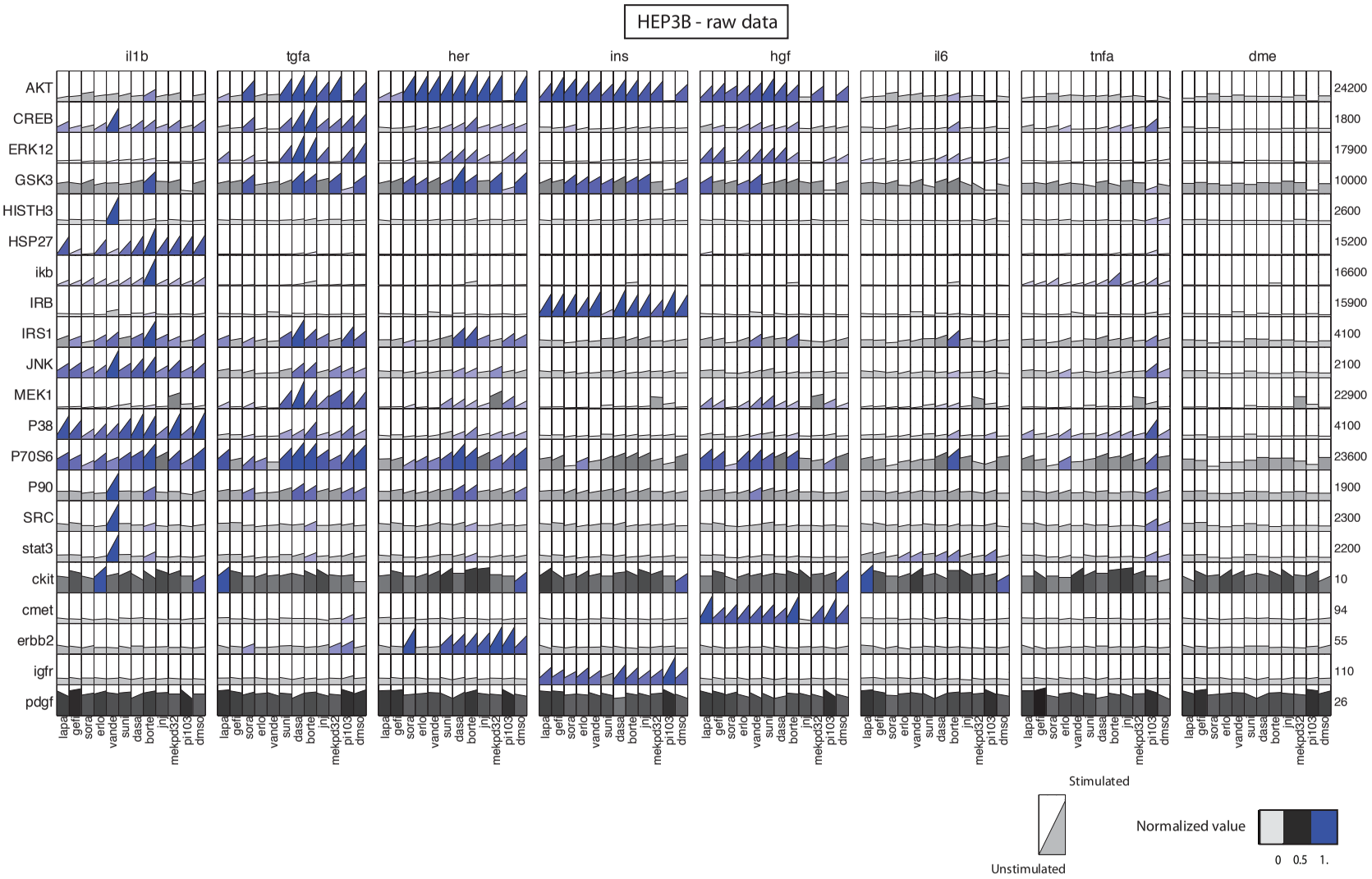


Figure 5.4: Raw phosphoproteomic data for the HEP3B cell line Phosphoproteomic data of the HEP3B cell type under 8 stimuli (including the no-stimuli treatment) and 12 molecular inhibitors (including the no-inhibitor treatment). The time course of the 16 phosphoprotein signals from the unstimulated state to the average early response is illustrated. The rows correspond to the 16 signals, the main columns to the 8 stimuli treatments and the 12 subcolumns to the inhibitors. In each subplot, the first point shows the unstimulated activity of the respective signal (zero time point) and the second point shows the raw value (in fluorescent units) of the signal 5+25 minutes after stimulation.

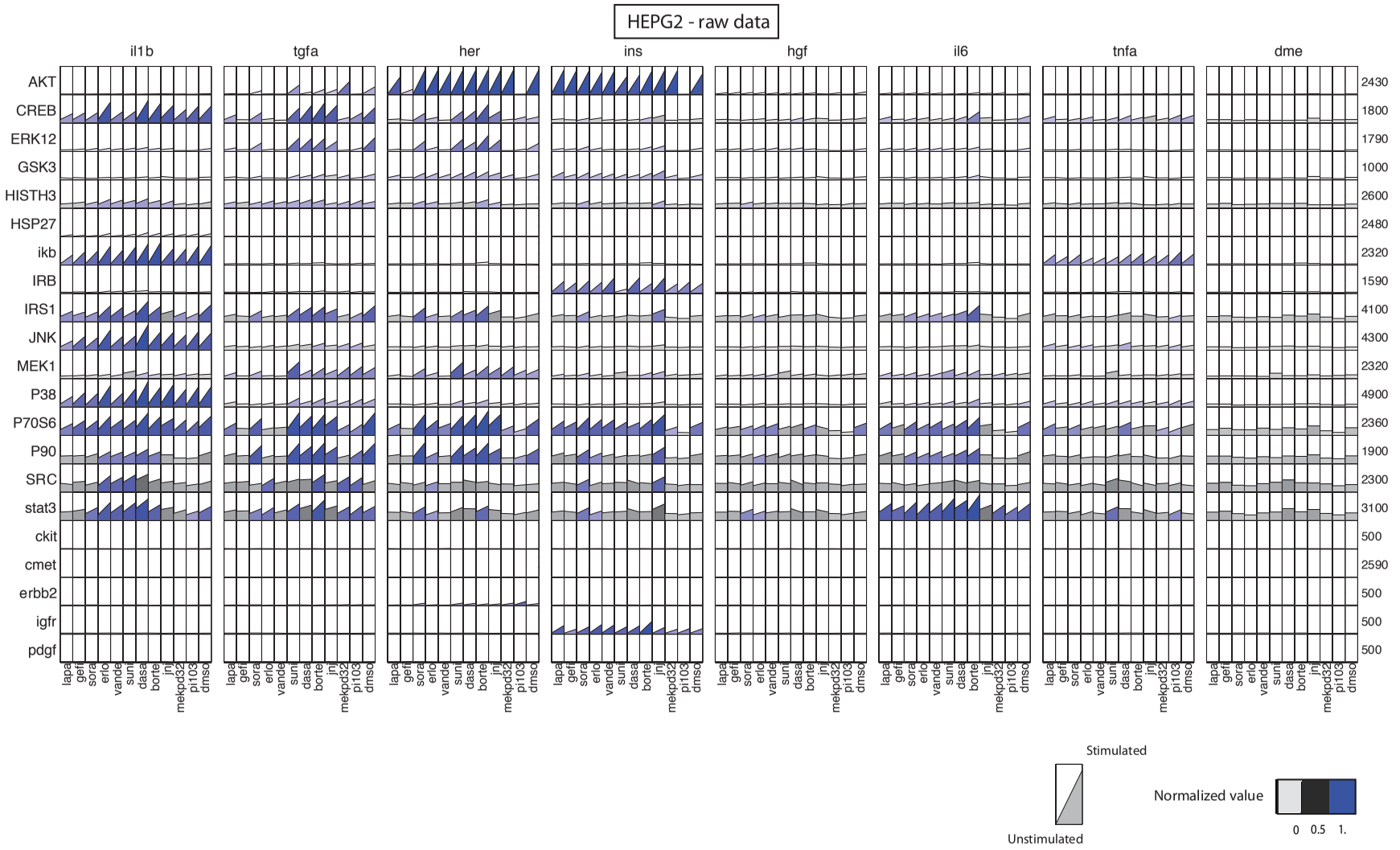


Figure 5.5: Raw phosphoproteomic data for the HEPG2 cell line Phosphoproteomic data of the HEPG2 cell type under 8 stimuli (including the no-stimuli treatment) and 12 molecular inhibitors (including the no-inhibitor treatment). The time course of the 16 phosphoprotein signals from the unstimulated state to the average early response is illustrated. The rows correspond to the 16 signals, the main columns to the 8 stimuli treatments and the 12 subcolumns to the inhibitors. In each subplot, the first point shows the unstimulated activity of the respective signal (zero time point) and the second point shows the raw value (in fluorescent units) of the signal 5+25 minutes after stimulation.

Regarding the effects of the interrogated stimuli on the average cancer cell type, TGFα and HER activated mostly AKT, CREB, MEK1 and ERK12, and partly activated P90, P70S6 and GSK3. TGFα, in contrast to HER, also activated P38 and partly IRS1. INS activated mostly AKT and IRB, also had medium effects on GSK3. HGF activated AKT, MEK1 and ERK12. On the other hand IL1β activated CREB, IKβ, HSP27, JNK, P38 and had medium effects on P70S6, STAT3 and IRS1. In similar fashion to IL1β, TNFα activated IKβ and P38. IL6 activated mostly STAT3 and partly affected ERK12.

Regarding the effects of the interrogated drugs / inhibitors on the average cancer cell type, Lapatinib inhibited AKT activation and partly CREB, MEK1 and ERK12 under TGFα . Also, inhibited CREB activation, MEK1, ERK12 and partly AKT under HER, while had no significant effects on any other pathway. Gefitinib had the same effects as Lapatinib under HER stimulation and also inhibited most signals under TGFα, such as AKT, CREB, MEK1, ERK12, P38, and IRS1. Gefitinib also partly inhibited HSP27 under IL1β. Sorafenib had no clear effects on the TGFα or HER pathway, but did inhibit MEK1, P70S6, and ERK12 under HGF and HSP27 in the IL1β pathway. Erlotinib and Vandetanib had very similar effects to Gefitinib under TGFα and on most of the signals under HER, apart from AKT (Gefitinib and Lapatinib both inhibited AKT under HER, while Erlotinib and Vandetanib left AKT unaffected). Sunitinib had no clear effects on any of the signals, apart from IRB under INS. Dasatinib also had no clear effects on the measured signals, indicating that the drug's mode of action is outside the observable part of the pathway. Bortezomib, being a proteasome inhibitor, increased activation of IKβ and IRS1 under all stimuli treatments. The rest of the inhibitors (JNJ, MEKPD32, PI103) had a much clearer mode of action affecting only the target kinase, that being cMET for JNJ, leading to the inhibition of most signals under HGF; MEK1 for MEKPD32, inhibiting signals downstream of MEK1 (e.g ERK12) under all stimuli treatments; PI3K for PI103, inhibiting AKT and GSK3 under all stimuli treatments.

Data Normalization

The phosphoproteomic data of 3 of the HCC cell lines (HUH7, HEP3B, and HEPG2) was averaged in an "average cancer cell type" and normalized using a linear regression model, that modeled the measured value of each signal as a linear function of the stimuli and inhibitor introduced, and analyte measured [145]. Thus, if $y_{n\times 1} \in \mathbb{R}$ is a vector of the response variables, where n is the number of data points (or observations); $i = 1, ..., n_{st}$ is an index set for the stimuli treatments ; $j = 1, ..., n_{inh}$ is an index set for the inhibitor (or drug) treatments and $k = 1, ..., n_{sig}$ is an index set for the analyte measured, then

$$y = \mu + x_{i,j,k} + \epsilon \tag{5.1}$$

where, μ is the grand mean; $x \in \mathbb{R}$ are variables modeling the effect of stimuli i, inhibitor j and signal k on the response variable and $\epsilon_{n\times 1} \in \mathbb{R}$ is the effect of technical factors. The technical factors may include batch effects, stimuli and inhibitor vehicle effects and so forth, here lumped together in ϵ since including all the individual terms would increase the number of model parameters significantly to the point were biological replicates must be introduced. To fit the model, the `lm` function was used, as implemented in `languageR`. The effect of the technical factors ϵ is computed as the residual of the model fitting.

To decrease the number of x variables so we avoid over fitting, as well as take advantage of the internal replicates present in the dataset, we identified all observations in the dataset where the measured signal under a specific stimuli treatment was significantly close to the no-stimuli treatment, over all drugs, and set that equal to the no-stimuli treatment, creating replicates of the no-stimuli treatment. In similar fashion, we identified all observations in the dataset where

the measured signal under a specific drug was significant close to the no-drug treatment, over all stimuli and set that equal to the no-drug treatment, creating replicates of the no-drug treatment. In this way a significant part of the dataset that did not respond to any stimuli or drugs was counted as a no-stimuli (or no-drug) treatment, both decreasing the number of x variables and creating replicates for the no-stimuli (or no-drug) treatment making the respective measurement more robust to experimental noise. Subsequently the normalized data was scaled between 0 and 1 by evaluating the fold change of the signal before and after stimulation and dividing it by the maximum fold change for that signal.

Pathway optimization to signaling data and identification of topology alterations

The Integer Linear Programming (ILP) formulation introduced in [23] was used to train a Prior Knowledge Network (PKN) to the signaling data and subsequently identify the drugs mode of action in terms of topology alterations of the PKN. First, the PKN was constructed by merging together canonical pathways obtained from literature (on line pathway databases, mostly: KEGG, pathway commons and Ingenuity) [13] and then an executable model of the pathway was constructed by modeling the signaling reactions using Boolean gates (AND/OR/NOT) [23, 13, 33, 51]. The model was subsequently trained to the subset of the data containing only the specific molecular inhibitors (JNJ, MEKPD32, PI103) to remove all reactions (in the PKN) that contradicted the data at hand, thus resulting in a cell-type specific model, predictive of the signaling response of this average cancer cell type. The drugs' mode of action, in terms of topology alterations, was identified by training the cell-type specific model to the data from the 8 cancer drugs. Reactions removed in this step of the training process are the reactions inhibited by the respective drug. The cell-type specific pathway is shown in figure 5.6A and the drug specific topology alterations in Figure 5.6B. Topology alterations plotted on the signaling pathway are shown in figures 5.7-5.10.

The signaling topology in Figure 5.6A is in good accordance to previous results [23, 51, 46]. In more detail, HER, TGFα and INS signaled through the several ERBB dimers and IRB and activated MEK1, ERK12, CREB and P70S6 via GRB2, SOS, RAS and RAF1. Moreover, they activated AKT via PI3K. HGF also signaled through PI3K to activate AKT. TNFα and IL1β signaled through partially overlapping pathways, activating IKB, JNK, P38 HSP27 and CREB. IL6, having only medium size effects on STAT3, was removed from the optimized topology. To evaluate the performance of the optimization algorithm and ensure the obtained model successfully fits the data at hand, the measurement - prediction mismatch (i.e. mean fitness error) before and after the optimization procedure is computed. The fitness error dropped from 38% (for the initial topology) to 11% (for the optimized topology), proving the ILP formulation removed the reactions conflicting with the signaling data and the optimized model successfully captures the signaling response of the average cancer cell type.

The drug induced topology alterations shown in Figure 5.6B and Figures 5.7-5.10, demonstrate what is widely known, that most of the HCC drugs are highly promiscuous, clearly inhibiting kinases other than the predefined targets. Lapatinib blocked the HER pathway from the receptor level (on target effect), and also blocked the AKT activation under TGFα, by removing the RAS $\rightarrow$ PI3K reaction (off target effect). Gefitinib inhibited both TGFα and HER pathways from the receptor level (on target effect). Sorafenib inhibited the HGF pathway from the receptor level and HSP27 activation via P38 (both off target effects). Also partly inhibited MEK1 and ERK12 under TGFα (on target effects) however MEK1 and ERK12 inhibition was not as clear as for other drugs (e.g. Gefitinib and Lapatinib) and did not lead to further alterations of the pathway. Erlotinib and Vandetanib caused the exact same topology alterations inhibiting TGFα from the receptor level (on target effect), and also inhibiting MEK1, ERK12

and P70S6 under HER, without affecting AKT, by removing RAS → RAF reaction (off target effect). Bortezomib's, Sunitinib's and Dasatinib's effects could not be modeled as alterations of the PKN either because they increased the activation level of signaling molecules instead of decreasing it, and that is not supported by the ILP formulation, or because were not strong enough, indicating that the drug's mode of action is outside the observable part of the pathway.

Recursive Feature extraction

Phosphoproteomic data.5 of the original 8 drugs (Lapatinib, Gefitinib, Sorafenib, Erlotinib and Vandetanib), that had clear effects on the measured phosphoproteins, were used in a recursive feature extraction procedure to identify the key phosphoprotein signatures that are predictive of drug efficacy in HCC.

Recursive feature extraction was implemented using the Matlab `classify` function within a custom Matlab script. Default values for the function parameters were used. Data was formatted as a 2-D matrix, with rows corresponding to the different drugs (here serving as samples) and columns corresponding to observations or features (different combinations of stimuli and signals). The differential response of the drugs was used, computed as follows. For every feature (column), the mean response over all drugs was evaluated and the differential response of each drug was obtained by subtracting the response of that drug from the mean. Before the feature extraction, observations with a constant value over all samples were removed as non-informative, moreover, replicate observations (sets of observations with the same value under all drugs) were removed after keeping record of them. Ultimately, every one of the remaining observations was used separately and a classifier was trained to distinguish the drugs between PASS or FAIL according to the clinical trial data available. This process was repeated as many times as the available drugs, every time leaving one of the drugs out as a test dataset to evaluate the classifiers performance. The average (over all drugs) performance of the classifier for each feature was evaluated. Results are shown in figure 5.11.

The most predictive features are the measurement of (i) AKT and CREB under TGFα, or ERK12 and MEK1 under HER, (ii) AKT under HER, and (iii) ERK12 and MEK1 under HGF. In more detail, the feature extraction dictates that inhibition of (i) and (ii) is indicative of a drug that failed in clinical trials, while inhibition of (iii), of a drug that succeeded in clinical trials. (accuracy 80%).

The same procedure was repeated after scrambling the clinical trial results (here serve as labels or classes). Results are shown in figure 5.12. The classification accuracy was significantly lower with the scrambled data, implying the significance of the extracted features.

Topology alterations. In similar fashion to the phosphoproteomic data, recursive feature extraction was applied on the topology alterations of 5 of the original 8 drugs (Lapatinib, Gefitinib, Sorafenib, Erlotinib and Vandetanib), that had clear effects on the measured phosphoproteins, to identify key alterations, predictive of drug efficacy in HCC. Results are shown in figure 5.13. The most predictive features are the removal of (i) ERK12 → MSK12 reaction, (ii) RAS → PI3K reaction, and (iii) PRAK → HSP27 reaction (accuracy 80%). In more detail, the feature extraction dictates that removal of (i) and (ii) is indicative of a drug that failed in clinical trials, while removal of (iii) is indicative of a drug that succeeded in clinical trials. The results are in good accordance to the extracted phosphoproteins, since removal of (i) ERK12 → MSK12 essentially inhibits the activation of CREB under TGFα and, as identified by the feature extraction on the phosphoproteomic data, inhibition of CREB is indicative of a drug that failed clinical trials. Moreover, removal of (ii) RAS → PI3K inhibits the activation of AKT under HER and TGFα, also identified by the feature extraction on the phosphoproteomic data to be predictive of drugs that failed clinical trials. To ensure the significance of these results, the same analysis

was performed after scrambling the classes (see supplementary figure 8).

Using topology alterations for the feature extraction offers a clear advantage over the use of signaling data, mostly regarding the way results are interpreted. E.g. Identifying the measurement of CREB under TGFα as being predictive of drug efficacy does not imply the CREB signal is predictive in general (CREB was also activated by IL1β, but these measurements are not predictive), it implies that a signaling mechanism upstream of CREB in the TGFα pathway is connected to drug efficacy in HCC. Using topology alterations in the feature extraction process enables us to identify this mechanism. In the current example, it is the inhibition of ERK12 $\rightarrow$ MSK12 reaction that is connected to drug efficacy (Inhibition of ERK12 $\rightarrow$ MSK12, induces the inhibition of CREB under TGFα). In similar fashion regarding the measurement of AKT under TGFα and HER, the AKT signal itself is not predictive (AKT was also activated by INS and HGF, however, these measurements are not predictive), the inhibition of RAS $\rightarrow$PI3K reaction is predictive of drug efficacy. Equivalently, the measurement of AKT under EGFR ligands (TGFα and HER) is predictive of drug efficacy, while its measurement under INS or HGF is not. This is in good context to the interrogated disease, since the EGFR is known to be over expressed in HCC (as in most cancers) [146].

Linking extracted features to significant genes in HCC

The phosphorotein signals and topology alterations, identified by the feature extraction process to be correlated to drug efficacy in HCC, are subsequently linked to significant genes in HCC. The gene expression data published in [126] was obtained from GEO (Gene Expression Omnibus - Series GSE3500 `http://www.ncbi.nlm.nih.gov/geo/`) and was used to infer a network, connecting the measured phosphoproteins to genes differentially expressed in normal versus cancer tissue (significant genes) (see Figure 5.1). To that effect the TRED database (Transcriptional Regulatory Element Database
`http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home`) was used to obtain the target genes of the transcription factors (TFs) present in the observable part of the pathway (P53, CREB, FOS, JUN, ATF2, ELK1, STAT1, STAT3 and NFKB). The connection of the measured phosphoproteins to the significant genes is accomplished through the TFs they affect. Thus, the AKT signal, for example, inhibiting P53 via MDM2 [147] affects the P53 target genes (as obtained from TRED). The expression values of the target genes were obtained from GSE3500 and a network was constructed connecting the target genes to the significant genes (see Figure 5.1). In this manner, we were able to correlate the phosphoprotein signals to the significant genes.

Network inference on the GSE3500 dataset: The GSE3500 dataset was downloaded from Gene Expression Omnibus (GEO) and parsed into a 2-D matrix (rows corresponding to samples, columns corresponding to genes) using a custom bash script (`http://www.gnu.org/software/bash/`). Genbank ids were mapped to gene names using the Clone/Gene ID Converter
(`http://idconverter.bioinfo.cnio.es/IDconverter.php`). Subsequently, the expression values of significant genes [126] or survival genes [124] and the TF target genes were extracted, and imported to `R` programming language (`http://www.r-project.org/`) where the ARACNE [148, 149] algorithm was run to infer a mutual information network. Every one of the target genes was scored based on the number of connections to significant genes, the 10 most highly scored target genes are shown in figure 5.14A. Subsequently, the connectivity of the target genes to random subsets of GSE3500 (of equal size to the significant genes) is examined to identify the target genes that are more strongly correlated to the significant genes than to any other gene set. Figure 5.14B, shows for the most highly scored target genes, the number of connections to the significant genes minus the maximum number of connections to any other gene set.

As shown in figure 5.14B, SIAH1 and NME1, the most highly scored genes (genes that are found to correlate more to the significant genes than to any other gene set), are both target genes of P53. P53 is inhibited by AKT via MDM2 [147]. This supports our previous finding that inhibition of AKT under TGFα and HER is predictive of drug efficacy in HCC, since inhibition of AKT will increase activity of P53, that will affect SIAH1 and NME1, that correlate strongly to genes differentially expressed in cancer versus normal tissue. Apart from P53, JUN is also correlated to significant genes through its target genes LOX and IL7R, and is too affected by AKT [150], supporting our speculation. Another TF that is correlated to significant genes, although not as strongly as P53 or JUN, is CREB though PGK1. CREB is also amongst the phosphoproteins that were identified as being correlated to drug efficacy, even though the feature extraction process applied on the topology alterations showed that inhibition of ERK12→MSK12 reaction is the truly predictive feature and it can also affect other TFs or phosphoproteins apart from CREB (e.g. ATF1) . The importance of P53 in HCC is also highlighted in [151]. Regarding the target genes, the implication of SIAH1 in tumor progression in HCC and gastric cancer is also supported by literature [152], NME1 gene is connected to patient survival in breast cancer [153], LOX is found to contribute to the progression of HCC [154], while il17R and PGK1 are also implicated in HCC [155, 156].

Linking extracted features to survival genes in HCC

To further solidify the role of AKT (and consequently of P53 and JUN) to mechanisms of drug efficacy in HCC, we use the gene set identified by Lee et al in [124] as being highly correlated to hazard ratios in HCC. The expression values for most of the survival genes were obtained from GSE3500 (as with the significant genes) and a network was constructed connecting the target genes of the TFs present in the observable part of the pathway to these survival genes. In similar fashion to before, the target genes were scored according to the number of connections to the survival genes. The most highly scored target genes are shown in figure 5.14C. Figure 5.14D, shows for the most highly scored target genes, the number of connections to the survival genes minus the maximum number of connections to any other gene set, to identify the ones that are more strongly connected to survival genes than to any other gene set. In similar fashion to the significant genes, the most highly scored gene (KLK3) is a P53 target gene, supporting our previous findings.

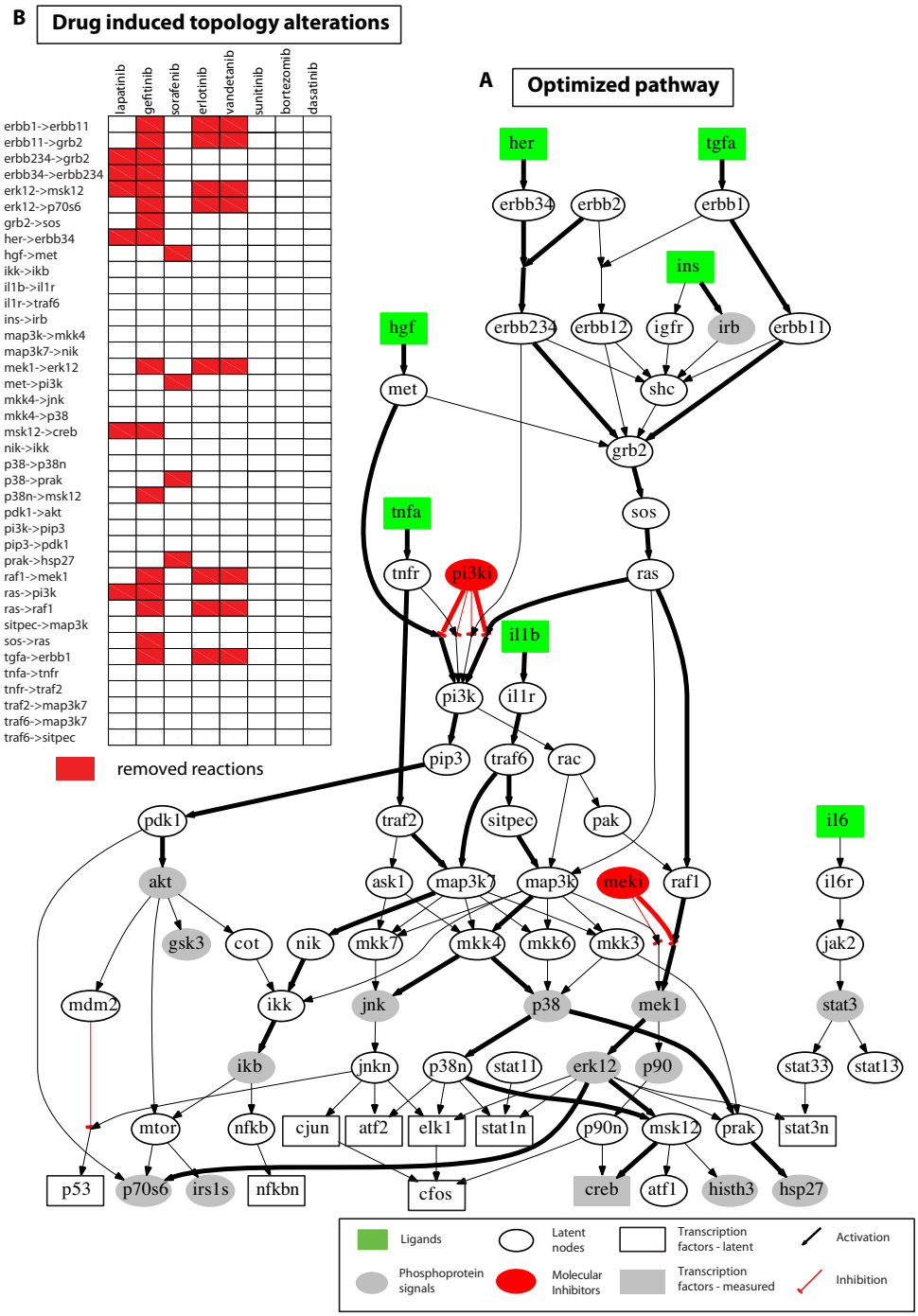


Figure 5.6: **Optimized pathway and drug induced topology alterations** A) The signaling pathway after applying the ILP formulation. Reactions that contradicted the data at hand were removed from the topology, the resulting pathway captures the signaling response of the average cancer cell type. Green nodes correspond to the imposed stimuli, grey nodes to the measured phosphoproteins, red nodes to the specific molecular inhibitors used in the dataset to constrain the optimization problem and clear (white) nodes to latent signaling proteins in the network. B) Drug induced topology alterations. The reactions removed by each drug are plotted in red, reactions that were not affected are plotted in white.

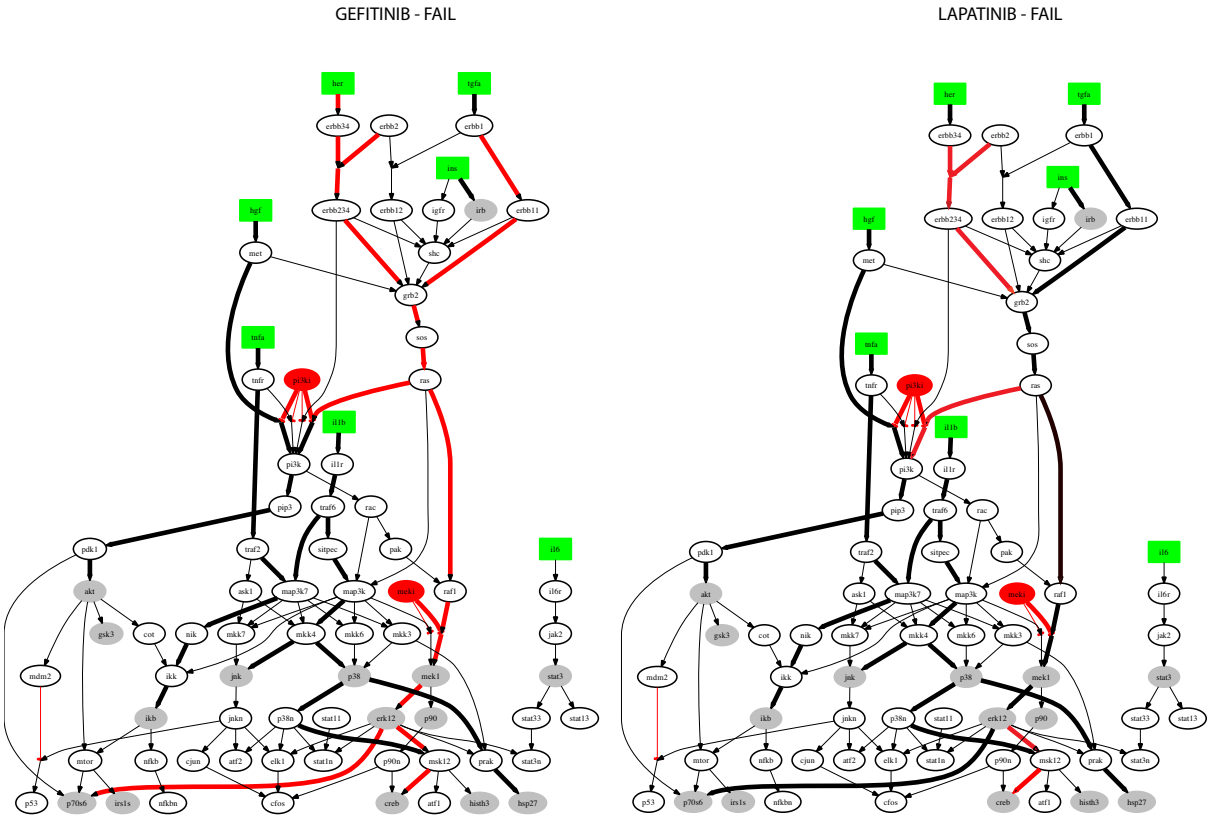


Figure 5.7: Drug induced topology alterations - gefitinib, lapantinib Topology alterations induced by gefitinib and lapatinib. Reactions that were inhibited by gefitinib or lapatinib are plotted in red, the rest are plotted in black. Green nodes correspond to the imposed stimuli, grey nodes to the measured phosphoproteins, red nodes to the specific molecular inhibitors used in the dataset to constrain the optimization problem and clear (white) nodes to latent signaling proteins in the network.

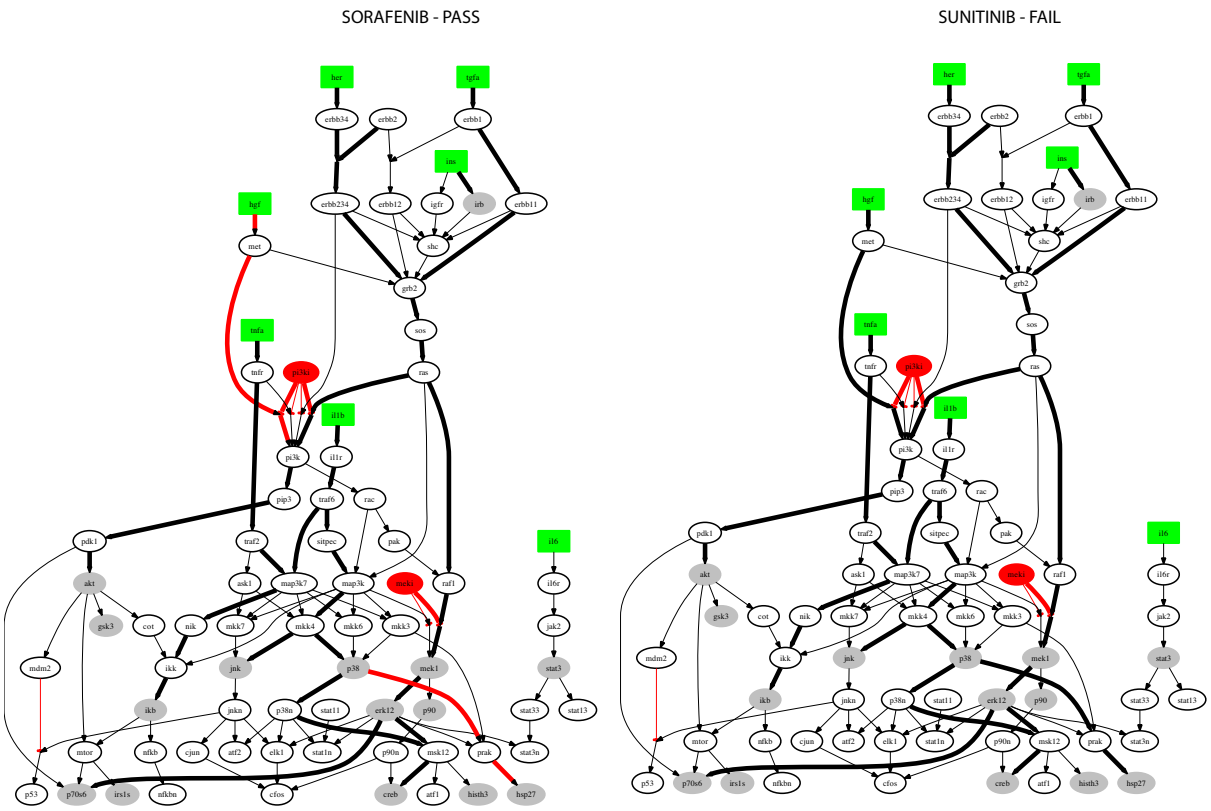


Figure 5.8: Drug induced topology alterations - sorafenib, sunitinib Topology alterations induced by sorafenib and sunitinib. Reactions that were inhibited by sorafenib or sunitinib are plotted in red, the rest are plotted in black. Green nodes correspond to the imposed stimuli, grey nodes to the measured phosphoproteins, red nodes to the specific molecular inhibitors used in the dataset to constrain the optimization problem and clear (white) nodes to latent signaling proteins in the network.

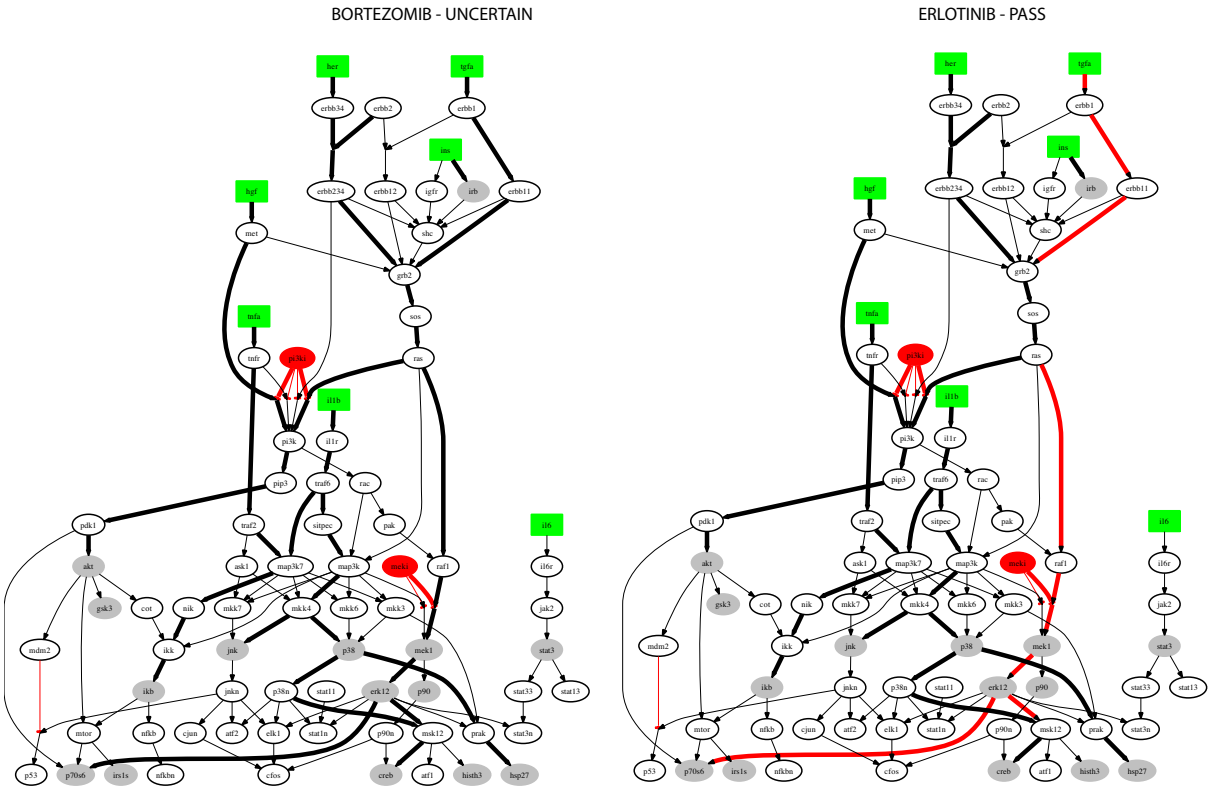


Figure 5.9: Drug induced topology alterations - bortezomib, erlotinib Topology alterations induced by bortezomib and erlotinib. Reactions that were inhibited by bortezomib or erlotinib are plotted in red, the rest are plotted in black. Green nodes correspond to the imposed stimuli, grey nodes to the measured phosphoproteins, red nodes to the specific molecular inhibitors used in the dataset to constrain the optimization problem and clear (white) nodes to latent signaling proteins in the network.

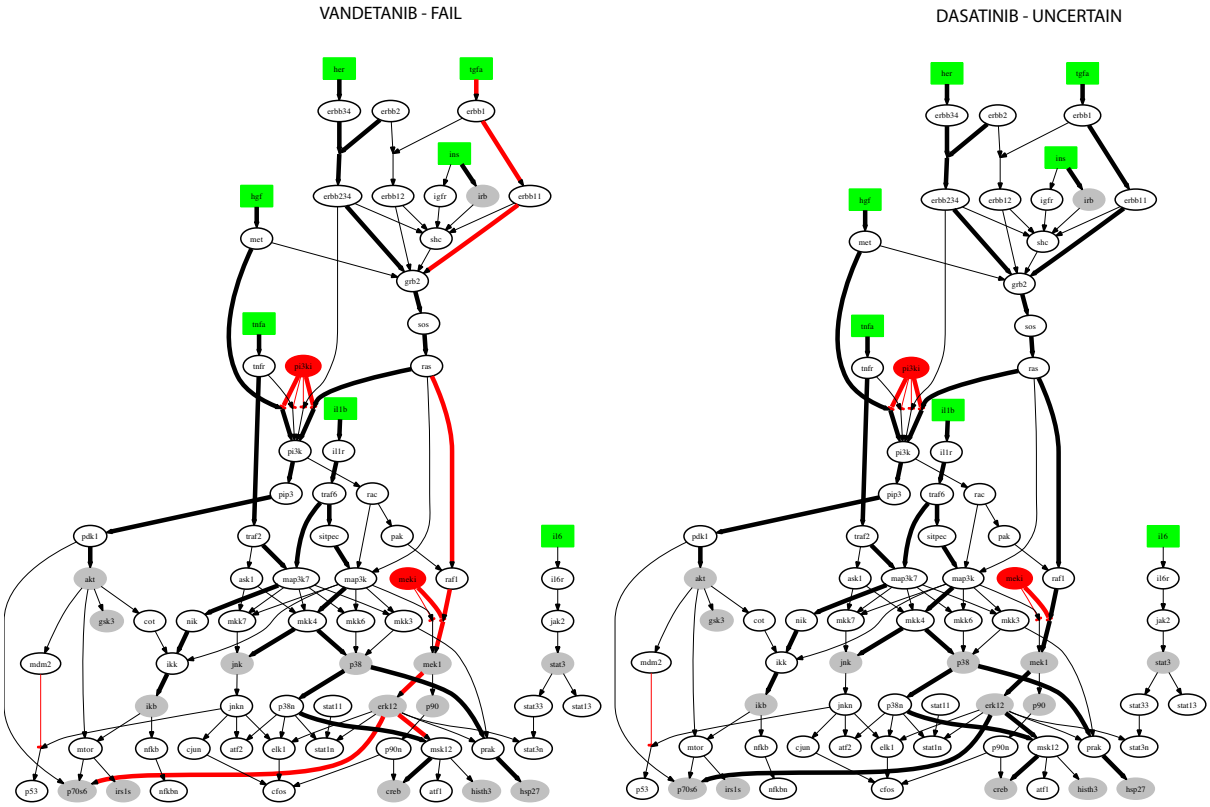


Figure 5.10: Drug induced topology alterations - vandetanib Topology alterations induced by vandetanib. Reactions that were inhibited by vandetanib are plotted in red, the rest are plotted in black. Green nodes correspond to the imposed stimuli, grey nodes to the measured phosphoproteins, red nodes to the specific molecular inhibitors used in the dataset to constrain the optimization problem and clear (white) nodes to latent signaling proteins in the network.

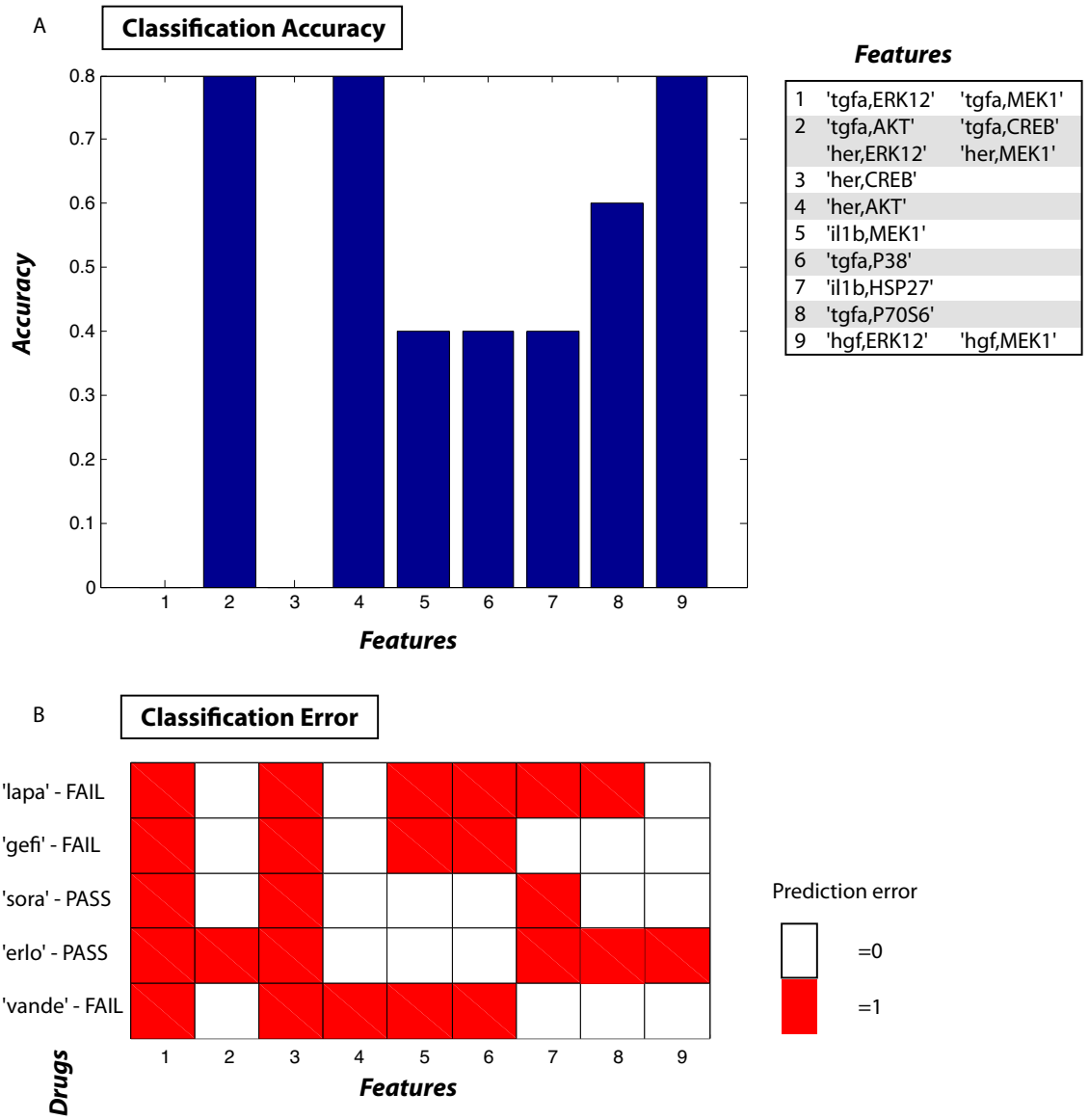


Figure 5.11: **figure 4 - Feature extraction - phosphoproteins** A) Classification accuracy for different phosphoprotein features. The y-axis corresponds to classification accuracy (out of 1.0), the x-axis corresponds to the different phosphoprotein features. Features are not unique since a number of observations (combinations of stimuli and signals) may be the same under all drugs e.g. measurement of AKT under TGFα, CREB under TGFα, ERK12 and MEK1 under HER are the same under all the interrogated drugs. B) Classification error for the different phosphoprotein features, for each drug. Prediction error is plotted in red.

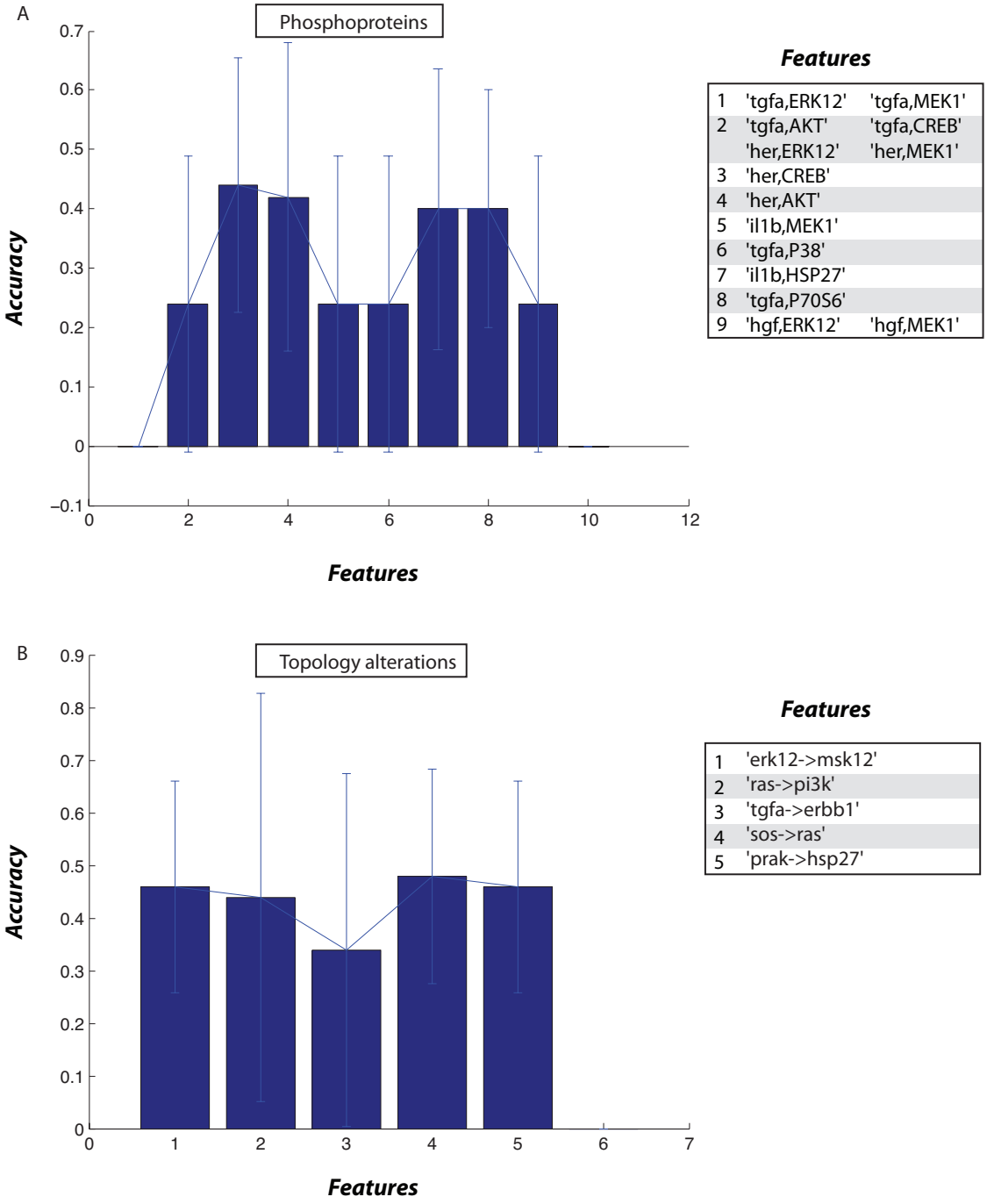


Figure 5.12: Validation of the feature extraction A) Validation of the feature extraction on the phosphoproteomic data. The feature extraction process, was repeated here after scrambling the clinical trials results. The x-axis corresponds to the phosphoprotein features, and the y-axis to the classification accuracy (out of 1.0). B) Validation of the feature extraction on the topology alterations. The x-axis corresponds to the drug induced topology alterations, and the y-axis to the classification accuracy (out of 1.0).

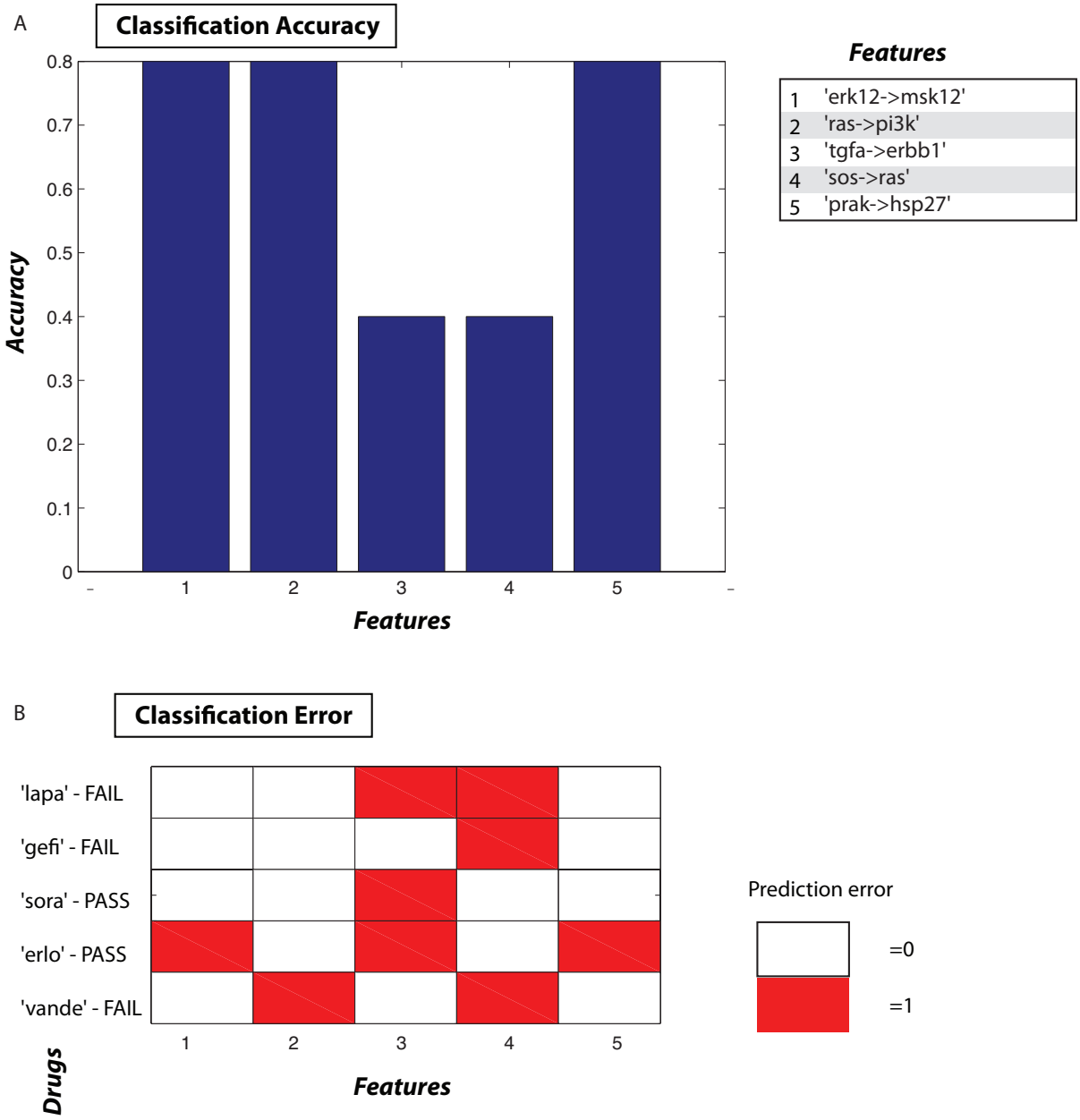


Figure 5.13: **Feature extraction - topology alterations** A) Classification accuracy under different topology alterations. The y-axis corresponds to classification accuracy (out of 1.0), the x-axis corresponds to reactions removed by certain drugs. B) Classification error for the different removed reactions, for each drug. Prediction error is plotted in red.

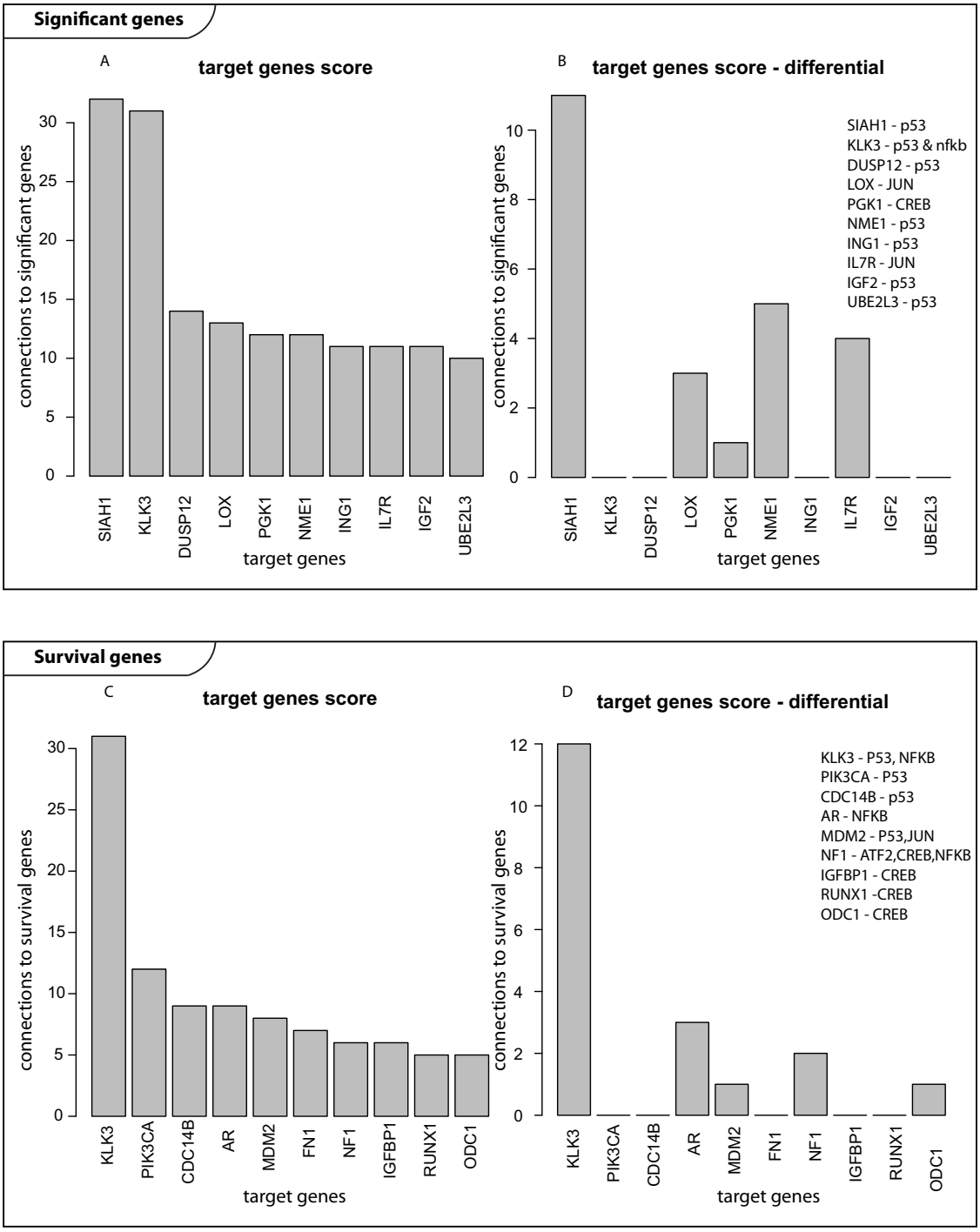


Figure 5.14: **Correlation of phosphoproteins to significant and survival genes in HCC** A) Correlation of the transcription factors' target genes to significant genes. The y-axis corresponds to the number of connections to significant genes, the x-axis corresponds to the most highly connected target genes. B) Correlation of the transcription factors' target genes to significant genes minus the maximum number of correlations to any other gene set. This figure serves to identify the target genes that are more closely connected to the significant genes than to any other gene set. The y-axis corresponds to the number of connections to significant genes minus the maximum number of correlations to any other gene set , the x-axis corresponds to the most highly connected target genes.

5.3 Discussion

Herein, we tried to identify signaling pathways implicated in drug efficacy in HCC by combining signaling, gene expression and clinical data within a consistent, mechanistic framework. Starting point of our approach was a phosphoproteomic dataset that captured the signaling response of 3 HCC cell types to environmental perturbations, and to 8 drugs for HCC of known clinical efficacy. A cell-type specific pathway was constructed using an Integer Linear Programming formulation on the phosphoproteomic data and the effects of the 8 cancer drugs on the signaling pathway were modeled as topology alterations of this pathway. Both on-target and off-target effects were uncovered. Using a standard recursive feature extraction procedure we were able to identify key phosphoprotein signatures (AKT activation under TGFα and HER; CREB activation under TGFα) and topology alterations (ERK12 $\rightarrow$ MSK12; RAS $\rightarrow$ PI3K) , predictive of drug efficacy in HCC. Subsequently, we translated the findings of this analysis to the gene expression level, by using the TRED database to obtain the target genes of the Transcription Factors directly correlated to the extracted phosphoproteins, and constructed regulatory networks between these target genes and significant genes in HCC. In this manner we were able to correlate key phosphoproteins to genes that are either differentially expressed in tumor versus normal tissue in HCC, or genes that are strongly correlated to hazard ratios in HCC. This analysis validated the significance of the extracted phosphoprotein features and identified mostly P53 as the key regulator of drug efficacy in HCC. The implication of p53 in HCC is also supported by literature [151, 152].

Unexpectedly, inhibition of AKT activation under the EGFR ligands, was found to be indicative of drugs that failed clinical trials, even though the PI3K/AKT/MTOR pathway is known to be correlated with HCC cell proliferation [157]. Our findings imply that inhibition of the PI3K/AKT/MTOR pathway upstream of PI3K by removing the RAS $\rightarrow$ PI3K reaction may have mixed (both positive and negative) effects. On the one hand (positive effect) MTOR activation is inhibited, known to correlate with reversal of gene expression and autophagy in HCC [158] and on the other hand (negative effect) the inhibition of P53 via MDM2 is blocked and P53 is known to induce apoptosis [159].

The major limitation of the methodology presented herein, is the small number of drugs in this study, which is a direct consequence of the limited number of drugs for HCC with known clinical efficacy. Even though the recursive feature extraction was accurate for 80% of the samples, with only 2 of the drugs having succeeded in clinical trials it is possible that key phosphoprotein signatures eluded as. On the plus side, by translating our findings to the gene expression level and linking the extracted features with 2 independent gene sets, known to correlate to HCC, we ensure the biological relevance of our results. Another limitation is regarding Dasatinib, Bortezomib, and Sunitinib and our inability to obtain a clear profile for these drugs, either because they don't affect the measured phosphoproteins, or because their mode of action is not supported by the Interger Linear Programming formulation we use (Bortezomib increased the activation level of most measured signals).

Overall, the present study has succeeded in identifying key signaling pathways implicated in drug efficacy in HCC, while most of these findings are also supported by previous studies. This is amongst the first attempts to combine signaling, gene expression and clinical data and integrate it all within a mechanistic, hypothesis generating/validating framework for drug targets in HCC.

Chapter 6

Summary

In this PhD, the author negotiated the development of a novel class of methodologies that model signal transduction networks as logic models, and using regular optimization formulations (Integer Linear Programming (ILP) and Non Linear Programming (NLP) formulations) cross reference them with high throughput phosphoproteomic data to construct predictive models of the signaling mechanisms of the interrogated cell type. In contrast to previously published approaches, the use of an ILP formulation to optimize signaling pathways to experimental data guarantees to return a global minimum of the objective function (mismatch between model predictions and experimental measurements) and speeds up the runtime significantly, thus allowing the interrogation of more complex signaling pathways and accompanying phosphoproteomic datasets.

In more detail, 3 different regular optimization formulations are introduced: **(i)** an ILP that models signal transduction networks as Boolean models and optimizes the network structure by pruning to minimize the mismatch between model predictions and experimental measurements. **(ii)** An NLP (Non-Linear Programming) formulation that models signal transduction networks as constrained fuzzy logic models and optimizes the parameters of the transfer functions to minimize the mismatch between model predictions and experimental measurements. **(iii)** An ILP that models the signal transduction networks as interaction graphs and optimizes their structure by removing (pruning) or adding de novo reactions to minimize the mismatch between model predictions and experimental data.

Additionally, 5 case studies were interrogated: **(i)** Identification of the effects of drugs for Hepatocellular carcinoma (i.e. liver cancer) on the signaling pathways of normal and cancer liver cells. **(ii)** Construction of a large scale signal transduction network for primary human hepatocytes (normal liver cells), based on prior knowledge of protein connectivity and high throughput phosphoproteomic data. **(iii)** Construction of an integrated, multi-layered signal transduction network starting at the receptor level, to intracellular signaling and activation of key phosphorylation pathways and ending at the expression of cytokines and their release in the supernatant. **(iv)** Construction of signaling pathways in chondrocytes (cell type found in articular cartilage) and identification of novel catabolic players that may affect cartilage physiology in osteoarthritis. **(v)** Integration of proteomic, genomic and clinical data to identify signaling pathways predictive of drug efficacy in hepatocellular carcinoma.

The work presented herein elucidated the complex mechanisms that orchestrate cellular response to environmental perturbations and shed light into the etiology underlying complex disease.

Chapter 7

Work produced within this PhD

A total of 12 papers were published to peer reviewed journals and conferences of the IEEE. These are the following (sorted by publication date):

Peer Reviewed papers

1. *Modeling of signaling pathways in chondrocytes based on phosphoproteomic and cytokine release data.* **Ioannis N. Melas***, Aikaterini D. Chairakaki*, Elisavet I. Chatzopoulou*, Dimitris E. Messinis, Alexander Mitsos, Zoe Dailiana, Panagoula Kollia, Leonidas G. Alexopoulos. Submitted, 2013. *Equal contributors

2. *Phosphoproteomics in drug discovery.* Melody K Morris, An Chi, **Ioannis N. Melas**, Leonidas G Alexopoulos. Drug Discovery Today, Accepted, 2013

3. *Leveraging systems biology approaches in clinical pharmacology.* **Ioannis N. Melas**, Kosmas Kretsos, Leonidas G Alexopoulos. Biopharm Drug Dispos. 2013 Aug 23. doi: 10.1002/bdd.1859

4. *Detecting and Removing Inconsistencies Between Experimental Data and Signaling Network Topologies using Integer Linear Programming on Interaction Graphs.* **Ioannis N. Melas***, Regina Samaga*, Leonidas G Alexopoulos, and Steffen Klamt. PLoS Comput Biol. 2013;9:e1003204. doi:10.1371/journal.pcbi.1003204, *Equal contributors

5. *Identification of signaling pathways related to drug efficacy in HCC via integration of phosphoproteomic, genomic and clinical data.* **Ioannis N. Melas**, Douglas A. Lauffenburger, Leonidas G. Alexopoulos. *Accepted*, 13th IEEE International Conference on BioInformatics and BioEngineering. 2013

6. *Construction of cell type-specific logic models of signaling networks using CellNetOptimizer.* M. K. Morris, **I. Melas**, J. Saez-Rodriguez. Methods in Molecular Biology:Computational Toxicology, Ed. B. Reisfeld and A. Mayeno, Humana Press. 2012

7. *Non Linear Programming (NLP) formulation for quantitative modeling of protein signal transduction pathways.* Alexander Mitsos*, **Ioannis N. Melas***, Melody K. Morris, Julio Saez-Rodriguez, Douglas A. Lauffenburger, Leonidas G. Alexopoulos. PLoS One. 2012;7(11):e50085. doi: 10.1371/journal.pone.0050085. Epub 2012 Nov 30 2012, *Equal contributors

8. *Construction of large signaling pathways using an adaptive perturbation approach with phosphoproteomic data.* **Ioannis N. Melas***, Alexander Mitsos*, Dimitris E. Messinis, Thomas S. Weiss, Julio Saez-Rodriguez, and Leonidas G. Alexopoulos, Mol. BioSyst., 10.1039/C2MB05482E, 2012. *Equal Contributors.

9. *Combined logical and data-driven models for linking signaling pathways to cellular response.* **I. Melas***, A. Mitsos*, D. Messinis, T. Weiss, L. Alexopoulos, Bmc Syst Biol 5, 107 (2011).(*equal contributors)

10. *Modeling signaling pathways in articular cartilage.* **Ioannis N. Melas**, Aikaterini D. Chairakaki, Alexander Mitsos, Zoe Dailiana, Christopher G. Provatidis, Leonidas G. Alexopoulos 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2011), Boston, MA, US

11. *Construction of Pathways and identification of drug effects in liver cancer cells via an Integer Linear Programming (ILP) formulation.* Leonidas G Alexopoulos, **Ioannis N. Melas**, Aikaterini D. Chairakaki, Julio Saez-Rodriguez, Alexander Mitsos. 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2010, Buenos Aires, Argentina

12. *Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data.* Alexander Mitsos, **Ioannis N. Melas**, Paraskevas S Siminelakis, Aikaterini D. Chairakaki, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. PloS Comp Biol. 2009. 5(12): e1000591. doi:10.1371/journal.pcbi.1000591

A total of 24 abstracts/short papers were published to refereed conferences, these are the following (sorted by publication date)

Refereed abstracts/conferences

1. Network analysis of signaling in chondrocytes Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Zoe H. Dailiana, Christopher Provatidis, Leonidas G. Alexopoulos. European Society of Biomechanics 2013.

2. Real time evaluation of the mechanical properties of articular cartilage during collagenase induced digestion. Evangelos Zympeloudis, Elisavet I. Chatzopoulou, Ioannis N. Melas, Zoe H. Dailiana, Christopher Provatidis, Leonidas G. Alexopoulos. European Society of Biomechanics 2013.

3. A device for simultaneous compression measurements of human cartilage. Nikolaos V. Georgiou , Nikolaos D. Nikolaou, Panagiotis D. Alevras, Elisavet I. Chatzopoulou, Ioannis N. Melas, Christopher G. Provatidis, Leonidas G. Alexopoulos European Society of Biomechanics 2013.

4. A comparison of chondrocyte response in cartilage 2d cultures. Elisavet I. Chatzopoulou, Ioannis N. Melas, Christoforos G. Provatidis , Zoe Dailiana, Leonidas G. Alexopoulos European Society of Biomechanics 2013.

5. Modeling signaling pathways in chondrocytes. Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Zoe Dailiana, Christopher Provatidis, Leonidas G. Alexopoulos. Fifth International Conference on Computational Bioengineering in Leuven, Belgium. 2013. Submitted

6. A Non Linear Programming (NLP) formulation for modeling signaling transduction pathways, Ioannis N. Melas*, Alexander Mitsos*, Melody K. Morris, Julio Saez-Rodriguez, Douglas A. Lauffenburger, Leonidas G. Alexopoulos (*eq contributors). 4th Hellenic Conference on Biomedical Technology 2012, Athens, Greece.

7. Monitoring cartilage degeneration in a high throughput format via its mechanical properties, Nikolaos Nikolaou, Panagiotis Alevras, Elisavet Chatzopoulou, Ioannis Melas, Zoe Dailiana, Christoforos Provatidis, Leonidas Alexopoulos. 4th Hellenic Conference on Biomedical Technology 2012, Athens, Greece.

8. Optimization of a large scale signaling network using an Integer Linear Programming formulation. Ioannis N. Melas*, Alexander Mitsos*, Julio Saez-Rodriguez, Leonidas G Alexopoulos, *Equal Contributors. 7th GRACM International Congress on Computational Mechanics 2011, Athens, Greece,

9. A device for multiple indentation tests of human cartilage. Nikolaos D. Nikolaou, Panagiotis D. Alevras, Ioannis N. Melas, Christopher P. Provatidis, Leonidas G Alexopoulos. 7th GRACM International Congress on Computational Mechanics 2011, Athens, Greece

10. Construction of large signaling pathways from phosphoproteomic data using an ILP computational approach, Dimitris E. Messinis, Ioannis N. Melas, Alexander Mitsos, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. 6th conference of the Hellenic Society for Computational Biology and Bioinformatics, 2011, Patra, Greece.

11. Construction of large signaling pathways from phosphoproteomic data, Dimitris E. Messinis, Ioannis N. Melas, Alexander Mitsos, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. Planet xMap Congress 2011, Vienna, Austria.

12. Construction of large signaling pathways from phosphoproteomic data, Leonidas G. Alexopoulos, Ioannis N. Melas, Julio Saez-Rodriguez, Thomas S. Weiss, Dimitris E. Messinis, Alexander Mitsos. ICSB, Heidelberg and Mannheim, 2011.

13. Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data. Leonidas G. Alexopoulos, Ioannis N. Melas, Aikaterini D. Chairakaki, Julio Saez-Rodriguez, Alexander Mitsos., BIO-IT world conference 2010, Boston, USA.

14. Extending logical models of signaling pathways to predict cytokine release. Ioannis N. Melas, Alexander Mitsos, Thomas S. Weiss, Leonidas G Alexopoulos. 10th International Conference on Systems Biology 2010, Edinburgh, Scotland.

15. A systems biology approach to modeling signaling pathways in cartilage degeneration. Ioannis N. Melas, Aikaterini D. Chairakaki, Zoe Dailiana, Panagoula Kolia, Leonidas G Alexopoulos. Gordon Research Conference on Musculoskeletal Biology & Bioengineering 2010, Andover NH, USA

16. Identifying Drug effects on HepG2 cells via pathway alterations using an Integer Linear Programming (ILP) formulation. Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Julio Saez-Rodriguez, Dimitris E. Messinis, Danai Kirli-Florou, Leonidas G. Alexopoulos. European Association for the study of the Liver monothematic conference, signaling in the liver 2010, Amsterdam, Netherlands

17. Biomechanical and Systems Biology Approach for Modelling Cartilage Degeneration. I.N. Melas, A. Chairakaki, A. Mitsos, C.P. Provatidis, Z. Dailiana, P. Kolia, L.G. Alexopoulos. 4th Conference of the Helenic Society of Biomechanics 2010, Athens, Greece

18. Linking signaling pathways to cellular behavior using proteomic data. Ioannis N. Melas, Alexander Mitsos, Thomas S. Weiss, Leonidas G Alexopoulos. Planet xMap Congress 2010, Vienna, Austria

19. Crosstalk between EGF, HGF and Insulin Signaling in Hepatocytes: A logical modeling approach. Regina Samaga, Ioannis N. Melas, Leonidas G. Alexopoulos, Steffen Klamt. Conference on Systems Biology of Mammalian Cells, SBMC 2010, Freiburg, Germany

20. Systems biology approach and high-throughput proteomic analysis identifies Toll-Like-Receptor activators as major players of cartilage degeneration. Leonidas G Alexopoulos, Aikaterini D. Chairakaki, Ioannis N. Melas, Christopher P Provatidis, Panagoula Kolia, Zoe Dailiana. Gordon Research Conference on Musculoskeletal Biology & Bioengineering 2010, Andover NH, USA

21. Systems biology approach and high throughput proteomic analysis identifies Toll-Like-Receptor acivators as major players of cartilage degeneration. Leonidas G Alexopoulos, Aikaterini D. Chairakaki, Ioannis N. Melas, Christopher P. Provatidis, Panagoula Kolia, Zoe Dailiana. OARSI conference 2010, Brussels, Belgium.

22. Drug Effects via Pathway Alterations using Integer Linear Programming Optimization on Phosphoproteomic Data. A Mitsos, IN Melas, P Siminelakis, AD Chairakaki, J Saez-Rodriguez, LG Alexopoulos. 13th Annual International Conference on Research in Computational Molecular Biology RECOMB 2009, Boston, USA

23. Drug Effects Identification using an Integer Linear Programming Optimization Formulation Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Julio Saez- Rodriguez, Leonidas G. Alexopoulos. Conference of Hellenic Society of Biomechanics and Systems Biology 2009, Ioannina, Greece

24. Systems Biology for phosphoproteomic-based Drug Targeting, Efficacy, and Safety. Aikaterini D. Chairakaki, Ioannis N. Melas, Georgios Manikis, Paraskevas S Siminelakis, Steffen Klamt, Alexander Mitsos, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. International Greek Biotechnology Forum, 2009, Athens, Greece.

Two awards were won for work presented in this thesis. These are the following (sorted by date)

Awards

1. "Best Poster First Prize" 9th Planet xMAP congress. 2011

2. "Best Practice award" Bio-IT world conference and expo. 2010

Chapter 8

Bibliography

[1] Hiroaki Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002.

[2] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson. *Molecular biology of the cell, 3rd edition.* 1994.

[3] Julian Downward. The ins and outs of signalling. *nature*, 411:759–762, 2001.

[4] Akhilesh Pandey and Matthias Mann. Proteomics to study genes and genomes. *Nature*, 405:837–846, 2000.

[5] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1):D535–D539, 2006.

[6] Alex Eccleston and Ritu Dhand. Signalling in cancer. *nature*, 411:423, 2006.

[7] Daniel Kirouac, Julio Saez-Rodriguez, Jennifer Swantek, John Burke, Douglas Lauffenburger, and Peter Sorger. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC Systems Biology*, 6(1):29, 2012.

[8] Bree B. Aldridge, John M. Burke, Douglas A. Lauffenburger, and Peter K. Sorger. Physicochemical modelling of cell signalling pathways. *Nat Cell Biol*, 8:1195–1203, 2006.

[9] Melody K. Morris, Julio Saez-Rodriguez, Peter K. Sorger, and Douglas A. Lauffenburger. Logic-based models for the analysis of cell signaling networks. *Biochemistry*, 49(15):3216–3224, 2010. PMID: 20225868.

[10] Melody K. Morris, Julio Saez-Rodriguez, David C. Clarke, Peter K. Sorger, and Douglas A. Lauffenburger. Training signaling pathway maps to biochemical data with constrained fuzzy logic: Quantitative analysis of liver cell responses to inflammatory stimuli. *PLoS Comput Biol*, 7(3):e1001099, 03 2011.

[11] Julio Saez-Rodriguez, Leonidas G. Alexopoulos, and Gustavo Stolovitzky. Setting the standards for signal transduction research. *Sci. Signal.*, 4(160):pe10, 2011.

[12] Leonidas G. Alexopoulos, Julio Saez Rodriguez, Benjamin D. Cosgrove, Douglas A. Lauffenburger, and Peter K. Sorger. Networks inferred from biochemical data reveal profound differences in tlr and inflammatory signaling between normal and transformed hepatocytes. *Molecular and Cellular Proteomics*, 2010.

[13] Julio Saez-Rodriguez, Leonidas G. Alexopoulos, Jonathan Epperlein, Regina Samaga, Douglas A. Lauffenburger, Steffen Klamt, and Peter K Sorger. Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular Systems Biology*, 5, 2009.

[14] RC Alves, D Alves, B Guz, C Matos, M Viana, M Harriz, D Terrabuio, M Kondo, O Gampel, and P Polletti. Advanced hepatocellular carcinoma. review of targeted molecular drugs. *Annals of Hepatology*, 10(1):21–27, 2011.

[15] Ahmedin Jemal, Freddie Bray, Melissa M. Center, Jacques Ferlay, Elizabeth Ward, and David Forman. Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61(2):69–90, 2011.

[16] Minoru Fujimori, Takashi Tokino, Okio Hino, Tomoyuki Kitagawa, Takayuki Imamura, Eizo Okamoto, Masao Mitsunobu, Takashi Ishikawa, Hitoshi Nakagama, Hideharu Harada, Michiyasu Yagura, Kenichi Matsubara, and Yusuke Nakamura. Allelotype study of primary hepatocellular carcinoma. *Cancer Research*, 51(1):89–93, 1991.

[17] Hitoshi Tsuda, Tatsuya Oda, Michiie Sakamoto, and Setsuo Hirohashi. Different pattern of chromosomal allele loss in multiple hepatocellular carcinomas as evidence of their multifocal origin. *Cancer Research*, 52(6):1504–1509, 1992.

[18] S Whittaker, R Marais, and A X Zhu. The role of signaling pathways in the development and treatment of hepatocellular carcinoma. *Oncogene*, 29:4989–5005, 2010.

[19] Francis Berenbaum. Signaling transduction: target in osteoarthritis. *Current Opinion in Rheumatology*, 16(5):616–622, 2004.

[20] Sajal Chakraborti, Malay Mandal, Sudip Das, Amritlal Mandal, and Tapati Chakraborti. Regulation of matrix metalloproteinases: An overview. *Molecular and Cellular Biochemistry*, 253:269–285, 2003. 10.1023/A:1026028303196.

[21] AS Baldwin. The nf-kappa b and i kappa b proteins: New discoveries and insights. *ANNUAL REVIEW OF IMMUNOLOGY*, 14:649–683, 1996.

[22] F Guilak, B Fermor, FJ Keefe, VB Kraus, SA Olson, DS Pisetsky, LA Setton, and JB Weinberg. The role of biomechanics and inflammation in cartilage injury and repair. *Clin Orthop Relat Res*, 423:17–26, 2004.

[23] Alexander Mitsos, Ioannis N. Melas, Paraskeuas Siminelakis, Aikaterini D. Chairakaki, Julio Saez-Rodriguez, and Leonidas G. Alexopoulos. Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS Comput Biol*, 5(12):e1000591, 12 2009.

[24] Rob M Ewing, Peter Chu, Fred Elisma, Hongyan Li, Paul Taylor, Shane Climie, Linda McBroom-Cerajewski, Mark D Robinson, Liam O'Connor, Michael Li, Rod Taylor, Moyez Dharsee, Yuen Ho, Adrian Heilbut, Lynda Moore, Shudong Zhang, Olga Ornatsky, Yury V Bukhman, Martin Ethier, Yinglun Sheng, Julian Vasilescu, Mohamed Abu Farha, Jean-Philippe Lambert, Henry S Duewel, Ian I Stewart, Bonnie Kuehl, Kelly Hogue, Karen Colwill, Katharine Gladwish, Brenda Muskat, Robert Kinach, Sally-Lin Adams, Michael F Moran, Gregg B Morin, Thodoros Topaloglou, and Daniel Figeys. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*, 3, 2007.

[25] Jean-Francois Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amelie Dricot, Ning Li, Gabriel F. Berriz, Francis D. Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S. Goldberg, Lan V. Zhang, Sharyl L. Wong, Giovanni Franklin, Siming Li, Joanna S. Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S. Sikorski, Jean Vandenhaute, Huda Y. Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E. Cusick, David E. Hill, Frederick P. Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173, 1178, 2005.

[26] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[27] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G.R. Gopinath, G.R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl 1):D428–D432, 2005.

[28] Ethan G. Cerami, Benjamin E. Gross, Emek Demir, Igor Rodchenkov, Ozgun Babur, Nadia Anwar, Nikolaus Schultz, Gary D. Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl 1):D685–D690, 2011.

[29] Albert-Laszlo Barabasi and Zoltan N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5:101,113, 2004.

[30] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, February 2006.

[31] Bree B. Aldridge, Julio Saez-Rodriguez, Jeremy L. Muhlich, Peter K. Sorger, and Douglas A. Lauffenburger. Fuzzy logic analysis of kinase pathway crosstalk in tnf/egf/insulin-induced signaling. *PLoS Comput Biol*, 5(4):e1000340, 04 2009.

[32] Regina Samaga, Julio Saez-Rodriguez, Leonidas G. Alexopoulos, Peter K. Sorger, and Steffen Klamt. The logic of egfr/erbb signaling: Theoretical properties and analysis of high-throughput data. *PLoS Comput Biol*, 5(8):e1000438, 08 2009.

[33] Steffen Klamt, Julio Saez-Rodriguez, and Ernst Gilles. Structural and functional analysis of cellular networks with cellnetanalyzer. *BMC Systems Biology*, 1(1):2, 2007.

[34] William Bosl. Systems biology by the rules: hybrid intelligent systems for pathway modeling and discovery. *BMC Systems Biology*, 1(1):13, 2007.

[35] Birgit Schoeberl, Claudia Eichler-Jonsson, Ernst Dieter Gilles, and Gertraud Muller. Computational modeling of the dynamics of the map kinase cascade activated by surface and internalized egf receptors. *Nat Biotech*, 20:370–375, 2004.

[36] Minh Quach, Nicolas Brunel, and Florence d'Alche Buc. Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. *Bioinformatics*, 23(23):3209–3216, 2007.

[37] Peng Qiu and S.K. Plevritis. Reconstructing directed signed gene regulatory network from microarray data. *Biomedical Engineering, IEEE Transactions on*, 58(12):3518 –3521, dec. 2011.

[38] Boris N. Kholodenko, Oleg V. Demin, Gisela Moehren, and Jan B. Hoek. Quantification of short term signaling by the epidermal growth factor receptor. *Journal of Biological Chemistry*, 274(42):30169–30181, 1999.

[39] Y. Chu, A. Jayaraman, and J. Hahn. Parameter sensitivity analysis of il-6 signalling pathways. *IET Systems Biology*, 1(6):342–352, 2007.

[40] I-Chun Chou and Eberhard O. Voit. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical Biosciences*, 219(2):57 – 83, 2009.

[41] SA Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3), 03 1969.

[42] J Thakar, M Pilione, G Kirimanjeswara, ET Harvill, and Albert R. Modeling systems-level regulation of host immune responses. *PLoS Comput Biol.*, 3:e109, 2007.

[43] Tomáš Helikar, John Konvalina, Jack Heidel, and Jim A. Rogers. Emergent decision-making in biological signal transduction networks. *Proceedings of the National Academy of Sciences*, 105(6):1913–1918, 2008.

[44] L Calzone, L Tournier, S Fourquet, D Thieffry, B Zhivotovsky, E Barillot, and A. Zinovyev. Mathematical modelling of cell-fate decision in response to death receptor engagement. *PLoS Comput Biol.*, 6:e1000702, 2010.

[45] Steven Watterson, Stephen Marshall, and Peter Ghazal. Logic models of pathway biology. *Drug Discovery Today*, 13:447 – 456, 2008.

[46] Ioannis N Melas, Alexander Mitsos, Dimitris Messinis, Thomas Weiss, Julio Saez-Rodriguez, and Leonidas G. Alexopoulos. Construction of large signaling pathways using an adaptive perturbation approach with phosphoproteomic data. *Molecular Biosystems*, 2012.

[47] Kunal Aggarwal and H. Kelvin Lee. Functional genomics and proteomics as a foundation for systems biology. *Briefings in Functional Genomics & Proteomics*, 2(3):175–184, 2003.

[48] J. Perren Cobb, Michael N. Mindrinos, Carol Miller-Graziano, Steve E. Calvano, Henry V. Baker, Wenzhong Xiao, Krzysztof Laudanski, Bernard H. Brownstein, Constance M. Elson, Douglas L. Hayden, David N. Herndon, Stephen F. Lowry, Ronald V. Maier, David A. Schoenfeld, Lyle L. Moldawer, Ronald W. Davis, Ronald G. Tompkins, Inflammation, and Host Response to Injury Large-Scale Collaborative Research Program. Application of genome-wide expression analysis to human health and disease. *Proceedings of the National Academy of Sciences of the United States of America*, 102(13):4801–4806, 2005.

[49] Bjorn Schwanhausser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473:337,342, 2011.

[50] Julio Saez-Rodriguez, Arthur Goldsipe, Jeremy Muhlich, Leonidas G. Alexopoulos, Bjorn Millard, Douglas A. Lauffenburger, and Peter K. Sorger. Flexible informatics for linking experimental data to mathematical models via datarail. *Bioinformatics*, 24(6):840–847, 2008.

[51] Ioannis Melas, Alexander Mitsos, Dimitris Messinis, Thomas Weiss, and Leonidas Alexopoulos. Combined logical and data-driven models for linking signalling pathways to cellular response. *BMC Systems Biology*, 5(1):107, 2011.

[52] Feng Zhou, Maureen N. Ajuebor, Paul L. Beck, Tai Le, Cory M. Hogaboam, and Mark G. Swain. Cd154–cd40 interactions drive hepatocyte apoptosis in murine fulminant hepatitis. *Hepatology*, 42(2):372–380, 2005.

[53] Eugene C. Butcher. Can cell systems biology rescue drug discovery? *Nat Rev Drug Discov*, 4(6):461–467, 06 2005.

[54] David M. Goldstein, Nathanael S. Gray, and Patrick P. Zarrinkar. High-throughput kinase profiling as a platform for drug discovery. *Nat Rev Drug Discov*, 7(5):391–397, 05 2008.

[55] Miles A Fabian, William H Biggs, Daniel K Treiber, Corey E Atteridge, Mihai D Azimioara, Michael G Benedetti, Todd A Carter, Pietro Ciceri, Philip T Edeen, Mark Floyd, Julia M Ford, Margaret Galvin, Jay L Gerlach, Robert M Grotzfeld, Sanna Herrgard, Darren E Insko, Michael A Insko, Andiliy G Lai, Jean-Michel Lelias, Shamal A Mehta, Zdravko V Milanov, Anne Marie Velasco, Lisa M Wodicka, Hitesh K Patel, Patrick P Zarrinkar, and David J Lockhart. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat Biotech*, 23(3):329–336, 03 2005.

[56] Kevin A. Janes, John G. Albeck, Lili X. Peng, Peter K. Sorger, Douglas A. Lauffenburger, and Michael B. Yaffe. A high-throughput quantitative multiplex kinase assay for monitoring information flow in signaling networks: Application to sepsis-apoptosis. *Molecular & Cellular Proteomics*, 2(7):463–473, 2003.

[57] Enrico Missner, Inke Bahr, Volker Badock, Ulrich Lücking, Gerhard Siemeister, and Peter Donner. Off-target decoding of a multitarget kinase inhibitor by chemical proteomics. *ChemBioChem*, 10(7):1163–1174, 2009.

[58] Steven E. Hall. Chemoproteomics-driven drug discovery: addressing high attrition rates. *Drug Discovery Today*, 11:495 – 502, 2006.

[59] Ramesh Ramanathan, Chandra Belani, Deepti Singh, Michael Tanaka, Heinz-Josef Lenz, Yun Yen, Hedy Kindler, Syma Iqbal, Jeff Longmate, Philip Mack, Georg Lurje, Regina Gandour-Edwards, Janet Dancey, and David Gandara. A phase ii study of lapatinib in patients with advanced biliary tree and hepatocellular cancer. *Cancer Chemotherapy and Pharmacology*, 64:777–783, 2009. 10.1007/s00280-009-0927-7.

[60] Melanie B. Thomas, Romil Chadha, Katrina Glover, Xuemei Wang, Jeffrey Morris, Thomas Brown, Asif Rashid, Janet Dancey, and James L. Abbruzzese. Phase 2 study of erlotinib in patients with unresectable hepatocellular carcinoma. *Cancer*, 110(5):1059–1067, 2007.

[61] Eduardo Schiffer, Chantal Housset, Wulfran Cacheux, Dominique Wendum, Christele Desbois-Mouthon, Colette Rey, Francois Clergue, Raoul Poupon, Veronique Barbu, and Olivier Rosmorduc. Gefitinib, an egfr inhibitor, prevents hepatocellular carcinoma development in the rat liver with cirrhosis. *Hepatology*, 41(2):307–314, 2005.

[62] Bernard Escudier, Tim Eisen, Walter M. Stadler, Cezary Szczylik, Stephane Oudard, Michael Siebels, Sylvie Negrier, Christine Chevreau, Ewa Solska, Apurva A. Desai, Frederic Rolland, Tomasz Demkow, Thomas E. Hutson, Martin Gore, Scott Freeman, Brian

Schwartz, Minghua Shan, Ronit Simantov, and Ronald M. Bukowski. Sorafenib in advanced clear-cell renal-cell carcinoma. *New England Journal of Medicine*, 356(2):125–134, 2007.

[63] Kegg: Kyoto encyclopedia of genes and genomes. www.genome.jp/kegg/.

[64] Pathway commons. www.pathwaycommons.org.

[65] Mazen W Karaman, Sanna Herrgard, Daniel K Treiber, Paul Gallant, Corey E Atteridge, Brian T Campbell, Katrina W Chan, Pietro Ciceri, Mindy I Davis, Philip T Edeen, Raffaella Faraoni, Mark Floyd, Jeremy P Hunt, Daniel J Lockhart, Zdravko V Milanov, Michael J Morrison, Gabriel Pallares, Hitesh K Patel, Stephanie Pritchard, Lisa M Wodicka, and Patrick P Zarrinkar. A quantitative analysis of kinase inhibitor selectivity. *Nat Biotech*, 26(1):127–132, 01 2008.

[66] Scott M. Wilhelm, Christopher Carter, LiYa Tang, Dean Wilkie, Angela McNabola, Hong Rong, Charles Chen, Xiaomei Zhang, Patrick Vincent, Mark McHugh, Yichen Cao, Jaleel Shujath, Susan Gawlak, Deepa Eveleigh, Bruce Rowley, Li Liu, Lila Adnane, Mark Lynch, Daniel Auclair, Ian Taylor, Rich Gedrich, Andrei Voznesensky, Bernd Riedl, Leonard E. Post, Gideon Bollag, and Pamela A. Trail. Bay 43-9006 exhibits broad spectrum oral antitumor activity and targets the raf/mek/erk pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis. *Cancer Research*, 64(19):7099–7109, 2004.

[67] Birgit Schoeberl, Emily A. Pace, Jonathan B. Fitzgerald, Brian D. Harms, Lihui Xu, Lin Nie, Bryan Linggi, Ashish Kalra, Violette Paragas, Raghida Bukhalid, Viara Grantcharova, Neeraj Kohli, Kip A. West, Magdalena Leszczyniecka, Michael J. Feldhaus, Arthur J. Kudla, and Ulrik B. Nielsen. Therapeutically targeting erbb3: A key node in ligand-induced activation of the erbb receptor-pi3k axis. *Sci. Signal.*, 2(77):ra31, 2009.

[68] Kevin A Janes and Douglas A Lauffenburger. A biological approach to computational models of proteomic networks. *Current Opinion in Chemical Biology*, 10(1):73 – 80, 2006. <ce:title>Proteomics and genomics</ce:title>.

[69] Trey Ideker and Douglas Lauffenburger. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends in Biotechnology*, 21(6):255 – 262, 2003.

[70] B.D. Cosgrove, L.G. Alexopoulos, J. Saez-Rodriguez, L.G. Griffith, and D.A. Lauffenburger. A multipathway phosphoproteomic signaling network model of idiosyncratic drug- and inflammatory cytokine-induced toxicity in human hepatocytes. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 5452–5455, Sept.

[71] Kevin A. Janes and Michael B. Yaffe. Data-driven modelling of signal-transduction networks. *Nat Rev Mol Cell Biol*, 7(11):820–828, 11 2006.

[72] Suzanne Gaudet, Kevin A. Janes, John G. Albeck, Emily A. Pace, Douglas A. Lauffenburger, and Peter K. Sorger. A compendium of signals and responses triggered by prodeath and prosurvival cytokines. *Molecular & Cellular Proteomics*, 4(10):1569–1590, 2005.

[73] Kevin A. Janes, John G. Albeck, Suzanne Gaudet, Peter K. Sorger, Douglas A. Lauffenburger, and Michael B. Yaffe. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science*, 310(5754):1646–1653, 2005.

[74] Shih Peng, David Wong, Kai Tung, Yan Chen, Chun Chao, Chien Peng, Yung Chuang, and Chuan Tang. Computational modeling with forward and reverse engineering links signaling network and genomic regulatory responses: Nf-kappab signaling-induced gene expression responses in inflammation. *BMC Bioinformatics*, 11(1):308, 2010.

[75] Hashem B. El-Serag and K. Lenhard Rudolph. Hepatocellular carcinoma: Epidemiology and molecular carcinogenesis. *Gastroenterology*, 132(7):2557 – 2576, 2007.

[76] Hidekazu Nakabayashi, Kazuhisa Taketa, Keiko Miyano, Takashi Yamane, and Jiro Sato. Growth of human hepatoma cell lines with differentiated functions in chemically defined medium. *Cancer Research*, 42(9):3858–3863, 1982.

[77] Toshitkatsu Hanada and Akihiko Yoshimura. Regulation of cytokine signaling and inflammation. *Cytokine & Growth Factor Reviews*, 13:413 – 421, 2002. <ce:title>Cytokines in Autoimmune Disease</ce:title>.

[78] Gary S. Firestein. Evolving concepts of rheumatoid arthritis. *Nature*, 423(6937):356–361, 05 2003.

[79] Mark A. Birrell, Elizabeth Hardaker, Sissie Wong, Kerryn McCluskie, Matthew Catley, Jorge De Alba, Robert Newton, Saleem Haj-Yahia, K. Tao Pun, Clarissa J. Watts, Robert J. Shaw, Tony J. Savage, and Maria G. Belvisi. Ik-b kinase-2 inhibitor blocks inflammation in human airway smooth muscle and a rat model of asthma. *American Journal of Respiratory and Critical Care Medicine*, 172(8):962–971, 2005.

[80] Akihiko Yoshimura. Signal transduction of inflammatory cytokines and tumor development. *Cancer Science*, 97(6):439–447, 2006.

[81] Sergei I. Grivennikov, Florian R. Greten, and Michael Karin. Immunity, inflammation, and cancer. *Cell*, 140(6):883 – 899, 2010.

[82] Marcin Kortylewski, Maciej Kujawski, Tianhong Wang, Sheng Wei, Shumin Zhang, Shari Pilon-Thomas, Guilian Niu, Heidi Kay, James Mule, William G Kerr, Richard Jove, Drew Pardoll, and Hua Yu. Inhibiting stat3 signaling in the hematopoietic system elicits multicomponent antitumor immunity. *Nat Med*, 11(12):1314–1321, 12 2005.

[83] Beverly A. Teicher. Transforming growth factor-b and the immune response to malignant disease. *Clinical Cancer Research*, 13(21):6247–6251, 2007.

[84] Tianhong Wang, Guilian Niu, Marcin Kortylewski, Lyudmila Burdelya, Kenneth Shain, Shumin Zhang, Raka Bhattacharya, Dmitry Gabrilovich, Richard Heller, Domenico Coppola, William Dalton, Richard Jove, Drew Pardoll, and Hua Yu. Regulation of the innate and adaptive immune responses by stat-3 signaling in tumor cells. *Nat Med*, 10(1):48–54, 01 2004.

[85] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57 – 70, 2000.

[86] A Mitsos, IN Melas, MK Morris, J Saez-Rodriguez, DA Lauffenburger, and LG Alexopoulos. Non linear programming (nlp) formulation for quantitative modeling of protein signal transduction pathways. *PLoS ONE*, 7:e50085, 2012.

[87] Andreas Wachter and Lorenz T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.*, 106(1):25–57, May 2006.

[88] JV Beck and JA Kenneth. *Parameter estimation in engineering and science*. John Wiley & Sons Inc, 1977.

[89] Federica Eduati, Javier De Las Rivas, Barbara Di Camillo, Gianna Toffolo, and Julio Saez-Rodriguez. Integrating literature-constrained and data-driven inference of signalling networks. *Bioinformatics*, 28(18):2311–2317, 2012.

[90] Aidan MacNamara, Camille Terfve, David Henriques, Beatriz Peñalver Bernabé, and Julio Saez-Rodriguez. State–time spectrum of signal transduction logic models. *Physical Biology*, 9(4):045003, 2012.

[91] I.N. Melas, A.D. Chairakaki, A. Mitsos, Z. Dailiana, C.G. Provatidis, and L.G. Alexopoulos. Modeling signaling pathways in articular cartilage. In *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*, pages 3712–3715, 30 2011-Sept. 3.

[92] Frank P. Luyten, Przemko Tylzanowski, and Rik J. Lories. Wnt signaling and osteoarthritis. *Bone*, 44(4):522 – 527, 2009.

[93] Rik Lories and Frank Luyten. Bone morphogenetic proteins in destructive and remodeling arthritis. *Arthritis Research & Therapy*, 9(2):207, 2007.

[94] Andrew L Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4:682 – 690, 2008.

[95] Ali Mobasheri. Applications of proteomics to osteoarthritis, a musculoskeletal disease characterized by aging. *frontiers in physiology*, 2:108, 2011.

[96] Richard Wilson, John M. Whitelock, and John F. Bateman. Proteomics makes progress in cartilage and arthritis research. *Matrix Biology*, 28(3):121 – 128, 2009.

[97] Daichi Shigemizu, Zhenjun Hu, Jui-Hung Hung, Chia-Ling Huang, Yajie Wang, and Charles DeLisi. Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer. *PLoS Comput Biol*, 8(2):e1002347, 02 2012.

[98] Benjamin A. Garcia, Mark D. Platt, Timothy L. Born, Jeffrey Shabanowitz, Norman A. Marcus, and Donald F. Hunt. Protein profile of osteoarthritic human articular cartilage using tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 20(20):2999–3006, 2006.

[99] Jiang Wu, Wei Liu, Amanda Bemis, Eunice Wang, Yongchang Qiu, Elisabeth A. Morris, Carl R. Flannery, and Zhiyong Yang. Comparative proteomic characterization of articular cartilage tissue from normal donors and patients with osteoarthritis. *Arthritis & Rheumatism*, 56(11):3675–3684, 2007.

[100] Vincourt Jean-Baptiste, Lionneton Frederic, Kratassiouk Gueorgui, Guillemin Francois, Netter Patrick, Mainard Didier, and Magdalou Jacques. Establishment of a reliable method for direct proteome characterization of human articular cartilage. *Mol Cell Proteomics*, 5:1984–95, 2006.

[101] Lambrecht S., Verbruggen G., Verdonk P.C.M., Elewaut D., and Deforce D. Differential proteome analysis of normal and osteoarthritic chondrocytes reveals distortion of vimentin network in osteoarthritis. *Osteoarthritis and Cartilage*, 16(2):163 – 173, 2008.

[102] Andrea Sinz, Marcus Bantscheff, Stefan Mikkat, Bruno Ringel, Susanne Drynda, Jorn Kekow, Hans-Jurgen Thiesen, and Michael O. Glocker. Mass spectrometric proteome analyses of synovial fluids and plasmas from patients suffering from rheumatoid arthritis and comparison to reactive arthritis or osteoarthritis. *ELECTROPHORESIS*, 23(19):3445–3456, 2002.

[103] Hua Liao, Jiang Wu, Eric Kuhn, Wendy Chin, Betty Chang, Michael D. Jones, Steve O'Neil, Karl R. Clauser, Johann Karl, Fritz Hasler, Ronenn Roubenoff, Werner Zolg, and Brad C. Guild. Use of mass spectrometry to identify protein biomarkers of disease severity in the synovial fluid and serum of patients with rheumatoid arthritis. *Arthritis & Rheumatism*, 50(12):3792–3803, 2004.

[104] J. B. Catterall, A. D. Rowan, S. Sarsfield, J. Saklatvala, R. Wait, and T. E. Cawston. Development of a novel 2d proteomics approach for the identification of proteins secreted by primary chondrocytes after stimulation by il-1 and oncostatin m. *Rheumatology*, 45(9):1101–1109, 2006.

[105] Lisbet Haglund, Suzanne M. Bernier, Patrik IDnnerfjord, and Anneliese D. Recklies. Proteomic analysis of the lps-induced stress response in rat chondrocytes reveals induction of innate immune response components in articular cartilage. *Matrix Biology*, 27(2):107 – 118, 2008.

[106] J Saez-Rodriguez, LG Alexopoulos, J Epperlein, R Samaga, DA Lauffenburger, S Klamt, and PK. Sorger. Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol Syst Biol.*, 5:331, 2009.

[107] Vere Beck James and J. Arnold Kenneth. *Parameter estimation in engineering and science*. Wiley, 1977.

[108] Xibin Wang, Paul A. Manner, Alan Horner, Lillian Shum, Rocky S. Tuan, and Glen H. Nuckolls. Regulation of mmp-13 expression by runx2 and fgf2 in osteoarthritic cartilage. *Osteoarthritis and Cartilage*, 12(12):963 – 973, 2004.

[109] Y. Zheng and A. Rundell. Comparative study of parameter sensitivity analyses of the tcr-activated erk-mapk signalling pathway. *Systems Biology, IEE Proceedings*, 153(4):201 –211, july 2006.

[110] A. Siegel, O. Radulescu, M. Le Borgne, P. Veber, J. Ouy, and S. Lagarrigue. Qualitative analysis of the relation between dna microarray data and behavioral models of regulation networks. *Biosystems*, 84(2):153 – 174, 2006. <ce:title>Dynamical Modeling of Biological Regulatory Networks</ce:title>.

[111] Carito Guziolowski, Annabel Bourde, Francois Moreews, and Anne Siegel. Bioquali cytoscape plugin: analysing the global consistency of regulatory networks. *BMC Genomics*, 10(1):244, 2009.

[112] MARTIN GEBSER, TORSTEN SCHAUB, SVEN THIELE, and PHILIPPE VEBER. Detecting inconsistencies in large biological networks with answer set programming. *Theory and Practice of Logic Programming*, 11:323–360, 2011.

[113] Rosa Maria Gutierrez-Rios, David A. Rosenblueth, Jose Antonio Loza, Araceli M. Huerta, Jeremy D. Glasner, Fred R. Blattner, and Julio Collado-Vides. Regulatory network of escherichia coli: Consistency between literature knowledge and microarray profiles. *Genome Research*, 13:24352443, 2003.

[114] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.

[115] Steffen Klamt, Julio Saez-Rodriguez, Jonathan Lindquist, Luca Simeoni, and Ernst Gilles. A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics*, 7(1):56, 2006.

[116] Rui-Sheng Wang, Assieh Saadatpour, and Reka Albert. Boolean modeling in systems biology: an overview of methodology and applications. *PHYSICAL BIOLOGY*, 9(5), OCT 2012.

[117] R Thomas. quantum noise. *Springer series in Synergetics*, pages 180–193, 1981.

[118] C Soule. Graphic requirements for multistationarity. *ComPlexUs*, pages 123–133, 2003.

[119] Steffen Klamt, Julio Saez-Rodriguez, Jonathan Lindquist, Luca Simeoni, and Ernst Gilles. A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics*, 7(1):56, 2006.

[120] Josep M. Llovet, Sergio Ricci, Vincenzo Mazzaferro, Philip Hilgard, Edward Gane, Jean-Frederic Blanc, Andre Cosme de Oliveira, Armando Santoro, Jean-Luc Raoul, Alejandro Forner, Myron Schwartz, Camillo Porta, Stefan Zeuzem, Luigi Bolondi, Tim F. Greten, Peter R. Galle, Jean-Francois Seitz, Ivan Borbath, Dieter Haussinger, Tom Giannaris, Minghua Shan, Marius Moscovici, Dimitris Voliotis, and Jordi Bruix. Sorafenib in advanced hepatocellular carcinoma. *New England Journal of Medicine*, 359(4):378–390, 2008.

[121] Yujin Hoshida, Augusto Villanueva, Masahiro Kobayashi, Judit Peix, Derek Y. Chiang, Amy Camargo, Supriya Gupta, Jamie Moore, Matthew J. Wrobel, Jim Lerner, Michael Reich, Jennifer A. Chan, Jonathan N. Glickman, Kenji Ikeda, Masaji Hashimoto, Goro Watanabe, Maria G. Daidone, Sasan Roayaie, Myron Schwartz, Swan Thung, Helga B. Salvesen, Stacey Gabriel, Vincenzo Mazzaferro, Jordi Bruix, Scott L. Friedman, Hiromitsu Kumada, Josep M. Llovet, and Todd R. Golub. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *New England Journal of Medicine*, 359(19):1995–2004, 2008.

[122] Yujin Hoshida, Sebastian M.B. Nijman, Masahiro Kobayashi, Jennifer A. Chan, Jean-Philippe Brunet, Derek Y. Chiang, Augusto Villanueva, Philippa Newell, Kenji Ikeda, Masaji Hashimoto, Goro Watanabe, Stacey Gabriel, Scott L. Friedman, Hiromitsu Kumada, Josep M. Llovet, and Todd R. Golub. Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Research*, 69(18):7385–7392, 2009.

[123] Xin Wei Wang and Snorri S. Thorgeirsson. Transcriptome analysis of liver cancer: Ready for the clinic? *Journal of Hepatology*, 50(5):1062 – 1064, 2009.

[124] Ju-Seog Lee, In-Sun Chu, Jeonghoon Heo, Diego F. Calvisi, Zongtang Sun, Tania Roskams, Anne Durnez, Anthony J. Demetris, and Snorri S. Thorgeirsson. Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology*, 40(3):667–676, 2004.

[125] Hyun Goo Woo, Eun Sung Park, Jae Hee Cheon, Ju Han Kim, Ju-Seog Lee, Bum Joon Park, Won Kim, Su Cheol Park, Young Jin Chung, Byeong Gwan Kim, Jung-Hwan Yoon,

Hyo-Suk Lee, Chung Yong Kim, Nam-Joon Yi, Kyung-Suk Suh, Kuhn Uk Lee, In-Sun Chu, Tania Roskams, Snorri S. Thorgeirsson, and Yoon Jun Kim. Gene expression based recurrence prediction of hepatitis b virus related human hepatocellular carcinoma. *Clinical Cancer Research*, 14(7):2056–2064, 2008.

[126] Xin Chen, Siu Tim Cheung, Samuel So, Sheung Tat Fan, Christopher Barry, John Higgins, Kin-Man Lai, Jiafu Ji, Sandrine Dudoit, Irene O.L. Ng, Matt van de Rijn, David Botstein, and Patrick O. Brown. Gene expression patterns in human liver cancers. *Molecular Biology of the Cell*, 13(6):1929–1939, 2002.

[127] Wei Sun, Baocai Xing, Yi Sun, Xiaojuan Du, Min Lu, Chunyi Hao, Zhuang Lu, Wei Mi, Songfeng Wu, Handong Wei, Xue Gao, Yunping Zhu, Ying Jiang, Xiaohong Qian, and Fuchu He. Proteome analysis of hepatocellular carcinoma by two-dimensional difference gel electrophoresis. *Molecular & Cellular Proteomics*, 6(10):1798–1808, October 2007.

[128] Ning Li, Yunzhu Long, Xuegong Fan, Hongbo Liu, Cui Li, Lizhang Chen, and Zhiming Wang. Proteomic analysis of differentially expressed proteins in hepatitis b virus-related hepatocellular carcinoma tissues. *Journal of Experimental & Clinical Cancer Research*, 28(1):122, 2009.

[129] Z Jiang and Y Zhou. Using gene networks to drug target identification. *Journal of Integrative Bioinformatics*, 2005.

[130] P.J O'Dwyer, B.J Giantonio, D.E Levy, J.S Kauh, D.B Fitzgerald, and A.B Benson. Gefitinib in advanced unresectable hepatocellular carcinoma: Results from the eastern cooperative oncology group's study e1203. *Journal of Clinical Oncology, 2006 ASCO Annual Meeting Proceedings Part I*, 24, 2006.

[131] Chiun Hsu, Tsai-Sheng Yang, Teh-Ia Huo, Ruey-Kuen Hsieh, Chih-Wei Yu, Wei-Shou Hwang, Tsai-Yuan Hsieh, Wen-Tsung Huang, Yee Chao, Robin Meng, and Ann-Lii Cheng. Vandetanib in patients with inoperable hepatocellular carcinoma: A phase ii, randomized, double-blind, placebo-controlled study. *Journal of Hepatology*, 56(5):1097 – 1103, 2012.

[132] A. Cheng, Y. Kang, J. Lin, J. Park, M. Kudo, S. Qin, M. Omata, Lowenthal S.W Pitman, S. Lanzalone, L. Yang, M. Lechuga, E. Raymod, SUN1170 HCC Study Group, Taipei Taiwan National Taiwan University Hospital, Seoul South Korea Asan Medical Center, Taoyuan Taiwan Chang Gung Memorial Hospital, Taoyuan Taiwan Chang Gung University, Goyang South Korea Center for Liver Cancer, Goyang South Korea National Cancer Center, Osaka Japan Kinki University Faculty of Medicine, Nanjing China Nanjing Bayi Hospital, Yamanashi Japan Yamanashi Prefecture Central Hospital, New York NY Pfizer Oncology, Pfizer Italia Srl Milan Italy Pfizer Oncology, La Jolla CA Pfizer Oncology, and Clichy France Beaujon University Hospital. Phase iii trial of sunitinib (su) versus sorafenib (so) in advanced hepatocellular carcinoma (hcc). *Journal of Clinical Oncology*, 29, 2011.

[133] Daniele Baiz, Gabriele Pozzato, Barbara Dapas, Rossella Farra, Bruna Scaggiante, Mario Grassi, Laura Uxa, Carlo Giansante, Cristina Zennaro, Gianfranco Guarnieri, and Gabriele Grassi. Bortezomib arrests the proliferation of hepatocellular carcinoma cells hepg2 and jhh6 by differentially affecting e2f1, p21 and p27 levels. *Biochimie*, 91(3):373 – 382, 2009.

[134] Jose Baselga, Ian Bradbury, Holger Eidtmann, Serena Di Cosimo, Evandro de Azambuja, Claudia Aura, Henry Gomez, Phuong Dinh, Karine Fauria, Veerle Van Dooren, Gursel

Aktan, Aron Goldhirsch, Tsai-Wang Chang, Zsolt Horvath, Maria Coccia-Portugal, Julien Domont, Ling-Min Tseng, Georg Kunz, Joo Hyuk Sohn, Vladimir Semiglazov, Guillermo Lerzo, Marketa Palacova, Volodymyr Probachai, Lajos Pusztai, Michael Untch, Richard D Gelber, and Martine Piccart-Gebhart. Lapatinib with trastuzumab for her2-positive early breast cancer (neoaltto): a randomised, open-label, multicentre, phase 3 trial. *The Lancet*, 379 Issue 9816:633–640, 2012.

[135] Ming-Sound Tsao, Akira Sakurada, Jean-Claude Cutz, Chang-Qi Zhu, Suzanne Kamel-Reid, Jeremy Squire, Ian Lorimer, Tong Zhang, Ni Liu, Manijeh Daneshmand, Paula Marrano, Gilda da Cunha Santos, Alain Lagarde, Frank Richardson, Lesley Seymour, Marlo Whitehead, Keyue Ding, Joseph Pater, and Frances A. Shepherd. Erlotinib in lung cancer $\tilde{\text{N}}$ molecular and clinical predictors of outcome. *New England Journal of Medicine*, 353(2):133–144, 2005.

[136] Malcolm J. Moore, David Goldstein, John Hamm, Arie Figer, Joel R. Hecht, Steven Gallinger, Heather J. Au, Pawel Murawa, David Walde, Robert A. Wolff, Daniel Campos, Robert Lim, Keyue Ding, Gary Clark, Theodora Voskoglou-Nomikos, Mieke Ptasynski, and Wendy Parulekar. Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: A phase iii trial of the national cancer institute of canada clinical trials group. *Journal of Clinical Oncology*, 25(15):1960–1966, May 20, 2007.

[137] Jean-Francois Chatal, Francoise Kraeber-Bodere, David M. Goldenberg, and Jacques Barbet. Treatment of metastatic medullary thyroid cancer with vandetanib: Need to stratify patients on basis of calcitonin doubling time. *Journal of Clinical Oncology*, 30(17):2165, 2012.

[138] Jin Soo Lee, Vera Hirsh, Keunchil Park, Shukui Qin, Cesar R. Blajman, Reury-Perng Perng, Yuh-Min Chen, Laura Emerson, Peter Langmuir, and Christian Manegold. Vandetanib versus placebo in patients with advanced non small-cell lung cancer after prior therapy with an epidermal growth factor receptor tyrosine kinase inhibitor: A randomized, double-blind phase iii trial (zephyr). *Journal of Clinical Oncology*, 2012.

[139] Motzer RJ, Rini BI, Bukowski RM, and et al. Sunitinib in patients with metastatic renal cell carcinoma. *JAMA: The Journal of the American Medical Association*, 295(21):2516–2524, 2006.

[140] George D Demetri, Allan T van Oosterom, Christopher R Garrett, Martin E Blackstein, Manisha H Shah, Jaap Verweij, Grant McArthur, Ian R Judson, Michael C Heinrich, Jeffrey A Morgan, Jayesh Desai, Christopher D Fletcher, Suzanne George, Carlo L Bello, Xin Huang, Charles M Baum, and Paolo G Casali. Efficacy and safety of sunitinib in patients with advanced gastrointestinal stromal tumour after failure of imatinib: a randomised controlled trial. *The Lancet*, 368:1329–1338, 2006.

[141] Andreas Hochhaus, Hagop M. Kantarjian, Michele Baccarani, Jeffrey H. Lipton, Jane F. Apperley, Brian J. Druker, Thierry Facon, Stuart L. Goldberg, Francisco Cervantes, Dietger Niederwieser, Richard T. Silver, Richard M. Stone, Timothy P. Hughes, Martin C. Muller, Rana Ezzeddine, Athena M. Countouriotis, and Neil P. Shah. Dasatinib induces notable hematologic and cytogenetic responses in chronic-phase chronic myeloid leukemia after failure of imatinib therapy. *Blood*, 109(6):2303–2309, March 15, 2007.

[142] R. S. Finn, A. Aleshin, D. Rivera, P. Yang, J. Dering, G. Bentley, A. Desai, D. J. Slamon, and R. W. Busuttil. Effect of dasatinib, an orally active small molecule inhibitor of both src and abl kinases, on growth of hepatic progenitor subtype human hepatocellular carcinoma cells in vitro. *Journal of Clinical Oncology*, 2009.

[143] Paul G. Richardson, Pieter Sonneveld, Michael W. Schuster, David Irwin, Edward A. Stadtmauer, Thierry Facon, Jean-Luc Harousseau, Dina Ben-Yehuda, Sagar Lonial, Hartmut Goldschmidt, Donna Reece, Jesus F. San-Miguel, Joan Blade, Mario Boccadoro, Jamie Cavenagh, William S. Dalton, Anthony L. Boral, Dixie L. Esseltine, Jane B. Porter, David Schenkein, and Kenneth C. Anderson. Bortezomib or high-dose dexamethasone for relapsed multiple myeloma. *New England Journal of Medicine*, 352(24):2487–2498, 2005.

[144] Julio Saez-Rodriguez, Leonidas Alexopoulos, MingSheng Zhang, Melody K. Morris, Douglas A. Lauffenburger, and Peter K. Sorger. Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer Research*, 2011.

[145] David C. Clarke, Melody K. Morris, and Douglas A. Lauffenburger. Normalization and statistical analysis of multiplexed bead-based immunoassay data using mixed-effects modeling. *Molecular & Cellular Proteomics*, 12(1):245–262, 2013.

[146] Y Ito, T Takeda, M Sakon, M Tsujimoto, S Higashiyama, K Noda, E Miyoshi, M Monden, and N Matsuura. Expression and clinical significance of erbb receptor family in hepatocellular carcinoma. *Br J Cancer*, 84:1377–1383, 2001.

[147] Yoko Ogawara, Shohei Kishishita, Toshiyuki Obata, Yuko Isazawa, Toshiaki Suzuki, Keiji Tanaka, Norihisa Masuyama, and Yukiko Gotoh. Akt enhances mdm2-mediated ubiquitination and degradation of p53. *Journal of Biological Chemistry*, 277(24):21843–21850, 2002.

[148] Adam Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Favera, and Andrea Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

[149] Patrick Meyer, Frederic Lafitte, and Gianluca Bontempi. minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9(1):461, 2008.

[150] Young-Mi Go, Yong Chool Boo, Heonyong Park, Matthew C. Maland, Rakesh Patel, Kirkwood A. Pritchard, Yasushi Fujio, Kenneth Walsh, Victor Darley-Usmar, and Hanjoong Jo. Protein kinase b/akt activates c-jun nh2-terminal kinase by increasing no production in response to shear stress. *Journal of Applied Physiology*, 91(4):1574–1581, 2001.

[151] Sarah Seton-Rogers. Hepatocellular carcinoma: Putting p53 in context. *Nat Rev Cancer*, 6:423, 2006.

[152] Koichi Matsuo, Seiji Satoh, Hiroshi Okabe, Akinari Nomura, Toshiki Maeda, Yoshio Yamaoka, and Iwao Ikai. Siah1 inactivation correlates with tumor progression in hepatocellular carcinomas. *Genes, Chromosomes and Cancer*, 36(3):283–291, 2003.

[153] Shimian Qu, Jirong Long, Qiuyin Cai, Xiao-Ou Shu, Hui Cai, Yu-Tang Gao, and Wei Zheng. Genetic polymorphisms of metastasis suppressor gene nme1 and breast cancer survival. *Clinical Cancer Research*, 14(15):4787–4793, 2008.

[154] Xi-Ming Xu, Jun-Jian Deng, Guang-Jin Yuan, Fang Yang, Hong-Ting Guo, Miao Xiang, Wei Ge, and Yao-Gui Wu. 5-lipoxygenase contributes to the progression of hepatocellular carcinoma. *Molecular Medicine Reports*, pages 1195–1200, 2011.

[155] Y Midorikawa, S Tsutsumi, H Taniguchi, M Ishii, Y Kobune, T Kodama, M Makuuchi, and H Aburatani. Identification of genes associated with dedifferentiation of hepatocellular carcinoma with expression profiling analysis. *Cancer Research*, 93:636–43, 2002.

[156] J Ai, H Huang, X Lv, Z Tang, M Chen, T Chen, W Duan, H Sun, Q Li, R Tan, Y Liu, J Duan, Y Yang, Y Wei, Y Li, and Q Zhou. Flna and pgk1 are two potential markers for progression in hepatocellular carcinoma. *Cell Physiol Biochem*, 27:207–216, 2011.

[157] ROBERTO GEDALY, PAUL ANGULO, JONATHAN HUNDLEY, MICHAEL F. DAILY, CHANGGUO CHEN, ALVARO KOCH, and B. MARK EVERS. Pi-103 and sorafenib inhibit hepatocellular carcinoma cell proliferation by blocking ras/raf/mapk and pi3k/akt/mtor pathways. *Anticancer Research*, 30(12):4951–4958, December 2010.

[158] Hala Elnakat Thomas, Carol A. Mercer, Larissa S. Carnevalli, Jongsun Park, Jesper B. Andersen, Elizabeth A. Conner, Kazuhiro Tanaka, Tomoo Matsutani, Akio Iwanami, Bruce J. Aronow, Liu Manway, S. Michel Maira, Snorri S. Thorgeirsson, Paul S. Mischel, George Thomas, and Sara C. Kozma. mtor inhibitors synergize on regression, reversal of gene expression, and autophagy in hepatocellular carcinoma. *Science Translational Medicine*, 4(139):139ra84, 2012.

[159] Ute M Moll, Sonja Wolff, Daniel Speidel, and Wolfgang Deppert. Transcription-independent pro-apoptotic functions of p53. *Current Opinion in Cell Biology*, 17(6):631 – 636, 2005. <ce:title>Cell division, growth and death / Cell differentiation</ce:title>.

Appendix A

SigNetTrainer - A toolbox for interrogating and training signaling networks to experimental data

A.1 Introduction

SigNetTrainer toolbox was put together for interrogating and training signaling networks (represented as interaction graphs) to experimental data from stimulus response experiments. The toolbox was implemented by I.N. Melas and uses the Integer Linear Programming (ILP) framework presented in the manuscript entitled "Detecting and Removing Inconsistencies Between Experimental Data and Signaling Network Topologies using Integer Linear Programming on Interaction Graphs". Given an interaction graph topology (stored in a file 'Network.sif') and a set of perturbation experiments (defined in 'inputs.txt') with associated measurements (defined in 'measurements.txt'), basically four different problems can be tackled by *SigNetTrainer*:

- (1) SCEN_FIT: Determine a causal explanation for the measured activation changes of readout nodes for one given perturbation scenario. If the measurements are inconsistent with the network topology, find the closest feasible scenario.
- (2) Minimal Correction Sets (MCoS): In case of an inconsistent scenario, what is a minimal set of nodes that need to be corrected to make a single inconsistent scenario consistent.
- (3) OPT_SUBGRAPH: Determine an optimal subgraph of a given network topology that can fit the measurements for a set of scenarios the best.
- (4) OPT_GRAPH: Identify edge candidate(s) whose insertion (in combination with removal of existing edges) would improve the consistency of the graph with respect to a set of experimental scenarios the most.

The first two optimization problems seek to match the network topology with measurements from a single stimulus-response experiment. In contrast, (3) and (4) operate on a set of scenarios and seek to optimize (train) the network structure over all scenarios, either by removing or by adding edges. For the first three problems *SigNetTrainer* also provides enumeration algorithms to find multiple or all solutions that solve the optimization problem equally well (e.g., for problem (3), all optimal subgraphs that minimize the number of inconsistencies between measurements and predictions).

Note that the manual will describe usage of the available algorithms rather than (re-) describing the theory behind the methods; we will refer to the manuscript mentioned above.

A.2 Files

SigNetTrainer consists of the following processing files.

- **"SigNetTrainer.c"**. C source code; it includes the **main()** function of the toolbox. Based on the type of analysis requested by the user, either runs the ILP formulation (as presented in *{ Melas et al., PLoS CB 2013}*), or parses a list of addable reactions to score them based on how much they would improve the goodness of fit to the data if added to the Prior Knowledge Network (PKN) (OPT_GRAPH problem).
- **"optimize.c"**. C source code; it implements the ILP formulation and employs it to solve problems (1)-(3). It is the function that calls GUROBI optimizer.
- **"import_data.h"**. h file; it includes functions for importing the signaling data (inputs.txt and measurements.txt files).
- **"data_preprocessor.h"**. h file; it includes functions for generating the inputs.txt and measurements.txt files from a tab delimited file, similar to that exported by the Luminex xMAP 200 system.
- **"import_pathway.h"**. h file; it parses the Network.sif file, containing the PKN.
- **"observable_controllable.h"**. h file; includes files for preprocessing the PKN by removing non-observable and non-controllable parts of it. Non-observable are defined all nodes downstream of which there are no measured signals, thus their activation state cannot be inferred. Non-controllable are defined all nodes upstream of which there are no stimuli, thus their activation state cannot be controlled.
- **"warshall.h"**. h file; implements the Floyd-Warshall algorithm for transitive closure. It is used by "observable_controllable.h" to identify the non-observable and non-controllable parts of the pathway.
- **"Makefile"**. C language makefile. It is included in the GUROBI installation. It differs across platforms. The user should edit it to include the path of the GUROBI installation (if that is modified from the default).

A.3 Installation

A.3.1 System Requirements

Requirements: You definitely need to install GUROBI on your computer and to set the path to this library (we used version 4.6.1). We precompiled executable versions for Mac OS X (SigNetTrainerMac) and Linux (SigNetTrainerLinux) and distribute it with the package (but note that GUROBI needs nevertheless to be installed on your system when using these versions).
If you have a different platform or if these versions do not run on your system you need to recompile it as explained in the next section.

A.3.2 Compiling C code

For recompiling the code you need to proceed / prepare the following:

- working C compiler. We compiled the ILP code using gcc version 4.2 (under Mac OS X).
- working GUROBI installation; we used version 4.6.1 (under Mac OS X).
- In file **optimize.c**, line 5, add absolute path to "gurobi_c.h" file (file installed to your system by GUROBI installer). For Mac OS X default path is "/Library/gurobi461/mac64/include/gurobi_c.h".
- In file optimize.c, lines 34-38 add absolute path to "observable_controllable.h", "import_data.h", "data_preprocessor.h", "warshall.h" and "import_pathway.h" files.
- In a terminal, navigate to the folder where the ILP code is located and run "`make SigNetTrainer`". If when installing GUROBI, the installation path was modified, please update the "Makefile" with the new path.

A.3.3 MATLAB files

Network compression (optionally) and the generation of a list of addable edges (required for the OPT_GRAPH problem) needs to be done via the two MATLAB script files "compressIG.m" and "addableEdges.m", respectively. These files, together with two further m-files for reading/writing ".sif" files (see below), are also distributed with *SigNetTrainer*. For using them, you first, have to install *CellNetAnalyzer* which can be downloaded (free for academic use) from
http://www.mpi-magdeburg.mpg.de/projects/cna/cna.html
After installation of *CellNetAnalyzer*, goto to its main directory, start MATLAB and enter "startcna(1)". Change then into the directory of *SigNetTrainer*.

For network compression do the following:

- (1) load the network from a ".sif" file (regarding sif format see section A.4.1) by running: "`cnap=CNAsif2SFNetwork('FullNetwork.sif')`". Here, "FullNetwork.sif" is the name of the file describing the initial (uncompressed) graph.
- (2) use MATLAB cell variables to specifiy the names of the (measured) readout nodes and of the nodes that are perturbed, for example "`readouts={'node1','node4'}`" and "`perturb={'node2','node4'}`".
- (3) Call the actual network compression routine by "`cnapnew=compressIG(cnap,readouts,perturb)`".
- (4) write the new network file which can then be used by *SigNetTrainer*:
  "`CNASFNetwork2sif(cnapnew,'Networknew.sif')`"
  Here, "Networknew.sif" is the name of the compressed file later to be used by *SigNetTrainer*.

For computing the list of addable reactions you need to run "`addableEdges('Network.sif')`". "Network.sif" is the name of the file describing the interaction graph. After running this command, a new file will be written ("addable_reactions.txt") containing all edges that can be considered for OPT_GRAPH (these edges do not induce a positive feedback loop).

A.4 Running the ILP code

The *SigNetTrainer* code may be used either for preprocessing the signaling dataset, or for executing the optimization procedure, depending on the user's input.

A.4.1 Files to be provided/edited by the user

To use the different functions of *SigNetTrainer* the user has to provide/edit four different files:

"ilp_options.txt" (specifies options for the different functions)

- Tab-delimited file
- **significant_increase** (required only for data preprocessing): corresponds to the threshold above which the fold increase of the signal, after versus before stimulation, is considered to be significant.
- **significant_decrease** (required only for data preprocessing): corresponds to the threshold bellow which the fold decrease of the signal, after versus before stimulation, is considered to be significant.
- **noise_threshold** (required only for data preprocessing): an absolute value in fluorescent units, bellow which the signal is considered to be insignificant.
- **mipgap**: the relative optimality GAP for the ILP solver (default value is 1E-06, smallest value is 1E-09)
- **timelimit**: Maximum time the ILP solver is allowed.
- **number_solutions**: The maximum number of solutions to be identified by the ILP algorithm
- **maximum_pathway**: If set to 1, then the solution with the maximum number of reactions is returned.
- **minimum_pathway**: If set to 1, then the solution with the minimum number of reactions is returned. If both **maximum_pathway** and **minimum_pathway** are set to 0, then the size of the solution will not be included in the objective value.
- **reactions_fixed**: If set to 1, then the interaction graph is simulated, evaluating the activation state of the included proteins to minimize the mismatch with the measured data. If set to 0, then the interaction graph is pruned, removing all reactions that contradict the data at hand.
- **minimum_corrections**: If set to 1 then the Minimum Corrections Sets (MCoS) are identified.
- **data_preprocessig**: If set to 1 then the ILP code receives as input a data file (e.g. "data.txt") containing the experimental dataset and prints two files named "measurements.txt" and "inputs.txt", specially formatted to be parsed by the optimizer.
- **best_scenario_fit**: If set to 1 then the best sign consistent solutions, for a given scenario, are enumerated. Note that if set to 1, then the inputs.txt and measurements.txt files may include only a single scenario.
- **add_new_reactions**: If set to 1 then *S*igNetTrainer parses a list of addable reactions provided by the users and scores them based on how much they improve the goodness of fit to the data if added to the PKN (OPT_GRAPH problem).

"Network.sif" (defines the interaction graph)

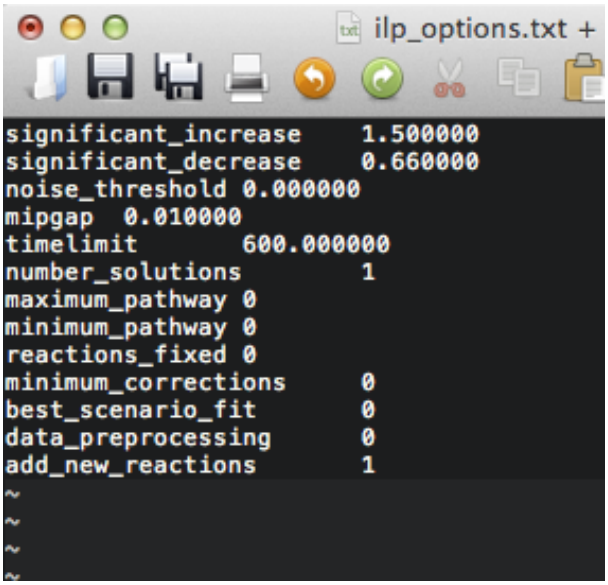


Figure A.1: Sample "ilp_options.txt" file

- Tab-delimited file
- Lists the reactions (edges) of the interaction graph
- First column contains the reactant in the corresponding edges.
- second column contains the type of the edges. Two types are supported, **"activations"** and **"inhibitions"**. If set to 1, then an activation reaction is assumed, if set to -1, then an inhibition is assumed. If set to 2, then an activation is assumed which cannot be removed by the algorithm (useful for when the user is positive the reaction is functional in the interrogated system). If set to -2, then an inhibition is assumed which cannot be removed by the algorithm.
- third column contains the product in the corresponding edges.
- No special characters are allowed in the names of the reactants or products.
- The names of signaling molecules must be the same with the stimuli, inhibitors, or signals included in the "data.txt", "inputs.txt" and "measurements.txt" files.

"inputs.txt" (defines the perturbations of the experimental scenarios)

- Tab-delimited file
- Rows correspond to different experimental scenarios, columns correspond to inputs of the interaction graph (i.e. perturbed nodes).
- first row contains the names of the corresponding input nodes.
- the rest of the entries correspond to the imposed activation values of the input nodes. "1" corresponds to up-regulation of the respective node, "-1" corresponds to down-regulation of the respective node, "0" corresponds to setting the respective node to inactive, "nan" implies the respective node is not perturbed in that scenario.

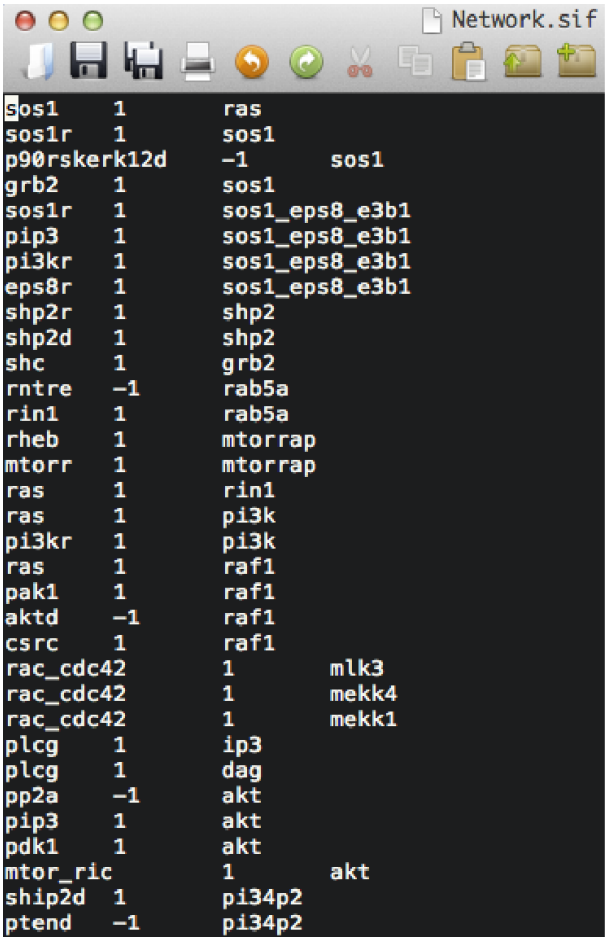


Figure A.2: Sample "Network.sif" file

- No special characters are allowed in the names of the input nodes.
- The names of the input nodes must be the same with the corresponding signaling molecules in the "Network.sif" file.

"measurements.txt" (specifies the resulting node changes (up, down, neutral) in the experiments)

- Tab-delimited file
- Rows correspond to different experimental scenarios, columns correspond to measured signals.
- first row contains the names of the measured signals.
- the rest of the entries correspond to the measured activation values of the signals. "1" corresponds to up-regulation of the respective signal, "-1" corresponds to down-regulation of the respective signal, "0" corresponds to setting the respective signal to inactive, "nan" implies the respective signal is not measured in that scenario.
- No special characters are allowed in the names of the measured signals.

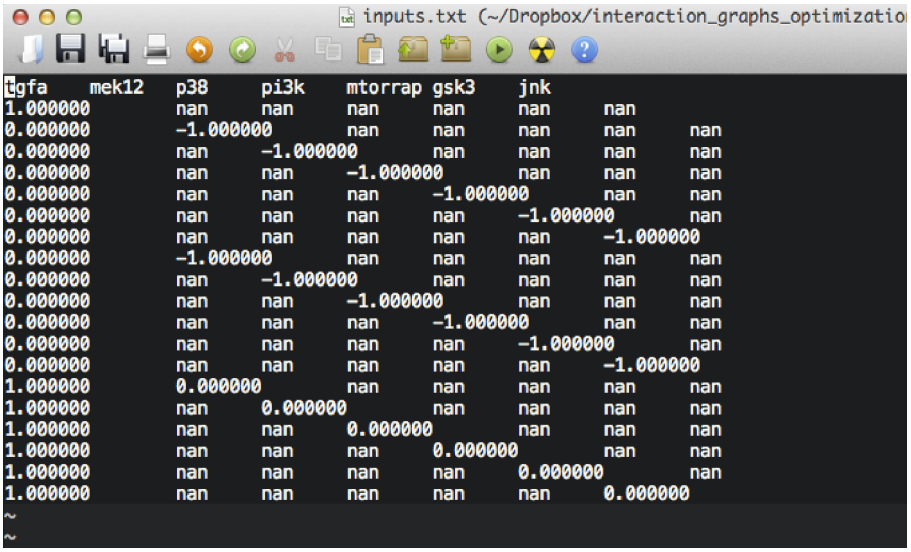


Figure A.3: Sample "inputs.txt" file

- The names of the measured signals must be the same with the corresponding signaling molecules in the "Network.sif" file.

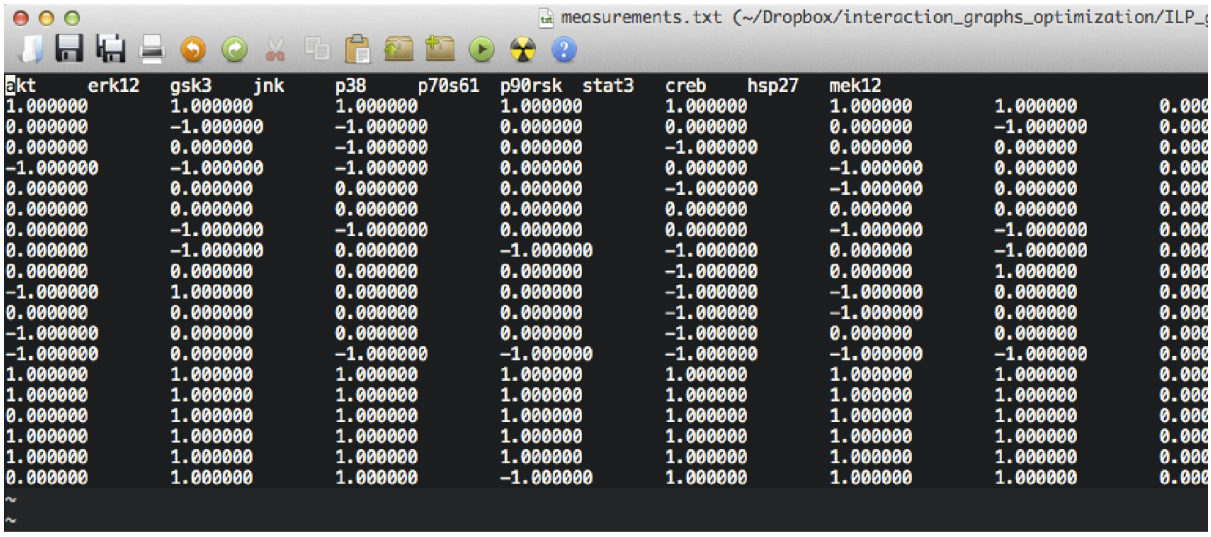
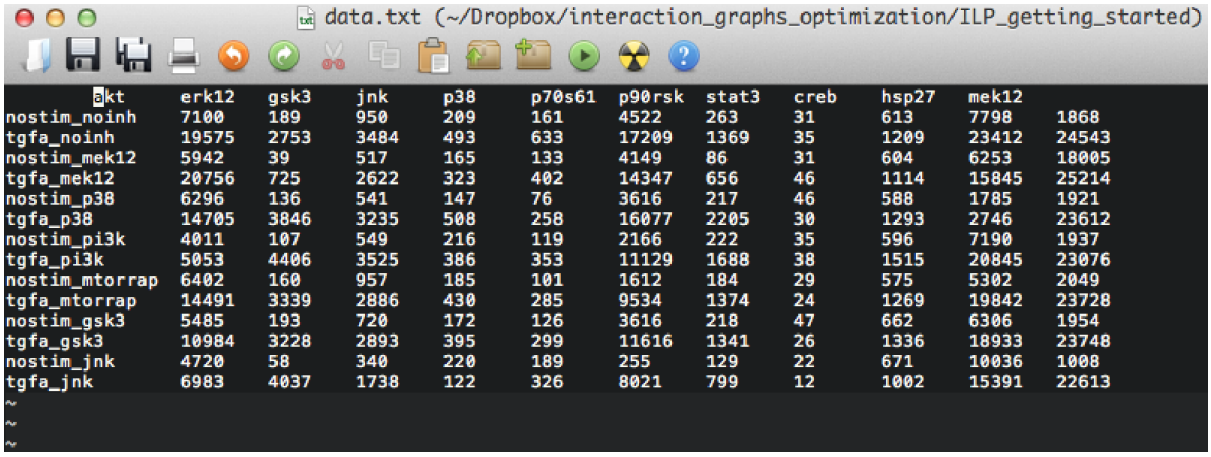


Figure A.4: Sample "measurements.txt" file

A.4.2 Data preprocessing

We implemented a data preprocessing which delivers the files "measurements.txt" and "input.txt" (described above) in an automatic way from raw data sets. This preprocessing is tailored to the type of data we are normally using (Luminex). Of course, the user may put together these two files on its own, if this is more convenient (see previous section A.4.3 how to format "measurements.txt" and "inputs.txt").

In order to use the data preprocessing feature of *SigNetTrainer* you must provide a file data.txt with the following structure:



	akt	erk12	gsk3	jnk	p38	p70s61	p90rsk	stat3	creb	hsp27	mek12	
nostim_noinh	7100	189	950	209	161	4522	263	31	613	7798	1868	
tgfa_noinh	19575	2753	3484	493	633	17209	1369	35	1209	23412	24543	
nostim_mek12	5942	39	517	165	133	4149	86	31	604	6253	18005	
tgfa_mek12	20756	725	2622	323	402	14347	656	46	1114	15845	25214	
nostim_p38	6296	136	541	147	76	3616	217	46	588	1785	1921	
tgfa_p38	14705	3846	3235	508	258	16077	2205	30	1293	2746	23612	
nostim_pi3k	4011	107	549	216	119	2166	222	35	596	7190	1937	
tgfa_pi3k	5053	4406	3525	386	353	11129	1688	38	1515	20845	23076	
nostim_mtorrap	6402	160	957	185	101	1612	184	29	575	5302	2049	
tgfa_mtorrap	14491	3339	2886	430	285	9534	1374	24	1269	19842	23728	
nostim_gsk3	5485	193	720	172	126	3616	218	47	662	6306	1954	
tgfa_gsk3	10984	3228	2893	395	299	11616	1341	26	1336	18933	23748	
nostim_jnk	4720	58	340	220	189	255	129	22	671	10036	1008	
tgfa_jnk	6983	4037	1738	122	326	8021	799	12	1002	15391	22613	

Figure A.5: Sample "data.txt" file

"data.txt"

- Tab-delimited file
- Rows correspond to different experimental scenarios, columns correspond to measured signals.
- first column contains a description for the corresponding scenarios in the following format: "Stimuli_Inhibitor". Where, "Stimuli" is the stimuli introduced in the current scenario, "inhibitor" is the inhibitor (or knock out) introduced in the current scenario.
- first row contains the names of the corresponding signals.
- No special characters are allowed in the labels of the scenarios or the signals.
- Names of the signals, stimuli or inhibitors must be the same with the ones included in the interaction graph.
- the remaining entries are floating point numbers that correspond to activation values, measured in fluorescent units.

After having compiled the C code, one may then run: `"./SigNetTrainer data.txt"`. "SigNetTrainer.c" will read the "ilp_options.txt" file, containing user-defined options for data preprocessing (see above), and the data file itself ("data.txt") containing the experimental dataset as described above. The two files named "measurements.txt" and "inputs.txt" will then be generated.

A.4.3 Running the optimization procedure

In a terminal, after having compiled the C code run: `"./SigNetTrainer Network.sif inputs.txt measurements.txt addable_edges.txt"`. "SigNetTrainer.c" receives as input the "ilp_options.txt" file, containing user defined options (for a detailed description of "ilp_options.txt" see section A.4.2) and **(i)** a file containing the interaction graph (e.g. "Network.sif", see example file in figure A.2), **(ii)** a file containing the experimental design (e.g. "inputs.txt", see example in

figure A.3), **(iii)** a file containing the measured data (e.g. "measurements.txt", see example in figure A.4), and returns the optimization results "objective_value.txt", "reactions_out.txt" (or "interventions_out.txt"), "network_out.txt", predicted values for the signaling molecules "predictions_out.txt" and auxiliary files "species_indices.txt", "solutions_history.txt", "reactions_indices.txt".

Output files

- **"objective_value.txt"**: The fitness error of the interaction graph to the measured data.
- **"reactions_out.txt"**: A list containing all reactions in the interaction graph and an identifier showing wether each reaction was removed during the optimization procedure or not. Tab delimited file. The first column contains the reactions ID. The third column, and every column after that, contains two numbers separated by a comma character. The first number in each column is used if the corresponding reaction is an activation, the second number is used if the corresponding reaction is an inhibition. An entry of 1.0 denotes the corresponding reaction was removed by the ILP, an entry of 0.0 denotes the reaction was conserved in the solution. Every column corresponds to a different solution in the solution pool (if "number_solutions" ≥ 2 in the "ilp_options.txt" file).
- **"network_out.txt"**: The network structure in .dot format ready to be parsed by graphviz.
- **"predictions_out.txt"**: A table containing the activation states of measured signals as predicted by the ILP algorithm. Rows correspond to different experimental scenarios, columns correspond to measured signals. First row contains the names of the measured signals. The rest of the entries correspond to the predicted activation values of the signals. "1" corresponds to up-regulation of the respective signal, "-1" corresponds to down-regulation of the respective signal, "0" corresponds to predicting the respective signal as inactive.
- **"species_indices.txt"**: A list containing all signaling nodes (i.e. species) of the interaction graph and their corresponding ID.
- **"reactions_indices.txt"**: A list containing all reactions of the interaction graph and their corresponding ID.
- **"solutions_history.txt"**: Includes the same information as "network_out.txt" formatted differently. Includes a list of the $y_{i,s}$ variables (see *{ Melas et al., PLoS CB 2013}*), where every column corresponds to a different solution in the solution pool. Every column numbers $(2 \times N)$ rows, where $N = Number\_of\_reactions$. The first N rows correspond to the $y^{+}_{i,s}$ variables, the last N rows correspond to the $y^{-}_{i,s}$ variables.
- **"interventions_out.txt"**: Is generated if "minimum_corrections" $= 1$ in the "ilp_options.txt" file. A list containing all species (i.e. signaling nodes) and experimental scenarios in the interaction graph and an identifier showing wether each species was perturbed (and how) for that given experiment. Tab delimited file. The first column contains the scenario ID. The second column contains the species ID. The third column, and every column after that, contains two numbers separated by a comma character. The first number in each column corresponds to positive perturbation (up-regulation) of the respective node, the second number corresponds to negative perturbation (down-regulation) of the respective node. An entry of 1.0 denotes the corresponding node was perturbed by the ILP, an entry of 0.0 denotes the node was not perturbed. Every column corresponds to a different solution in the solution pool (if "number_solutions" ≥ 2 in the "ilp_options.txt" file).

Appendix B

Acknowledgements

The author would like to acknowledge the contribution of the following people. From the NTUA, school of Mechanical Engineering, Prof. Christopher Provatidis and Ioannis Antoniadis for advising the author and for a lot of fruitful conversations. From Aachen University, department of Mechanical Engineering, Prof. Alexander Mitsos for an outstanding collaboration throughout the authors' PhD studies and for mentoring on regular optimization. From the European Bioinformatics Institute, Dr. Julio Saez-Rodriguez for hosting the author in the summer of 2010 and actively advising the author during that time. From the Massachusetts Institute of Technology, department of Biological Engineering, Prof. Douglas A. Lauffenburger for hosting the author two consecutive summers (2011, 2012) and for actively advising during that time. From the Max Planck Institute, Department for complex technical systems, Dr. Steffen Klamt for hosting the author in autumn 2012 and for actively advising during that time. Prof. Socrates Tsangaris, Prof. Nikolaos Chronis, Prof. Aristotelis Chatziioannou and Prof. Fragkiskos Kolisis, members of the author's thesis committee for reviewing the writeup and providing useful insight. All members of the Systems Biology and Bioengineering group at the NTUA, school of Mechanical Engineering, summer of 2008 until summer of 2013 for a very fruitful collaboration and with a special mention to Aikaterini D. Chairakaki (current address: Division of Immunogenetics, Center of Immunology and Transplantation, Biomedical Research Foundation of the Academy of Athens), Dimitris E. Messinis, Danai Kirli-Florou, Theodore Sakellaropoulos, Evangelos Zymbeloudis and Elisavet I. Chatzopoulou for collaborating closer. Finally, Prof. Leonidas Alexopoulos for mentoring the author throughout his PhD studies and stimulating his interest in the understanding of biological systems.

Παράρτημα Γ΄

Εκτεταμενη Ελληνικη Περιληψη

Η μοντελοποίηση των ενδοκυτταρικών σηματοδοτικών μονοπατιών είναι υψίστης σημασίας για την βαθύτερη κατανόηση της λειτουργίας και συμπεριφοράς των κυττάρων. Τα σηματοδοτικά μονοπάτια απεικονίζουν αλληλεπιδράσεις μεταξύ πρωτεϊνών και περιγράφουν πως τα κύτταρα αποκρίνονται σε ερεθίσματα του εξωτερικού τους περιβάλλοντος. Τα μονοπάτια αυτά είναι διαθέσιμα στην βιβλιογραφια, σε διαδικτυακές βάσεις δεδομένων. Τα τελευταία χρόνια η διεθνής κοινότητα κάνει προσπάθειες να τα μοντελοποιήσει, υιοθετώντας μεθοδολογίες από την θεωρία συστημάτων, προς την δημιουργία εκτελέσιμων μοντέλων που θα δίνουν την δυνατότητα προσομοίωσης σημαντικών κυτταρικών διεργασιών. Η μοντελοποίηση σηματοδοτικών μονοπατιών αποτελεί κύριο ενδιαφέρον της βιολογίας συστημάτων και αναμένεται να βελτιώσει τις διαδικασίες ανάπτυξης φαρμάκων, όπως αυτές εφαρμόζονται τώρα στην φαρμακευτική βιομηχανία.

Στην παρούσα διδακτορική διατριβή, ο υποψήφιος διδάκτορας εφαρμόζει μεθόδους Ακέραιου Γραμμικού (και μή γραμμικού) Προγραμματιμού (Integer Linear Programming - ILP, Non Linear Programming - NLP) για την μοντελοποίηση ενδοκυτταρικών σηματοδοτικών μονοπατιών και την εκπαίδευση των εν λόγω μοντέλων σε πειραματικά δεδομένα με σκοπό την πιστή απεικόνιση των σηματοδοτικών μηχανισμών στις υπο εξέταση κυτταρικές σειρές.

1 Εισαγωγή

Το κεφάλαιο αυτό αποσκοπεί σε μια μη τεχνική εισαγωγή στην μοντελοποίηση ενδοκυτταρικών σηματοδοτικών μονοπατιών συνδυάζοντας πρωτεομικά δεδομένα και την a priori γνώση αλληλεπιδράσεων μεταξύ πρωτεϊνών. Επίσης εξετάζεται η σημασία της ενδοκυτταρικής σηματοδότησης σε περίπλοκες ασθένειες όπως ο καρκίνος του ήπατος. Ο αναγνώστης μπορεί να παρακάμψει αυτό το κεφάλαιο αν είναι γνώριμος με αυτές τις έννοιες.

1.1 Περί Συστημικής Βιολογίας

Συστημική βιολογία (ΣΒ) ορίζεται η εφαρμογή μεθόδων από την θεωρία συστημάτων στην μοριακή βιολογία [1]. Η ΣΒ αποσκοπεί στην μελέτη των αλληλεπιδράσεων μεταξύ των συνιστωσών βιολογικών συστημάτων και πώς αυτές οι αλληλεπιδράσεις επηρεάζουν την κυτταρική λειτουργία και απόκριση. Η ΣΒ δημιουργήθηκε από την ανάγκη να ερμηνευτούν τα ολοένα αυξανόμενα σε όγκο βιολογικά δεδομένα (συμπεριλαμβανομένων πρωτεομικών, γενομικών και μεταβολομικών δεδομένων), καθώς οι αντίστοιχες τεχνολογίες ωρίμασαν και ο όγκος και πολυδιαστατικότητα των δεδομένων αυξήθηκε εκθετικά τα τελευταία δέκα χρόνια. Η συστηματική μοντελοποίηση των δεδομένων αυτών προσφέρει μια ποιοτική και ποσοτική ερμηνεία των κύριων διεργασιών που επιτελούνται στα κύτταρα και οι οποίες συχνά διαταράσσονται σε περίπλοκες ασθένειες.

1.2 Μοντελοποίηση σηματοδοτικών μονοπατιών

Γενικά

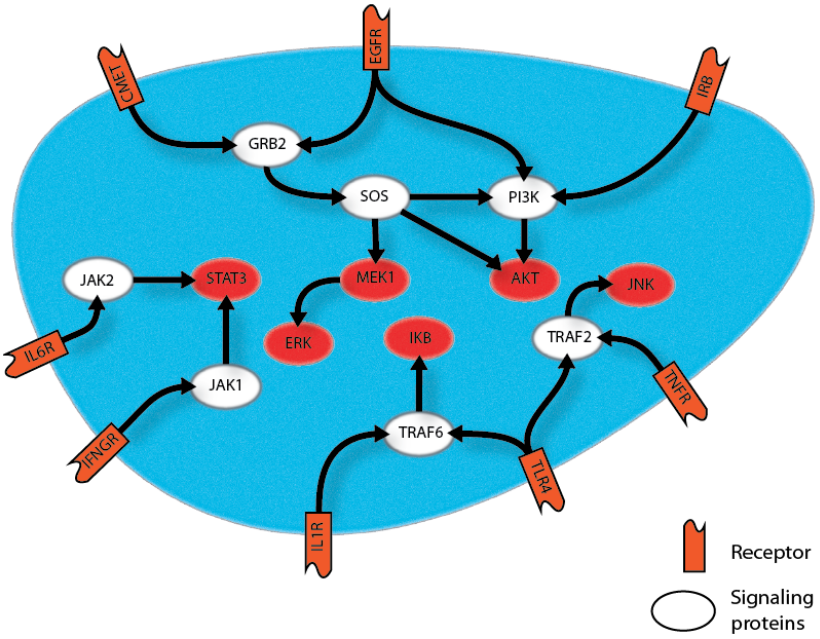
Η μελέτη των σηματοδοτικών μονοπατιών είναι ένα απο τα πεδία έρευνας της μοριακής βιολογίας που η ΣΒ έχει αλλάξει δραστικά. Τα σηματοδοτικά μονοπάτια περιγράφουν πώς το κύτταρο αποκρίνεται σε ερεθίσματα του εξωτερικού του περιβάλλοντος.

Τα κύτταρα υπάρχουν σε ένα πολύ περίπλοκο βιοχημικό περιβάλλον. Αντιλαμβάνονται το περιβάλλον αυτό μέσω εξειδικευμένων μορίων στην κυτταρική τους μεμβράνη, ονόματι υποδοχείς. Οι υποδοχείς δεσμεύουν τα ερεθίσματα του εξωκυτταρικού περιβάλλοντος (κάθε ερέθισμα έχει εξειδικευμένο υποδοχέα στο οποίο μπορεί να προσδεθεί). Κατόπιν πρόσδεσης του ερεθίσματος στον αντίστοιχο υποδοχέα, το ενδοκυτταρικό τμήμα του υποδοχέα αλλάζει την στερεοχημική του διάταξη (ενεργοποιείται) και άλλες ενδοκυτταρικές πρωτεϊνες μπορούν να προσδεθούν πάνω του, οι οποίες με την σειρά τους αλλάζουν την στερεοχημική τους διάταξη (ενεργοποιούνται / φωσφορυλιώνονται). Με αυτόν τον τρόπο διαμορφώνεται η σηματοδοτική διαδικασία και κάθε πρωτεϊνη ενεργοποιεί (φωσφορυλιώνει) την κάτωθί της στο μονοπάτι [2] (σχήμα 3.1). Ο τρόπος με τον οποίον η μία πρωτεϊνη αλληλεπιδρά με τις υπόλοιπες είναι ιδιαίτερα περίπλοκος και εξαρτάται απο την στερεοχημική της διάταξη και πολλές άλλες βιοχημικές ιδιότητες των εμπλεκόμενων μορίων. Τέλος, κάποιες από τις πρωτεϊνες στο σηματοδοτικό μονοπάτι έχουν την ιδιότητα να περνάνε στον πυρήνα του κυττάρου και να εκκινούν την έκφραση γονιδίων (μεταγραφικοί παράγοντες), ρυθμίζοντας έτσι την κυτταρική συμπεριφορά. Ενδεικτικές κυτταρικές συμπεριφορές είναι: **(i)** ο κυτταρικός διπλασιασμός, **(ii)** κυτταρικός θάνατος (απόπτωση) **(iii)** έκλυση εξωκυτταρικών πρωτεϊνων (κυτοκινών) οι οποίες θα αποτελέσουν ερεθίσματα για άλλα κύτταρα (εξωκυτταρική σηματοδότηση) και **(iv)** κυτταρική μετανάστευση (cell migration).

Λόγω της μείζονος σημασίας της ενδοκυτταρικής σηματοδότησης, η μοντελοποίηση σηματοδοτικών μονοπατιών είναι από τους κύριους στόχους της ΣΒ [3].

Κατασκεύη σηματοδοτικών μονοπατιών

Τα σηματοδοτικά μονοπάτια αποτελούνται από την συννένωση μεμονομένων πρωτεϊνικών αλληλεπιδράσεων. Οι εν λόγω αλληλεπιδράσεις αναγνωρίζονται μέσω πρωτεομικών πειραμάτων. Οι δύο πιο



Σχήμα 3.1: Διαγραμματική απεικόνιση απλοποιημένου σηματοδοτικού μονοπατιού

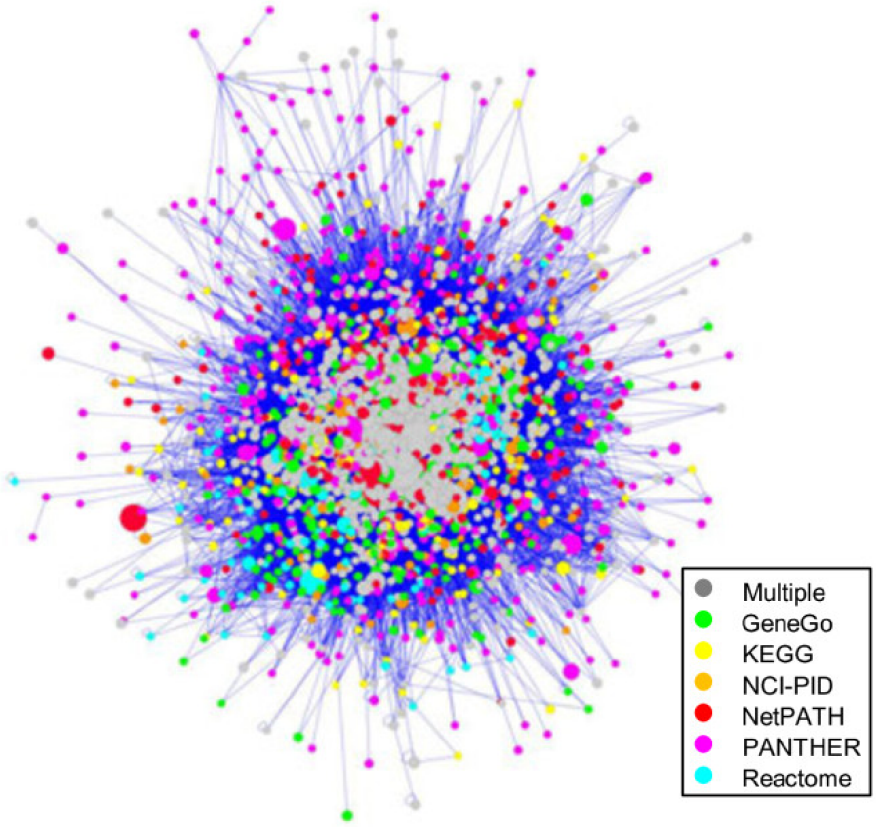
διαδεδομένες τεχνικές για την αναγνώριση πρωτεϊνικών αλληλεπιδράσεων είναι η φασματογραφία μάζας (Mass Spectrometry - MS) και το Yeast-two-hybrid (Y2H) screen [4]. Τα τελευταία δέκα χρόνια περισσότερες από 100.000 αλληλεπιδράσεις μεταξύ πρωτεϊνών έχουν αναγνωριστεί και είναι διαθέσιμες σε διαδικτυακές βιβλιοθήκες [5]. Στό σχήμα 3.2 φαίνεται η ένωση 7 από τις μεγαλύτερες βιβλιοθήκες. Η κατανόηση της δομής τέτοιων δικτύων θα αποσαφηνίσει τον τρόπο με τον οποίο τα κύτταρα αποκρίνοται σε παράγοντες του βιοχημικού τους περιβάλλοντος και θα συνδράμει στην αναγνώριση των μηχανισμών που ευθύνονται για περίπλοκες ασθένειες [6].

Μοντελοποίηση σηματοδοτικών μονοπατιών

Η μοντελοποίηση σηματοδοτικών μονοπατιών περιλαμβάνει την χρήση μαθηματικών φορμαλισμών για την ποσοτικοποίηση της μετάδοσης σήματος από την μία πρωτεϊνη στην άλλη και την κατασκευή εκτελέσιμων μοντέλων των σηματοδοτικών μηχανισμών των υπο εξέταση κυττάρων [8]. Προσομοιώνοντας τα μοντέλα αυτά μπορούμε να αποσαφηνίσουμε τους μηχανισμούς κυτταρικής απόκρισης σε ερεθίσματα του εξωκυτταρικού περιβάλλοντος και να εξάγουμε συμπεράσματα για την σημαντικότητα του κάθε κόμβου (πρωτεϊνης) στο δίκτυο. Οι πιο δημοφιλείς φορμαλισμοί περιλαμβάνουν την Boolean λογική, την fuzzy λογική και τις συνήθεις διαφορικές εξισώσεις [9, 8].

Στην περίπτωση των συνήθων διαφορικών εξισώσεων (ODEs), η μετάδοση σήματος μοντελοποιείται με βάση τον νόμο της χημικής κινητικής: Δεδομένης μιας αντίδρασης $A \rightarrow B$, ο ρυθμός σύνθεσης του προϊόντος (B) είναι ανάλογος της συγκέντρωσης του αντιδρώντος (A) και αντιστρόφος ανάλογος της ιδίας συγκέντρωσης του. Με αυτόν τον τρόπο οι αντιδράσεις του δικτύου μοντελοποιούνται με διαφορικές εξισώσεις. Εν συνεχεία, επιβάλλοντας σαν αρχικές - συνοριακές συνθήκες τις συγκεντρώσεις των εμπλεκόμενων μορίων την χρονική στιγμή 0 (πριν την εισαγωγή του ερεθίσματος) μπορεί να υπολογιστεί η συγκέντρωση σε κάθε επόμενη στιγμή.

Στην Boolean λογική οι πρωτεϊνες μπορούν να πάρουν μόνο δύο τιμές (ON/OFF) και η συνδεσμολογία μεταξύ τους μοντελοποιείται με χρήση λογικών πυλών. Εν συνεχεία επιβάλλονται οριακές συνθήκες στους κόμβους όπου εισάγεται το εξωτερικό ερέθισμα και η ροή του σήματος προσομοι-



Σχήμα 3.2: Η συννένωση 7 των μεγαλύτερων διαδικτυακών βιβλιοθηκών πρωτεϊνικών αλληλεπιδράσεων. Δείτε επίσης [7]

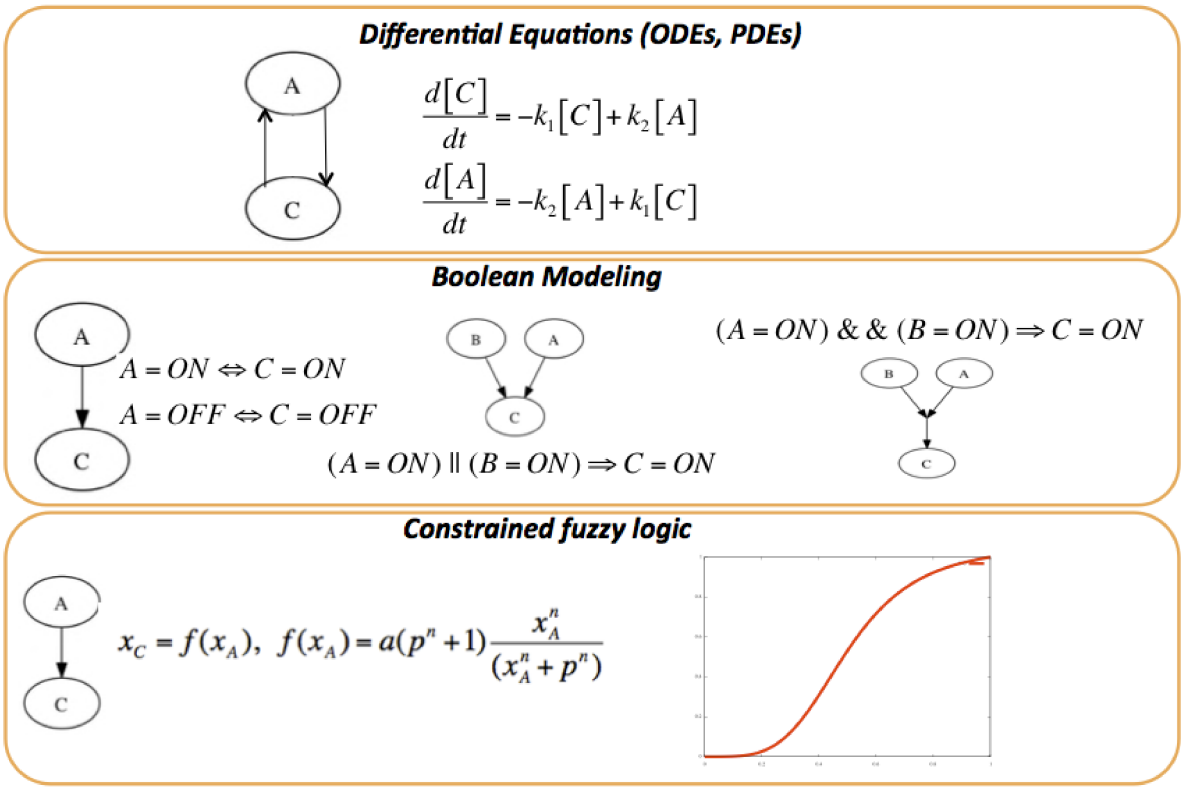
.

ώνεται από την μια πρωτεϊνη στην άλλη χρησιμοποιώντας την Boolean λογική. Σε αντίθεση με τις συνήθεις διαφορικές εξισώσεις, η Boolean λογική υποθέτει ότι έχει επέλθει ισορροπία στο σηματοδοτικό μονοπάτι (steady-state assumption). Αντί της Boolean λογικής συχνά χρησιμοποιείται fuzzy λογική καθώς επιτρέπει την ποσοτική μοντελοποίηση της σηματοδοτικής διαδικασίας, που είναι πιο κοντά στην βιολογία των εν λόγω συστημάτων [10]. Στην fuzzy λογική χρησιμοποιείται συνάρτηση μεταφοράς για να ποσοτικοποιήσει την μεταγωγή σήματος από την μια πρωτεϊνη στην άλλη (δείτε σχήμα 3.3).

Βελτιστοποίηση σηματοδοτικών μονοπατιών σε φωσφοπρωτεομικά δεδομένα

Παρότι η μοντελοποίση σηματοδοτικών μονοπατιών είτε μεσω συνήθων διαφορικών εξισώσεων είτε μέσω Boolean και fuzzy λογικής επιτυγχάνει την κατασκευή εκτελέσιμών μοντέλων των σηματοδοτικών μηχανισμών των υπο εξέταση κυττάρων, η πιστότητά τους εξαρτάται σε μεγάλο βαθμό από την ακρίβεια των δικτύων που χρησιμοποιήθηκαν σαν βάση. Οι αντιδράσεις που περιλαμβάνονται στα δίκτυα αυτά μετρήθηκαν πειραματικά είτε μέσω Yeast-two-hybrid είτε μέσω φασματογραφίας μάζας. Ωστόσο οι μέθοδοι αυτές χρησιμοποιούν purified protein και τα αποτελέσματα τους δεν αναφέρονται σε κάποιο συγκεκριμένο τύπο κυττάρων. Ενώ διαφορετικοί κυτταρικοί τύποι έχουν διαφορετικούς σηματοδοτικούς μηχανισμούς (επομένως και διαφορετικά μονοπάτια) αναλόγως με την λειτουργία που επιτελούν [2].

Επομένως, για την κατασκευή σηματοδοτικών μονοπατιών εξειδικευμένων για τον υπο εξέταση



Σχήμα 3.3: Επισκόπηση των κυριότερων μεθόδων για την μοντελοποίηση σηματοδοτικών μονοπατιών.

κυτταρικό τύπο εκτός από τα δίκτυα που υπάρχουν διαθέσιμα σε βιβλιοθήκες, πειραματικά δεδομένα πρέπει επίσης να χρησιμοποιηθούν που θα αιχμαλωτίζουν το βιολογικό υπόβαθρο των εν λόγω κυττάρων. Εν συνεχεία τα πειραματικά δεδομένα θα πρέπει να συνδυαστούν με τα βιβλιογραφικά δίκτυα μέσω αλγορίθμων βελτιστοποίησης που θα εκπαιδεύσουν τα υπολογιστικά μοντέλα στα πειραματικά δεδομένα, καταλήγοντας σε μοντέλα που πολύ στενά αιχμαλωτίζουν τους σηματοδοτικούς μηχανισμούς των υπο εξέταση κυττάρων.

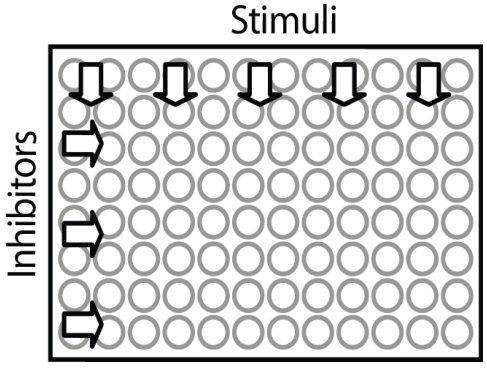
Μετρηση φωσφοπρωτεομικών δεδομένων

Η τεχνολογία Luminex xMAP είναι από τις πιο καινοτόμες τεχνολογίες, κατάλληλες για την κατασκευή σηματοδοτικών μονοπατιών [11]. Η πειραματική διαδικασία είναι η ακόλουθη:

Τα υπο εξέταση κύτταρα καλλιεργούνται σε 96αρες πλάκες και σε κάθε πηγάδι εισάγεται διαφορετικός συνδυασμός ερεθισμάτων ενώ μετράται η ενεργοποίηση ενός συνόλου φωσφοπρωτεϊνών [12] (δείτε επίσης σχήμα 3.4 για μια αναπαράσταση της πειραματικής διαδικασίας). Από τα ερεθίσματα που εισήχθησαν, κάποια ενεργοποιούν σηματοδοτικά μονοπάτια (ονόματι κυττοκίνες) ενώ κάποια μπλοκάρουν σηματοδοτικά μονοπάτια (ονόματι αναστολείς) σε συγκεκριμένες πρωτεϊνες. Επι παραδείγματι ο αναστολέας MEKi μπλοκάρει το σηματοδοτικό μονοπάτι στην πρωτεϊνη MEK. Κάποια ενδεικτικά δεδομένα φαίνονται στο σχήμα 3.5 [12]. Στα δεδομένα του σχήματος 3.5 17 σήματα (φωσφοπρωτεϊνες) μετρώνται σε 3 χρονικές στιγμές, σε χρόνο t=0 (πριν την εισαγωγή των ερεθισμάτων), σε χρόνο t=30min κατόπιν της ενεργοποίησης και σε χρόνο t=3h.

Τα δεδομένα του σχήματος 3.5 αντιπροσωπεύουν την απόκριση του υπο εξέταση κυτταρικού τύπου σε συνδυαστικά ερεθίσματα με 8 κυττοκίνες και 8 αναστολείς. Ο συνδυασμός των δεδομένων αυτών με τα βιβλιογραφικά δίκτυα μεταγωγής σήματος θα οδηγήσει στην κατασκευή υπολογιστικού

μοντέλου που αιχμαλωτίζει τους σηματοδοτικούς μηχανισμούς του υπο εξέταση κυτταρικού τύπου [13].



Σχήμα 3.4: Αναπαράσταση της πειραματικής διαδικασίας

Βελτιστοποίηση σηματοδοτικών μονοπατιών σε φωσφοπρωτεομικά δεδομένα – με χρήση Εξελικτικών Αλγορίθμων (ΕΑ)

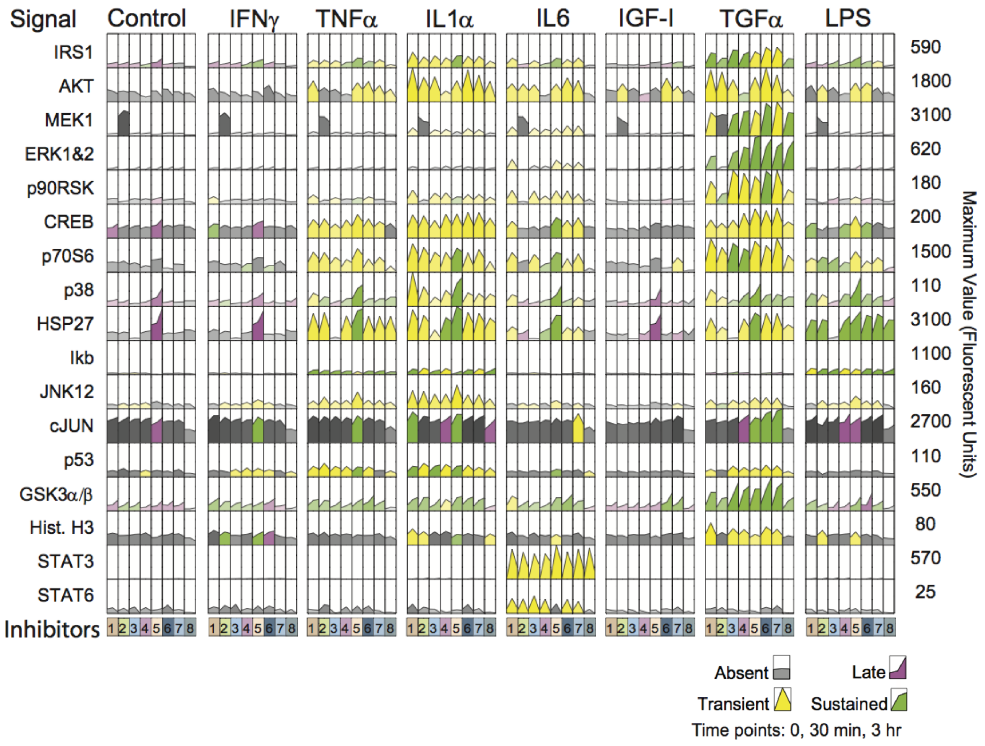
Στην εργασία των Saez-Rodriguez et al. [13], ένας ΕΑ χρησιμοποιήθηκε για να βελτιστοποιήσει σηματοδοτικό μονοπάτι σε φωσφοπρωτεομικά δεδομένα των Alexopoulos et al. [12] (τα δεδομένα που παρουσιάστηκαν στην προηγούμενη παράγραφο), σε μια απόπειρα να κατασκευαστεί υπολογιστικό μοντέλο που αναπαριστά τους σηματοδοτικούς μηχανισμούς καρκινικών ηπατοκυττάρων. Ο ηπατικός καρκίνος συνδέεται με πλήθος αλλαγών στα σηματοδοτικά μονοπάτια των ηπατοκυττάρων [14]. Η ανακάλυψη των αλλαγών αυτών θα οδηγήσει στην βαθύτερη κατανόηση της ασθένειας και στην πρόταση καινούριων φαρμακευτικών στόχων.

Στην εργασία των Saez-Rodriguez et al. [13], πρώτα κατασκευάστηκε σηματοδοτικό μονοπάτι με βάση βιβλιογραφικές πηγές και εν συνεχεία μοντελοποιήθηκε με χρήση Boolean λογικής για την μαθηματικοποίηση της μετάδοσης σήματος από την μια πρωτεΐνη στην άλλη μέσα στο δίκτυο. Όπως συζητήθηκε στην προηγούμενη παράγραφο, το σηματοδοτικό μονοπάτι δεν είναι εξειδικευμένο για τον υπο εξέταση κυτταρικό τύπο. Επομένως τα φωσφοπρωτεομικά δεδομένα των Alexopoulos et al. [12] χρησιμοποιήθηκαν και σε συνδυασμό με το Boolean μοντέλο του σηματοδοτικού μονοπατιού, μέσω ενός ΕΑ, οδήγησαν στην αφαίρεση των αντιδράσεων εκείνων του δικτύου που τα αντέκρουαν. Τα φωσφοπρωτεομικά δεδομένα εκφράζουν την απόκριση καρκινικών ηπατοκυττάρων κατόπιν ενεργοποίησης με 8 κυττοκίνες και 8 αναστολείς (σε συνδυαστικά πειράματα), ενώ μετρήθηκαν 17 φωσφοπρωτεΐνες. Τα δεδομένα φαίνονται στο σχήμα 3.5.

Πριν την διαδικασία βελτιστοποίησης, η προσομοίωση του Boolean μοντέλου με τα ίδια ερεθίσματα που εισήχθησαν στο πείραμα, αποκαλύπτει μια ασυμφωνία (fitness error) με τα πειραματικά δεδομένα. Απο το Boolean μοντέλο αφαιρούνται εν συνεχεία, μέσω του ΕΑ, αντιδράσεις με σκοπό την ελαχιστοποίηση της ασυμφωνίας αυτής με τα πειραματικά δεδομένα. Η βελτιστοποιημένη τοπολογία περιλαμβάνει μόνο ένα υποσύνολο των αρχικών αντιδράσεων (αυτό το υποσύνολο που φαίνεται λειτουργικό με βάση τα πειραματικά δεδομένα) και αντιπροσωπεύει πιο πιστά τους σηματοδοτικούς μηχανισμούς των καρκινικών ηπατοκυττάρων.

Βελτιστοποίηση σηματοδοτικών μονοπατιών σε φωσφοπρωτεομικά δεδομένα – με χρήση αλγόριθμου Ακέραιου Γραμμικού Προγραμματισμού (Integer Linear Programming - ILP)

Η μέθοδος ΕΑ που παρουσιάστηκε προηγουμένως εκπαίδευσε ένα Boolean μοντέλο σηματοδοτικού δικτύου σε φωσφοπρωτεομικά δεδομένα. Οι κανόνες της Boolean λογικής κωδικοποιήθηκαν στο μοντέλο το οποίο καλούσε επαναληπτικά ο ΕΑ, υπολόγιζε σε κάθε επανάληψη την ασυμφωνία του με τα πειραματικά δεδομένα και σταδιακά τροποποιούσε την συνδεσμολογία του ώστε να



Σχήμα 3.5: Ενδεικτικά πειραματικά δεδομένα που δημοσιεύτηκαν στην εργασία [12]. Οι γραμμές αντιστοιχούν στα μετρούμενα σήματα, οι κύριες στήλες αντιστοιχούν στις κυττοκίνες που εισήχθησαν στο πείραμα για να ενεργοποιήσουν τα μετρούμενα σηματοδοτικά μονοπάτια και οι δευτερεύουσες στήλες αντιστοιχούν στους αναστολείς. Οι αριθμοί στα δεξιά αντιστοιχούν στην μέγιστη μέτρηση του κάθε σήματος. Το σχήμα αυτό περιλαμβάνει συνολικά 64 πειράματα (8 κυττοκίνες x 8 αναστολείς). Κάθε τετραγωνάκι αναπαριστά την πορεία ενεργοποίησης του αντίστοιχου σήματος στον χρόνο, από την στιγμή t=0 (πριν την εισαγωγή των ερεθισμάτων), στην στιγμή t=30min κατόπιν της ενεργοποίησης και στην στιγμή t=3h. Με βάση την ποιοτική απόκριση του σήματος διακρίνοται τεσσάρων ειδών αποκρίσεις, 'απούσα', 'αργή', 'προσωρινή', και 'παρατεταμένη'.

ελαχιστοποιήσει την ασυμφωνία αυτή.

Η διδακτορική διατριβή αυτή πραγματεύεται την βελτιστοποίηση σηματοδοτικών μονοπατιών με χρήση μεθόδων regular optimization, δηλαδή Ακέραιου γραμμικού ή μη γραμμικού προγραμματισμού. Κατα την διαδικασία αυτή οι κανόνες της Boolean λογικής κωδικοποιούνται με την μορφή περιορισμών ενώ ελαχιστοποιείται η ασυμφωνία με πειραματικά δεδομένα. Οι μεθοδολογίες που δημοσιεύτηκαν σαν αποτέλεσματα της εν λόγω διατριβής υπολογίζουν global minimum της αντικειμενικής συνάρτησης και απαιτούν τάξεις μεγέθους λιγότερο υπολογιστικό χρόνο, καθιστόντας δυνατή την κατασκευή σηματοδοτικών μονοπατιών που αντιπροσωπεύουν πιστά τα πειραματικά δεδομένα και που περιλαμβάνουν εκατοντάδες πρωτεϊνες.

Στα επόμενα κεφάλαια της διατριβής τρείς διαφορετικές μεθοδολογίες παρουσιάζονται με λεπτομέρεια

1. Μέθοδος Ακέραιου Γραμμικού Προγραμματισμού για την μοντελοποίηση σηματοδοτικών μονοπατιών με χρήση Boolean λογικής.

2. Μέθοδος Μη Γραμμικού Προγραμματισμού για την μοντελοποίηση σηματοδοτικών μονοπατιών με χρήση fuzzy λογικής.

3. Μέθοδος Ακέραιου Γραμμικού Προγραμματισμού για την μοντελοποίηση σηματοδοτικών μονοπατιών ως άκυκλους γράφους και την εξάλειψη ασυμφωνιών με πειραματικά δεδομένα μέσω πλήθους διαφορετικών στρατηγικών

Οι μέθοδοι αυτοί χρησιμοποιούνται για την αντιμετώπιση προβλημάτων συνδεδεμένων με τον ηπατικό καρκίνο και την οστεοαρθρίτιδα.

2 Μέθοδος Ακέραιου Γραμμικού Προγραμματισμού για την βελτιστοποίηση σηματοδοτικών μονοπατιών σε φωσφοπρωτεομικά δεδομένα

Σε αυτήν την έρευνα που δημοσιεύτηκε στο έγκριτο περιοδικό PLoS Computational Biology [23], προτάθηκε μέθοδος Ακέραιου Γραμμικού Προγραμματισμού για την βελτιστοποίηση σηματοδοτικών μονοπατιών σε φωσφοπρωτεομικά δεδομένα και την κατασκευή εκτελέσιμων μοντέλων που αντιπροσωπεύουν τους σηματοδοτικούς μηχανισμούς των υπο εξέταση κυτταρικών σειρών. Η έρευνα αυτή έγινε σε συνεργασία με τον Αλέξανδρο Μητσό (την στιγμή της συγκεκριμένης δημοσίευσης επίκουρο καθηγητή στο τμήμα μηχανολόγων μηχανικών του MIT, Cambridge, MA, USA, την παρούση στιγμή καθηγητή στο RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany). Σε αντίθεση με προηγούμενες μεθόδους, η συγκεκριμένη προσέγγιση εγγυάται τον προσδιορισμό του global minimum της αντικειμενικής συνάρτησης (ασυμφωνία μεταξύ υπολογιστικού μοντέλου και πειραματικών δεδομένων) και μειώνει σημαντικά τον υπολογιστικό χρόνο, επιτρέποντας έτσι την κατασκευή εκτεταμένων δικτύων που περιλαμβάνουν εκατοντάδες πρωτεΐνες.

Δύο εφαρμογές εξετάστηκαν με την μέθοδο αυτή, (i) η βελτιστοποίηση ενός εκτεταμένου δικτύου μεταγωγής σήματος, εκφράζοντας την απόκριση ηπατικών κυττάρων σε περισσότερα απο 80 ερεθίσματα του εξωκυτταρικού περιβάλλοντος [46], (ii) η αναγνώριση των επιδράσεων 4 αντικαρκινικών φαρμάκων στο δίκτυο σηματοδότησης καρκινικών ηπατοκυττάρων [23]. Επίσης σε επόμενο βήμα, επεκτείναμε την μεθοδολογία αυτή για να περιλαμβάνει και εξωκυτταρική σηματοδότηση. Σαν εφαρμογή κατασκευάστηκε το εκτεταμένο σηματοδοτικό δίκτυο καρκινικών και φυσιολογικών ηπατοκυττάρων και μελετήθηκαν οι διαφορές τους [51].

2.1 Μαθηματική διατύπωση

Έστω δίκτυο μεταγωγής σήματος που ορίζεται ως ένα σύνολο αντιδράσεων $i = 1, ..., n_r$ και κόμβων $j = 1, ..., n_s$. Για κάθε αντίδραση ορίζονται 3 σύνολα: το σύνολο των αντιδρώντων R_i, το σύνολο των αναστολέων I_i και το σύνολο των προϊόντων P_i. Όπου $R_i, I_i, P_i \subset \{1, ..., n_s\}$. Μία αντίδραση θα πραγματοποιηθεί αν και μόνον αν όλα τα αντιδρώντα είναι παρόντα και κανένας αναστολέας δεν είναι παρών. Άν η αντίδραση πραγματοποιηθεί τότε όλα τα προϊόντα θα παραχθούν. Σκοπός της προτεινόμενης μεθόδου είναι ο προσδιορισμός των αντιδράσεων εκείνων του δικτύου i που δεν αντικρούουν τα πειραματικά δεδομένα και αφαίρεση των υπολοίπων. Για τον λόγο αυτό ορίζονται μεταβλητές y_i που αντιπροσωπεύουν αν η αντίδραση i είναι λειτουργική και άρα παρούσα στον υπο εξέταση κυτταρικό τύπο. ($y_i = 0$ αν ή αντίδραση δεν είναι παρούσα, $y_i = 1$ αν η αντίδραση είναι παρούσα).

Ένα σύνολο πειραμάτων $k = 1, ..., n_e$ πραγματοποιούνται. Σε κάθε πείραμα επιβάλλεται η ενεργοποίηση ενός συνόλου πρωτεϊνών και η απενεργοποίηση ενός άλλος συνόλου στο δίκτυο (αρχικές-οριακές συνθήκες). Αυτά περιγράφονται από τις μεταβλητές $M^{k,1}$ (πρωτεΐνες που ενεργοποιούνται) και $M^{k,0}$ (πρωτεΐνες που απενεργοποιούνται). Επίσης μεταβλητές $x_j^k = 1$ ή $x_j^k = 0$ εισάγωνται που εκφράζουν την φωσφορυλίωση (ενεργοποίηση) ή αποφωσφορυλίωση (απενεργοποίηση) αντίστοιχα των πρωτεϊνών j στο πείραμα k σύμφωνα με το υπολογιστικό μοντέλο. Επίσης ορίζονται μεταβλητές z_i^k που εκφράζουν την ενεργοποίηση της αντίδρασης i στο πείραμα k ($z_i^k = 1$ αν η αντίδραση i είναι ενεργή στο πείραμα k, αλλιώς $z_i^k = 0$). Σε περίπτωση που η φωσφορυλίωση της πρωτεΐνης j μετράται στο πείραμα k τότε η αντίστοιχη μέτρηση συμβολίζεται με $x_j^{k,m}$. $x_j^{k,m} = 1$ αν η πρωτεΐνη j μετράται ενεργοποιημένη, αλλιώς $x_j^{k,m} = 0$. Στόχος της προτεινόμενης μεθοδολογίας είναι η ελαχιστοποίηση της ασυμφωνίας μεταξύ προσομοιούμενων και μετρούμενων τιμών των πρωτεϊνών j στο πείραμα k: $\sum_{j,k} a_j^k \mid x_j^k - x_j^{k,m} \mid$. Η απόλυτη τιμή ξαναγράφεται $x_j^{k,m} + (1 - 2x_j^{k,m})x_j^k$. Το προτεινόμενο Ακέραιο Γραμμικό Πρόγραμμα διαμορφώνεται ως ακολούθως:

$$min \sum_{k=1}^{n_e} \sum_{j \in M^{k,2}} a_j^k (x_j^{k,m} + (1 - 2\, x_j^{k,m}) x_j^k); \quad \sum_{i=1}^{n_r} \beta_i\, y_i \tag{3.1}$$

s.t.

$$\sum_{i=1}^{n_r} a_i^l\, y_i \leq b^l, \quad l = 1, ..., n_e \tag{3.2}$$

$$z_i^k \leq y_i; \quad i = 1, ..., n_r, \quad k = 1, ..., n_e. \tag{3.3}$$

$$z_i^k \leq x_j^k; \quad i = 1, ..., n_r, \quad j \in R_i \quad k = 1, ..., n_e. \tag{3.4}$$

$$z_i^k \leq 1 - x_j^k; \quad i = 1, ..., n_r, \quad j \in I_i \quad k = 1, ..., n_e. \tag{3.5}$$

$$z_i^k \geq y_i + \sum_{j \in R_i} (x_j^k - 1) - \sum_{j \in I_i} (x_j^k), \quad i = 1, ..., n_r, \quad k = 1, ..., n_e. \tag{3.6}$$

$$x_j^k \geq z_i^k; \quad i = 1, ..., n_r, \quad j \in P_i \quad k = 1, ..., n_e. \tag{3.7}$$

$$x_j^k \leq \sum_{i=1,...,n_r : j \in P_i} z_i^k, \quad j = 1, ..., n_s, \quad k = 1, ..., n_e. \tag{3.8}$$

$$x_j^k = 0, \quad k = 1, ..., n_e \quad j \in M^{k,0} \tag{3.9}$$

$$x_j^k = 1, \quad k = 1, ..., n_e \quad j \in M^{k,1} \tag{3.10}$$

$$X \in \{0,1\}^{n_e \times n_s}, \quad y \in \{0,1\}^{n_r}, \quad Z \in \{0,1\}^{n_e \times n_r} \tag{3.11}$$

Όπου η εξίσωση 3.1 αποτελεί την αντικειμενική συνάρτηση. Ο πρώτος όρος είναι η ασυμφωνία μεταξύ προσομοιούμενων και μετρούμενων τιμών των πρωτεϊνών j στο πείραμα k και ο δεύτερος όρος είναι το μέγεθος της λύσης. Δηλαδή από όλες τις λύσεις που ελαχιστοποιούν την ασυμφωνία μεταξύ προσομοιούμενων και μετρούμενων τιμών, επιλέγεται αυτή που αριθμεί το ελάχιστο σύνολο αντιδράσεων.

Η λογική των εξισώσεων 3.1 - 3.11 εξηγείται παρακάτω

- Ο περιορισμός 3.3 εκφράζει ότι μια αντίδραση μπορεί να πραγματοποιηθεί μόνο αν είναι παρούσα ($y_i = 1$).
- Οι περιορισμοί 3.4 - 3.5 εκφράζουν ότι μια αντίδραση μπορεί να πραγματοποιηθεί μόνο αν όλα τα αντιδρόντα είναι παρόντα και κανένας αναστολέας δεν είναι παρών.
- Ο περιορισμός 3.6 εκφράζει ότι αν μια αντίδραση μπορεί να πραγματοποιηθεί ($y_i = 1$) και όλα τα αντιδρόντα είναι παρόντα και κανένας αναστολέας δεν είναι παρών τότε η αντίδραση θα πραγματοποιηθεί ($z_i^k = 1$).
- Ο περιορισμός 3.7 εκφράζει ότι μια πρωτεϊνη θα ενεργοποιηθεί αν κάποια αντίδραση στην οποία είναι προϊόν πραγματοποιηθεί.
- Ο περιορισμός 3.8 εκφράζει ότι μια πρωτεϊνη δεν θα ενεργοποιηθεί αν καμία αντίδραση στην οποία είναι προϊόν δεν πραγματοποιηθεί.

2.2 1η εφαρμογή: Βελτιστοποίηση εκτεταμένου δικτύου μεταγωγής σήματος σε φυσιολογικά ηπατοκύτταρα

Σε αυτήν την ενότητα χρησιμοποιούμε την μεθοδολογία που περιγράφηκε στην ενότητα 2.1 για την βελτιστοποίηση ενός εκτεταμένου δικτύου μεταγωγής σήματος σε φωσφοπρωτεομικά δεδομένα. Το εν λόγω δίκτυο θα περιγράφει τους σηματοδοτικούς μηχανισμούς φυσιολογικών ηπατοκυττάρων και την απόκρισή τους σε 81 ερεθίσματα του εξωκυτταρικού περιβάλλοντος (κυττοκίνες). Η έρευνα αυτή δημοσιεύτηκε από τους Melas et al. [46] και πραγματοποιήθηκε σε συνεργασία με τον Αλέξανδρο Μητσό (την στιγμή της συγκεκριμένης δημοσίευσης επίκουρο καθηγητή στο τμήμα μηχανολόγων μηχανικών του MIT, Cambridge, MA, USA, την παρούση στιγμή καθηγητή στο RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany) και τον Julio Saez-Rodriguez, group leader στο European Bioinformatics Institute (EBI), Cambridge, UK. Αυτή είναι από τις πρώτες απόπειρες για την κατασκευή σηματοδοτικού δικτύου, εξειδικευμένου για ένα τύπο κυττάρων και προσφέρει την βαθύτερη κατανόηση των σηματοδοτικών μηχανισμών των φυσιολογικών ηπατοκυττάρων.

Πιο συγκεκριμένα, η προτεινόμενη προσέγγιση αποτελείται από τα εξής βήματα (δείτε επίσης σχήμα 3.6): (i) Έλεγχος των 81 πιο σημαντικών κυττοκινών για τον υπο εξέταση κυτταρικό τύπο, (ii) Επιλογή των πιο ενεργών κυττοκινών, (iii) εισαγωγή τους σε συνδυαστικά πειράματα με σκοπό την δημιουργία ενός συνόλου φωσφοπρωτεομικών δεδομένων που αντιπροσωπεύουν τους σηματοδοτικούς μηχανισμούς των φυσιολογικών ηπατοκυττάρων, (iv) βελτιστοποίηση ενός εκτεταμένου βιβλιογραφικού δικτύου στα φωσφοπρωτεομικά δεδομένα με χρήση της μεθόδου Ακέραιου Γραμμικού Προγραμματισμού που περιγράφηκε στην ενότητα 2.1.

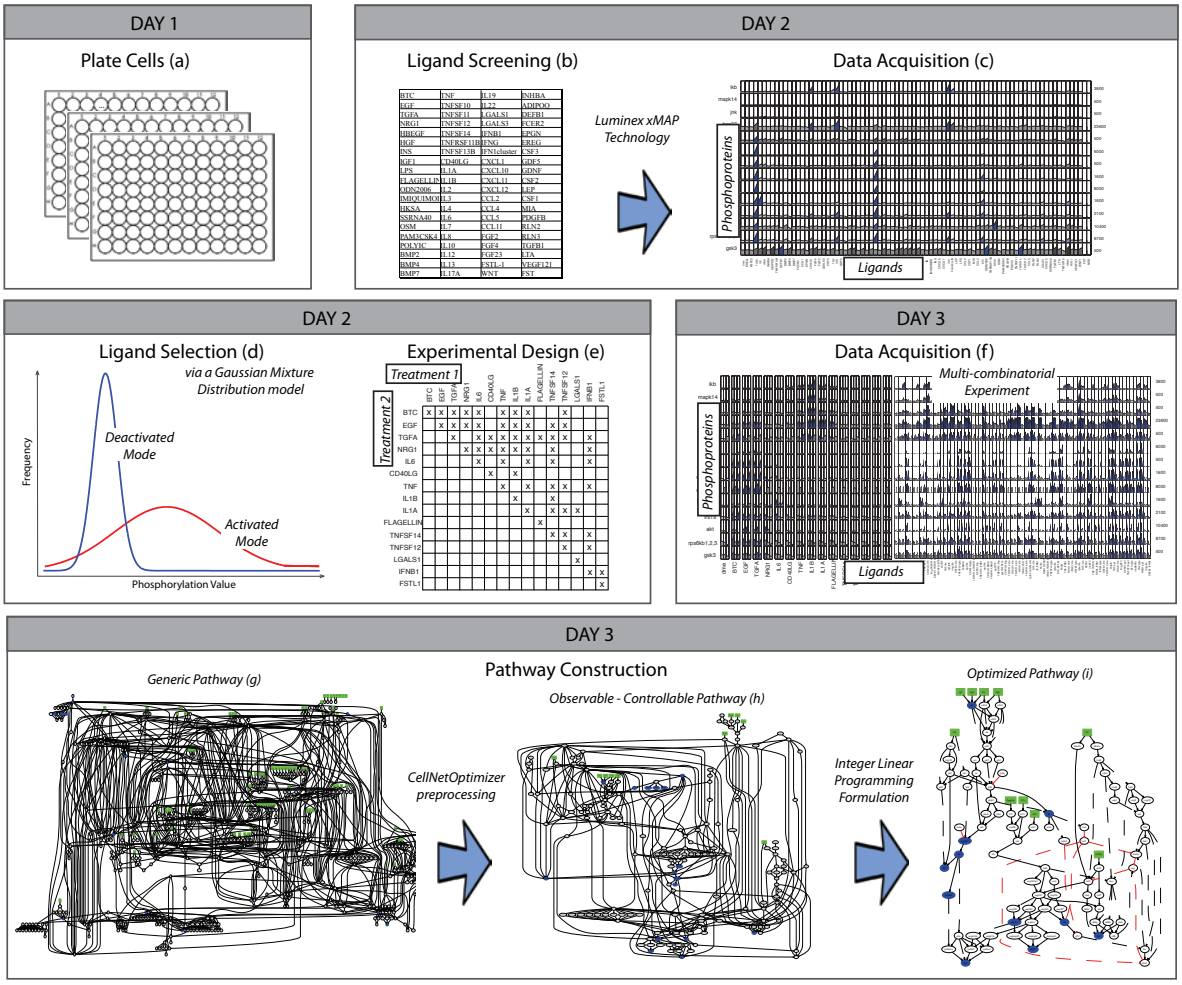
Έλεγχος των 81 πιο σημαντικών κυττοκινών

Οι κυττοκίνες του πίνακα 3.1 εξετάστηκαν ως προς τις επιδράσεις τους σε φυσιολογικά ηπατοκύτταρα. Τα φωσφοπρωτεομικά δεδομένα φαίνονται στο σχήμα 3.7.

66 από τα 81 ερεθίσματα δεν οδήγησαν στην ενεργοποίηση των μετρούμενων φωσφοπρωτεϊνών (δείτε σχήμα 3.7a). Τα 15 ερεθίσματα που είχαν κάποια επίδραση στα υπο εξέταση κύτταρα είναι: IL6, FLAGELLIN, TNF, IFNB1, TGFA, TNFSF14, TNFSF12, IL1A, EGF, IL1B, NRG1, BTC, CD40LG, DEFB1 και LGALS1. Σαν επαλήθευση των αποτελεσμάτων, παρατηρούμε οτι ερεθίσματα του EGFR όπως τα TGFA, EGF, NRG1 και BTC ενεργοποίησαν σήματα που συνδέονται με κυτταρικό διπλασιασμό όπως τα AKT, MAP2K1 (MEK), ERK, RPS6KB1,2,3 και RPS6KA1. Επίσης, ερεθίσματα που συνδέονται με την φλεγμονώδη αντίδραση του οργανισμού: IL1A, IL1B και TNF ενεργοποίησαν τα αντίστοιχα σήματα, όπως τα IKB και HSP27. Τέλος το IL6 ενεργοποίησε το STAT3, κάτι που επίσης είναι σύμφωνο με την βιβλιογραφία. Εκτός από τα ερεθίσματα των οποίων η δράση είναι σαφώς καταγεγραμμένη στην βιβλιογραφία αναγνωρίσαμε και άλλα όπως το CD40LG, μέλος της οικογένειας του TNF, το οποίο οδήγησε στην ενεργοποίηση των IKB, HSP27, IRS1S και RPS6KB1,2,3. Επίσης το FLAGELLIN, ερέθισμα του TLR5, ενεργοποίησε το IKB και το HSP27. Το LGALS1, επίσης γνωστό σαν GALECTIN-1, ενεργοποίησε τα RPS6KA1, RPS6KB1,2,3 και είχε μικρή επίδραση στα STAT3, IKB και IRS1S. Τα TNFSF12 και TNFSF14, επίσης μέλη της οικογένειας του TNF είχαν μικρή επίδραση στα RPS6KB1,2,3, IKB, IRS1S, HSP27 GSK3 και AKT. Τα ερεθίσματα αυτά μαζί με το IFNB1 που είναι αρκετά σημαντικό στην φυσιολογία του ήπατος χρησιμοποιήθηκαν και στο κύριο πείραμα.

Σχεδιασμός και εκτέλεση του συνδυαστικού πειράματος

Στο δεύτερο μέρος της πειραματικής διαδικασίας, τα 15 ερεθίσματα που βρέθηκαν να έχουν έντονη επίδραση στις 14 μετρούμενες πρωτεϊνες χρησιμοποιήθηκαν σε συνδυασμούς των δύο για την κατασκεύη ενός εκτεταμένου συνόλου φωσφοπρωτεομικών δεδομένων που θα αντιπροσωπεύει τους



Σχήμα 3.6: Πειραματική και υπολογιστική διαδικασία: (a) Τα κύτταρα καλλιεργούνται σε 96αρες πλάκες. (b) Εισάγονται τα 81 πιο κύρια ερεθίσματα για τον υπο εξέταση κυτταρικό τύπο. (c) Για κάθε ερέθισμα μετράται η φωσφορυλίωση 14 σημάτων στα 5 και 25 λεπτά (μετά την εισαγωγή του ερεθίσματος). (d) Αναγνωρίζονται οι πιο ενεργές κυττοκίνες. (e) Οι επιλεχθήσες κυττοκίνες εισάγωνται στα κύτταρα σε συνδυαστικά πειράματα. (f) Πραγματοποιείται το κύριο πείραμα. (h) Προεπεξεργαζόμαστε το βιβλιογραφικό δίκτυο προς την αφαίρεση περιττών κόμβων και αντιδράσεων. (i) Το βιβλιογραφικό δίκτυο βελτιστοποιείται με χρήση της μεθόδου Ακέραιου Γραμμικού Προγραμματισμού που περιγράφηκε στην ενότητα 2.1.

σηματοδοτικούς μηχανισμούς του υπο εξέταση κυτταρικού τύπου. Το σύνολο δεδομένων φαίνεται στο σχήμα 3.8, ενώ στο σχήμα3.9 τα εν λόγω δεδομένα έχουν ομαδοποιηθεί ιεραρχικά.

Η πολυπλοκότητα των δεδομένων όπως φαίνονται στο σχήμα 3.8 δεν επιτρέπει την λεπτομερή τους εποπτεία. Ωστόσο μπορούμε να διακρίνουμε κάποια βασικά πρότυπα. Οι αναστολείς MEKi και PI3Ki μπλόκαραν τους στόχους τους: ο αναστολέας MEKi μπλόκαρε την ενεργοποίηση του ERK παρουσία όλων των κυττοκινών. Επίσης ο αναστολέας PI3Ki μπλόκαρε την ενεργοποίηση του AKT παρουσία όλων των κυττοκινών. Σαν περαιτέρω επαλήθευση των αποτελεσμάτων, όλα τα ερεθίσματα που σχετίζονται με κυτταρικό διπλασιασμό ενεργοποίησαν τα AKT, MAP2K1 (MEK), ERK, RPS6KB1,2,3 και RPS6KA1, σε συμφωνία και με το πρώτο πείραμα. Επίσης κυττοκίνες που συνδέονται με φλεγμονώδη αντίδραση ενεργοποίησαν τα αντίστοιχα σήματα, όπως τα IKB,

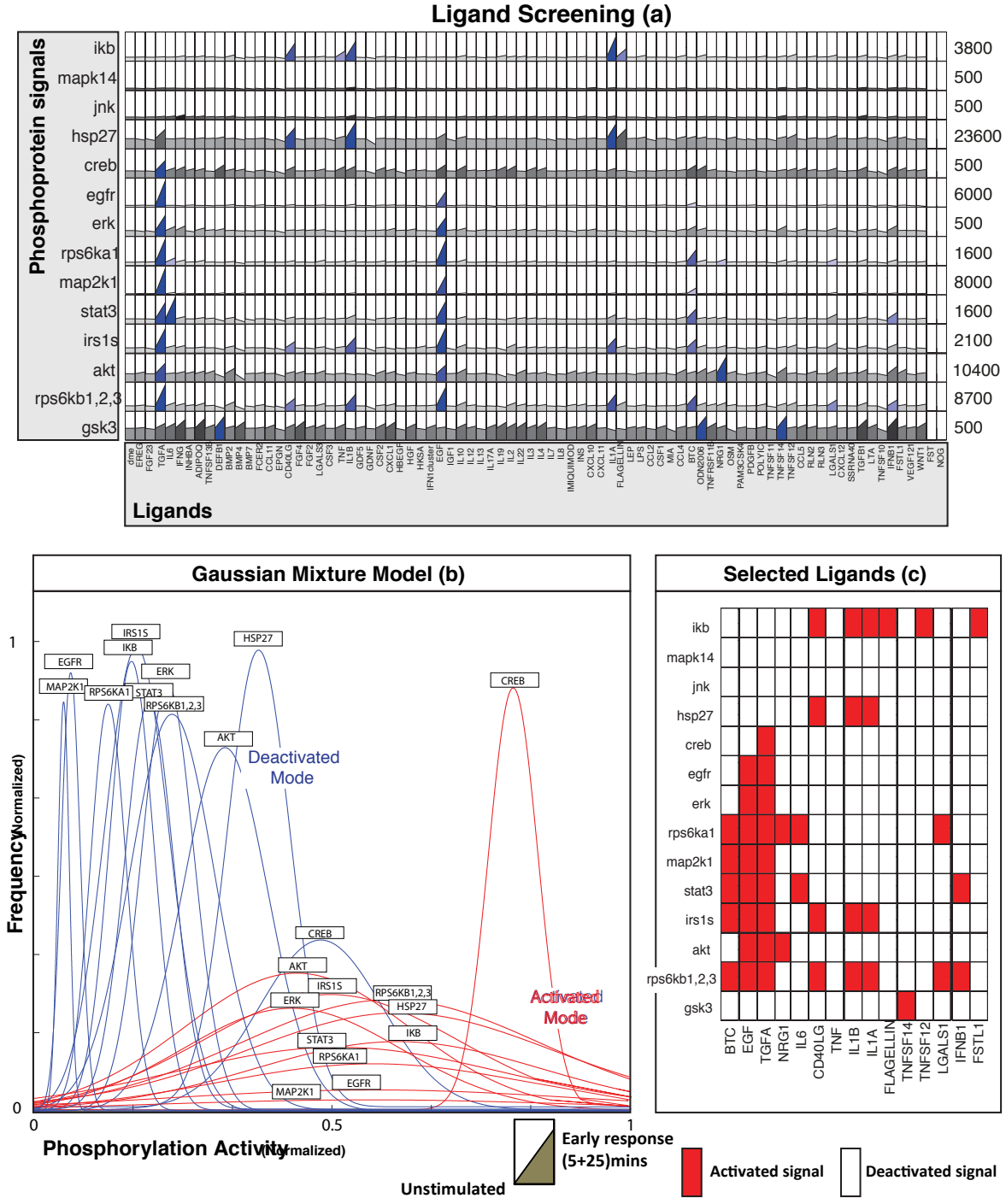
Acronym	Full name (HUGO Nomeclature)		Acronym	Full name HUGO nomeclature
BTC	Betacellulin	hspace10pt	IL22	Interleukin 22
EGF	epidermal growth factor	hspace10pt	LGALS1	lectin, galactoside-binding, soluble, 1
TGFA	transforming growth factor, alpha	hspace10pt	LGALS3	lectin, galactoside-binding, soluble, 3
NRG1	neuregulin 1	hspace10pt	IFNB1	interferon, beta 1, fibroblast
HBEGF	heparin-binding EGF-like growth factor	hspace10pt	IFNG	interferon, gamma
HGF	hepatocyte growth factor	hspace10pt	IFN1cluster	interferon, type 1, cluster
INS	Insulin	hspace10pt	CXCL1	chemokine (C-X-C motif) ligand 1
IGF1	insulin-like growth factor 1	hspace10pt	CXCL10	chemokine (C-X-C motif) ligand 10
LPS	interferon regulatory factor 6	hspace10pt	CXCL11	chemokine (C-X-C motif) ligand 11
FLAGELLIN	component of the bacterial flagellar filament	hspace10pt	CXCL12	chemokine (C-X-C motif) ligand 12
ODN2006	synthetic oligonucleotides	hspace10pt	CCL2	chemokine (C-C motif) ligand 2
IMIQUIMOD	imidazoquinoline amine analogue to guanosine	hspace10pt	CCL4	chemokine (C-C motif) ligand 4
HKSA	preparation of Listeria monocytogenes	hspace10pt	CCL5	chemokine (C-C motif) ligand 5
SSRNA40	20-mer single-stranded RNA oligo	hspace10pt	CCL11	chemokine (C-C motif) ligand 11
OSM	oncostatin M	hspace10pt	FGF2	fibroblast growth factor 2 (basic)
PAM3CSK4	synthetic tripalmitoylated lipopeptide	hspace10pt	FGF4	fibroblast growth factor 4
POLYIC	synthetic analog of double-stranded RNA	hspace10pt	FGF23	fibroblast growth factor 23
BMP2	bone morphogenetic protein 2	hspace10pt	FSTL1	follistatin-like 1
BMP4	bone morphogenetic protein 4	hspace10pt	WNT	wingless-type MMTV integration site family, mem.1
BMP7	bone morphogenetic protein 7	hspace10pt	INHBA	inhibin, beta A
TNF	tumor necrosis factor	hspace10pt	ADIPOQ	adiponectin, C1Q and collagen domain containing
TNFSF10	tumor necrosis factor lig. superfamily, 10	hspace10pt	DEFB1	defensin, beta 1
TNFSF11	tumor necrosis factor lig. superfamily, 11	hspace10pt	FCER2	Fc fragment of IgE, low affinity II, receptor for (CD23)
TNFSF12	tumor necrosis factor lig. superfamily, 12	hspace10pt	EPGN	epithelial mitogen homolog
TNFSF14	tumor necrosis factor lig. superfamily, 14	hspace10pt	EREG	Epiregulin
TNFRSF11B	tumor necrosis factor rec. superfamily, 11b	hspace10pt	CSF3	colony stimulating factor 3 (granulocyte)
TNFSF13B	tumor necrosis factor lig. superfamily, 13b	hspace10pt	GDF5	growth differentiation factor 5
CD40LG	CD40 ligand	hspace10pt	GDNF	glial cell derived neurotrophic factor
IL1A	Interleukin 1 alpha	hspace10pt	CSF2	colony stimulating factor 2
IL1B	Interleukin 1 beta	hspace10pt	LEP	Leptin
IL2	Interleukin 2	hspace10pt	CSF1	colony stimulating factor 1 (macrophage)
IL3	Interleukin 3	hspace10pt	MIA	melanoma inhibitory activity
IL4	Interleukin 4	hspace10pt	PDGFB	platelet-derived growth factor alpha polypeptide
IL6	Interleukin 6	hspace10pt	RLN2	relaxin 2
IL7	Interleukin 7	hspace10pt	RLN3	relaxin 3
IL8	Interleukin 8	hspace10pt	TGFB1	transforming growth factor, beta 1
IL10	Interleukin 10	hspace10pt	LTA	lymphotoxin alpha (TNF superfamily, member 1)
IL12	Interleukin 12	hspace10pt	VEGF121	vascular endothelial growth factor
IL13	Interleukin 13	hspace10pt	FST	follistatin
IL17A	Interleukin 17 alpha	hspace10pt	NOG	noggin
IL19	Interleukin 19	hspace10pt		

Table 3.1: Συντομογραφία των ερεθισμάτων που χρησιμοποιήθηκαν.

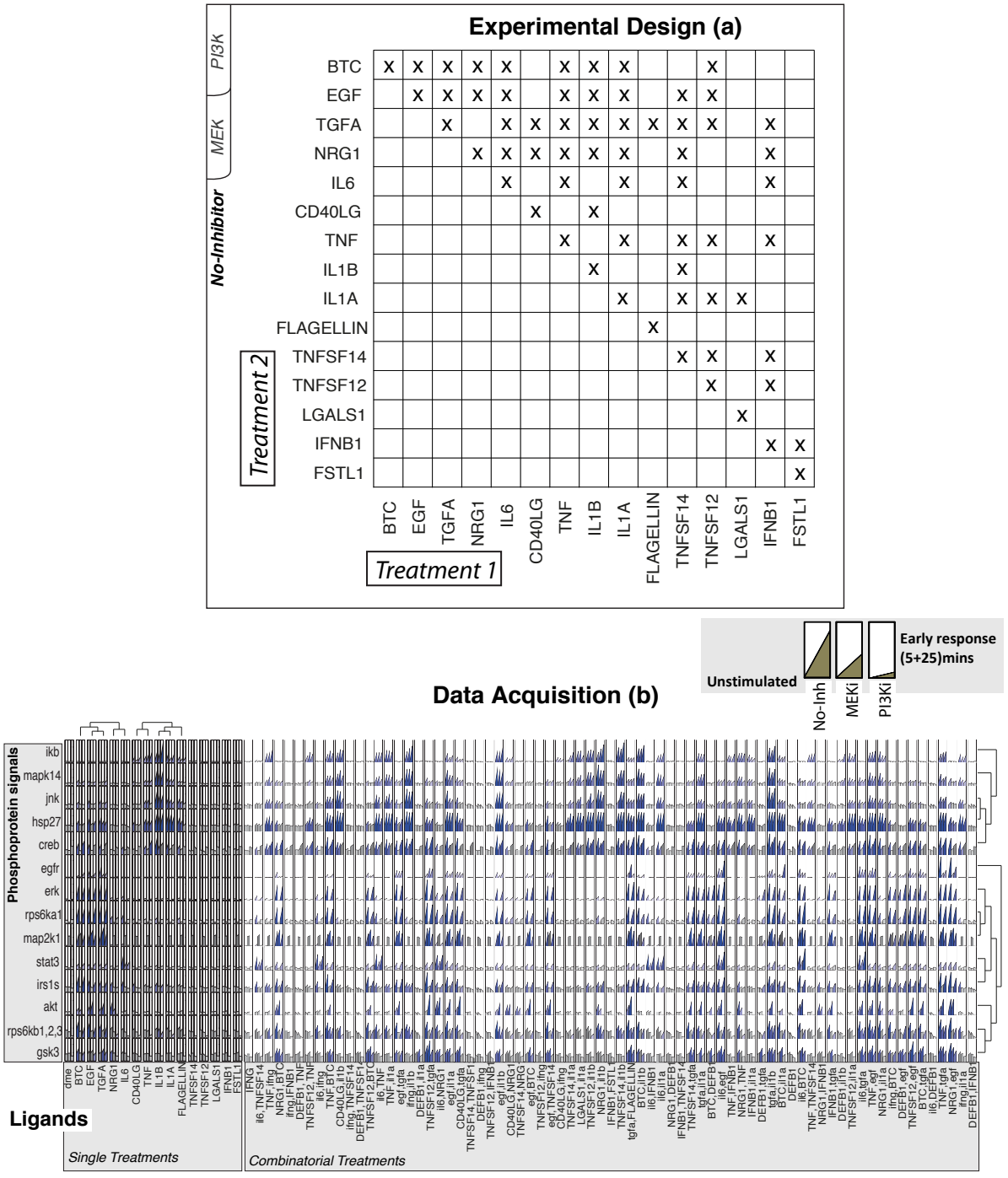
HSP27, MAPK14 και JNK. Παρά το γεγονός ότι τα περισσότερα ερεθίσματα είχαν παρόμοιες επιδράσεις στα υπο εξέταση κύτταρα στα δύο πειράματα, παρατηρήθηκαν κάποιες διαφορές όπως: (i) τα σήματα JNK και MAPK14 ενεργοποιήθηκαν κατόπιν εισαγωγής των IL1A, IL1B και TNF σε αντίθεση με το πρώτο πείραμα όπου υπήρξε μόνο οριακή ενεργοποίηση του JNK και MAPK14. (ii) Τα TGFA, EGF και BTC εμφανίζουν έντονες επιδράσεις και στα δύο πειράματα ωστόσο στο δεύτερο πείραμα ενεργοποιούν εκτός από τα σήματα που σχετίζονται με κυτταρικό διπλασιασμό και τα MAPK14, JNK, HSP27. Η εμπειρία μας είναι ότι μικρές ασυμφωνίες όπως αυτές που παρατηρήθηκαν εδώ είναι αρκετά συχνές και οφείλονται σε θόρυβο της πειραματικής διαδικασίας. Αυτό που είναι μείζονος σημασίας είναι να διατηρούνται οι ποιοτικές τάσεις των σημάτων, καθώς ο αλγόριθμος βελτιστοποίησης που θα χρησιμοποιηθεί σε επόμενο βήμα δεν επιρεάζεται τόσο πολύ από τις απόλυτες τιμές φωσφορυλίωσης όσο από τις ποιοτικές τις τάσεις. Επίσης ποσοστό του πειραματικού θορύβου μπορεί να εξουδετερωθεί με την διαδικασία κανονικοποίησης των δεδομένων που υιοθετούμε.

Κατασκεύη σηματοδοτικού μονοπατιού με βάση την βιβλιογραφία

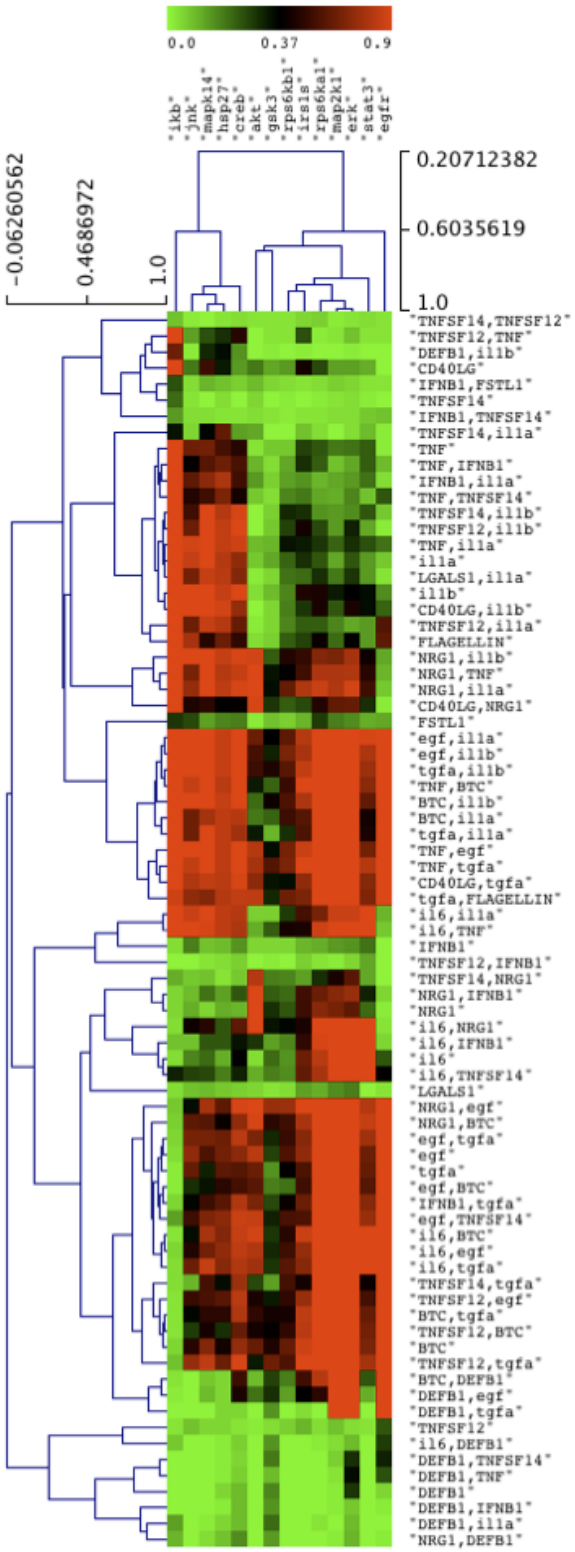
Το σηματοδοτικό δίκτυο κατασκευάζεται από την βιβλιογραφία και αναπαριστά την κυτταρική απόκριση σε 81 ερεθίσματα. Χρησιμοποιήθηκαν οι κυριότερες διαδικτυακές βιβλιοθήκες πρωτεϊνικών



Σχήμα 3.7: Έλεγχος των 81 πιο σημαντικών κυττοκινών: (α) Εισήχθησαν οι 81 κυττοκίνες και μετρήθηκε η ενεργοποίηση 14 φωσφοπρωτεϊνών. (β) Χρησιμοποιήθηκε Γκαουσιανό μοντέλο μίξης για την κατηγοριοποίηση των σημάτων σε ενεργά και ανενεργά. (γ) Ερεθίσματα που ενεργοποίησαν έστω και ένα σήμα επιλέχθηκαν για το επόμενο πείραμα.



Σχήμα 3.8: Συνδυαστικό πείραμα: (α) Σχεδιασμός του συνδυαστικού πειράματος. Τα επιλεχθέντα ερεθίσματα εισάγονται 3 φορές: 1. χωρίς την παρουσία κάποιου αναστολέα, 2. παρουσία αναστολέα της πρωτεΐνης MEK (MEKi), και 3. παρουσία αναστολέα της πρωτεΐνης PI3K(PI3Ki). (β) Τα φωσφοπρωτεομικά δεδομένα ανακτώνται και απεικονίζονται μέσω του λογισμικού DataRail [50]



Σχήμα 3.9: Ιεραρχική ομαδοποίηση του δεύτερου συνδυαστικού πειράματος.

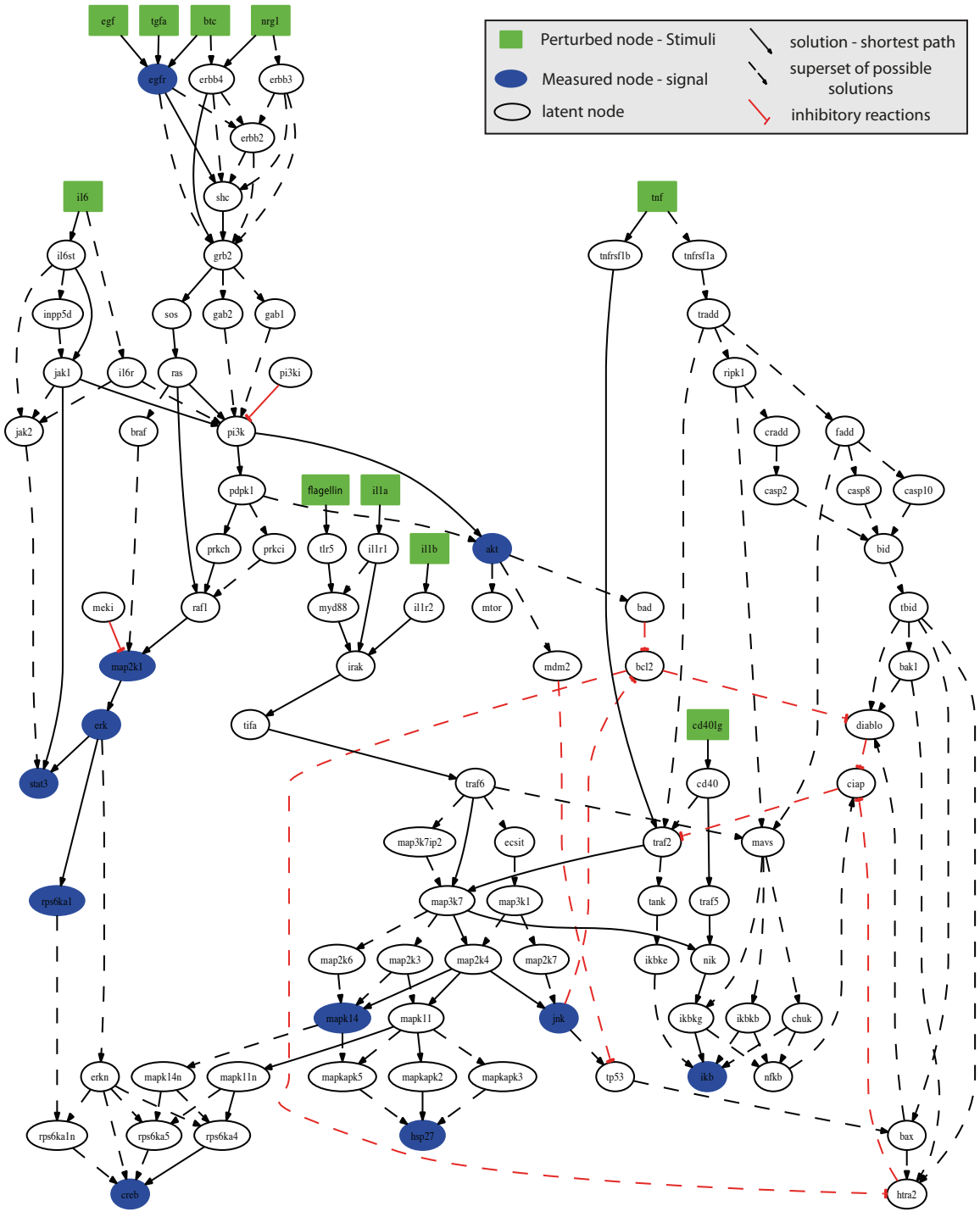
αλληλεπιδράσεων, ωστόσο οι περισσότερες αντιδράσεις προέρχονται από το KEGG - (http://www.genome.jp/kegg/) και το Ingenuity - (http://www.ingenuity.com/). Το δίκτυο περιλαμβάνει περισσότερους από 500 κόμβους και 1000 αντιδράσεις.

Βελτιστοποίηση του σηματοδοτικού μονοπατιού

Η διαδικασία βελτιστοποίησης βασίζεται στην μέθοδο της ενότητας 2.1. Δέχεται σαν είσοδο το βιβλιογραφικό σηματοδοτικό δίκτυο και το σύνολο φωσφοπρωτεομικών δεδομένων και αφαιρεί τις αντιδράσεις εκείνες του δικτύου που αντικρούουν τα πειραματικά δεδομένα. Εξ αιτίας του μεγέθους του εν λόγω δικτύου, μπορεί να υπάρχουν πολλές λύσεις με την ίδια (βέλτιστη) ασυμφωνία με τα πειραματικά δεδομένα, επομένως τρέχουμε τον αλγόριθμο επαναληπτικά και σε κάθε εκτέλεση υπολογίζουμε και μια διαφορετική λύση με σκοπό την αρίθμηση όλων των πιθανών λύσεων. Για την επίλυση του Ακέραιου Γραμμικού Προγράμματος χρησιμοποιήθηκε ο επιλύτης GUROBI (Gurobi library version 3.0.1. Houston, Texas: Gurobi Optimization, Inc., http://www.gurobi.com/) μέσω του λογισμικού GAMS. Τα αποτελέσματα φαίνονται στο σχήμα3.10.

Από τις αντιδράσεις που υπήρχαν στην αρχική τοπολογία οι περισσότερες αφαιρέθηκαν από τον αλγόριθμο βελτιστοποίησης. 161 αντιδράσεις διατηρήθηκαν συνολικά, ενώ η μικρότερη σε μέγεθος λύση απαρτίζεται από 53 αντιδράσεις. Σαν επαλήθευση των αποτελεσμάτων, παρατηρούμε ότι τα μοτίβα ενεργοποίσης που διακρίνονται στα πρωτεομικά δεδομένα διακρίνονται και στην βελτιστοποιημένη τοπολογία. Τα ερεθίσματα που σχετίζονται με κυτταρικό διπλασιασμό όπως τα EGF, TGFA, NRG1, BTC και IL6, σηματοδοτούν από παρόμοια μονοπάτια και ενεργοποιούν τα σήματα EGFR, AKT, MAP2K1, ERK, STAT3 και RPS6KA1. Από την άλλη, ερεθίσματα όπως τα TNF, CD40LG, IL1A, IL1B και FLAGELLIN που σχετίζονται με την φλεγμονώδη αντίδραση του οργανισμού, σηματοδοτούν μέσω του TRAF6, TRAF2 και των MAPKs και ενεργοποιούν τα MAPK14, IKB, JNK, HSP27 και CREB. Πιο συγκεκριμένα, σχετικά με το μονοπάτι του EGFR: Τα NRG1, BTC, EGF και TGFA σηματοδοτούν μέσω του GRB2 στο RAS και έπειτα είτε στα (i) PI3K - AKT ή στά (ii) RAF - MAP2K1 - ERK - RPS6KA - CREB, ERK - STAT3. Επίσης, ένας άλλος κλάδος εκκινά από το IL6, ενεργοποιεί το JAK1 και τα (i) PI3K - AKT και το (ii) STAT3. Σχετικά με τα προ-φλεγμονώδη ερεθίσματα IL1A, IL1B, FLAGELLIN, CD40LG και TNF: Τα IL1A, IL1B και FLAGELLIN σηματοδοτούν μέσω του TRAF6 στο MAP3K7 και έπειτα στα (i) NIK - IKB, και στα (ii) MAP2K4 - JNK, MAP2K4 - MAPK14, MAP2K4 - MAPK11 - HSP27, MAPK11 - RPS6KA4 - CREB. Το CD40LG σηματοδοτεί μέσω του TRAF5 στο NIK και ενεργοποιεί το IKB. Το TNF σηματοδοτεί μέσω του TRAF2 στο MAP3K7 και ενεργοποιεί τα MAPK14, CREB, HSP27, JNK και IKB. Είναι ξεκάθαρο ότι η ελάχιστη σε μέγεθος λύση (με τις έντονες γραμμές) και η υπέρθεση όλων των υπολοίπων λύσεων (διακεκομένες γραμμές) υλοποιούν την ίδια ακριβώς συνδεσμολογία μεταξύ ερεθισμάτων και φωσφοπρωτεϊνών. Η διαφορά τους έγγυται μόνο στο γεγονός ότι οι λύσεις με διακεκομένες περιλαμβάνουν όλους τους πιθανούς τρόπους να υλοποιηθεί αυτή η συνδεσμολογία, ενώ η λύση με έντονες γραμμές περιλαμβάνει μόνο τον τρόπο με το ελάχιστο πλήθος αντιδάσεων.

Η ασυμφωνία μεταξύ υπολογιστικού μοντέλου και πειραματικών δεδομένων μειώνεται από 31% στο 7%, επαληθεύοντας ότι το βιβλιογραφικό δίκτυο από μόνο του δεν ήταν αντιπροσωπευτικό των σηματοδοτικών μηχανισμών των φυσιολογικών ηπατοκυττάρων. Η προτεινόμενη μέθοδος Ακέραιου Γραμμικού Προγραμματισμού συνδυάζοντας το βιβλιογραφικό δίκτυο με τα πειραματικά δεδομένα, αφαίρεσε τις αντιδράσεις εκείνες που αντέκρουαν τα δεδομένα οδηγόντας σε ένα μοντέλο, εξειδικευμένο, για τον υπο εξέταση κυτταρικό τύπο.

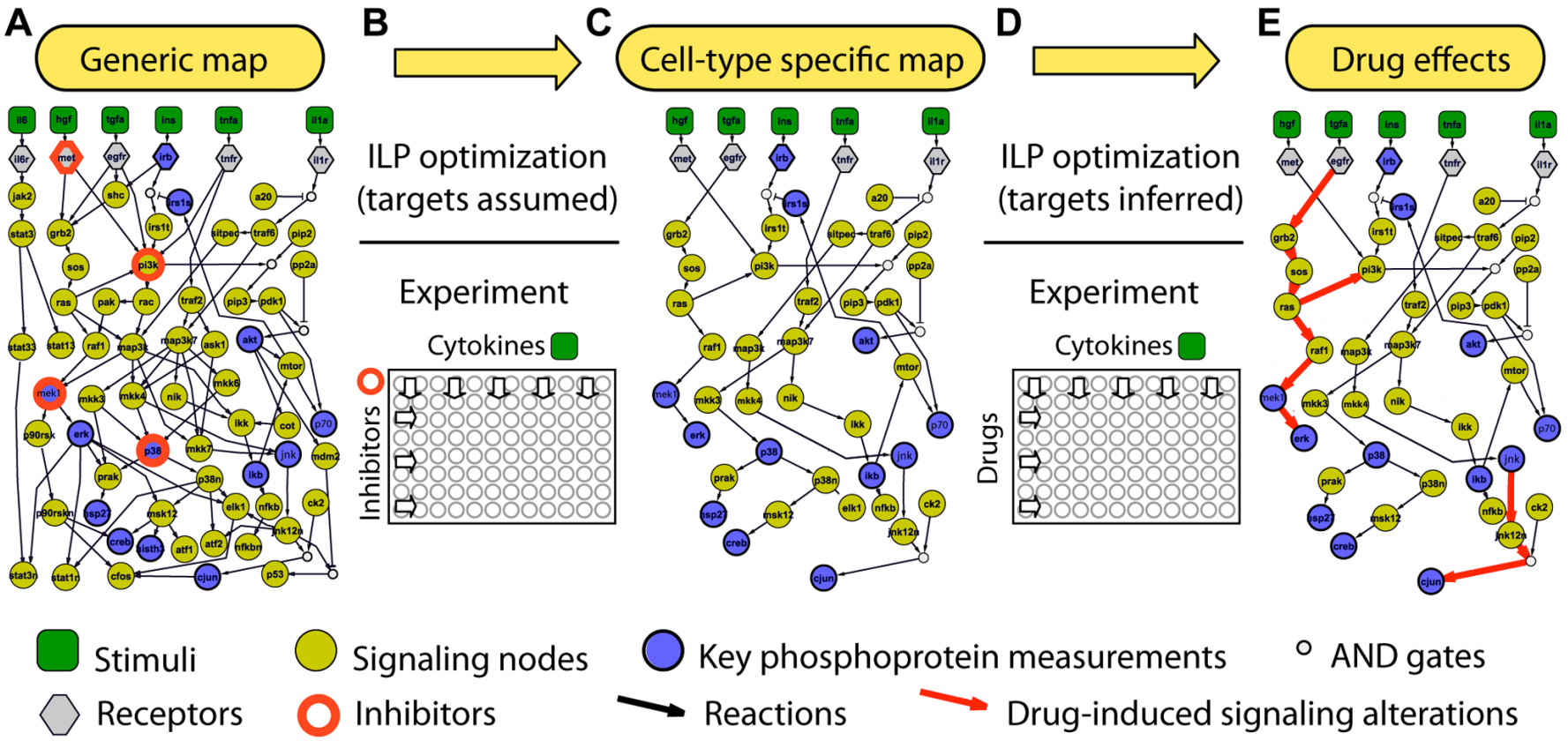


Σχήμα 3.10: Βελτιστοποιημένο δίκτυο: Το βελτιστοποιημένο δίκτυο αποτελείται από ολές τις αντιδράσεις και κόμβους που δεν αντικρούουν τα πειραματικά δεδομένα. Με έντονες γραμμές φαίνονται οι αντιδράσεις που απαρτίζουν την μικρότερη σε μέγεθος λύση.

2.3 2η εφαρμογή: Αναγνώριση των επιδράσεων αντικαρκινικών φαρμάκων στο δίκτυο σηματοδότησης καρκινικών ηπατοκυττάρων

Σε αυτήν την ενότητα χρησιμοποιούμε την μεθοδολογία που περιγράφηκε στην ενότητα 2.1 για την αναγνώριση των επιδράσεων 4 αντικαρκινικών φαρμάκων στο δίκτυο σηματοδότησης καρκινικών ηπατοκυττάρων. Η έρευνα αυτή δημοσιεύτηκε από τους Mitsos et al. [23] και πραγματοποιήθηκε σε συνεργασία με τον Αλέξανδρο Μητσό (την στιγμή της συγκεκριμένης δημοσίευσης επίκουρο καθηγητή στο τμήμα μηχανολόγων μηχανικών του MIT, Cambridge, MA, USA, την παρούση στιγμή καθηγητή στο RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany). Η προτεινόμενη μεθοδολογία οδήγησε στην ανακάλυψη παρενεργειών (off target effects) για τα υπο εξέταση φάρμακα που μέχρι εκείνη την στιγμή ήταν άγνωστες, συνδράμοντας έτσι στην κατανόηση του τρόπου δράσης των εν λόγω φαρμάκων και των χαρακτηριστικών που τα καθιστούν αποτελεσματικά ή όχι.

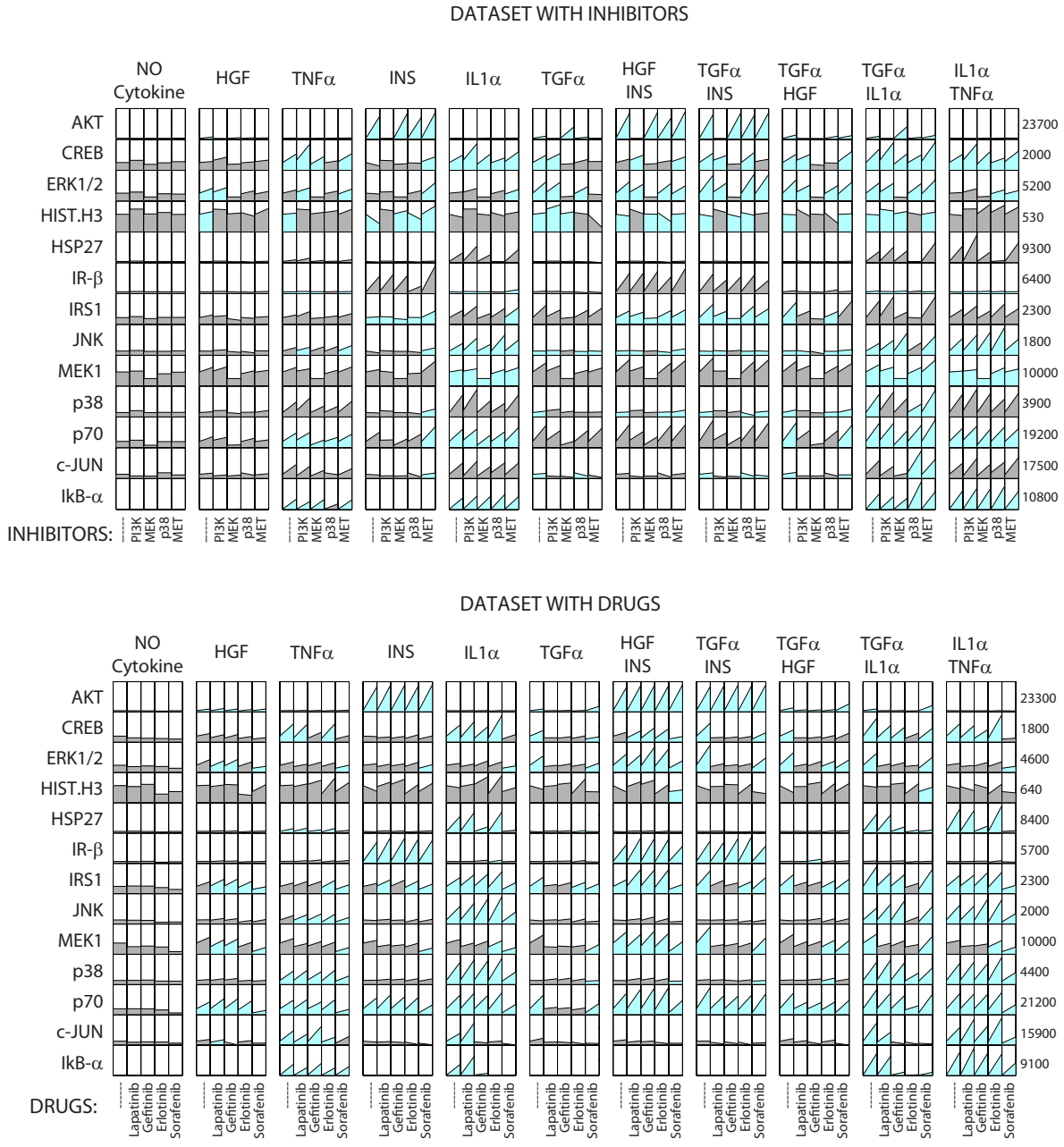
Πιο συγκεκριμένα, η προτεινόμενη προσέγγιση αποτελείται από τα εξής βήματα (δείτε επίσης σχήμα 3.11): (i) Εκτέλεση φωσφοπρωτεομικών πειραμάτων για την ποσοτικοποίηση των σηματοδοτικών μηχανισμών καρκινικών ηπατοκυττάρων. (ii) Κατασκεύη σηματοδοτικού μοντέλου, εξειδικευμένου για τα καρκινικά ηπατοκύτταρα. (iii) Εκτέλεση δεύτερης σειράς πειραμάτων για την μέτρηση των επιδράσεων των υπο εξέταση φαρμάκων στα καρκινικά ηπατοκύτταρα. (iv) Κατασκευή σηματοδοτικών μοντέλων για το εκάστοτε φάρμακο. Οι επιδράσεις του φαρμάκου εκφράζονται σαν αντιδράσεις που έχουν μπλοκαριστεί από την εισαγωγή του. Για την εκτέλεση των φωσφοπρωτεομικών πειραμάτων χρησιμοποιείται η τεχνολογία xMAP της Luminex [12], και μετρώνται τα επίπεδα ενεργοποίησης 13 φωσφοπρτεϊνων σε περισσότερους από 50 συνδυασμούς κυττοκινών και των 4 φαρμάκων. Για την κατασκεύη σηματοδοτικών μονοπατιών χρησιμοποιείται η μέθοδος Ακέραιου Γραμμικού Προγραμματισμού που περιγράφηκε στην ενότητα 2.1, η οποία συνδυάζει τα πρωτεομικά δεδομένα με βιβλιογραφικά σηματοδοτικά δίκτυα και αφαιρεί τις αντιδράσεις εκείνες απο το δίκτυο που αντικρούουν τα δεδομένα. Τα ακόλουθα φάρμακα χρησιμοποιήθηκαν: (1) Lapatinib (αναστολέας του (EGFR/ErbB-2)) [59], (2) Erlotinib (Αναστολέας της κινάσης του EGFR [60]), (3) Gefitinib (Αναστολέας της κινάσης του EGFR [61]), (4) Sorafenib (Αναστολέας του RAF [62]).



Σχήμα 3.11: Πειραματική και υπολογιστική διαδικασία. (Α) Κατασκευάζεται το Boolean μοντέλο του βιβλιογραφικού δικτύου στην γειτονιά των χρησιμοποιούμενων ερεθισμάτων (πράσινα παραλληλόγραμμα) και των μετρούμενων φωσφοπρωτεϊνών (κίτρινοι κυκλικοί κόμβοι). (Β) Εισάγωνται συνδυασμοί των κυττοκινών με αναστολείς για την δημιουργία συνόλου φωσφοπρωτεομικών δεδομένων που αντιπροσωπεύουν τους σηματοδοτικούς μηχανισμούς των καρκινικών ηπατοκυττάρων. (Γ) Βελτιστοποιείται το βιβλιογραφικό δίκτυο στα πειραματικά δεδομένα. (Δ) Εκτέλεση δεύτερου φωσφοπρωτεομικού πειράματος κατά το οποίο εισάγωνται συνδυασμοί κυττοκινών με τα 4 υπο εξέταση φάρμακα σε καρκινικά ηπατοκύτταρα. (Ε) Βελτιστοποίηση του σηματοδοτικού δικτύου στα δεδομένα από το εκάστοτε φάρμακο για την αναγνώριση των επιδράσεων των 4 φαρμάκων στο εξειδικευμένο σηματοδοτικό μονοπάτι. Οι αντιδράσεις που μπλοκάρονται από το εκάστοτε φάρμακο φαίνονται με κόκκινο.

Μέτρηση φωσφοπρωτεομικών δεδομένων

Στο πρώτο σκέλος της πειραματικής διαδικασίας μετράται η απόκριση καρκινικών ηπατοκυττάρων σε συνδυασμούς ερεθισμάτων (κυττοκίνες και αναστολείς) με σκοπό την δημιουργία ενός συνόλου δεδομένων που αντιπροσωπεύει τους σηματοδοτικούς μηχανισμούς του υπο εξέταση κυτταρικού τύπου. Τα δεδομένα φαίνονται στο σχήμα 3.12



Σχήμα 3.12: Φωσφοπρωτεομικά δεδομένα για την κατασκευή εξειδικευμένου σηματοδοτικού δικτύου.

Κατασκεύη σηματοδοτικού μονοπατιού με βάση την βιβλιογραφία

Το σηματοδοτικό δίκτυο κατασκευάζεται από την βιβλιογραφία και αναπαριστά την κυτταρική απόκριση σε 5 ερεθίσματα. Χρησιμοποιήθηκαν οι κυριότερες διαδικτυακές βιβλιοθήκες πρωτεϊνικών αλληλεπιδράσεων, ωστόσο οι περισσότερες αντιδράσεις προέρχονται από το KEGG - (http://www.genome.jp/kegg/) και το Ingenuity - (http://www.ingenuity.com/).

Βελτιστοποίηση σηματοδοτικού δικτύου σε πρωτεομικά δεδομένα

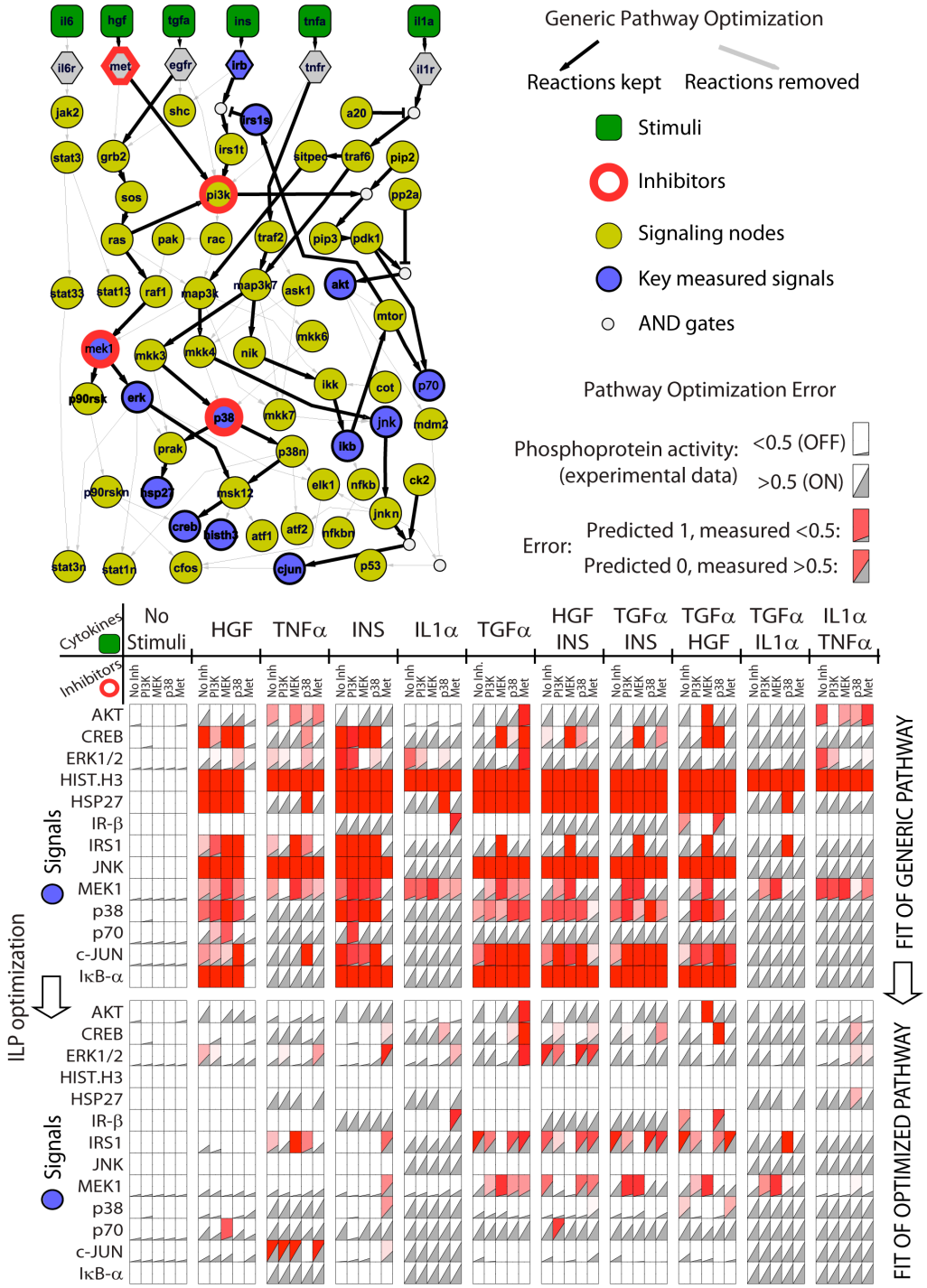
Για την βελτιστοποίηση του βιβλιογραφικού σηματοδοτικού δικτύου σε πρωτεομικά δεδομένα χρησιμοποιήθηκε η μεθοδος Ακέραιου Γραμμικού Προγραμματισμού που περιγράφηκε στην ενότητα 2.1. Η βελτιστοποιημένη τοπολογία φαίνεται στο σχήμα 3.13 και αντιπροσωπεύει τους σηματοδοτικούς μηχανισμούς των καρκινικών ηπατοκυττάρων.

Ο αλγόριθμος βελτιστοποίησης αφαιρεί τις αντιδράσεις του βιβλιογραφικού δικτύου που αντικρούουν τα πειραματικά δεδομένα ενώ διατηρεί μόνο το ελάχιστο σύνολο αντιδράσεων που χρειάζονται για να εξηγήσουν τα δεδομένα (δείτε επίσης σχήμα 3.13). Επι παραδείγματι, το μονοπάτι INS → IRb → IRS1t → PI3K → PIP3 → PDK1 → AKT (σχήμα 3.13) έχει διατηρηθεί στην λύση καθώς στα φωσφοπρωτεομικά δεδομένα, το ερέθισμα INS ενεργοποιεί το σήμα AKT. Αντιθέτως η αντίδραση TNFR → PI3K έχει αφαιρεθεί από το δίκτυο καθώς στα πρωτεομικά δεδομένα το AKT δεν ενεργοποιείται κατόπιν εισαγωγής του TNFa. Δείτε επίσης το μονοπάτι TNFR → PI3K → ... → AKT στο σχήμα 3.13. Εκτός από τις αντιδρασεις που αντικρούουν τα πειραματικά δεδομένα αφαιρούνται και αυτές για τις οποίες δεν υπάρχουν στοιχεία που να στηρίζουν την λειτουργικότητά τους, όπως αντιδράσεις χωρίς σήματα κάτωθεν τους, ή αντιδράσεις χωρίς ερεθίσματα άνωθεν τους. Σαν αποτέλεσμα η λύση αποτελείται μόνο από το ελάχιστο σύνολο αντιδράσεων που απαιτούνται για να εξηγήσουν τα δεδομένα.

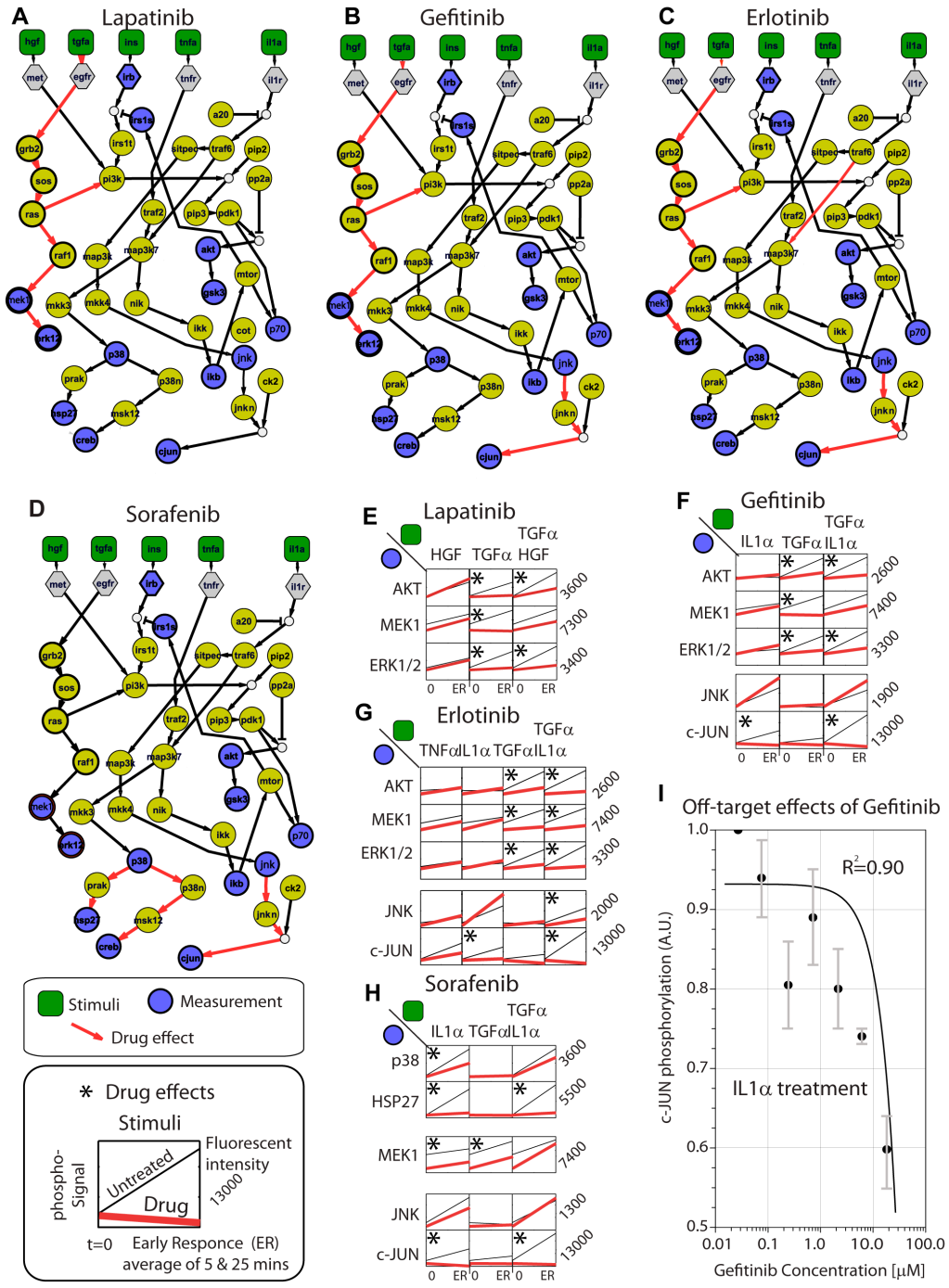
Αναγνώριση των επιδράσεων των 4 υπό εξέταση φαρμάκων στο σηματοδοτικό μονοπάτι των καρκινικών ηπατοκυττάρων

Για την αναγνώριση των επιδράσεων των 4 υπό εξέταση φαρμάκων στο σηματοδοτικό μονοπάτι των καρκινικών ηπατοκυττάρων, πραγματοποιείται δεύτερο φωσφοπρωτεομικό πείραμα κατά το οποίο εισάγωνται συνδυασμοί των 4 φαρμάκων με 5 κυττοκίνες σε καρκινικά ηπατοκύτταρα, ενώ μετράμε την ενεργοποίηση 13 φωσφοπρωτεϊνών. Ακολούθως το σηματοδοτικό μονοπάτι που υπολογίστηκε στο προηγούμενο βήμα (εξειδικευμένο για τα καρκινικά ηπατοκύτταρα) βελτιστοποιείται στα δεδομένα του εκάστοτε φαρμάκου και αναγνωρίζονται οι αντιδράσεις που μπλοκάρονται από την εισαγωγή του. Οι επιδράσεις των 4 αντικαρκινικών φαρμάκων φαίνονται στο σχήμα 3.14.

Αναφορικά με το Lapatinib (αναστολέας του (EGFR/ErbB-2)), μπλοκάρει το μονοπάτι κάτωθεν του TGFa: TGFa → GRB2 → SOS → RAS → PI3K και RAS → RAF1 → MEK1/2 → ERK1/2 (δείτε σχήμα 3.14A), καθώς τα πρωτεομικά δεδομένα δείχνουν αναστολή του MEK1/2, ERK1/2 και AKT παρουσία όλων των ερεθισμάτων (δείτε σχήμα 3.14E). Αναφορικά με το Gefitinib (αναστολέας της κινάσης του EGFR), μπλοκάρει παρόμοιες αντιδράσεις με το Lapatinib, ωστόσο μπλοκάρει επίσης τον κλάδο JNK → c-JUN (δείτε σχήμα 3.14B). Αυτό εξηγείται από την αναστολή του c-JUN κατόπιν εισαγωγής του IL1a και του IL1a+TGFa. Το εύρημα αυτό επαληθεύτηκε και σε ανεξάρτητο πείραμα (δείτε σχήμα 3.14I). Πιστεύουμε ότι η αναστολή του c-JUN από το Gefitinib δεν προκαλείται από την πρόσδεση του Gefitinib στο JNK αλλά από αδύναμες αναστολές σε άλλες πρωτεϊνες που μετέχουν στην μεταγωγή σήματος από το IL1a στο cJUN. Αναφορικά με το Erlotinib (επίσης αναστολέας της κινάσης του EGFR), έχει τις ίδιες επιδράσεις με το Gefitinib (δείτε σχήμα 3.14Γ), ενώ επιπρόσθετα μπλοκάρει τον κλάδο TRAF6 → MAP3K7. Αυτό οφείλεται πιθανότατα στην αναστολή του IkB-a κατόπιν εισαγωγής του IL1a, αλλά όχι κατόπιν εισαγωγής του TNFa (δείτε σχήμα 3.12). Ο μοναδικός τρόπος να εξηγηθεί αυτή η επίδραση



Σχήμα 3.13: Σηματοδοτικό δίκτυο, εξειδικευμένο για τα καρκινικά ηπατοκύτταρα. Με μαύρες έντονες ακμές φαίνονται οι αντιδράσεις που είναι σύμφωνες με τα πειραματικά δεδομένα και έχουν διατηρηθεί στην λύση. Οι αντιδράσεις που αντέκρουαν τα πειραματικά δεδομένα φαίνονται με λεπτές γκρί ακμές. Στο κάτω μέρος της εικόνας φαίνεται το σύνολο φωσφοπρωτεομικών δεδομένων. Με κόκκινο χρώμα φαίνεται η ασυμφωνία του με τα πειραματικά δεδομένα. Παρατηρούμε ότι κατόπιν της διαδικασίας βελτιστοποίησης, η ασυμφωνία έχει μειωθεί σημαντικά.



Σχήμα 3.14: Επιδράσεις των υπο εξέταση φαρμάκων στο σηματοδοτικό μονοπάτι των καρκινικών ηπατοκυττάρων. (A-D) Οι κόκκινες ακμές αντιπροσωπεύουν αντιδράσεις που μπλοκάρει το εκάστοτε φάρμακο. (E-H) Τα φωσφοπρωτεομικά δεδομένα που αντιστοιχούν στις επιδράσεις των φαρμάκων. (I) Επιδράσεις εκτός στόχου του Gefitinib. Η καμπύλη δείχνει ότι το Gefitinib μπλοκάρει την ενεργοποίηση του cJUN κατόπιν εισαγωγής του IL1a. Το R^2 αντιστοιχεί σε γραμμική προσέγγιση.

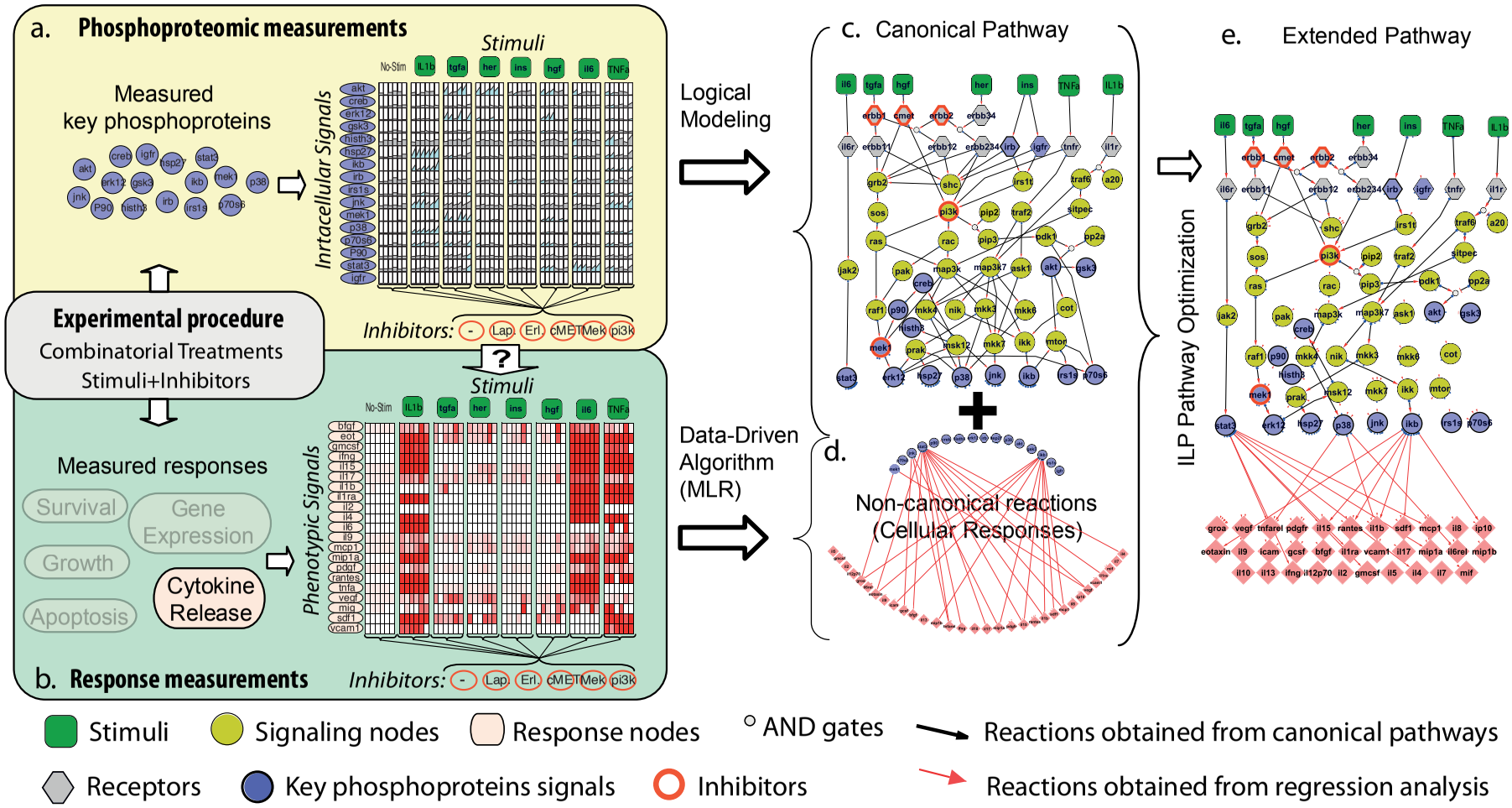
είναι μέσω αφαίρεσης της αντίδρασης TRAF6 → MAP3K. Το Sorafenib (αναστολέας του RAF), έχει πολύ διαφορετικές επιδράσεις από τα υπόλοιπα φάρμακα: Μπλοκάρει τον κλάδο JNK → c-JUN

και επίσης μπλοκάρει το μονοπάτι του p38 (δείτε σχήμα 3.14H). Μια ενδιαφέρουσα παρατήρηση είναι πως το συγκεκριμένο φάρμακο δεν μπλοκάρει τις αντιδράσεις κάτωθεν του RAF που είναι ο στόχος για τον οποίον αναπτύχθηκε. Η παρατήρηση αυτή είναι σύμφωνη με την βιβλιογραφία όπου διατυπώνεται ότι το Sorafenib δεν μπλοκάρει το RAF σε όλες τις κυτταρικές σειρές [66].

3 Κατασκευή εκτεταμένων σηματοδοτικών μονοπατιών ώστε να περιλαμβάνουν ενδοκυτταρική και εξωκυτταρική σηματοδότηση

Σε αυτήν την ενότητα επεκτείνουμε την μεθοδολογία που περιγράφηκε στην ενότητα 2.1 για την κατασκευή εκτεταμένων σηματοδοτικών δικτύων που θα περιλαμβάνουν ενδοκυτταρική και εξωκυτταρική σηματοδότηση. Θα εκκινούν δηλαδή από τα εισαχθέντα ερεθίσματα, θα προχωρούν στο φωσφοπρωτεομικό επίπεδο και την ενεργοποίηση μεταγραφικών παραγόντων και θα τερματίζουν με την έκκριση κυτοκινών στο εξωκυτταρικό περιβάλλον. Η έρευνα αυτή δημοσιεύτηκε από τους Melas et al. [51] και πραγματοποιήθηκε σε συνεργασία με τον Αλέξανδρο Μητσό (την στιγμή της συγκεκριμένης δημοσίευσης επίκουρο καθηγητή στο τμήμα μηχανολόγων μηχανικών του MIT, Cambridge, MA, USA, την παρούση στιγμή καθηγητή στο RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany). Σαν εφαρμογή κατασκευάστηκαν τα εκτεταμένα σηματοδοτικά μονοπάτια σε φυσιολογικά και καρκινικά ηπατοκύτταρα και αναγνωρίστηκαν οι μεταξύ τους διαφορές.

Βάση της προτεινόμενης μεθοδολογίας αποτελεί η μέτρηση φωσφοπρωτεϊνων και εκκρινόμενων κυτοκινών κάτω από τις ίδιες πειραματικές συνθήκες. Εν συνεχεία, χρησιμοποιήθηκε σαν βάση ο αλγόριθμος Ακέραιου Γραμμικού Προγραμματισμού που περιγράφηκε στην ενότητα 2.1, για την περιγραφή της σηματοδοτικής διαδικασίας στο φωσφοπρωτεομικό επίπεδο, ενώ αλγόριθμος γραμμικής παρεκβολής συσχέτισε τις μετρούμενες φωσφοπρωτεΐνες με τις εκκρινόμενες κυτοκίνες. Το εκτεταμένο σηματοδοτικό δίκτυο αντιπροσωπεύει τους σηματοδοτικούς μηχανισμούς της υπο εξέταση κυτταρικής σειράς και στα δύο επίπεδα. Η πειραματική και υπολογιστική διαδικασία συνοψίζεται στο σχήμα 3.15.



Σχήμα 3.15: Πειραματική και υπολογιστική διαδικασία. (a) Το σύνολο φωσφοπρωτεομικών δεδομένων. Περιλαμβάνει δεδομένα για την ενεργοποίηση 16 φωσφοπρωτεϊνων κατόπιν εισαγωγής 8 ερεθισμάτων και 6 αναστολέων. (b) Το σύνολο των εκκρινόμενων κυτοκινών. Περιλαμβάνει μετρήσεις εκκρινόμενων κυτοκινών κατόπιν των ίδιων ερεθισμάτων και αναστολέων που εισάγονται στο φωσφοπρωτεομικό πείραμα. (c) Σηματοδοτικό μονοπάτι με βάση την βιβλιογραφία που περιγράφει την διαδικασία σηματοδότησης κατόπιν εισαγωγής των 7 ερεθισμάτων. (d) Μέθοδος γραμμικής παρεκβολής που χρησιμοποιείται για την συσχέτιση των μετρούμενων φωσφοπρωτεϊνων και εκκρινόμενων κυττοκινών. (e) Το εκτεταμένο δίκτυο βελτιστοποιείται από αλγόριθμο Ακέραιου Γραμμικού Προγραμματισμού.

Κατασκεύη εκτεταμένων σηματοδοτικών μονοπατιών

Η κατασκευή των εκτεταμένων σηματοδοτικών μονοπατιών αποτελείται από τρία διακριτά βήματα. (a) Κατασκευή ενδοκυτταρικών σηματοδοτικών μονοπατιών (εκκινούν από τα 7 ερεθίσματα και τερματίζουν στο φωσφοπρωτεομικό επίπεδο). (b) Υπολογισμός συσχετίσεων μεταξύ των μετρούμενων φωσφοπρωτεϊνων και των εκκρινόμενων κυτοκινών. (c) Βελτιστοποίηση του εκτεταμένου σηματοδοτικού μονοπατιού με χρήση αλγόριθμου Ακέραιου Γραμμικού Προγραμματισμού.

Το σηματοδοτικό μονοπάτι κατασκευάζεται με βάση την βιβλιογραφία στην περιοχή των 7 ερεθισμάτων και 16 φωσφοπρωτεομικών σημάτων. Χρησιμοποιήθηκαν πολλές διαδικτυακές βιβλιοθήκες, ωστόσο οι περισσότερες αντιδράσεις προήλθαν απο το Ingenuity (Redwood City, California) (δείτε σχήμα 3.17a). Οι συσχετίσεις με τις εκκρινόμενες κυτοκίνες υπολογίστηκαν χρησιμοποιόντας μέθοδο γραμμικής παρεκβολής κατά την οποία υπολογίστηκαν βάρη W τέτοια ώστε αν με Y συμβολίσουμε τις εκκρινόμενες κυτοκίνες και με X συμβολίσουμε την ενεργοποίηση των μετρούμενων φωσφοπρωτεϊνων, τοτε να ισχύει: $Y = W \cdot X$ (δείτε σχήμα 3.17c). Οι συσχετίσεις αυτές εισάγονται στο βιβλιογραφικό σηματοδοτικό μονοπάτι προς κατασκευή του εκτεταμένου δικτύου. Τελικά εφαρμόζεται μέθοδος Ακέραιου Γραμμικού Προγραμματισμού για την βελτιστοποίηση του εκτεταμένου σηματοδοτικού μονοπατιού και στα δύο είδη δεδομένων. Ο αλγόριθμος Ακέραιου Γραμμικού Προγραμματισμού κατα αντιστοιχία με την μέθοδο που περιγράφηκε στην ενότητα 2.1 ελαχιστοποιεί την ασυμφωνία μεταξύ πειραματικών δεδομένων και υπολογιστικού μοντέλου, ενώ εισάγει περιορισμούς για να μοντελοποιήσει την μεταγωγή σήματος με όρους Boolean λογικής.

Αλγόριθμος Ακέραιου Γραμμικού Προγραμματισμού

Η ακόλουθη αντικειμενική συνάρτηση ελαχιστοποιείται:

$$\sum_{j,k} a_j^k \mid x_j^k - x_j^{k,m} \mid + \sum_k \sum_{j_{res}} a_{j_{res}}^k \mid x_{j_{res}}^{k,m} - \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res} j_{res}} \mid + \sum_i \beta_i \, y_i \tag{3.12}$$

Ο πρώτος όρος ελαχιστοποιεί την ασυμφωνία μεταξύ φωσφοπρωτεομικών δεδομένων και υπολογιστικού μοντέλου. Ο δεύτερος όρος ελαχιστοποιεί την ασυμφωνία μεταξύ των μετρούμενων τιμών των εκκρινόμενων κυτοκινών και υπολογιστικού μοντέλου. Ο τρίτος όρος ελαχιστοποιεί το μέγεθος της λύσης, αφαιρώντας όλες τις αντιδράσεις που δεν είναι απαραίτητες για να προσομοιώσουν τα πειραματικά δεδομένα. Πιο συγκεκριμένα,

- Το σύνολο $j_{res} = \{1, ..., n_{s,res}\}$ αντιπροσωπεύει τις εκκρινόμενες κυτοκίνες.
- Το σύνολο $i_{res} = \{1, ..., n_{r,res}\}$ αντιπροσωπεύει τις αντιδράσεις που συνδέουν φωσφοπρωτεϊνες (j) και εκκρινόμενες κυτοκίνες (j_{res}).
- Τα a_j^k, $a_{j_{res}}^k, \beta_i \geq 0$ είναι βάρη που ορίζονται από τον χρήστη.
- Τα x_j^k είναι οι τιμές των πρωτεϊνών j στο πείραμα k όπως προβλέπονται από το υπολογιστικό μοντέλο.
- Τα $x_j^{k,m}$, $x_{j_{res}}^{k,m}$ είναι οι μετρούμενες τιμές των πρωτεϊνών j στο πείραμα k.
- Τα $z_i^k \in \{0, 1\}$ αντιπροσωπεύουν την ενεργοποίηση (ή όχι) της αντίδρασης i στο πείραμα k.
- Τα $y_i \in \{0, 1\}$ αντιπροσωπεύουν εάν η αντίδραση i είναι παρούσα στην λύση η όχι.
- Τα $\sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res} j_{res}}$ αντιστοιχούν στις μεταβλητές $x_{j_{res}}^k$ και αντιπροσωπεύουν την τιμή των εκκρινόμενων κυτοκινών j_{res} που προβλέπεται από το υπολογιστικό μοντέλο στο πείραμα k. Ισούται με το άθροισμα όλων των αντιδράσεων i_{res} που οδηγούν στην κυτοκίνη j_{res} σταθμισμένες με βάρη $w_{i_{res} j_{res}}$. Αυτά είναι τα βάρη που υπολογίστηκαν από την γραμμική παρεκβολή σε προηγούμενο βήμα.

Επομένως ο όρος της αντικειμενικής συνάρτησης $\sum_{j,k} a_j^k \mid x_j^k - x_j^{k,m} \mid$ αντιστοιχεί στην ασυμφωνία πειραματικών δεδομένων και υπολογιστικού μοντέλου για όλες τις φωσφοπρωτεΐνες (j) και πειράματα k. Ο δεύτερος όρος $\sum_k \sum_{j_{res}} a_{j_{res}}^k \mid x_{j_{res}}^{k,m} - \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res}j_{res}} \mid$ αντιστοιχεί στην ασυμφωνία πειραματικών δεδομένων και υπολογιστικού μοντέλου για όλες τις εκκρινόμενες κυτοκίνες (j_{res}) και πειράματα k. Ο τρίτος όρος $\sum_i \beta_i \, y_i$ αντιστοιχεί στην ποινή που επιβάλλεται στο μέγεθος της λύσης.

Για την μοντελοποίηση της σηματοδοτικής διαδικασίας στο εκτεταμένο σηματοδοτικό μονοπάτι εισάγωνται γραμμικοί περιορισμοί. Για πλήρη περιγραφή της αρχικής μεθόδου Ακέραιου Γραμμικού Προγραμματισμού δείτε την ενότητα 2.1 ή την δημοσίευση [23]. Εδώ θα συζητηθούν μόνο οι περιορισμοί που αναφέρονται στις εκκρινόμενες κυτοκίνες.

Αναφορικά με τον όρο $\sum_k \sum_{j_{res}} a_{j_{res}}^k \mid x_{j_{res}}^{k,m} - \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res}j_{res}} \mid$, θέτοντας $a_{j_{res}}^k \geq 0, \quad \mid x_{j_{res}}^{k,m} - \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res}j_{res}} \mid \in [0,1]$ να αντιστοιχεί στην σταθμισμένη ασυμφωνία μεταξύ πειραματικών δεδομένων και υπολογιστικού μοντέλου, η ελάχιστη και η μέγιστη τιμή του είναι

$$v^{min} = \sum_{i_{res}, w_{i_{res}j_{res}} < 0} z_{i_{res}}^k \, w_{i_{res}j_{res}} \tag{3.13}$$

$$v^{max} = \sum_{i_{res}, w_{i_{res}j_{res}} \geq 0} z_{i_{res}}^k \, w_{i_{res}j_{res}} \tag{3.14}$$

Σκοπός μας είναι να ελαχιστοποιήσουμε το σταθμισμένο άθροισμα των απόλυτων διαφορών $\hat{d}_{j_{res}}^k = \mid x_{j_{res}}^{k,m} - \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res}j_{res}} \mid$. Υποθέτοντας ότι οι μετρήσεις είναι συνεπείς με τα βάρη θα είχαμε $x_{j_{res}}^{k,m} \in [v^{min}, v^{max}]$ που συνεπάγεται $\hat{d}_{j_{res}}^k \in [0,\, v^{max} - v^{min}]$. Ωατόσο αυτό δεν ισχύει πάντα επομένως υποθέτουμε την πιο γενική περίπτωση

$$\hat{d}_{j_{res}}^k \in [0,\, \hat{d}_{j_{res}}^{k,max}] \tag{3.15}$$

όπου

$$\hat{d}_{j_{res}}^{k,max} = max(v^{max}, x_{j_{res}}^{k,m}) - min(v^{min}, x_{j_{res}}^{k,m}) \tag{3.16}$$

Μπορούμε να σταθμίσουμε επομένως ως εξής

$$\hat{d}_{j_{res}}^k \, \hat{d}_{j_{res}}^{k,max} = \mid x_{j_{res}}^{k,m} - \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res}j_{res}} \mid \tag{3.17}$$

σε

$$\hat{d}_{j_{res}}^k \, \hat{d}_{j_{res}}^{k,max} = x_{j_{res}}^{k,m} - \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res}j_{res}} \tag{3.18}$$

$$\hat{d}_{j_{res}}^k \, \hat{d}_{j_{res}}^{k,max} = -x_{j_{res}}^{k,m} + \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res}j_{res}} \tag{3.19}$$

και έτσι να υπολογίσουμε το επιθυμητό διάστημα $(d_j^k \in [0,1])$. Για να εξασφαλίσουμε την γραμμικότητα επιβάλλουμε δύο περιορισμούς που είναι ισοδύναμοι με $a_j^k \geq 0$,

$$-\hat{d}_{j_{res}}^k \, \hat{d}_{j_{res}}^{k,max} - \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res}j_{res}} \leq -x_{j_{res}}^{k,m} \tag{3.20}$$

$$-\hat{d}_{j_{res}}^k \, \hat{d}_{j_{res}}^{k,max} + \sum_{i_{res}} z_{i_{res}}^k \, w_{i_{res}j_{res}} \leq -x_{j_{res}}^{k,m} \tag{3.21}$$

Οι παραπάνω περιορισμοί ολοκληρώνουν την μέθοδο.

Το παραπάνω Ακέραιο Γραμμικό Προγραμμα επιλύεται με την βοήθεια του CPLEX, εμπορικού κώδικα γραμμικής βελτιστοποίησης μέσω του GAMS. Επειδή είναι πιθανό να υπάρχουν περισσότερες από μία βέλτιστες λύσεις ο κώδικας CPLEX υπολογίζει μέχρι 100 λύσεις με ίδια τιμή της αντικειμενικής συνάρτησης. Το βελτιστοποιημένο μονοπάτι φαίνεται στο σχήμα 2.20b.

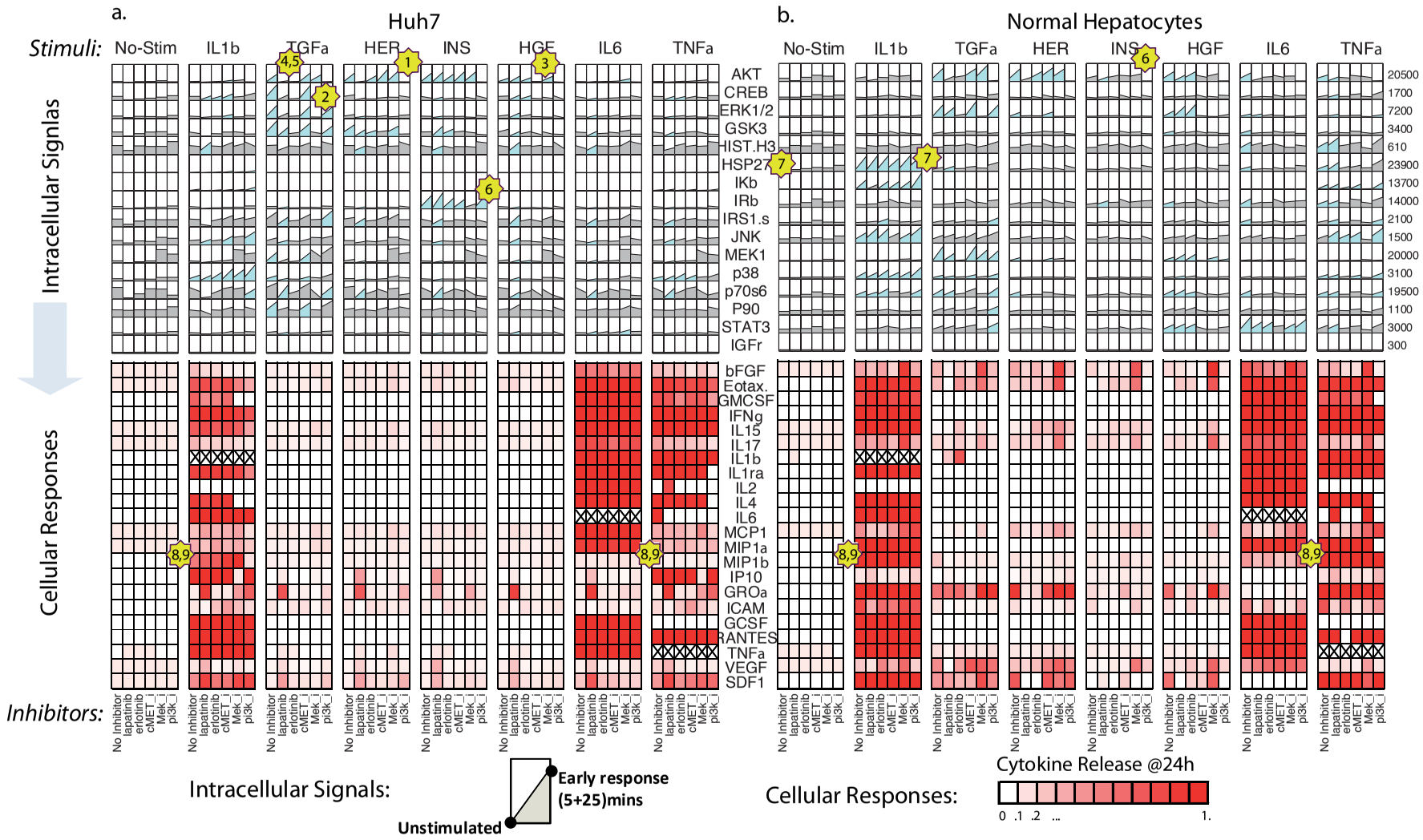
Αναγώριση διαφορών στα εκτεταμένα σηματοδοτικά μονοπάτια φυσιολογικών και καρκινικών ηπατοκυττάρων

Σαν εφαρμογή κατασκευάζουμε εκτεταμένα σηματοδοτικά μονοπάτια για φυσιολογικά και καρκινικά ηπατοκύτταρα και αναγνωρίζουμε τις μεταξύ τους διαφορές. Για την μέτρηση των πειραματικών δεδομένων χρησιμοποιούμε την τεχνολογία xMAP της Luminex. Εισάγουμε συνδυασμούς 7 διαφορετικών ερεθισμάτων και 5 αναστολέων ενώ μετράμε 16 σήματα στο φωσφοπρωτεομικό επίπεδο και 22 εκκρινόμενες κυττοκίνες. Τα πειραματικά δεδομένα (φωσφοπρωτεϊνες και εκκρινόμενες κυττοκίνες) φαίνονται στο σχήμα 3.16.

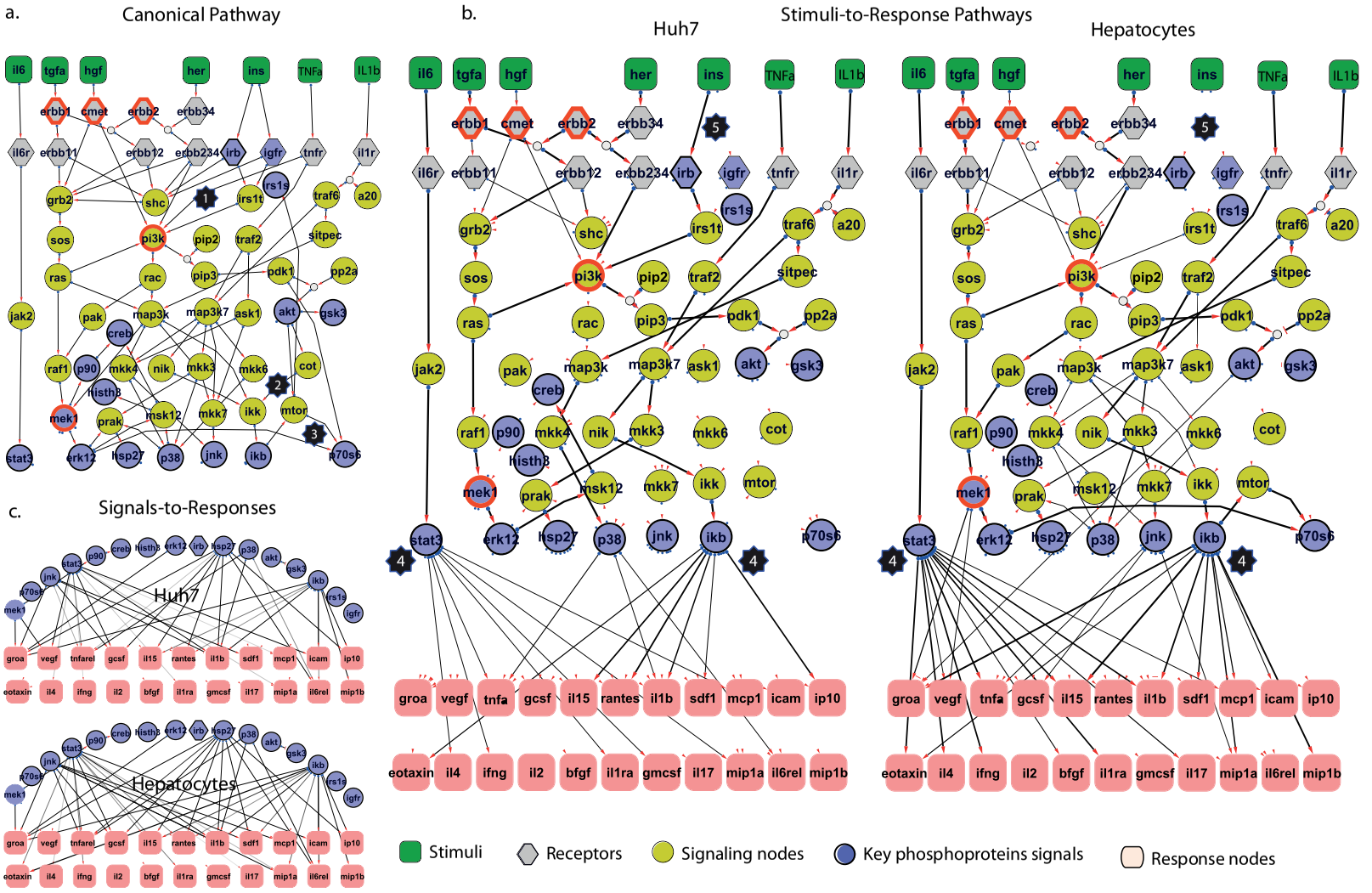
Τα βελτιστοποιμένα σηματοδοτικά μονοπάτια φαίνονται στο σχήμα 3.17. Σημαντικές διαφορές μεταξύ των δύο κυτταρικών τύπών (όπως αυτές επαληθεύονται και από τα δεδομένα) είναι πρώτον: η πιο έντονη απόκριση των καρκινικών ηπατοκυττάρων Huh7 στην Ινσουλίνη (INS), ενεργοποιώντας το AKT και το IRb σε αντίθεση με τα φυσιολογικά ηπατοκύτταρα που φαίνεται να μην επηρεάζονται από την Ινσουλίνη (δείτε σχήμα 2.19, star6). Δεύτερον: η πιο έντονη απόκριση των φυσιολογικών κυττάρων σε IL1b με την ενεργοποίηση του HSP27. Σχετικά με τις εκκρινόμενες κυτοκίνες, εκεί υπάρχουν πολύ εντονότερες διαφορές καθώς τα φυσιολογικά κύτταρα αποκρίνονται πολύ πιο έντονα απελευθερώνοντας κυττοκίνες όπως MIP1a και MIP1b παρουσία TNFa και IL1b, επαληθεύοντας προηγούμενες παρατηρήσεις ότι τα καρκινικά κύτταρα αποφεύγουν τους αμυντικούς μηχανισμούς του οργανισμού χάρη στην απουσία εκκρινόμενων φλεγμονοδών κυτοκινών (δείτε σχήμα 3.16 star8, star9).

Τα εκτεταμένα σηματοδοτικά δίκτυα όπως φαίνονται στο σχήμα 3.17 επιτυγχάνουν την αναγνώριση των φωσφοπρωτεϊνων εκείνων που ευθύνονται για την έκκριση κυτοκινών και συνεπώς την φλεγμονώδη αντίδραση των εν λόγω κυττάρων. Έτσι σαν κύρια μονοπάτια για την έκκριση κυτοκινών αναγνωρίζονται τα IL6 $\rightarrow$ STAT3, IL1b $\rightarrow$ NFkB/p38, και TNFa $\rightarrow$ NFkB/p38. Τα ευρήματα αυτά είναι σύμφωνα και με την βιβλιογραφία [77]. Μόλις τρεις φωσφοπρωτεϊνες (STAT3, Ikb και λιγότερο το p38) φαίνονται να είναι υπεύθυνες για την έκκριση των περισσότερων φλεγμονωδών σημάτων συμπεριλαμβανομένων των TNFa, GROa, RANTES, MCP1, ILb, και EOTAXIN (δείτε σχήμα 3.17, star4), μια παρατήρηση που επίσης είναι σύμφωνη με μεγάλο μέρος της βιβλιογραφίας. Αυτό που ωστόσο δεν είναι γνωστό είναι τα διαφορετικά μονοπάτια που οδηγούν στην ενεργοποίηση των φωσφοπρωτεϊνών αυτών και συνεπώς στην έκκριση κυτοκινών. Τα αποτελέσματα μας δείχνουν ότι υπάρχουν το πολύ τρία τέτοια μονοπάτια (όπως φαίνεται στο σχήμα 3.17) διευκολύνοτας έτσι τον σχεδιασμό φαρμάκων που θα αποσκοπούν στην αναστολή των εκκρινόμενων κυτοκινών.

Η παρούσα έρευνα είναι από τις πρώτες προσπάθειες να αναγωριστούν οι μηχανισμοί που συνδέουν το φωσφοπρωτεομικό επίπεδο με την έκκριση κυτοκινών οδηγώντας στην κατασκευή εκτεταμένων μοντέλων που ενσωματώνουν και τις δύο πτυχές της κυτταρικής σηματοδότησης (ενδοκυτταρική και εξωκυτταρική).



Σχήμα 3.16: Σύνολο πειραματικών δεδομένων για (a) την καρκινική σειρά Huh7 και (b) φυσιολογικά ηπατοκύτταρα. Τα πρώτα 2 σχήματα αντιστοιχούν στα φωσφοπρωτεομικά δεδομένα, τα επόμενα στις εκκρινόμενες κυτοκίνες.



Σχήμα 3.17: Εκτεταμένα σηματοδοτικά μονοπάτια για φυσιολογικά και καρκινικά ηπατοκύτταρα τύπου Huh7. (a) Το βιβλιογραφικό σηματοδοτικό μονοπάτι. (b) Συσχετίσεις μεταξύ φωσφοπρωτεϊνών και εκκρινόμενων κυτοκινών όπως υπολογίστηκαν από μέθοδο γραμμικής παρεκβολής. (c) Βελτιστοποιημένα σηματοδοτικά μονοπάτια για φυσιολογικά και καρκινικά ηπατοκύτταρα τύπου Huh7

4 Μέθοδος Μη Γραμμικου Προγραμματισμού για την ποσοτική μοντελοποίηση σηματοδοτικών δικτύων

Στην παρούσα ενότητα εισάγουμε μέθοδο Μη Γραμμικού Προγραμματισμού για την ποσοτική μοντελοποίηση σηματοδοτικών δικτύων. Η έρευνα αυτή δημοσιεύτηκε από τους Mitsos et al., [86] και πραγματοποιήθηκε σε συνεργασία με τον Αλέξανδρο Μητσό (την στιγμή της συγκεκριμένης δημοσίευσης επίκουρο καθηγητή στο τμήμα μηχανολόγων μηχανικών του MIT, Cambridge, MA, USA, την παρούση στιγμή καθηγητή στο RWTH Aachen University, AVT Process Systems Engineering (SVT), Germany) και τον Douglas A. Lauffenburger (head of the Biological Engineering department of MIT, Cambridge, MA, USA). Σε αντίθεση με προηγούμενες μεθόδους η εφαρμογή μεθόδων Μη Γραμμικου Προγραμματισμού επιταχύνει σημαντικά την βελτιστοποίηση σηματοδοτικών δικτύων σε πρωτεομικά δεδομένα επιτρέποντας έτσι την δημιουργία εκτεταμένων μοντέλων που περιγράφουν πιο πιστά τους σηματοδοτικούς μηχανισμούς του υπο εξέταση κυτταρικού τύπου.

Πιο συγκεκριμένα, η προτεινόμενη μέθοδος υιοθετεί την fuzzy [10] λογική για την ποσοτική μοντελοποίηση της μεταγωγής σήματος από την μια πρωτεΐνη στην άλλη μέσα στο δίκτυο, και εν συνεχεία εισάγει μια αντικειμενική συνάρτηση που αντιπροσωπεύει την ασυμφωνία μεταξύ πειραματικών δεδομένων και υπολογιστικού μοντέλου και μη γραμμικούς περιορισμούς με σκοπό την κατασκευή μοντέλου που περιγράφει πιστά τους σηματοδοτικούς μηχανισμούς του υπο εξέταση κυτταρικού τύπου. Κατα την fuzzy λογική χρησιμοποιείται μια συνάρτηση μεταφοράς για την μοντελοποίηση της μεταγωγής σήματος της μορφής

$$f(x) = a(p^n + 1)\frac{x^n}{x^n + p^n} \tag{3.22}$$

Όπου, με x αντιπροσωπεύεται η ενεργοποίηση του αντιδρώντος, n ο εκθέτης Hill, p είναι μια σταθερά που ορίζει το σημείο καμπής της συνάρτησης, το a είναι συντελεστής στάθμισης και με $f(x)$ αντιπροσωπεύεται η ενεργοποίηση του προϊόντος της εν λόγω αντίδρασης. Η συνάρτηση μεταφοράς φαίνεται στο σχήμα 3.18. Στόχος της προτεινόμενης μεθόδου είναι δοθέντων φωσφοπρωτεομικών δεδομένων να υπολογιστούν οι παράμετροι a, n και p ώστε να ελαχιστοποιηθεί η ασυμφωνία μεταξύ υπολογιστικού μοντέλου και πειραματικών δεδομένων.

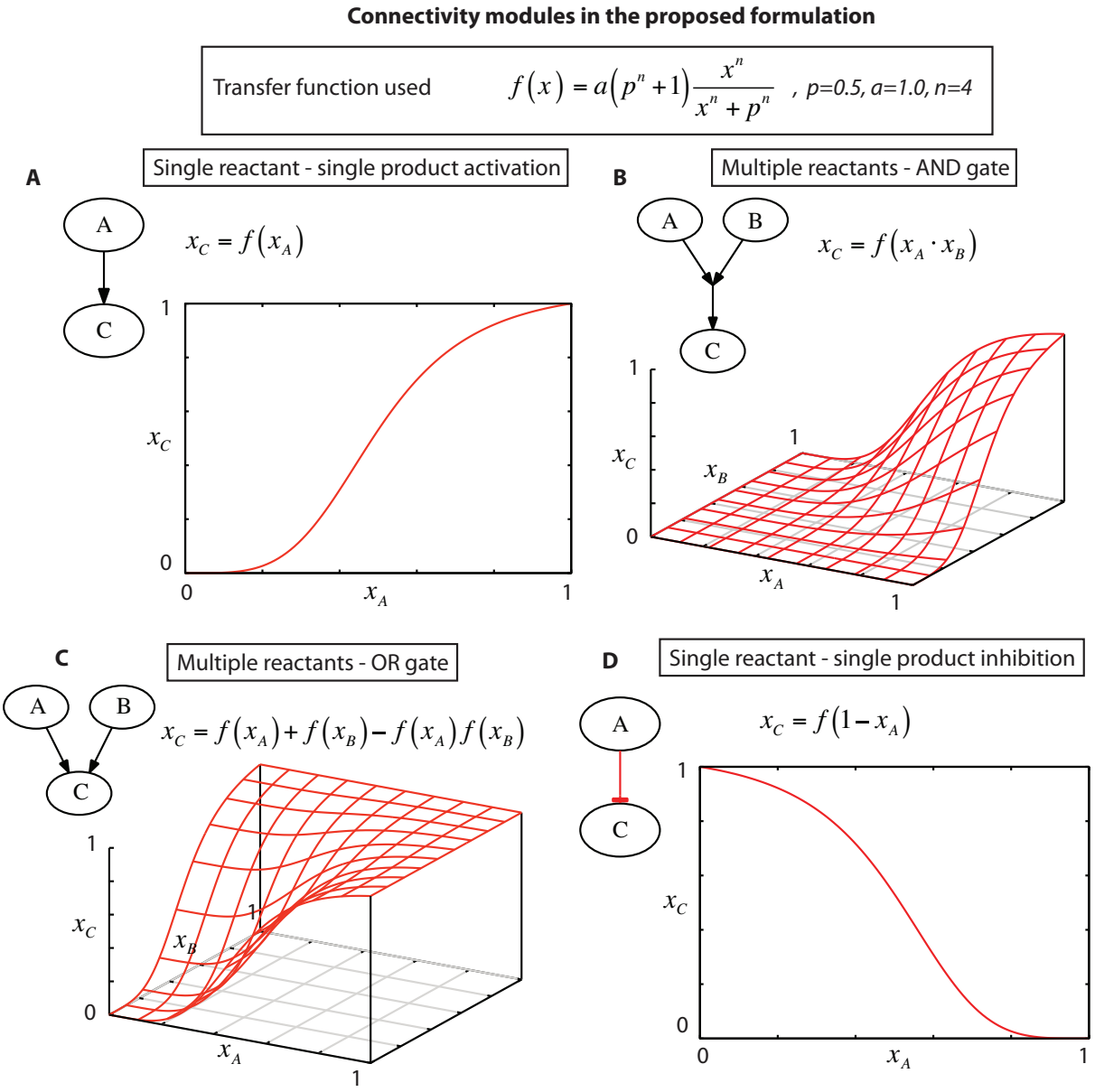
4.1 Μέθοδος Μη Γραμμικού Προγραμματισμού - μαθηματική διατύπωση

Ορίζεται σηματοδοτικό μονοπάτι ως ένα σύνολο αντιδράσεων $i = 1, ..., n_r$; και κόμβων (πρωτεϊνών) $j = 1, ..., n_s$. Σε κάθε αντίδραση αντιστοιχούν τρία σύνολα. Το σύνολο των αντιδρώντων R_i το σύνολο των αναστολέων I_i και το σύνολο των προϊόντων P_i, με $R_i, P_i, I_i \subset \{1, ..., n_s\}$. Πραγματοποιείται σύνολο προσομοιώσεων (εικονικών πειραμάτων) $k = 1, ..., n_e$. Σε κάθε πείραμα επιβάλλεται η ενεργοποίηση ενός συνόλου κόμβων και η απενεργοποίηση ενός άλλου συνόλου κόμβων. Η προβλεπόμενη από το μοντέλο ενεργοποίηση του κόμβου j στο πείραμα k συμβολίζεται με $x_j^k \in [0, 1]$. Αν είναι διαθέσιμη η μετρούμενη τιμή της ενεργοποίησης του αντίστοιχου κόμβου τότε αυτή συμβολίζεται με $x_j^{k,m} \in [0, 1]$. Το τελευταιο σύνολο μεταβλητών που εισάγωνται είναι οι $z_i^{k,m} \in [0, 1]$, και εκφράζουν την ενεργοποίηση της αντίδρασης i στο πείραμα k.

Εισάγεται η ακόλουθη αντικειμενική συνάρτηση προς ελαχιστοποίηση

$$\sum_{j,k} a_j^k \mid x_j^k - x_j^{k,m} \mid \tag{3.23}$$

και αντιπροσωπεύει την σταθμισμένη ασυμφωνία μεταξύ υπολογιστικού μοντέλου και πειραματικών δεδομένων. Με $a_j^k \in [0, 1]$ συμβολίζονται όροι στάθμισης που ορίζονται από τον χρήστη.



Σχήμα 3.18: Συναρτήσεις μεταφοράς για αντιδράσεις διαφορετικών τύπων. (A) Αντίδραση ενεργοποίησης μεταξύ ενός αντιδρώντος και ενός προϊόντος. (B) Αντίδραση μεταξύ δύο αντιδρώντων και ενός προϊόντος, ισοδύναμη με λογική πύλη AND. (C) Αντίδραση μεταξύ δύο αντιδρώντων και ενός προϊόντος, ισοδύναμη με λογική πύλη OR. (D) Αντίδραση απενεργοποίησης μεταξύ ενός αντιδρώντος και ενός προϊόντος, ισοδύναμη με λογική πύλη NOT.

Αντίδραση ενεργοποίησης μεταξύ ενός αντιδρώντος και ενός προϊόντος

Για αντιδράσεις ενεργοποίησης μεταξύ ενός αντιδρώντος και ενός προϊόντος χρησιμοποιείται η ακόλουθη συνάρτηση μεταφοράς:

$$f(x) = a(p^n + 1)\frac{x^n}{x^n + p^n} \tag{3.24}$$

Η ενεργοποίηση της αντίδρασης i ισούται με: $z_i^k = f(x_j^k)$, όπου $j \in R_i$. Η τιμή ενεργοποίησης του προϊόντος ισούται με: $x_j^k = z_i^k$, όπου $j \in P_i$. Σε περίπτωση που ο κόμβος j είναι αναστολέας τότε: $z_i^k = f(1 - x_j^k)$, όπου $j \in R_i$.

Αντίδραση μεταξύ δύο αντιδρώντων και ενός προϊόντος - λογική πύλη AND

Σε περίπτωση που περισσότερα από ένα αντιδρώντα απαιτούνται για την μετάδοση του σήματος, τότε η αντίδραση i μοντελοποιείται ως ακολούθως:

$$z_i^k = f(\prod_{j \in R_i} x_j^k \times \prod_{j \in I_i} (1 - x_j^k)) \tag{3.25}$$

Η τιμή ενεργοποίησης του προϊόντος ισούται με: $x_j^k = z_i^k$, όπου $j \in P_i$. Σε περίπτωση που η ενεργοποίηση των κόμβων j, μπορεί να πάρει μόνο τις τιμές 0 και 1, τότε η εξίσωση 3.25 ισοδυναμεί με πύλη AND.

Αντίδραση μεταξύ δύο αντιδρώντων και ενός προϊόντος - λογική πύλη OR

Σε περίπτωση που περισσότερες των μία αντιδράσεων οδηγούν στο ίδιο προϊόν και έστω και μία είναι αρκετή για την ενεργοποίηση του, τότε η αντίδραση i μοντελοποιείται ως ακολούθως:

$$x_j^k = b_{|T_j|}^k \tag{3.26}$$

όπου,

$$T_j = \{i \in \{1, ..., n_r\} : j \in P_i\} \tag{3.27}$$

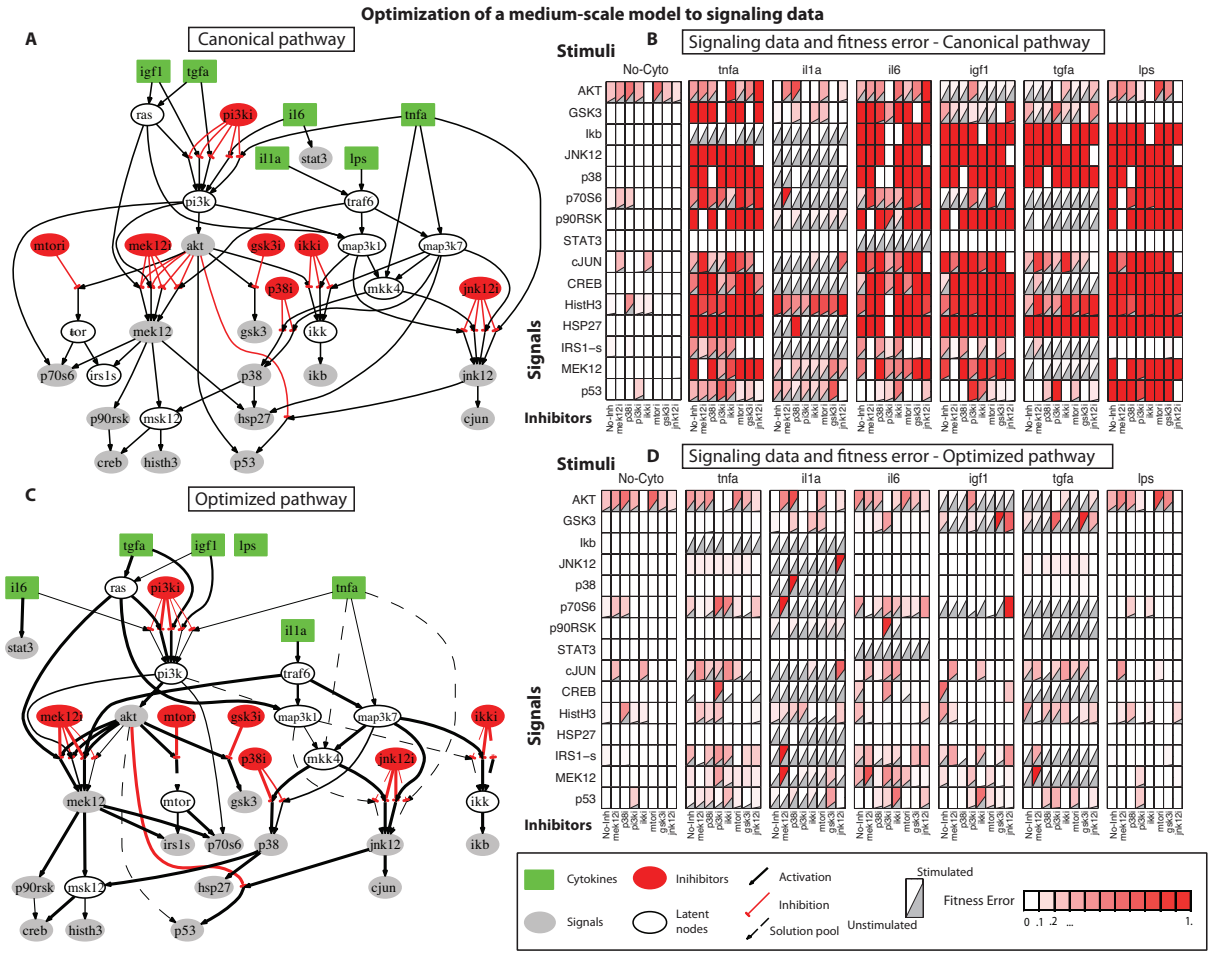
T_j είναι το σύνολο όλων των αντιδράσεων που έχουν τον κόμβο j σαν προϊόν τους. Εστω $i_1, i_2, ..., i_{|T_j|}$ είναι τα στοιχεία του T_j. Τότε τα b_m^k υπολογίζονται ως ακολούθως:

$$b_m^k = b_{m-1}^k + z_{i_m}^k - b_{m-1}^k \, z_{i_m}^k; \quad 2 < m \leq \mid T_j \mid \tag{3.28}$$

$$b_2^k = z_{i_1}^k + z_{i_2}^k - z_{i_1}^k \, z_{i_2}^k \tag{3.29}$$

Βελτιστοποίηση σηματοδοτικού δικτύου σε πειραματικά δεδομένα

Σαν εφαρμογή της μεθόδου Μη Γραμμικού Προγραμματισμού που περιγράφηκε παραπάνω, βελτιστοποιήσαμε το σηματοδοτικό μονοπάτι που χρησιμοποίησαν και οι Morris et al. [10] στο σύνολο πρωτεομικών δεδομένων απο καρκινικά ηπατικά κύτταρα των Alexopoulos et al. [12]. Η αρχική τοπολογία και τα πειραματικά δεδομένα φαίνονται στο σχήμα 3.19A και 3.19B, ενώ η βελτιστοποιημένη τοπολογία στο σχήμα 3.19C. Οι έντονες γραμμές αντιπροσωπεύουν τις αντιδράσεις που είναι απαραίτητες για την σωστή αναπαραγωγή των πειραματικών δεδομένων. Οι διακεκομένες γραμμές αντιπροσωπεύουν αντιδράσεις οι οποίες παρότι δεν είναι απαραίτητες να διατηρηθούν στην λύση, δεν αντικρούουν τα πειραματικά δεδομένα. Στο σχήμα 3.19D με κόκκινο χρώμα φαίνεται η ασυμφωνία του βελτιστοποιημένου μοντέλου με τα πειραματικά δεδομένα και ταυτόχρονα παρατηρείται σημαντική πτώση σε σχέση με την ασυμφωνία του αρχικού μοντέλου που φαίνεται στο σχήμα 3.19B. Τα αποτελέσματα της προτεινόμενης μεθοδολογίας είναι πολύ κοντά σε αυτά προηγούμενων μεθόδων [10], ενώ ο υπολογιστικός χρόνος μειώθηκε σημαντικά (στο 10% των προηγούμενων μεθόδων).



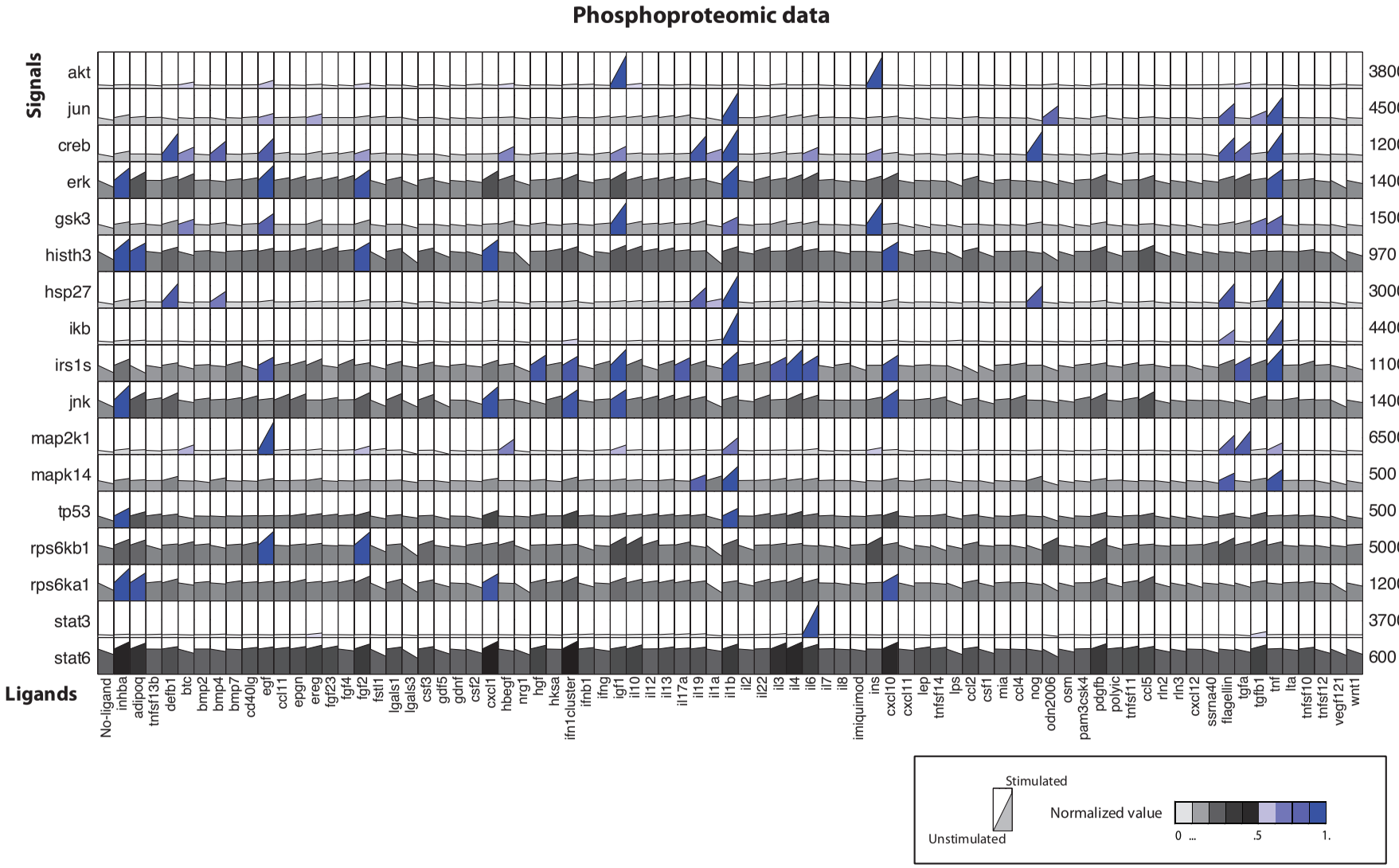
Σχήμα 3.19: Βελτιστοποίηση σηματοδοτικού δικτύου σε πειραματικά δεδομένα. (A) Αρχική τοπολογία όπως χρησιμοποιήθηκε από τους Morris et al. [10]. (B) Πειραματικά δεδομένα για καρκινικά ηπατοκύτταρα κατόπιν εισαγωγής 6 ερεθισμάτων σε συνδυασμό με 7 αναστολείς. Με κόκκινο φαίνεται η ασυμφωνία με το αρχικό υπολογιστικό μοντέλο. (C) Βελτιστοποιημένο δίκτυο. Οι έντονες γραμμές αντιπροσωπεύουν τις αντιδράσεις που διατηρήθηκαν στο δίκτυο κατόπιν της διαδικασίας βελτιστοποίησης. (D) Τα πειραματικά δεδομένα σε αντιδιαστολή με το βελτιστοποιημένο σηματοδοτικό δίκτυο. Η μείωση του κόκκινου χρώματος (από 46% σε 8%) δείχνει πως το βελτιστοποιημένο δίκτυο εκφράζει πολύ καλύτερα τα πειραματικά δεδομένα σε σχέση με το αρχικό μοντέλο.

4.2 Μοντελοποίηση των σηματοδοτικών μηχανισμών στα χονδροκύτταρα χρησιμοποιώντας αλγόριθμο Μη Γραμμικού Προγραμματισμού σε πρωτεομικά δεδομένα

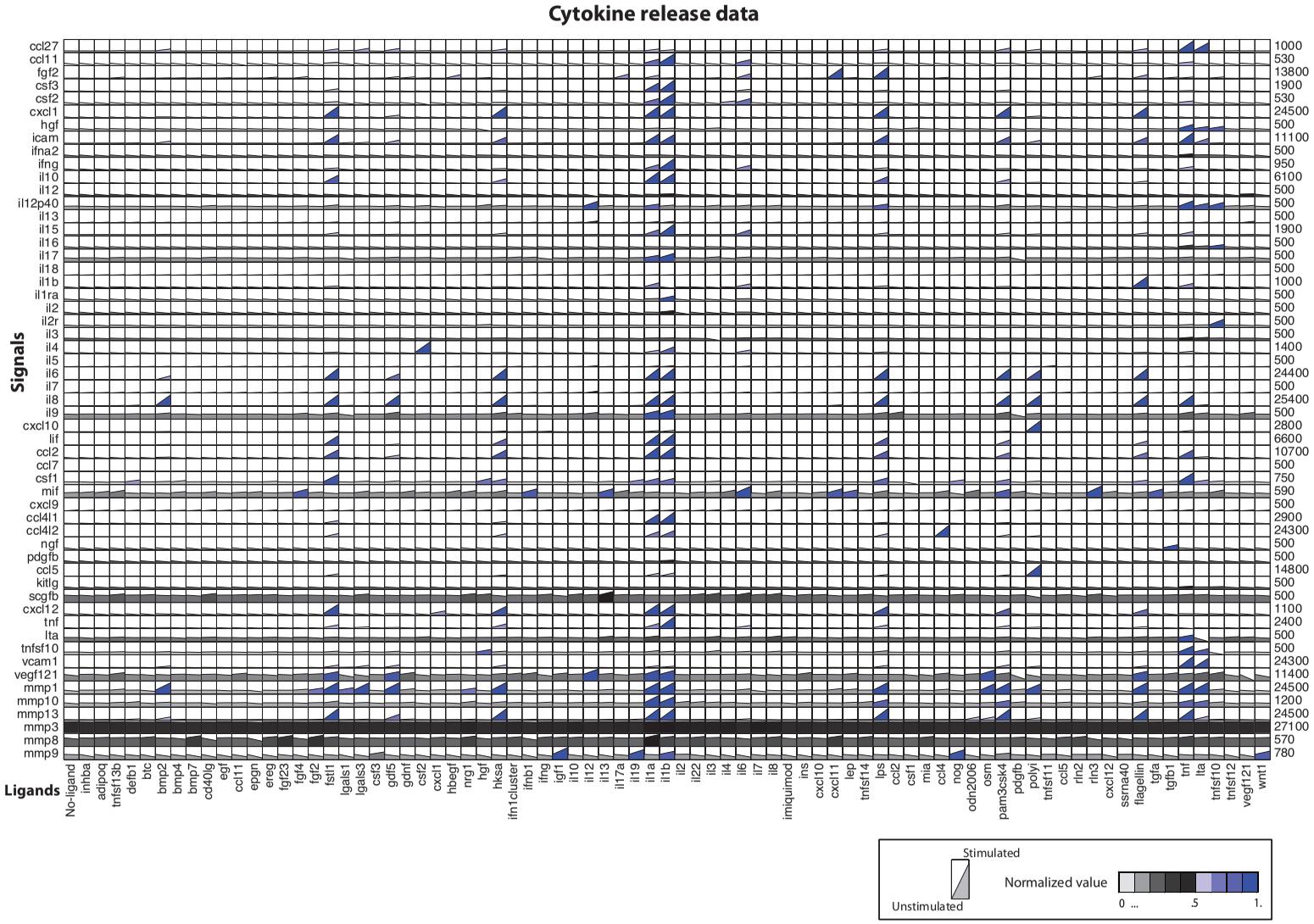
Στην παρούσα ενότητα εφαρμόζουμε την μέθοδο Μη Γραμμικού Προγραμματισμού που περιγράφηκε στην ενότητα 4 για την ποσοτική μοντελοποίηση των σηματοδοτικών μηχανισμών στα χονδροκύτταρα. Η έρευνα αυτή δημοσιεύτηκε από τους Melas et al., [91]. Πιο συγκεκριμένα, αρχικά κατασκευάστηκε σηματοδοτικό μονοπάτι με βάση την βιβλιογραφία που να περιγράφει την μεταγωγή σήματος κάτωθεν 78 ερεθισμάτων που διαδραματίζουν σημαντικό ρόλο στην ομοιόσταση του αρθρικού χόνδρου. Εν συνεχεία πραγματοποιήθηκαν δύο πειράματα όπου εισήχθησαν τα 78 ερεθίσματα και μετρήθηκαν, στο πρώτο πείραμα η ενεργοποίηση 17 φωσφοπρωτεϊνων στο κυτταρόπλασμα, στο δεύτερο πείραμα η έκκριση 55 κυτοκινών στο εξωκυτταρικό περιβάλλον. Τελικά εφαρμόστηκε ο αλγόριθμος Μη Γραμμικού Προγραμματισμού που περιγράφηκε στην ενότητα 4 για την κατασκευή εκτεταμένου σηματοδοτικού δικτύου και την ποσοτική μοντελοποίηση των σηματοδοτικών μηχανισμών των υπο εξέταση κυττάρων. Τα αποτελέσματα της έρευνας αυτής αποσαφηνίζουν τους ενδοκυτταρικούς μηχανισμούς που οδηγούν στην εκφύλιση του αρθρικού χόνδρου σε ασθένειες όπως η οστεοαρθρίτιδα και η ρευματοειδής αρθρίτιδα [19, 20, 21, 22].

Πειραματικό σκέλος

Χονδροκύτταρα καλλιεργήθηκαν σε 96αρες πλάκες και κατόπιν εισήχθησαν 78 ερεθίσματα που με βάση την βιβλιογραφία επηρεάζουν σημαντικά την ομοιόσταση του αρθρικού χόνδρου, ή είναι σημαντικά για άλλους κυτταρικούς τύπους. Σε φωσφοπρωτεομικό επίπεδο μετρήθηκε η ενεργοποίηση 17 φωσφοπρωτεϊνικών σημάτων, ενώ σε εξωκυτταρικό επίπεδο μετρήθηκε η έκκριση 55 κυτοκινών. Για τις πρωτεομικές μετρήσεις χρησιμοποιήθηκε η τεχνολογία xMAP της Luminex. Τα πειραματικά δεδομένα φαίνονται στα σχήματα 3.20 και 3.21.



Σχήμα 3.20: Φωσφοπρωτεομικά δεδομένα: Οι γραμμές αντιστοιχούν στα 17 μετρούμενα φωσφοπρωτεομικά σήματα, οι στήλες αντιστοιχούν στα 79 ερεθίσματα (μαζί με το control). Σε κάθε τετραγωνάκι φαίνεται η πορεία του σήματος στον χρόνο από την στιγμή t=0 στην στιγμή t=25min. Το χρώμα σε κάθε τετραγωνάκι αντιστοιχεί στην κανονικοποιημένη τιμή του σήματος (το μπλέ χρώμα υποδηλώνει ενεργοποίηση).



Σχήμα 3.21: Εκκρινόμενες κυτοκίνες: Οι γραμμές αντιστοιχούν στις 55 εκκρινόμενες κυτοκίνες που μετρώνται στο υπερκείμενο υγρό, οι στήλες αντιστοιχούν στα 79 ερεθίσματα. Σε κάθε τετραγωνάκι φαίνεται η πορεία του σήματος στον χρόνο από την στιγμή t=0 στην στιγμή t=24h. Το χρώμα σε κάθε τετραγωνάκι αντιστοιχεί στην κανονικοποιημένη τιμή του σήματος (το μπλέ χρώμα υποδηλώνει ενεργοποίηση).

Σαν επαλήθευση των αποτελεσμάτων μας, παρατηρούμε ότι ερεθίσματα όπως τα IL1α, IL1β και TNF γνωστά για την φλεγμονώδη τους δράση ενεργοποίησαν τα αντίστοιχα σήματα. Επι παραδείγματι: Το IL1β ενεργοποίησε τα περισσότερα από τα σήματα που σχετίζονται με φλεγμονή όπως τα JUN, CREB, IκB, MAPK14, HSP27 και TP53. Επίσης ενεργοποίησε τα ERK, IRS1S, MAP2K1 και GSK3. Το TNF παραπλήσια με το IL1β, ενεργοποίησε φλεγμονώδη σήματα εκτός από το TP53 και ταυτόχρονα τα ERK, GSK3, IRS1S και MAP2K1. Το IL1α, από την άλλη ενεργοποίησε μόνο το CREB ενώ απερνεργοποίησε τα HISTH3, JNK, TP53, RPS6KB1 και RPS6KA1. Σχετικά με τα ερεθίσματα που σχετίζονται με κυτταρικο διπλασιασμό, το EGF και το TGFα ενεργοποίησαν τις σχετικές φωσφοπρωτεϊνες, όπως τις MAP2K1, ERK, GSK3 και μερικώς την AKT, ενώ δεν ενεργοποίησαν τις πρωτεϊνες που σχετίζονται με φλεγμονή. Παρόμοια επίδραση είχαν τα INS και IGF1.

Εκτός από τα προαναφερθέντα ερεθίσματα που έχουν καταγραφεί και στην βιβλιογραφία, αναγνωρίσαμε και καινούρια όπως τα: INHBA, DEFB1, BTC, FGF2, CXCL1, HBEGF, IL19, CXCL10, NOG, ODN2006 και FLAGELLIN, τα οποία βρέθηκαν να έχουν σημαντική επίδραση στο φωσφοπρωτεομικό επίπεδο των ανθρώπινων χονδροκυττάρων.

Σχετικά με τις εκκρινόμενες κυτοκίνες, παρατηρούμε ότι τα προ-φλεγμονώδη ερεθίσματα όπως IL1α, IL1β και TNF οδηγούν σε πολύ έντονη έκκριση κυτοκινών, απελευθερώνοντας IL6, IL17, TNF, IFNg, LIF, ICAM, VCAM1, IL8, CCL2, CCL4L1, CCL4L2, CXCL1, CCL11, CCL5 και CXCL12. Επίσης απελευθερώνουν MMPs όπως τα MMP1, MMP9, MMP10 και MMP13, τα οποία ευθύνονται για την αποδόμηση του ιστού σε παθολογικές καταστάσεις.

Κατασκευή σηματοδοτικού δικτύου

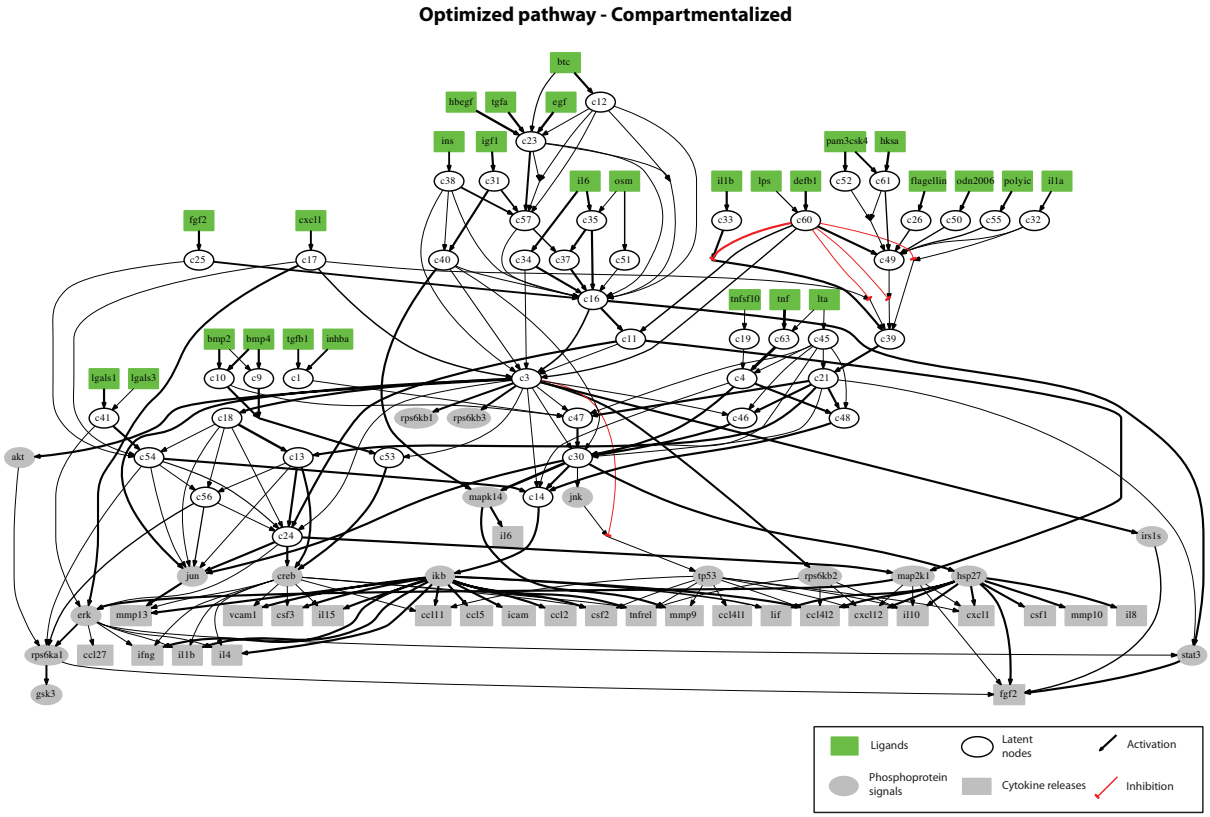
Τα δεδομένα των σχημάτων 3.20 και 3.21 χρησιμοποιήθηκαν μέσω του αλγόριθμου Μη Γραμμικού Προγραμματιμού που περιγράφηκε στην ενότητα 4, για την βελτιστοποίηση ενός βιβλιογραφικού σηματοδοτικού δικτύου και την κατασκευή υπολογιστικού μοντέλου που θα αντιπροσωπεύει πιστά τους σηματοδοτικούς μηχανισμούς του υπο εξέταση κυτταρικού τύπου. Το βιβλιογραφικό σηματοδοτικό δίκτυο συντέθηκε κυρίως απο το KEGG και το Ingenuity στην γειτονιά των 78 επιλεχθέντων ερεθισμάτων και 17 μετρούμενων φωσφοπρωτεϊνων, και αποτελείται από περισσότερους από 500 κόμβους και 1000 αντιδράσεις. Για να βελτιστοποιηθεί από τον αλγόριθμο Μη Γραμμικού Προγραμματισμού συμπιέζεται πρώτα από κατάλληλο αλγόριθμο προεπεξεργασίας, ο οποίος προσομοιώνει το δίκτυο χρησιμοποιώντας την Boolean λογική σε εικονικά πειράματα στα οποία εισάγονται τα 78 επιλεχθέντα ερεθίσματα και αναγνωρίζει τους κόμβους που έχουν την ίδια προβλεπόμενη τιμή ενεργοποίησης σε όλα τα πειράματα. Αυτοί οι κόμβοι συμπιέζονται σε διαμερίσματα (compartments) και έτσι απλοποιείται σημαντικά το σηματοδοτικό μονοπάτι. Παράλληλα με αυτόν τον τρόπο αφαιρούνται όλοι οι κόμβοι που δεν είναι κάτωθεν κάποιου ερεθίσματος. Στον πίνακα 3.22 φαίνονται τα διαμερίσματα του εν λόγω σηματοδοτικού δικτύου. Κατόπιν βελτιστοποίησης τα διαμερίσματα αντικαθίστονται με τους αντίστοιχους κόμβους. Το βελτιστοποιημένο σηματοδοτικό μονοπάτι στην συμπιεσμένη του μορφή φαίνεται στο σχήμα 3.23, ενώ στην ασυμπίεστή του μορφή στο σχήμα 3.24.

Μελετώντας την τοπολογία του βελτιστοποιημένου σηματοδοτικού μονοπατιού, παρατηρούμε οτι αντικατοπτρίζει σε μεγάλο βαθμό τα πρότυπα των πρωτεομικών δεδομένων των σχημάτων 3.20 και 3.21. Το εν λόγω σηματοδοτικό μονοπάτι είναι από τις πρώτες απόπειρες για την δημιουργία εκτεταμένου υπολογιστικού μοντέλου που να αναπαριστά πιστά τους σηματοδοτικούς μηχανισμούς των χονδροκυττάρων. Μελέτη της συνδεσμολογίας του μπορεί να αποκαλύψει βέλτιστους στόχους φαρμακευτικής παρέμβασης με στόχο την αναχαίτιση της έκκρισης φλεγμονώδων κυτοκινών και MMPs, παράγοντες που συνδέονται στενά με την παθογένεια εκφυλιστικών νόσων των αρθρώσεων.

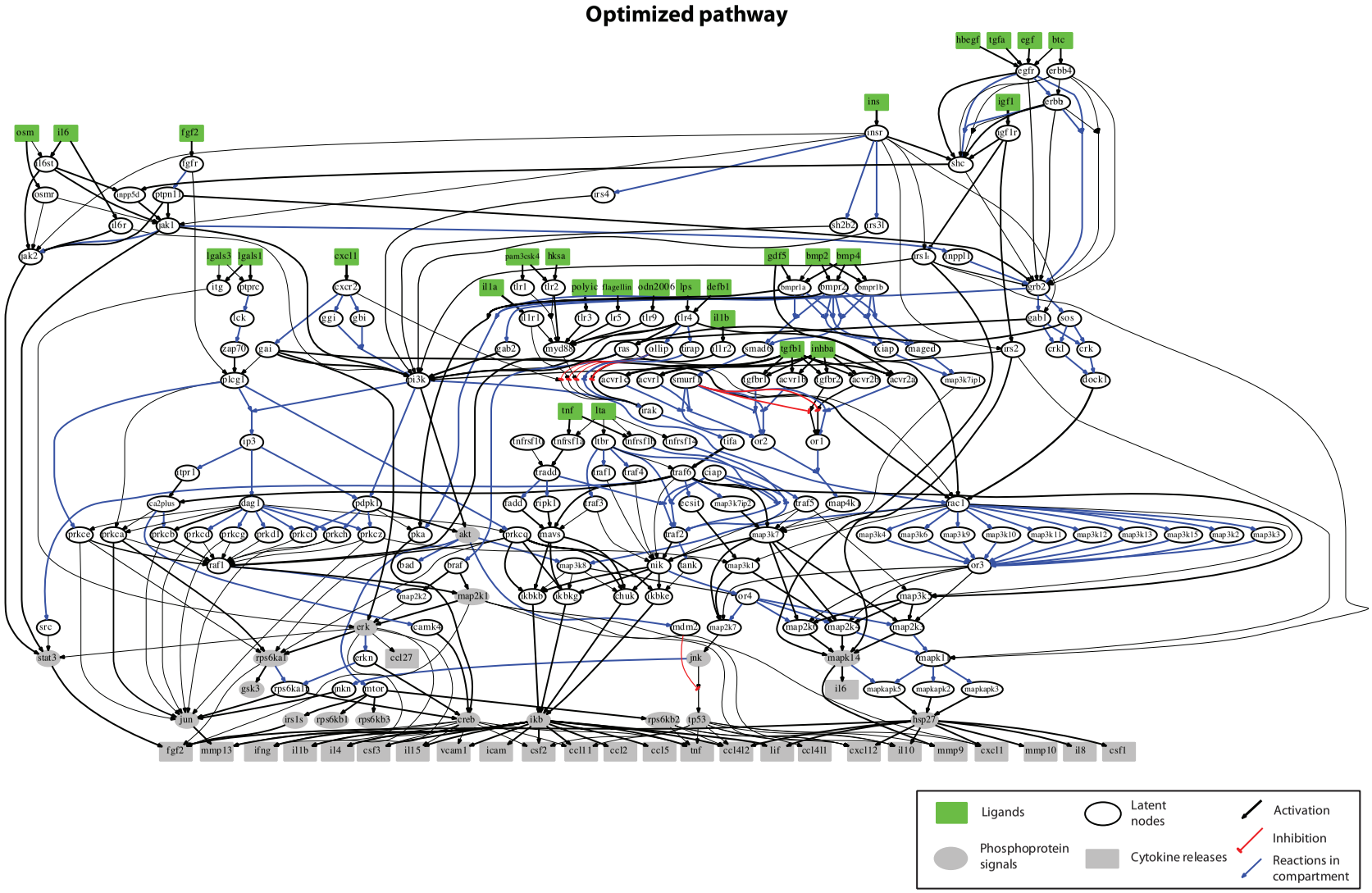
Compartments of the PKN

c1	acvr1, acvr1b, acvr1c, acvr2b, map4k1, tgfbr1, tgfbr2	c30	hsp27, jnk, jnkn, map2k3, map2k4, map2k6, map2k7, mapk11, mapk14, mapkapk2, mapkapk3, mapkapk5, nik
c2	acvr2a		
c3	akt, bad, irs1s, map3k10, map3k11, map3k12, map3k13, map3k15, map3k2, map3k3, map3k4, map3k6, map3k8, map3k9, mdm2, mtor, pdpk1, pi3k, prkch, prkci, prkcz, rac1, rps6kb1, rps6kb2, rps6kb3	c31	igf1, igf1r
		c32	il1a, il1r1
		c33	il1b, il1r2
		c34	il6, il6r
c4	bak1, bid, casp10, casp2, casp8, cradd, fadd, ripk1, tank, tbid, tradd, traf2	c35	il6st
		c36	inhba
c5	bax, diablo, htra2, tp53	c37	inpp5d
c6	bcl2, ciap	c38	ins, insr, irs3l, irs4, sh2b2
c7	bmp2	c39	irak, tifa
c8	bmp4	c40	irs1t, irs2
c9	bmpr1a	c41	itg, lck, ptprc, zap70
c10	bmpr1b, bmpr2, maged1, map3k7ip1, smad6, smurf1, xiap	c42	lgals1
c11	braf, ras	c43	lgals3
c12	btc, erbb4	c44	lps
c13	ca2plus, camk4, prkcb	c45	lta, ltbr, tnfrsf14, traf1, traf3, traf4, traf5
c14	chuk, ikb, ikbkb, ikbke, ikbkg, jun	c46	map3k1, map3k5
c15	creb	c47	map3k7
c16	crk, crkl, dock1, gab1, gab2, grb2, inppl1, jak1, jak2, sos	c48	mavs
c17	cxcl1, cxcr2, gai, gbi, ggi	c49	myd88
c18	dag1, ip3, itpr1, prkcd, prkcg, prkd1	c50	odn2006, tlr9
c19	dap3, tnfrsf10, tnfsf10	c51	osm, osmr
c20	defb1	c52	pam3csk4, tlr1
c21	ecsit, map3k7ip2, src, traf6	c53	pka
c22	egf	c54	plcg1, prkce, prkcq
c23	egfr, erbb2	c55	polyic, tlr3
c24	erk, erkn, gsk3, map2k1, map2k2, raf1, rps6ka1, rps6ka1n, stat3	c56	prkca
		c57	shc
c25	fgf2, fgfr, ptpn11	c58	tgfa
c26	flagellin, tlr5	c59	tgfb1
c27	gdf5	c60	tirap, tlr4, tollip
c28	hbegf	c61	tlr2
c29	hksa	c62	tnf
		c63	tnfrsf1a, tnfrsf1b

Σχήμα 3.22: Διαμερίσματα του σηματοδοτικού δικτύου. Το σηματοδοτικό δίκτυο περιλαμβάνει συνολικά 63 διαμερίσματα



Σχήμα 3.23: Βελτιστοποιημένο σηματοδοτικό μονοπάτι στην συμπιεσμένη του μορφή.



Σχήμα 3.24: Βελτιστοποιημένο σηματοδοτικό μονοπάτι.

5 Μοντελοποίηση σηματοδοτικών δικτύων με την μορφή άκυκλων γράφων και ελαχιστοποίηση των ασυμφωνιών τους με πειραματικά δεδομένα μέσω αλγορίθμου Ακέραιου Γραμμικού Προγραμματισμού

Στην παρούσα ενότητα εισάγουμε μέθοδο Ακέραιου Γραμμικού Προγραμματισμού για την μοντελοποίηση σηματοδοτικών δικτύων με την μορφή άκυκλων γράφων και την ελαχιστοποίηση των ασυμφωνιών τους με πειραματικά δεδομένα. Η έρευνα αυτή κατατέθηκε προς δημοσίευση από τους Melas et al., τον Ιανουάριο του 2013 στο έγκριτο περιοδικό PLoS Computational Biology, και πραγματοποιήθηκε σε συνεργασία με τον Dr. Steffen Klamt (group leader in the Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany).

Πιο συγκεκριμένα, σε αντίθεση με προηγούμενες μεθόδους όπου υιοθετούν έναν αυστηρό μαθηματικό φορμαλισμό για την μοντελοποίηση της μεταγωγής σήματος από την μια πρωτεΐνη στην άλλη μέσα στο δίκτυο (Boolean, fuzzy λογική, συνήθεις διαφορικές εξισώσεις), η προτεινόμενη μέθοδος μοντελοποιεί το σηματοδοτικό δίκτυο σαν προσημασμένο άκυκλο γράφο και η μόνη παραδοχή που επιβάλει στην σηματοδοτική διαδικασία είναι τα πρόσημα της ενεργοποίησης στους κόμβους να είναι συμβατά με τα πειραματικά δεδομένα. Σε αντίθετη περίπτωση εισάγει στρατηγικές για την αφαίρεση των ασυμφωνιών αυτών, που περιλαμβάνουν την αφαίρεση αντιδράσεων από το δίκτυο, η την διόρθωση της τιμής ενεργοποίησης συγκεκριμένων κόμβων. Σαν εφαρμογή εξετάζουμε το σηματοδοτικό δίκτυο που δημοσιεύτηκε από τους Samaga et al. [32] αντιπαραβάλλοντας το με τα δεδομένα των Alexopoulos et al. [12], και συγκρίνουμε την λύση μας με δημοσιευμένες λύσεις που στηρίζονται στην Boolean λογική. Ο κώδικας που υλοποιεί την συγκεκριμένη μέθοδο είναι διαθέσιμος για το κοινό με την εμπορική ονομασία *SigNetTrainer* (`http://www.mpi-magdeburg.mpg.de/projects/cna/etcdownloads.html`). Για μια λεπτομερή περιγραφή του λογισμικού κοιτάξτε το παράρτημα Α.

5.1 Μοντελοποίηση του σηματοδοτικού δικτύου με την μορφή προσημασμένου άκυκλου γράφου και ορισμός της συνέπειας με πειραματικά δεδομένα

Ο προσημασμένος άκυκλος γράφος είναι πιθανότατα η πιο απλή μορφή μοντελοποίησης σηματοδοτικών δικτύων. Το δίκτυο G ορίζεται ως ένα σύνολο κόμβων και ακμών $G = (V, E, \sigma)$, όπου V είναι το σύνολο των κόμβων και E είναι το σύνολο των ακμών. Επίσης ορίζεται σ το σύνολο των προσήμων των ακμών E ($\sigma_e \in \{-1, 1\}$, $e \in E$). Στόχος μας είναι να αναλύσουμε και να βελτιώσουμε την συνέπεια του δικτύου G ως προς ένα σύνολο πειραματικών δεδομένων.

Ορισμός 1:
Ένα δίκτυο G ορίζεται συνεπές με δοσμένα πειραματικά δεδομένα **s**, τα οποία περιλαμβάνουν για κάθε κόμβο X ένα πρόσημο $s_X \in \{-1, 0, 1\}$ ανάλογα με την ποιοτική συμπεριφορά του κόμβου, όταν για όλους τους κόμβους ισχύουν τα παρακάτω:
(α) Αν $s_X = -1$: τότε είτε στο s_X επιβλήθηκε εκείνη η τιμή (με βάση τις πειραματικές συνθήκες), ή υπάρχει κόμβος Y άνωθεν του X και ακμή $e : Y \to X$ τέτοια ώστε $\sigma_e \cdot s_Y = -1$.
(β) Αν $s_X = 1$: τότε είτε στο s_X επιβλήθηκε εκείνη η τιμή (με βάση τις πειραματικές συνθήκες), ή υπάρχει κόμβος Y άνωθεν του X και ακμή $e : Y \to X$ with $\sigma_e \cdot s_Y = 1$.
(γ) Αν $s_X = 0$: τότε είτε (i) στο s_X επιβλήθηκε εκείνη η τιμή (με βάση τις πειραματικές συνθήκες) ή (ii) το X δεν έχει κόμβο άνωθεν του, ή (iii) για όλες τις ακμές $Y \to X$ ισχύει $s_Y = 0$, ή (iv) υπάρχει ακμή $e : Y \to X$ με $\sigma_e \cdot s_Y = -1$ και άλλη ακμή $h : Z \to X$ με $\sigma_h \cdot s_Z = 1$.

Στην ανάλυση που ακολουθεί υποθέτουμε ότι τα πειραματικά δεδομένα αντιπροσωπεύονται από πρόσημα για ένα σύνολο κόμβων και εκφράζουν την ποιοτική συμπεριφορά των κόμβων αυτών στο

εκάστοτε πειραματικό σενάριο. Άν τα πρόσημα αυτά είναι σύμφωνα με το σηματοδοτικό δίκτυο για όλους τους κόμβους του δικτύου τότε ορίζουμε το δίκτυο συνεπές με τα δεδομένα.

Ακολούθως ορίζουμε τα παρακάτω προβλήματα συμφωνίας μεταξύ δικτύου και πειραματικών δεδομένων.

(1) SCEN_FIT
Δοσμένου ενός πειραματικού σεναρίου κατά το οποίο εισάγονται κάποια ερεθίσματα και ίσως κάποιοι αναστολείς στο δίκτυο, και μετρώνται τα πρόσημα ενός συνόλου κόμβων (σημάτων) στο δίκτυο, να βρεθούν τα πρόσημα των υπολοίπων κόμβων ώστε να ελαχιστοποιηθεί η ασυμφωνία μεταξύ πειραματικών δεδομένων και υπολογιστικού μοντέλου σύμφωνα με τον ορισμό 1.
(2) Minimal Correction Sets (MCoS)
Δοσμένου ενός πειραματικού σεναρίου κατά το οποίο εισάγονται κάποια ερεθίσματα και ίσως κάποιοι αναστολείς στο δίκτυο, και μετρώνται τα πρόσημα ενός συνόλου κόμβων (σημάτων) στο δίκτυο, εφόσον τα πειραματικά δεδομένα δεν είναι συνεπή με το δίκτυο, να βρεθεί το ελάχιστο σύνολο κόμβων των οποίων το πρόσημο πρέπει να διορθωθεί (και πώς να διορθωθεί αυτό) ώστε να εξαλειφθεί αυτή η ασυνέπεια.
(3) OPT_SUBGRAPH
Δοσμένου συνόλου πειραματικών σεναρίων κατά τα οποία εισάγονται κάποια ερεθίσματα και ίσως κάποιοι αναστολείς στο δίκτυο, και μετρώνται τα πρόσημα ενός συνόλου κόμβων (σημάτων) στο δίκτυο, εφόσον το δίκτυο δεν είναι συνεπές με τα πειραματικά δεδομένα σε όλα τα σενάρια, να βρεθούν αντιδράσεις από το δίκτυο που άν αφαιρεθούν θα ελαχιστοποιηθεί αυτή η ασυνέπεια.

5.2 Αλγόριθμος Ακέραιου Γραμμικού Προγραμματισμού - Μαθηματική διατύπωση

Έστω σηματοδοτικό δίκτυο στην μορφή άκυκλου προσημασμένου γράφου $G = (V, E, \sigma)$ και ένα σύνολο πειραματικών σεναρίων, όπου στο κάθε σενάριο εισάγωνται κάποια ερεθίσματα, κάποιοι αναστολείς, ενώ μετράται το πρόσημο ενός συνόλου κόμβων (σήματα), περιγράφοντας την ποιοτική συμπεριφορά των κόμβων αυτών. Οι ακμες συμβολίζονται με $i = 1, \ldots, n_E, n_E =\mid E \mid$, οι κόμβοι συμβολίζονται με $j = 1, \ldots, n_V, n_V =\mid V \mid$, και τα σενάρια με $k = 1, \ldots, n_S$. Τα πειραματικά σενάρια προσδιορίζονται από 2 πίνακες (i) τον πίνακα των ερεθισμάτων και αναστολέων $\mathbf{p} = n_V \times n_S$ με $p_{j,k} \in \{-1, 0, 1\}$ τέτοιο ώστε να εκφράζει την επιβαλλόμενη τιμή του κόμβου j στο σενάριο k, και (ii) τον πίνακα μετρήσεων $\mathbf{m} = n_V \times n_S$ με $m_{j,k} \in \{-1, 0, 1\}$ τέτοιο ώστε να εκφράζει το μετρούμενο πρόσημο του κόμβου j στο πείραμα k. Ολες οι τιμές (και οι επιβαλλόμενες και οι μετρούμενες) αντικατοπτρίζουν την ποιοτική συμπεριφορά του κόμβου j και μπορεί να είναι είτε (1) αντικατοπτρίζοντας αύξηση της ενεργοποίησης του j, είτε (-1) αντικατοπτρίζοντας μείωση της ενεργοποίησης του j, ή (0) αντικατοπτρίζοντας ότι δεν αλλάζει η τιμή ενεργοποίησης του j.

Ορίζουμε μεταβλητές $x_{j,k} \in \{-1, 0, 1\}$ που αντιπροσωπεύουν την προβλεπόμενη τιμή ενεργοποίησης του κόμβου j στο πειραμα k. Ορίζουμε με μεταβλητή i την αντίδραση $S_i \rightarrow P_i$, όπου $S_i \in V$ είναι το αντιδρόν της αντίδρασης και $P_i \in V$ είναι το προϊον. Επίσης το πρόσημο της αντίδρασης i αντιπροσωπεύεται από τις μεταβλητές σ_i^+ και σ_i^-, τέτοιες ώστε: Αν i είναι αντίδραση ενεργοποίησης τότε $\sigma_i^+ = 1$ & $\sigma_i^- = 0$, αλλιώς αν είναι αντίδραση απενεργοποίησης τότε $\sigma_i^- = 1$ & $\sigma_i^+ = 0$. Επίσης ορίζονται μεταβλητές $u_{i,k}^+$ και $u_{i,k}^-$ για να αντιπροσωπεύουν την δυνατότητα της αντίδρασης i να ενεργοποιήσει ή απενεργοποιήσει τον κάτωθεν της κόμβο P_i στο πείραμα k. Η αντίδραση i με αντιδρόν $j = S_i$ έχει την δυνατότητα να ενεργοποιήσει το προϊον της P_i στο πείραμα k (i.e., $u_{i,k}^+ = 1$) αν και μόνο αν $\sigma_i^+ = 1$ και $x_{j,k} = 1$ ή $\sigma_i^- = 1$ και $x_{j,k} = -1$. Σε κάθε άλλη περίπτωση $u_{i,k}^+ = 0$. Το αντίστροφο ισχύει για τις μεταβλητές $u_{i,k}^-$. Τα παραπάνω με την μορφή εξισώσεων γράφονται:

$$u_{i,k}^+ = \max(0,\, x_{j,k} + \sigma_i^+ - 1,\, -x_{j,k} + \sigma_i^- - 1) \tag{3.30}$$

$$u^-_{i,k} = \max(0,\, -x_{j,k} + \sigma^+_i - 1,\, x_{j,k} + \sigma^-_i - 1) \tag{3.31}$$

Εισάγοντας βοηθητικές μεταβλητές $d1_{i,k}, d2_{i,k}, \ldots, d6_{i,k}$ οι παραπάνω περιορισμοί γίνονται:

$$x_{j,k} + \sigma^+_i - 1 = aux1_{i,k} \tag{3.32}$$

$$-x_{j,k} + \sigma^-_i - 1 = aux2_{i,k} \tag{3.33}$$

$$-x_{j,k} + \sigma^+_i - 1 = aux3_{i,k} \tag{3.34}$$

$$x_{j,k} + \sigma^-_i - 1 = aux4_{i,k} \tag{3.35}$$

$$u^+_{i,k} \geq aux1_{i,k} \tag{3.36}$$

$$u^+_{i,k} \geq aux2_{i,k} \tag{3.37}$$

$$u^-_{i,k} \geq aux3_{i,k} \tag{3.38}$$

$$u^-_{i,k} \geq aux4_{i,k} \tag{3.39}$$

$$u^+_{i,k} \leq 1 - d1_{i,k} \tag{3.40}$$

$$u^+_{i,k} \leq aux1_{i,k} + 3 - 3\, d2_{i,k} \tag{3.41}$$

$$u^+_{i,k} \leq aux2_{i,k} + 3 - 3\, d3_{i,k} \tag{3.42}$$

$$u^-_{i,k} \leq 1 - d4_{i,k} \tag{3.43}$$

$$u^-_{i,k} \leq aux3_{i,k} + 3 - 3\, d5_{i,k} \tag{3.44}$$

$$u^-_{i,k} \leq aux4_{i,k} + 3 - 3\, d6_{i,k} \tag{3.45}$$

$$d1_{i,k} + d2_{i,k} + d3_{i,k} = 1 \tag{3.46}$$

$$d4_{i,k} + d5_{i,k} + d6_{i,k} = 1 \tag{3.47}$$

Τέλος εισάγονται άλλες δύο μεταβλητές οι $x^+_{j,k}$ και $x^-_{j,k}$ για να αντιπροσωπεύουν την δυνατότητα του κόμβου j να ενεργοποιηθεί ή να απενεργοποιηθεί, αναλόγως την δραστηριότητα των άνωθεν αντιδράσεων. Ο κόμβος j έχει την δυνατότητα να ενεργοποιηθεί ($x^+_{j,k} = 1$) αν και μόνον αν υπάρχει ακμή i τέτοια ώστε $j = P_i$ και $u^+_{i,k} = 1$. Το αντίστροφο ισχύει για την μεταβλητή $x^-_{j,k}$. Επομένως,

$$x^+_{j,k} \geq u^+_{i,k}, \quad \forall i \text{ with } P_i = j \tag{3.48}$$

$$x^-_{j,k} \geq u^-_{i,k}, \quad \forall i \text{ with } P_i = j \tag{3.49}$$

$$x^+_{j,k} \leq \sum_{i:j=P_i} u^+_{i,k}, \quad \forall i \text{ with } P_i = j \tag{3.50}$$

$$x^-_{j,k} \leq \sum_{i:j=P_i} u^-_{i,k}, \quad \forall i \text{ with } P_i = j \tag{3.51}$$

Η τιμή που τελικά παίρνει ο κόμβος j στο πείραμα k ($x_{j,k}$) περιορίζεται από τις τιμές των $x^+_{j,k}$ και $x^-_{j,k}$ σύμφωνα με τον οριμό 1. (i) Ο κόμβος j έχει την δυνατότητα να ενεργοποιηθεί ($x_{j,k} = 1$) άν $x^+_{j,k} = 1$. (ii) Ο κόμβος j έχει την δυνατότητα να απενεργοποιηθεί ($x_{j,k} = -1$) άν $x^+_{j,k} = -1$. (iii) Ο κόμβος j έχει την δυνατότητα να παραμείνει ανενεργός ($x_{j,k} = 0$) αν έχει την

δυνατότητα να ενεργοποιηθεί και να απενεργοποιηθεί ($x^-_{j,k} = x^+_{j,k} = 1$) ή κανένα από τα παραπάνω ($x^-_{j,k} = x^+_{j,k} = 0$). Αυτοί οι κανόνες μοντελοποιούνται ως ακολούθως:

$$x_{j,k} \leq x^+_{j,k} \tag{3.52}$$

$$x_{j,k} \geq -x^-_{j,k} \tag{3.53}$$

$$x_{j,k} \leq 2\, x^+_{j,k} - x^-_{j,k} \tag{3.54}$$

$$x_{j,k} \geq -2\, x^-_{j,k} + x^+_{j,k} \tag{3.55}$$

Οι εξισώσεις της παραγράφου αυτής θα αποτελέσουν το πλαίσιο για την επίλυση των 3 βασικών προβλημάτων (1) SCEN_FIT, (2) Minimal Correction Sets (MCoS), και (3) OPT_SUBGRAPH.

SCEN_FIT

Στόχος της συγκεκριμένης μεθόδου είναι η ελαχιστοποίηση της ασυμφωνίας μεταξύ υπολογιστικού μοντέλου και πειραματικών δεδομένων. Επομένως εισάγεται προς ελαχιστοποίηση η ακόλουθη αντικειμενική συνάρτηση.

$$\sum_{j:m_{j,k} \neq NaN} a_{j,k} |m_{j,k} - x_{j,k}| = \min! \tag{3.56}$$

Τα $a_{j,k}$ είναι συντελεστές στάθμισης που ορίζονται από τον χρήστη. Οι απόλυτες τιμές διαμορφώνονται ως ακολούθως εισάγοντας βοηθητικές μεταβλητές.

$$abs_{j,k} \geq m_{j,k} - x_{j,k} \tag{3.57}$$

$$abs_{j,k} \geq x_{j,k} - m_{j,k} \tag{3.58}$$

Η ελαχιστοποίηση της εν λόγω αντικειμενικής συνάρτησης υπό τους περιορισμούς που διατυπώθηκαν παραπάνω οδηγεί στον υπολογισμό των βέλτιστων τιμών για τα $x_{j,k}$, ώστε να ελαχιστοποιηθεί η ασυμφωνία υπολογιστικού μοντέλου και πειραματικών δεδομένων σύμφωνα με το πρόβλημα SCEN_FIT.

Επειδή είναι πιθανό να υπάρχουν περισσότερες των μία λύσεων με την ίδια βέλτιστη τιμή της αντικειμενικής συνάρτησης, τρέχουμε τον αλγόριθμο επαναληπτικά και για κάθε λύση εισάγουμε περιορισμούς ώστε να μην είναι ίδια με καμία προηγούμενη λύση.

$$\sum_j |x_{j,k} - x_{j,k,s}| \geq 1, \tag{3.59}$$

Όπου το $x_{j,k,s}$ αντιπροσωπεύει την τιμή του $x_{j,k}$ στην λύση s. Θέτοντας $|x_{j,k} - x_{j,k,s}| = dx_{j,k,s}$. Ο περιορισμός 3.59 γίνεται ως ακολούθως:

$$\sum_{j,k} dx_{j,k,s} \geq 1 \tag{3.60}$$

$$-x_{j,k} + dx_{j,k,s} - 4\, dx2_{j,k,s} \leq x_{j,k,s} \tag{3.61}$$

$$x_{j,k} + dx_{j,k,s} - 4\, dx1_{j,k,s} \leq -x_{j,k,s} \tag{3.62}$$

$$dx1_{j,k,s} + dx2_{j,k,s} = 1 \tag{3.63}$$

Οπου $dx1_{j,k,s}$ & $dx2_{j,k,s}$ είναι βοηθητικές μεταβλητές. Επίσης, σε κάθε επόμενη λύση επιβάλλουμε να έχει την ίδια τιμή της αντικειμένικης συνάρτησης όπως η πρώτη λύση.

$$\sum_{j:m_{j,k} \neq NaN} a_{j,k} |m_{j,k} - x_{j,k}| = objval \tag{3.64}$$

Όπου, $objval$ είναι η τιμή της αντικειμενικής συνάρτησης στην πρώτη λύση.

Minimal Correction Sets - MCoS

Ακολούθως προσπαθούμε να αναγνωρίσουμε το ελάχιστο σύνολο κόμβων των οποίων το πρόσημο πρέπει να διορθωθεί (και πώς να διορθωθεί αυτό) ώστε να εξαλειφθεί η ασυνέπεια μεταξύ υπολογιστικού μοντέλου και πειραματικών δεδομένων. Έστω οτι συμβολίζουμε με $B^+_{j,k}$ και and $B^-_{j,k}$ τέτοιες διορθώσεις. Η μεταβλητή $B^+_{j,k} = 1$ επιβάλλει στον κόμβο j, στο πείραμα k την τιμή $x_{j,k} = 1$, ενώ η μεταβλητή $B^-_{j,k} = 1$ επιβάλλει στον κόμβο j, στο πείραμα k την τιμή $x_{j,k} = -1$. Για να συνυπολογίσουμε τις διορθώσεις αυτές αλλάζουμε τους περιορισμούς 3.52 – 3.55 ως ακολούθως:

$$x_{j,k} \leq x^+_{j,k} + B^+_{j,k} \tag{3.65}$$

$$x_{j,k} \geq -x^-_{j,k} - B^-_{j,k} \tag{3.66}$$

$$x_{j,k} \leq 2\, x^+_{j,k} - x^-_{j,k} + 2\, B^+_{j,k} \tag{3.67}$$

$$x_{j,k} \geq -2\, x^-_{j,k} + x^+_{j,k} - 2\, B^-_{j,k} \tag{3.68}$$

Στην συνέχεια θέτουμε σαν επιπλέον περιορισμό την τέλεια συμφωνία των πειραματικών δεδομένων με το υπολογιστικό μοντέλο

$$\sum_{j:m_{j,k} \neq NaN} |m_{j,k} - x_{j,k}| = 0 \tag{3.69}$$

Η απόλυτη τιμή μετασχηματίζεται όπως στην παράγραφο “SCEN_FIT”.

Επειδή ενδιαφερόμαστε για το ελάχιστο σύνολο διορθώσεων εφαρμόζουμε την ακόλουθη αντικειμενική συνάρτηση:

$$\sum_j B^+_{j,k} + \sum_j B^-_{j,k} = \min! \tag{3.70}$$

Για τον υπολογισμό όλων των πιθανών συνόλων απο διορθώσεις που απαιτούνται για την τέλεια συμφωνία μεταξύ υπολογιστικού μοντέλου και πειραματικών δεδομένων λύνουμε το ίδιο πρόβλημα επαναληπτικά και σε κάθε επανάληψη απαιτούμε λύση που να διαφέρει από όλες τις προηγούμενες. Ετσι εισάγονται οι παρακάτω περιορισμοί.

$$\sum_{j,k} (|B^+_{j,k} - B^+_{j,k,s}| + |B^-_{j,k} - B^-_{j,k,s}|) \geq 1, \tag{3.71}$$

Οπου $B^+_{j,k,s}$ και $B^-_{j,k,s}$ αντιπροσωπεύουν την τιμή του $B^+_{j,k}$ και $B^-_{j,k}$ στην λύση s. Η απόλυτη τιμή μετασχηματίζεται με τον ακόλουθο τρόπο.

$$\sum_{j,k} dB^+_{j,k,s} + dB^-_{j,k,s} \geq 1 \tag{3.72}$$

$$-B^+_{j,k} + dB^+_{j,k,s} - 2\, dB2^+_{j,k,s} \leq B^+_{j,k,s} \tag{3.73}$$

$$B^+_{j,k} + dB^+_{j,k,s} - 2\, dB1^+_{j,k,s} \leq -B^+_{j,k,s} \tag{3.74}$$

$$-B^-_{j,k} + dB^-_{j,k,s} - 2\, dB2^-_{j,k,s} \leq B^-_{j,k,s} \tag{3.75}$$

$$B^-_{j,k} + dB^-_{j,k,s} - 2\, dB1^-_{j,k,s} \leq -B^-_{j,k,s} \tag{3.76}$$

$$dB1^+_{j,k,s} + dB2^+_{j,k,s} = 1 \tag{3.77}$$

$$dB1^-_{j,k,s} + dB2^-_{j,k,s} = 1, \tag{3.78}$$

Έτσι μπορούμε να υπολογίσουμε καινούριες διορθώσεις που θα έχουν την ίδια τιμή της αντικειμενικής συνάρτησης αρκεί να θέσουμε τον παρακάτω περιορισμό.

$$\sum_{j,k} B^+_{j,k} + \sum_{j,k} B^-_{j,k} = objval. \tag{3.79}$$

Οπου $objval$ είναι η τιμή της αντικειμενικής συνάρτησης στην πρώτη λύση.

OPT_SUBGRAPH

Σε αυτήν την παράγραφο υπολογίζουμε σύνολα αντιδράσεων που πρέπει να αφαιρεθούν από το δίκτυο ώστε να ελαχιστοποιηθεί η ασυμφωνία πειραματικών δεδομένων και υπολογιστικού μοντέλου. Η αφαίρεση αντιδράσεων πραγματοποιείται με την βοήθεια των μεταβλητών y^+_i και y^-_i. Ο αλγόριθμος θα θέσει $y^+_i = 1/\ y^-_i = 1$ αν η θετική/αρνητική αντίδραση i πρέπει να αφαιρεθεί για την ελαχιστοποίηση της ασυμφωνίας πειραματικών δεδομένων και υπολογιστικού μοντέλου (αλλιώς $y^+_i = 0/\ y^-_i = 0$). Χρησιμοποιούμε τους ίδιους περιορισμούς που περιγράφηκαν παραπάνω με μοναδική τροποποίηση αυτή των περιορισμών 3.30 και 3.31 ως ακολούθως:

$$u^+_{i,k} = \max(0,\ x_{j,k} + \sigma^+_i - y^+_i - 1,\ -x_{j,k} + \sigma^-_i - y^-_i - 1) \tag{3.80}$$

$$u^-_{i,k} = \max(0,\ -x_{j,k} + \sigma^+_i - y^+_i - 1,\ x_{j,k} + \sigma^-_i - y^-_i - 1) \tag{3.81}$$

Το max μετασχηματίζεται με τον ακόλουθο τρόπο.

$$x_{j,k} + \sigma^+_i - y^+_i - 1 = aux1_{i,k} \tag{3.82}$$

$$-x_{j,k} + \sigma^-_i - y^-_i - 1 = aux2_{i,k} \tag{3.83}$$

$$-x_{j,k} + \sigma^+_i - y^+_i - 1 = aux3_{i,k} \tag{3.84}$$

$$x_{j,k} + \sigma^-_i - y^-_i - 1 = aux4_{i,k} \tag{3.85}$$

$$u^+_{i,k} \geq aux1_{i,k} \tag{3.86}$$

$$u^+_{i,k} \geq aux2_{i,k} \tag{3.87}$$

$$u^-_{i,k} \geq aux3_{i,k} \tag{3.88}$$

$$u^-_{i,k} \geq aux4_{i,k} \tag{3.89}$$

$$u^+_{i,k} \leq 1 - d1_{i,k} \tag{3.90}$$

$$u^+_{i,k} \leq aux1_{i,k} + 4 - 4\,d2_{i,k} \tag{3.91}$$

$$u^+_{i,k} \leq aux2_{i,k} + 4 - 4\,d3_{i,k} \tag{3.92}$$

$$u^-_{i,k} \leq 1 - d4_{i,k} \tag{3.93}$$

$$u^-_{i,k} \leq aux3_{i,k} + 4 - 4\,d5_{i,k} \tag{3.94}$$

$$u^-_{i,k} \leq aux4_{i,k} + 4 - 4\,d6_{i,k} \tag{3.95}$$

$$d1_{i,k} + d2_{i,k} + d3_{i,k} = 1 \tag{3.96}$$

$$d4_{i,k} + d5_{i,k} + d6_{i,k} = 1 \tag{3.97}$$

Οι παρακάτω περιορισμοί εισάγωνται επιπρόσθετα:

$$y_i^+ \leq \sigma_i^+ \tag{3.98}$$

$$y_i^- \leq \sigma_i^- \tag{3.99}$$

Η αντικειμενική συνάρτηση παραμένει η ίδια (εξίσωση 3.56), ωστόσο τώρα ελαχιστοποιούμε για όλα τα πειραματικά σενάρια:

$$\sum_{j,k:m_{j,k}\neq NaN} a_{j,k}|m_{j,k} - x_{j,k}| = \min! \tag{3.100}$$

Μπορούμε να υπολογίσουμε όλες τις πιθανές λύσεις με την ίδια βέλτιστη τιμή της αντικειμενικής συνάρτησης εισάγοντας τους παρακάτω περιορισμούς.

$$\sum_i (|y_i^+ - y_{i,s}^+| + |y_i^- - y_{i,s}^-|) \geq 1, \tag{3.101}$$

Οπου $y_{i,s}^+$ και $y_{i,s}^-$ αντιπροσωπεύουν τις τιμές των y_i^+ και y_i^- στην λύση s. Θέτοντας $|y_i^+ - y_{i,s}^+| = dy_{i,s}^+$ και $|y_i^- - y_{i,s}^-| = dy_{i,s}^-$. Ο περιορισμός 3.101 μετασχηματίζεται ως ακολούθως.

$$\sum_i (dy_{i,s}^+ + dy_{i,s}^-) \geq 1 \tag{3.102}$$

$$-y_i^+ + dy_{i,s}^+ - 2\,dy2_{i,s}^+ \leq y_{i,s}^+ \tag{3.103}$$

$$y_i^+ + dy_{i,s}^+ - 2\,dy1_{i,s}^+ \leq -y_{i,s}^+ \tag{3.104}$$

$$-y_i^- + dy_{i,s}^- - 2\,dy2_{i,s}^- \leq y_{i,s}^- \tag{3.105}$$

$$y_i^- + dy_{i,s}^- - 2\,dy1_{i,s}^- \leq -y_{i,s}^- \tag{3.106}$$

$$dy1_{i,s}^+ + dy2_{i,s}^+ = 1 \tag{3.107}$$

$$dy1_{i,s}^- + dy2_{i,s}^- = 1, \tag{3.108}$$

Επίσης για κάθε λύση μετά την πρώτη απαιτούμε την τιμή της αντικειμενικής συνάρτησης να παραμένει βέλτιστη. Ετσι εισάγουμε τον παρακάτω περιορισμό.

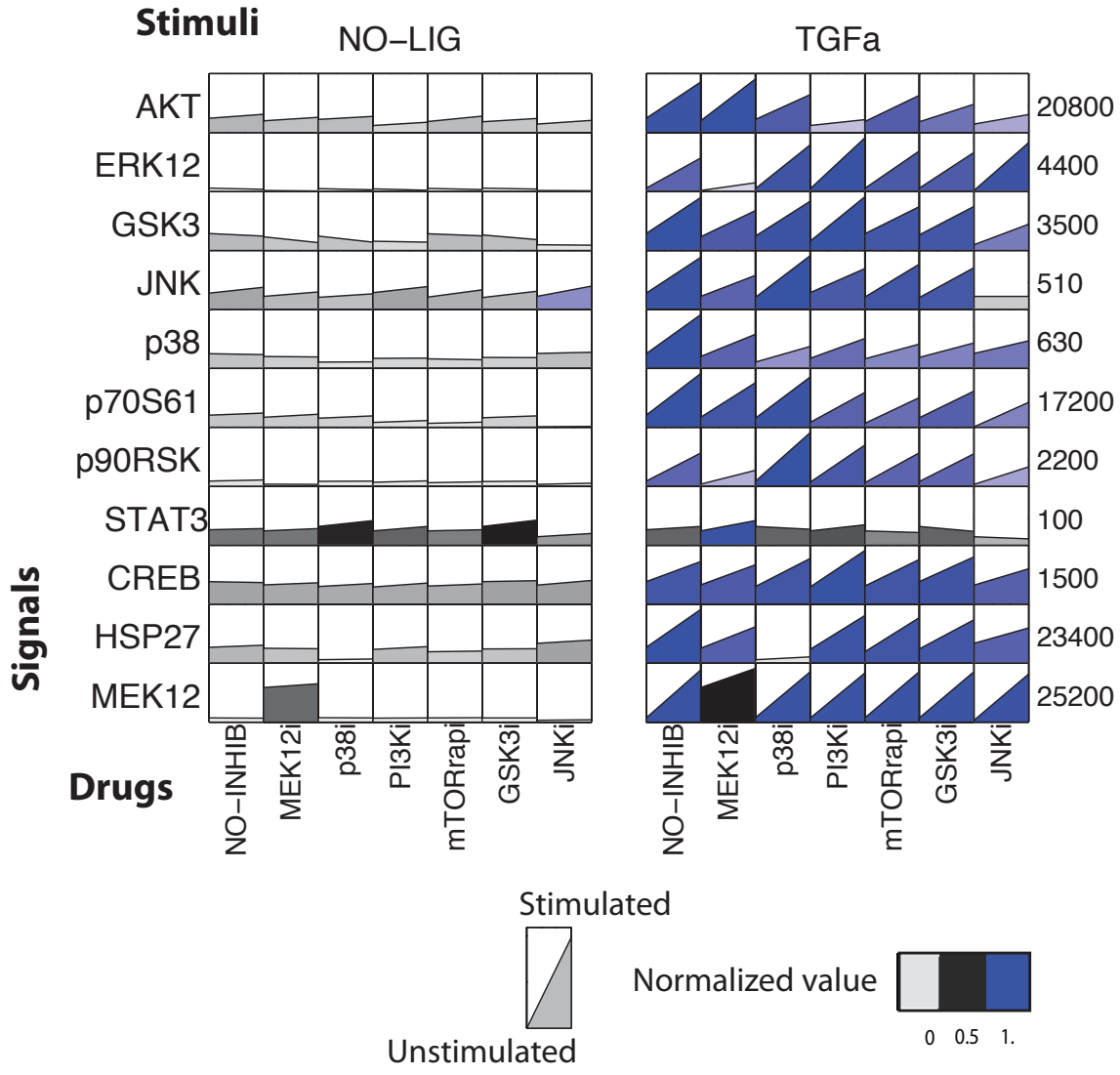
$$\sum_{j,k:x_{j,k,m}\neq NaN} a_{j,k}|x_{j,k,m} - x_{j,k}| = objval, \tag{3.109}$$

5.3 Λογισμικό

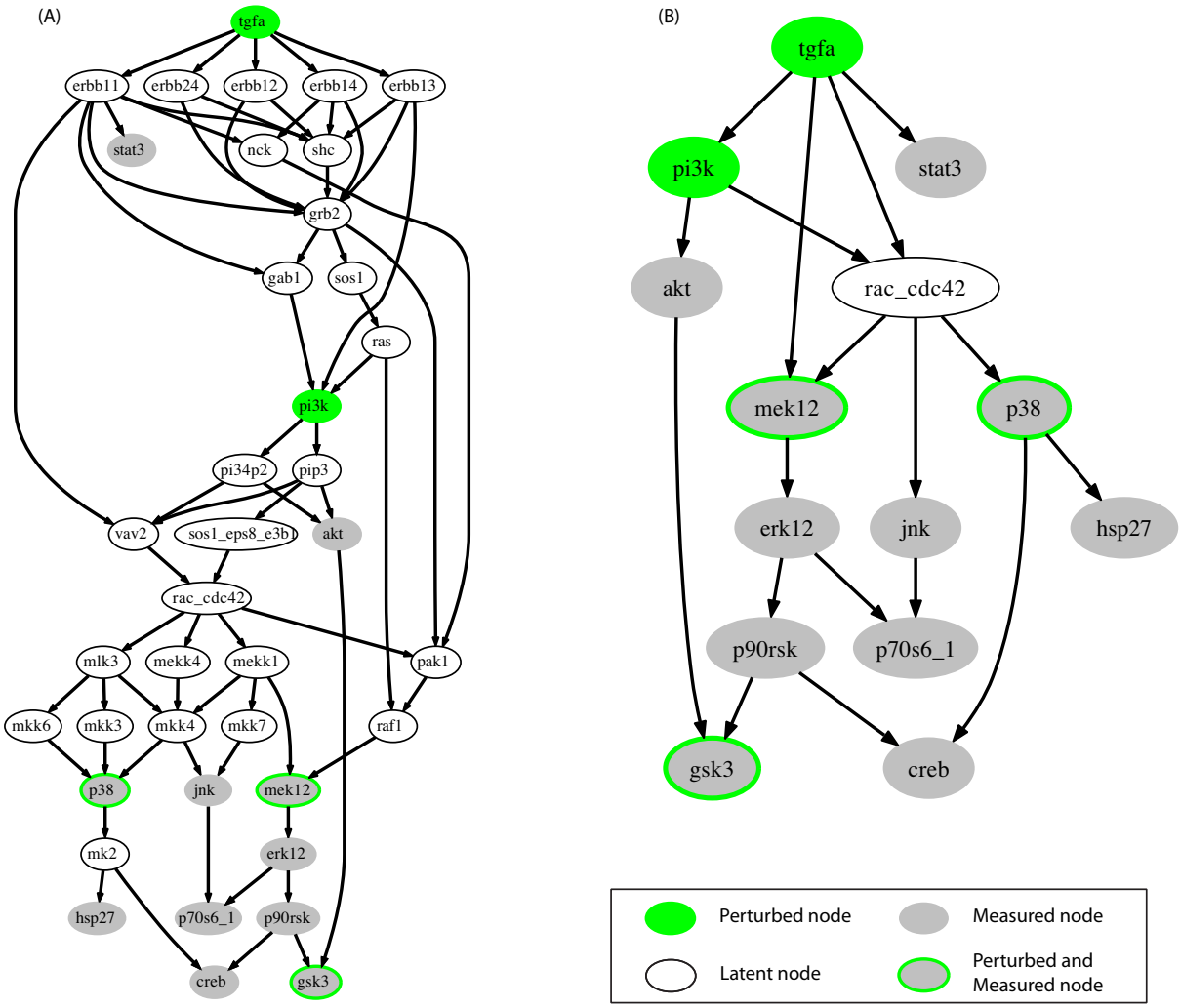
Οι αλγόριθμοι που περιγράφηκαν στην ενότητα αυτή είναι διαθέσιμοι στο κοινό μέσω του λογισμικού *SigNetTrainer*. Το λογισμικό είναι γραμμένο σε C και χρησιμοποιεί τον GUROBI (`http://www.gurobi.com`) για την επίλυση των εν λόγω προβλημάτων Ακέραιου Γραμμικού Προγραμματισμού. Το λογισμικό είναι δωρεάν για ακαδημαϊκή χρήση. Διατίθεται στην διεύθυνση `http://www.mpi-magdeburg.mpg.de/project/cna/etcdownloads.html`

5.4 Εφαρμογή

Για να δείξουμε την επίδοση των προτεινόμενων αλγορίθμων σε ρεαλιστικά προβλήματα βελτιστοποίησης σηματοδοτικών δικτύων σε πειραματικά δεδομένα, χρησιμοποιούμε το δίκτυο που δημοσιεύτηκε από τους Samaga et al. [32] αντιπαραβάλλοντας το με τα δεδομένα των Alexopoulos et al. [12]. Τα πειραματικά δεδομένα φαίνονται στο σχήμα 3.25. Το σηματοδοτικό δίκτυο φαίνεται στο σχήμα 3.26. Η ασυμφωνία των πειραματικών δεδομένων και υπολογιστικού μοντέλου που δεν ήταν δυνατόν να εξαλειφθεί με τον αλγόριθμο SCEN_FIT φαίνεται στο σχήμα 3.27. Στον πίνακα 3.2 φαίνονται τα ελάχιστα σύνολα διορθώσεων των τιμών ενεργοποίησης κόμβων του δικτύου (Minimal Correction Sets (MCoS)) ώστε να εξαλειφθεί η ασυμφωνία των πειραματικών δεδομένων και υπολογιστικού μοντέλου. Τα αποτελέσματα του αλγορίθμου OPT_SUBGRAPH φαίνονται στα σχήματα 3.28 και 3.29. Στο σχήμα 3.28 φαίνεται η υπέρθεση όλων των τοπολογιών που ελαχιστοποιούν την ασυμφωνία μεταξύ πειραματικών δεδομένων και υπολογιστικού μοντέλου ενώ στο σχήμα 3.29 φαίνονται οι προβλεπόμενες τιμές ενεργοποίησης των κόμβων του βελτιστοποιημένου δικτύου.

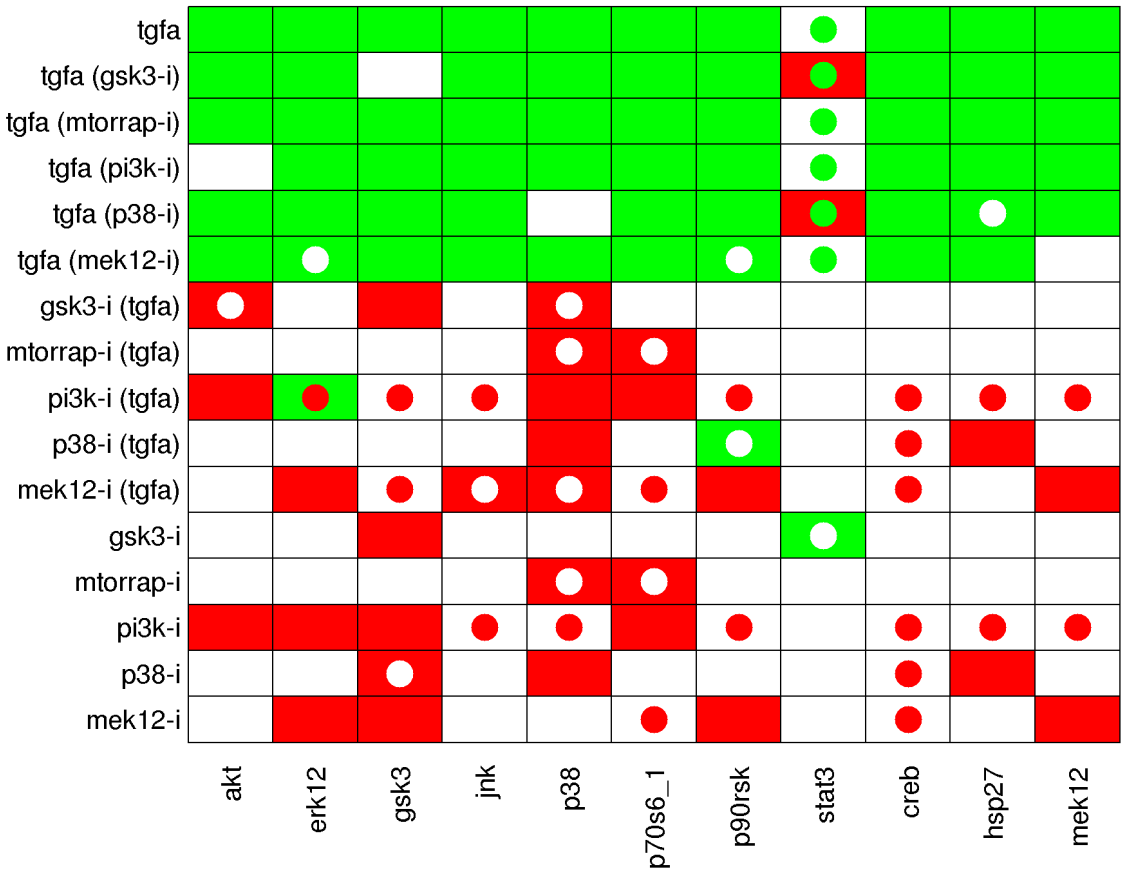


Σχήμα 3.25: Πειραματικά δεδομένα όπως μετρήθηκαν με την τεχνολογία xMAP της Luminex



Σχήμα 3.26: Σηματοδοτικό δίκτυο όπως δημοσιέυτηκε από τους Samaga et al. [32]. (A) το αρχικό σηματοδοτικό δίκτυο. (B) συμπιεσμένο σηματοδοτικό δίκτυο.

Σε αυτήν την ενότητα παρουσιάσαμε ένα καινούριο πλαίσιο για την εξέταση και ελαχιστοποίηση της ασυμφωνίας μεταξύ πειραματικών δεδομένων και σηματοδοτικού δικτύου. Σε αντίθεση με προηγούμενες μεθόδους, δεν υιοθετήσαμε κάποιον αυστηρο μαθηματικό φορμαλισμό όπως η Boolean, fuzzy λογική ή οι συνήθεις διαφορικές εξισώσεις, για την μοντελοποίηση της σηματοδοτικής διαδικασίας από τον έναν κόμβο του δικτύου στον άλλον. Αλλα μοντελοποιώντας το δίκτυο σαν άκυκλο προσημασμένο γράφο εισάγαμε τις ελάχιστες δυνατές παραδοχές, καταλήγοντας σε μία μέθοδο που μπορεί να εφαρμοστεί σε πληθώρα προβλημάτων και που πάντα παρέχει ερμηνεύσιμα και διαισθητικά αποτελέσματα.

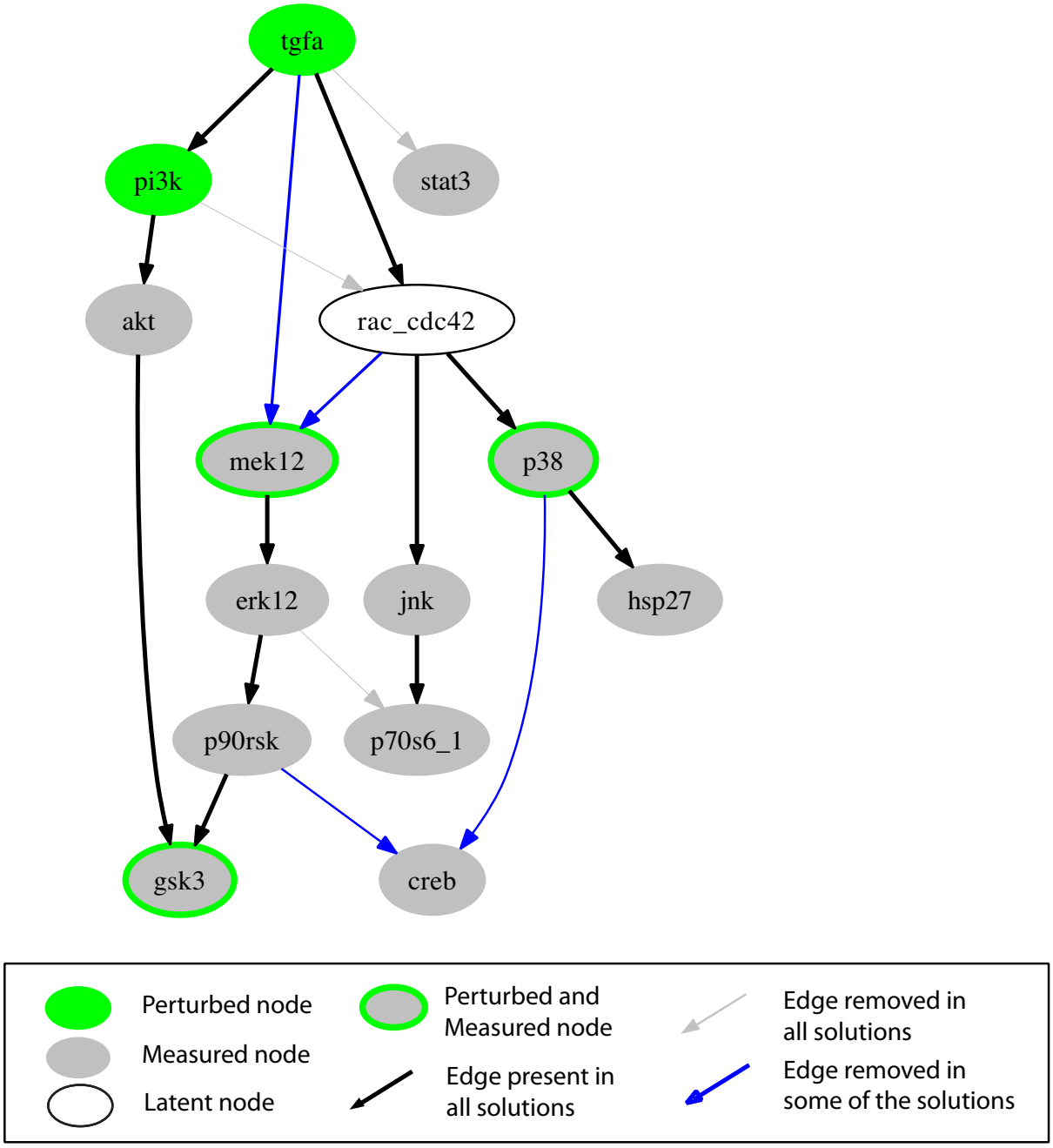


Σχήμα 3.27: Η ασυμφωνία των πειραμάτων δεδομένων και υπολογιστικού μοντέλου που δεν ήταν δυνατόν να εξαλειφθεί με τον αλγόριθμο SCEN_FIT. Οι γραμμές αντιστοιχούν στα διαφορετικά πειραματικά σενάρια, οι στήλες αντιστοιχούν στα μετρούμενα σήματα. Οι διακριτοποιημένες τιμές των μετρούμενων σημάτων αντιστοιχούν στο χρώμα των παραλληλόγραμμων σχημάτων. Αν η ενεργοποίηση ενός κόμβου αυξάνεται τότε το χρώμα του είναι πράσσινο, αν η ενεργοποίηση ενός κόμβου μειώνεται τότε είναι κόκκινο, αν η ενεργοποίηση ενός κόμβου ούτε αυξάνεται ούτε μειώνεται αισθητά τότε το χρώμα του είναι άσπρο. Το χρώμα των κουκίδων στο εσωτερικό των παραλληλογράμμων αντιστοιχεί στην προβλεπόμενη τιμή του εκάστοτε σήματος. Ετσι, πράσσινες κουκίδες εκφράζουν προβλεπόμενη τιμή ίση με 1, κόκκινες κουκίδες εκφράζουν προβλεπόμενη τιμή ίση με -1 και άσπρες κουκίδες εκφράζουν προβλεπόμενη τιμή ίση με 0.

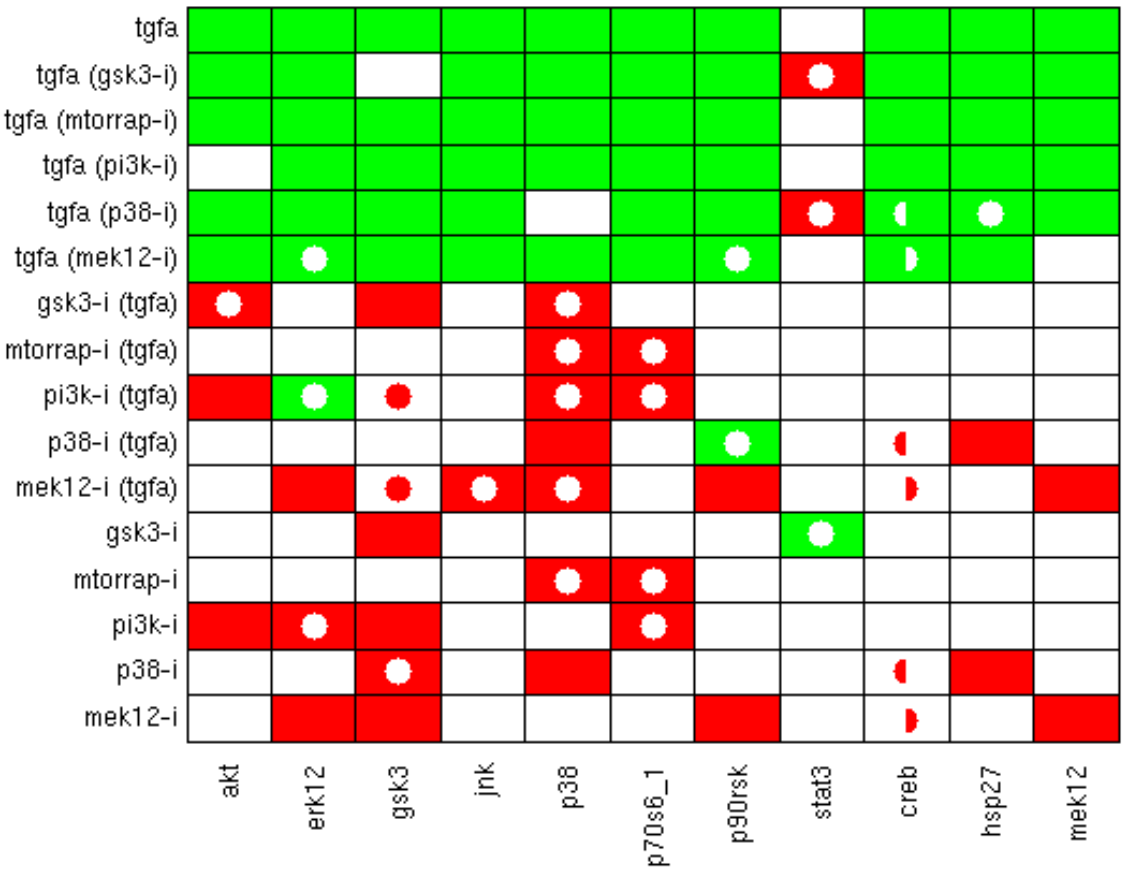
Table 3.2: Ελάχιστα σύνολα διορθώσεων για το σενάριο 14 του σχήματος 3.27.

	MCoS 1			**MCoS 2**			**MCoS 3**			**MCoS 4**			**MCoS 5**		
Node id	B_i^+	B_i^-	**Val**	B_i^+	B_i^-	**Val**	B_i^+	B_i^-	**Val**	B_i^+	B_i^-	**Val**	B_i^+	B_i^-	**Val**
rac_cdc42	1		**0**												
p90rsk	1		**0**	1		**0**	1		**0**	1		**0**	1		**0**
erk12		1	**-1**		1	**-1**		1	**-1**		1	**-1**		1	**-1**
sos1_eps8_e3b1				1		**1**									
vav2							1		**1**						
pi34p2										1		**1**			
pip3													1		**1**

Πέντε σύνολα διορθώσεων αναγνωρίζονται για το δίκτυο του EGFR (σχήμα 3.26) σχετικά με το σενάριο 14 του σχήματος 4.4. Κάθε σύνολο οδηγεί στην εξάλειψη της ασυμφωνίας μεταξύ πειραματικών δεδομένων και υπολογιστικού μοντέλου για το σενάριο 14. Οι διορθώσεις στους κόμβους p90rsk, erk12 είναι κοινές για όλα τα σύνολα. Οι διορθώσεις στους κόμβους rac_cdc42, sos1_eps8_e3b1, vav2, pi34p2, pip3 εναλλάσσονται διαδοχικά στα σύνολα MCoS 1–5. Στις στήλες MCoS 1–5, τρεις υποστήλες φαίνονται: η υποστήλη "Val" δείχνει την διορθωμένη τιμή του εκάστοτε κόμβου, η υποστήλη "B_i^+" δείχνει θετική διόρθωση του αντίστοιχου κόμβου, και η υποστήλη "B_i^-" δείχνει αρνητική διόρθωση του αντίστοιχου κόμβου.



Σχήμα 3.28: Υπέρθεση όλων των τοπολογιών που ελαχιστοποιούν την ασυμφωνία μεταξύ πειραματικών δεδομένων και υπολογιστικού μοντέλου. Οι μαύρες αντιδράσεις διατηρούνται σε όλες τις λύσεις, οι μπλέ αντιδράσεις διατηρούνται σε κάποιες από τις λύσεις, ενώ οι γκρί αντιδράσεις αφαιρούνται από όλες τις λύσεις.



Σχήμα 3.29: Διακριτοποιημένα δεδομένα και ασυμφωνία με το βελτιστοποιημένο υπολογιστικό μοντέλο (κατόπιν εφαρμογής του αλγορίθμου OPT_SUBGRAPH).

6 Αναγνώριση σηματοδοτικών μονοπατιών που σχετίζονται με την κλινική αποτελεσματικότητα φαρμάκων στον ηπατικό καρκίνο χρησιμοποιώντας φωσφοπρωτεομικά, γενομικά και κλινικά δεδομένα

Στην παρούσα ενότητα προτείνουμε μια καινοτόμα μέθοδο για την αναγνώριση σηματοδοτικών μονοπατιών που σχετίζονται με την κλινική αποτελεσματικότητα φαρμάκων στον ηπατικό καρκίνο χρησιμοποιώντας φωσφοπρωτεομικά, γενομικά και κλινικά δεδομένα. Η έρευνα αυτή κατατέθηκε προς δημοσίευση από τους Melas et al., τον Φεβρουάριο του 2013, και πραγματοποιήθηκε σε συνεργασία με τον Prof. Douglas A. Lauffenburger (head of the Biological Engineering department of MIT, Cambridge, MA, USA).

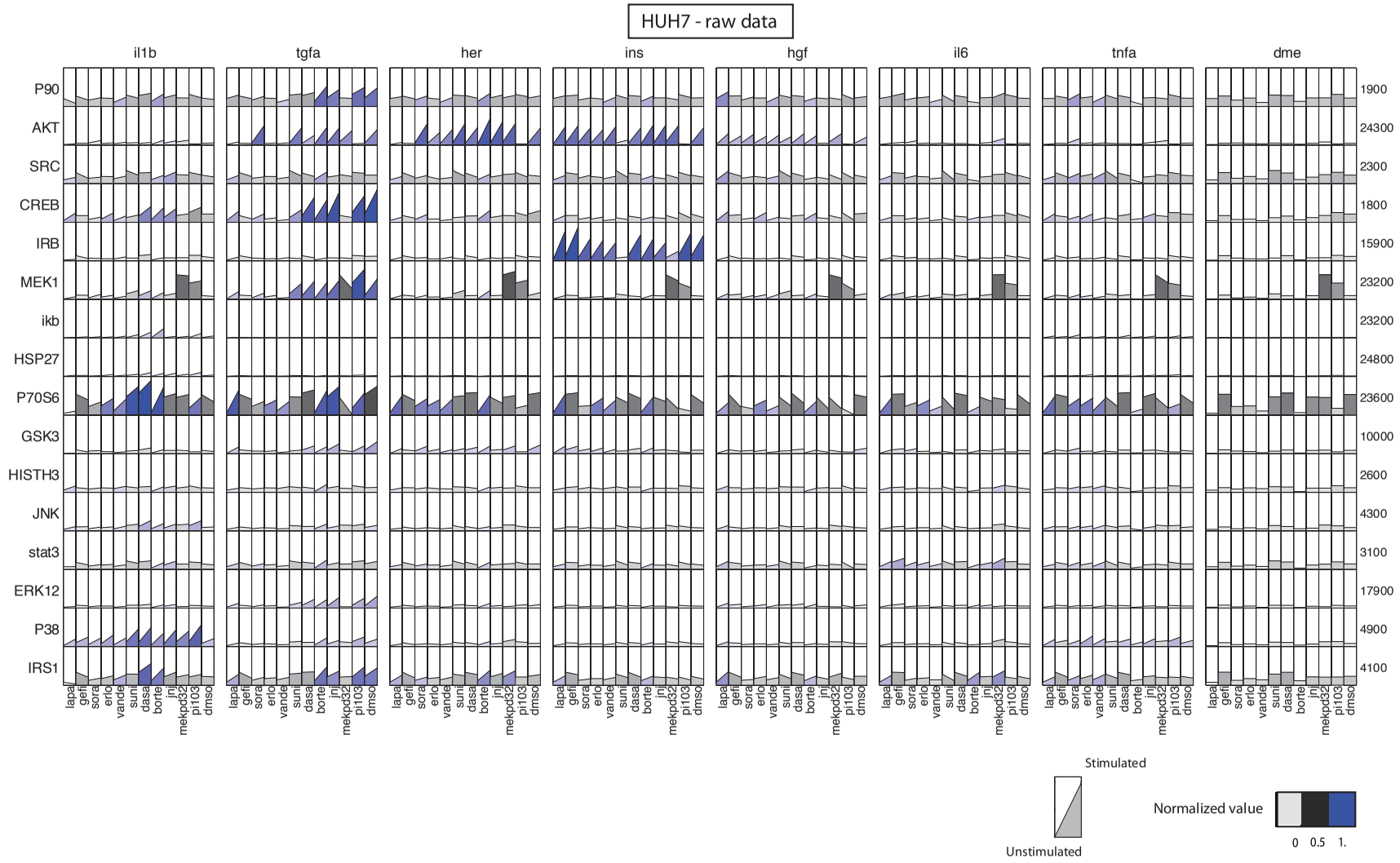
Πιο συγκεκριμένα εξετάζουμε τις επιδράσεις στο φωσφοπρωτεομικό επίπεδο 3 καρκινικών ηπατικών σειρών (HUH7, HEPG2, HEP3B), 8 φαρμάκων για μη εγχειρήσιμο ηπατικό καρκίνο, και στην συνέχεια χρησιμοποιούμε αλγόριθμο μηχανικής μάθησης (SVM) για την εξαγωγή εκείνων των επιδράσεων που είναι ενδεικτικές της κλινικής αποτελεσματικότητας των εν λόγω φαρμάκων. Εξετάζουμε τα παρακάτω φάρμακα:

- Lapatinib, αναστολέας του epidermal growth factor receptor (EGFR), tyrosine kinase 1 and 2 (Her2/Neu). Δεν αποδείχθηκε αποτελεσματικός κατά του μη εγχειρήσιμου ηπατικού καρκίνου σε κλινικές δοκιμές φάσης 2.
- Gefitinib, επίσης αναστολέας του EGFR ινηβιτορ. Δεν αποδείχθηκε αποτελεσματικός κατά του μη εγχειρήσιμου ηπατικού καρκίνου σε κλινικές δοκιμές φάσης 2.
- Sorafenib, μικρός αναστολέας του VEGFR, PDGFR και του Raf. Έχει εγκριθεί για την αγωγή μη εγχειρήσιμου ηπατικού καρκίνου [62].
- Erlotinib, επίσης αναστολέας του EGFR. Έχει εγκριθεί για την αγωγή μη εγχειρήσιμου ηπατικού καρκίνου [135, 136].
- Vandetanib, ανταγωνιστής του VEGFR και EGFR. Δεν αποδείχθηκε αποτελεσματικός κατά του μη εγχειρήσιμου ηπατικού καρκίνου σε κλινικές δοκιμές φάσης 3.
- Sunitinib, αναστολέας του PDGFR και VEGFR. Δεν αποδείχθηκε αποτελεσματικός κατά του μη εγχειρήσιμου ηπατικού καρκίνου σε κλινικές δοκιμές φάσης 3.
- Dasatinib, αναστολέας του BCR/ABL και Src. Βρίσκεται ακόμα σε δοκιμές τύπου 2 για μη εγχειρήσιμο ηπατικό καρκίνο.
- Bortezomib, αναστολέας πρωτεασών. Βρίσκεται ακόμα σε δοκιμές τύπου 2 για μη εγχειρήσιμο ηπατικό καρκίνο [133].

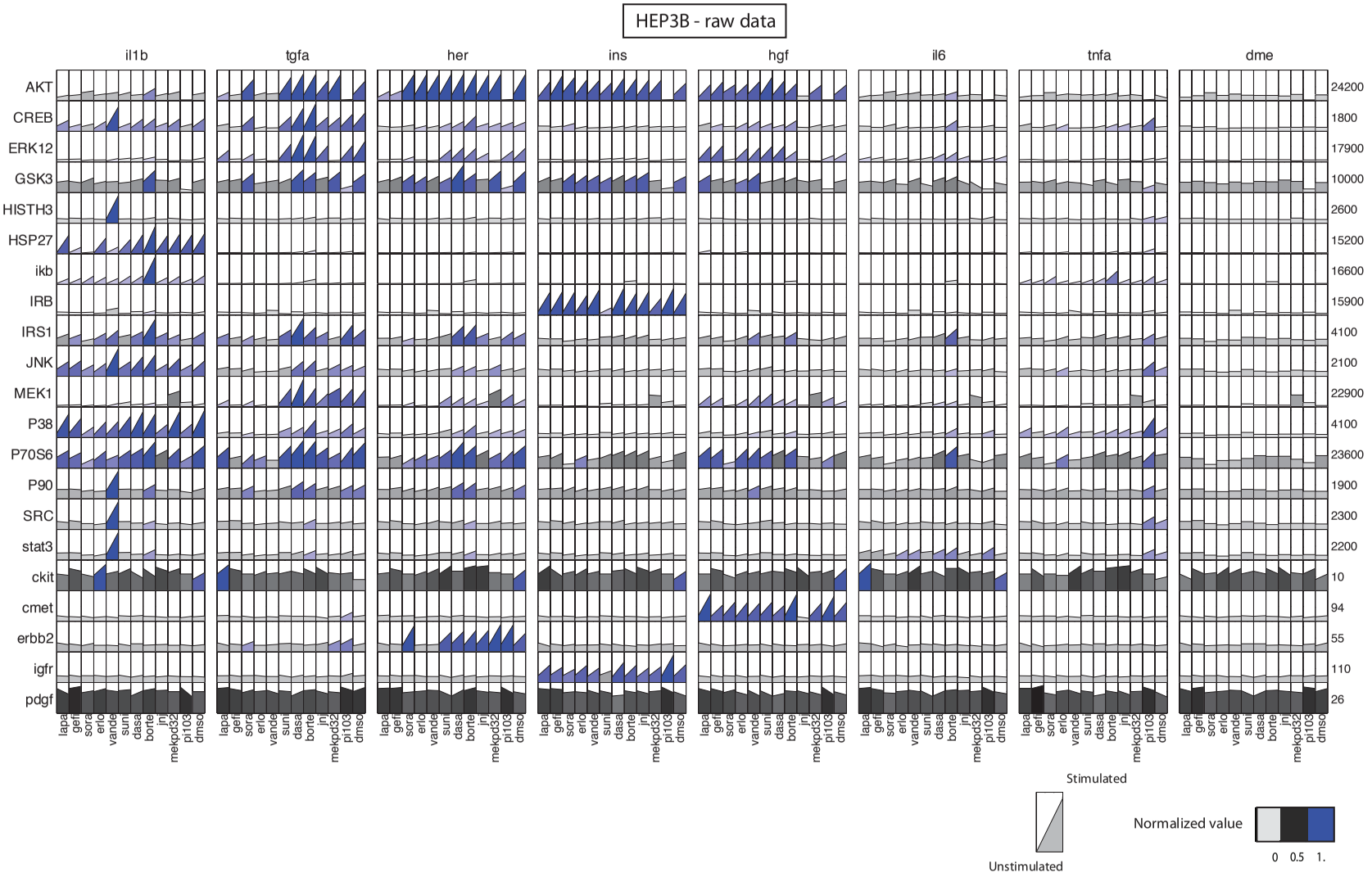
Τα συγκεκριμένα φάρμακα εισάγουμε στις τρεις υπό εξέταση κυτταρικές σειρές σε συνδυασμό με 6 ερεθίσματα (κυτοκίνες): IL1β, TGFα, Heregulin (HER), Insulin (INS), IL6 and TNFα και τρεις εξειδικευμένους αναστολείς: MEKi, PI3Ki, cMETi, ενώ μετράμε την ενεργοποίηση 16 φωσφοπρωτεϊνικών σημάτων. Στην συνέχεια κατασκευάζονται σηματοδοτικά μονοπάτια με χρήση αλγορίθμου Ακέραιου Γραμμικού Προγραμματισμού προς αναγνώριση των αντιδράσεων που μπλοκαρίστηκαν από τα εν λόγω φάρμακα (κατα αντιστοιχία με την μέθοδο που περιγράφηκε στην ενότητα 2.3), και τέλος, εφαρμόζεται αλγόριθμος SVM για την εξαγωγή των επιδράσεων εκείνων που είναι ενδεικτικές της κλινικής αποτελεσματικότητας των 8 αντικαρκινικών φαρμάκων.

6.1 Φωσφοπρωτεομικά δεδομένα

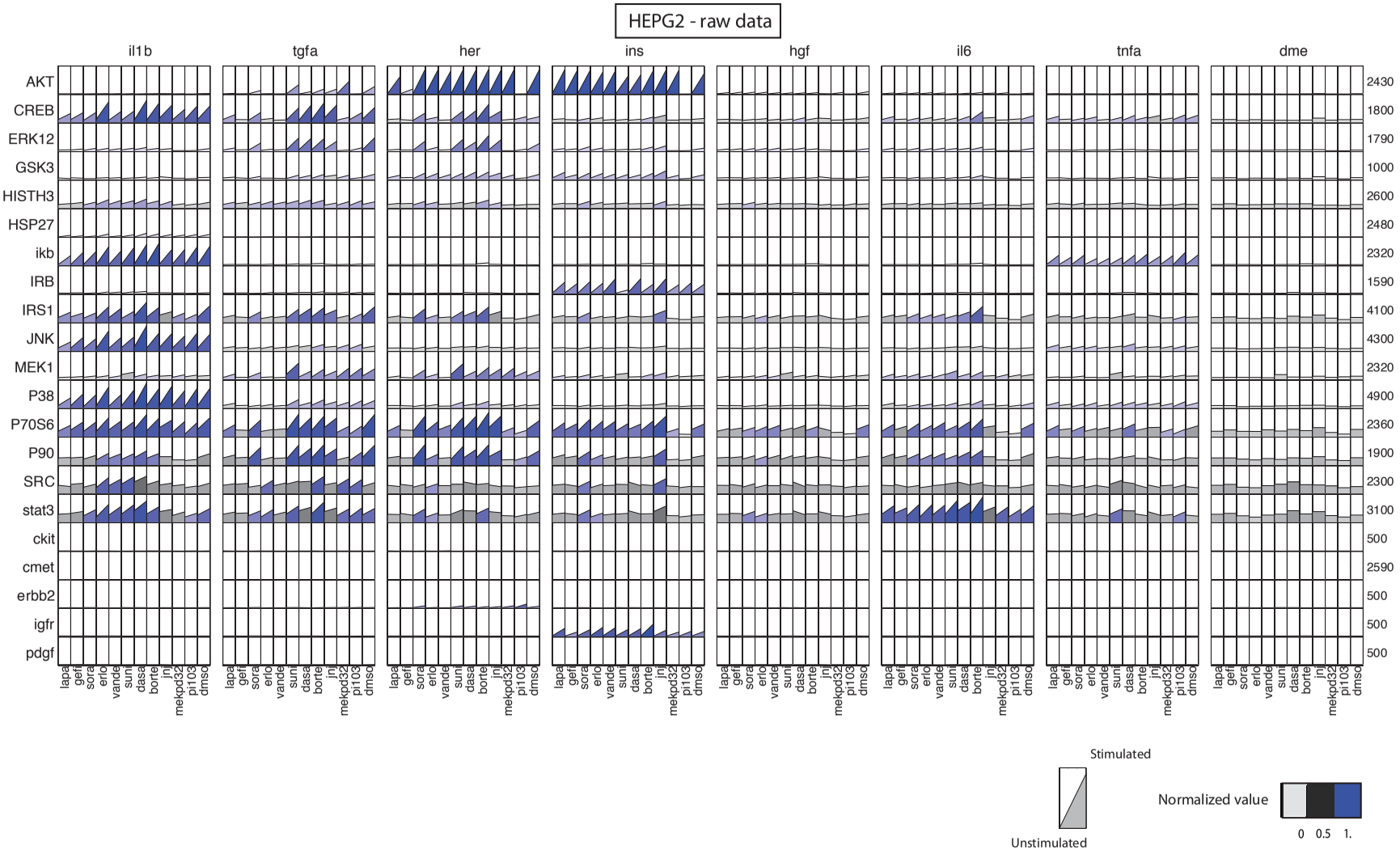
Τα φωσφοπρωτεομικά δεδομένα για τις 3 καρκινικές σειρές φαίνονται στα σχήματα 3.30, 3.31, 3.32. Στο σχήμα 3.33 φαίνεται η υπέρθεση των 3 ηπατικών σειρών σε μία μέση καρκινική κυτταρική σειρά.



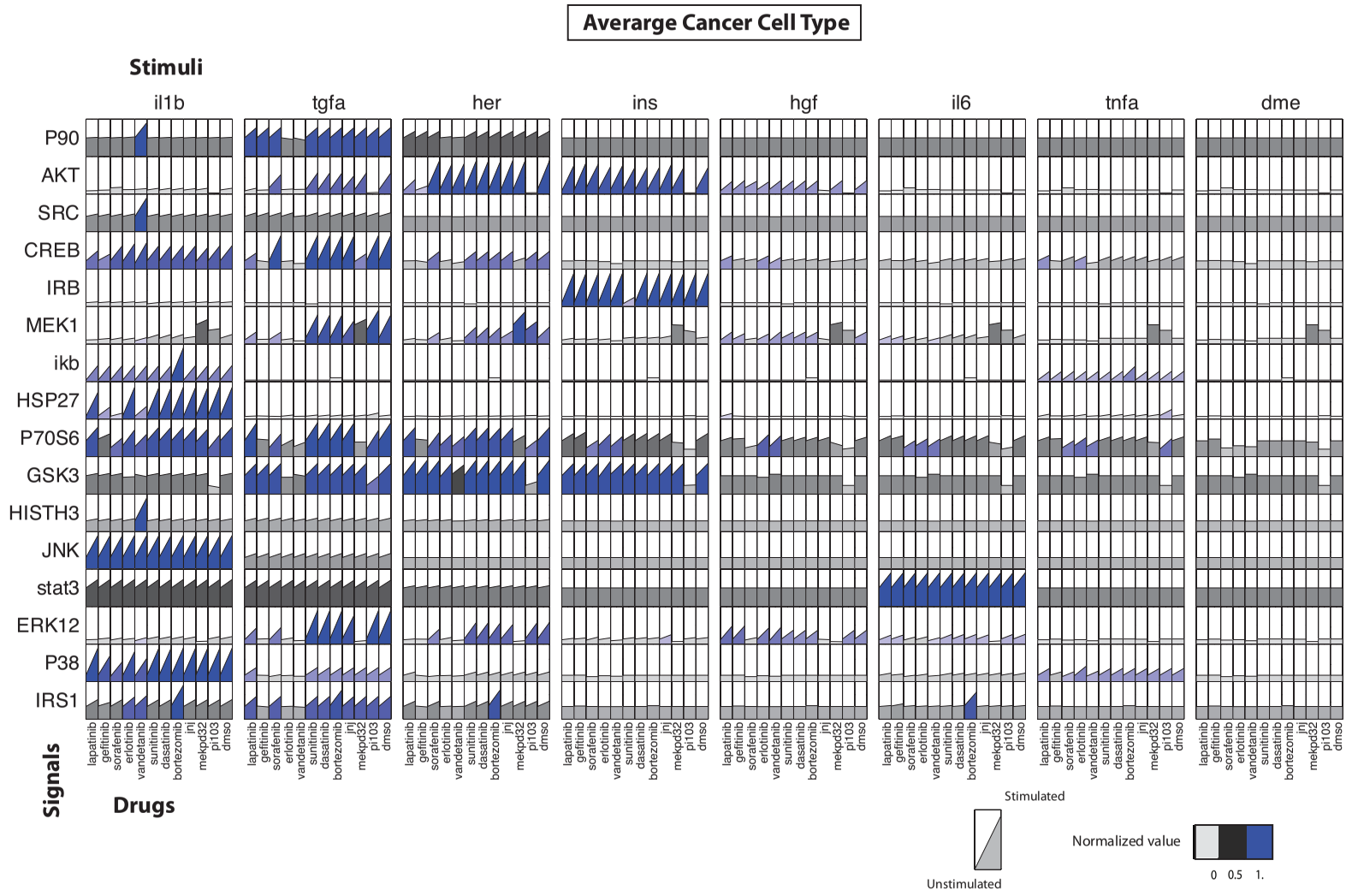
Σχήμα 3.30: Φωσφοπρωτεομικά δεδομένα για την κυτταρική σειρά HUH7.



Σχήμα 3.31: Φωσφοπρωτεομικά δεδομένα για την κυτταρική σειρά HEP3B.



Σχήμα 3.32: Φωσφοπρωτεομικά δεδομένα για την κυτταρική σειρά HEPG2.



Σχήμα 3.33: Φωσφοπρωτεομικά δεδομένα για την μέση καρκινική κυτταρική σειρά.

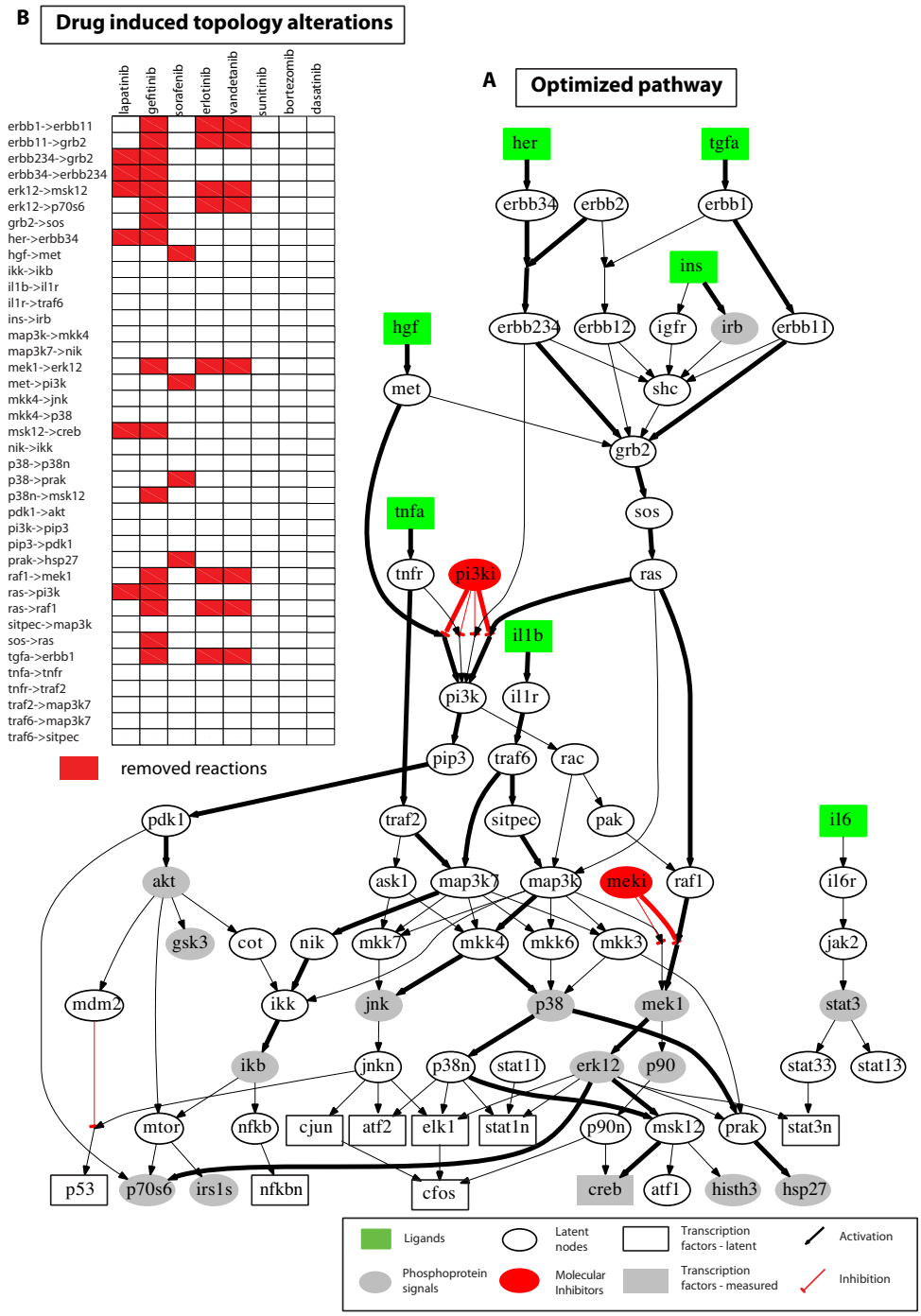
Σχετικά με τις επιδράσεις των φαρμάκων στις μετρούμενες φωσφοπρωτεϊνες (μέση καρκινική κυτταρική σειρά), παρατηρούμε ότι το Lapatinib μπλόκαρε την ενεργοποίηση του AKT και μερικώς την ενεργοποίηση του CREB, MEK1 και ERK12 κατόπιν εισαγωγής του TGFα. Επίσης μπλόκαρε την ενεργοποίηση του CREB, MEK1, ERK12 και μερικώς του AKT κατόπιν εισαγωγής του HER, ενώ δεν είχε σημαντικές επιδράσεις σε κανένα άλλο μονοπάτι. Το Gefitinib είχε τις ίδιες επιδράσεις με το Lapatinib παρουσία του HER και επίσης μπλόκαρε τα περισσότερα σήματα παρουσία του TGFα, όπως τα AKT, CREB, MEK1, ERK12, P38, και IRS1. Το Gefitinib μπλόκαρε επίσης μερικώς την ενεργοποίηση του HSP27 παρουσία του IL1β. Το Sorafenib δεν είχε σημαντικές επιδράσεις στο μονοπάτι του TGFα ή του HER, ωστόσο μπλόκαρε την ενεργοποίηση των MEK1, P70S6, και ERK12 παρουσία HGF και το HSP27 παρουσία IL1β. Το Erlotinib και το Vandetanib είχαν πολύ παραπλήσιες επιδράσεις με το Gefitinib κατόπιν εισαγωγής του TGFα και στα περισσότερα από τα μετρούμενα σήματα κατόπιν εισαγωγής του HER, εκτός από το AKT (το Gefitinib και το Lapatinib μπλόκαραν την ενεργοποίηση του AKT παρουσία HER, ενώ το Erlotinib και το Vandetanib δεν το μπλόκαραν). Το Sunitinib δεν είχε σημαντική επίδραση σε κανένα από τα μετρούμενα σήματα εκτός από το IRB παρουσία INS. Το Dasatinib επίσης δεν είχε σημαντική επίδραση σε κανένα από τα μετρούμενα σήματα υποδηλώνοντας ότι το πεδίο δράσης του είναι έξω από την γειτονιά του σηματοδοτικού δικτύου που εξετάζουμε. Το Bortezomib, αύξησε την τιμή ενεργοποίησης του IKβ και του IRS1 σε όλα τα πειραματικά σενάρια. Οι υπόλοιποι εξειδικευμένοι αναστολείς (JNJ, MEKPD32, PI103) είχαν πολύ πιο καθαρή δράση, επηρεάζοντας μόνο την πρωτεϊνη στόχο τους.

6.2 Βελτιστοποίηση σηματοδοτικών μονοπατιών στα φωσφοπρωτεομικά δεδομένα

Για την καταγραφή των επιδράσεων των 8 υπο εξέταση φαρμάκων στο σηματοδοτικό δίκτυο της μέσης καρκινικής κυτταρικής σειράς, χρησιμοποιήθηκε η μεθοδολογία που περιγράφηκε στην ενότητα 2.3 και επίσης δημοσιεύτηκε από τους Mitsos et al. [23]. Συνοπτικά, πρώτα κατασκευάσαμε ένα σηματοδοτικό μονοπάτι με βάση την βιβλιογραφία και στην συνέχεια το βελτιστοποιήσαμε χρησιμοποιώντας τα δεδομένα από τα 7 ερεθίσματα και τους 3 εξειδικευμένους αναστολείς με στόχο την δημιουργία δικτύου που θα αντιπροσωπεύει πιστά τους σηματοδοτικούς μηχανισμούς των υπό εξέταση κυττάρων. Ακολούθως, βελτιστοποιήσαμε το εξειδικευμένο δίκτυο στα φωσφοπρωτεομικά δεδομένα από το εκάστοτε φάρμακο ώστε να αναγνωριστούν οι αντιδράσεις που μπλοκάρει το κάθε φάρμακο. Το εξειδικευμένο δίκτυο μαζί με τις αντιδράσεις που μπλοκάρει το κάθε φάρμακο φαίνεται στο σχήμα 3.34. Τα σηματοδοτικά δίκτυα για το εκάστοτε φάρμακο φαίνονται στα σχήματα 3.35 - 3.38

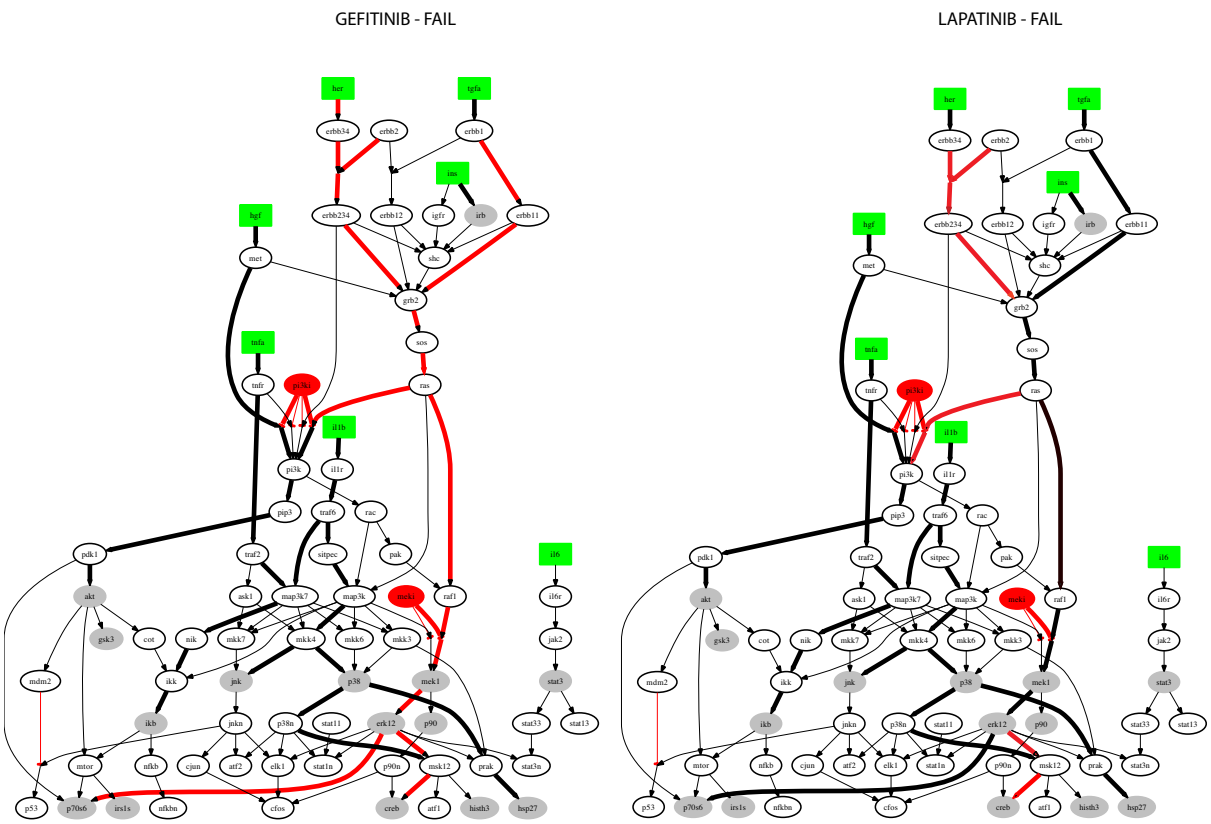
6.3 Εξαγωγή των επιδράσεων που είναι ενδεικτικές για την κλινική αποτελεσματικότητα των υπο εξέταση φαρμάκων

Με βάση τα φωσφοπρωτεομικά δεδομένα. Δεδομένα για 5 από τα αρχικά 8 φάρμακα (Lapatinib, Gefitinib, Sorafenib, Erlotinib and Vandetanib), που είχαν σημαντικές επιδράσεις στις μετρούμενες φωσφοπρωτεϊνες χρησιμοποιήθηκαν από ένα SVM για την αναγνώριση των φωσφοπρωτεομικών μετρήσεων που είναι ενδεικτικές για την κλινική αποτελεσματικότητα των υπο εξέταση φαρμάκων. Το SVM εφαρμόστηκε μέσω του `Matlab`, συνάρτηση `classify`. Τα αποτελέσματα φαίνονται στο σχήμα 3.39. Οι πιο ενδεικτικές μετρήσεις είναι (i) η μέτρηση του AKT και του CREB κατόπιν εισαγωγής του TGFα, και η μέτρηση του ERK12 και του MEK1 κατόπιν εισαγωγής του HER, (ii) η μέτρηση του AKT παρουσία HER, και (iii) η μέτρηση του ERK12 και του MEK1 παρουσία HGF. Συγκεκριμένα, το SVM αναγνώρισε ότι το μπλοκάρισμα των (i) και (ii) είναι ενδεικτικό ενός φαρμάκου που απέτυχε σε κλινικές δοκιμές ενώ το μπλοκάρισμα του (iii), είναι ενδεικτικό ενός φαρμάκου που πέτυχε στις κλινικές δοκιμές. (ακρίβεια 80%).



Σχήμα 3.34: Το εξειδικευμένο δίκτυο και οι αντιδράσεις που μπλοκάρει το κάθε φάρμακο.

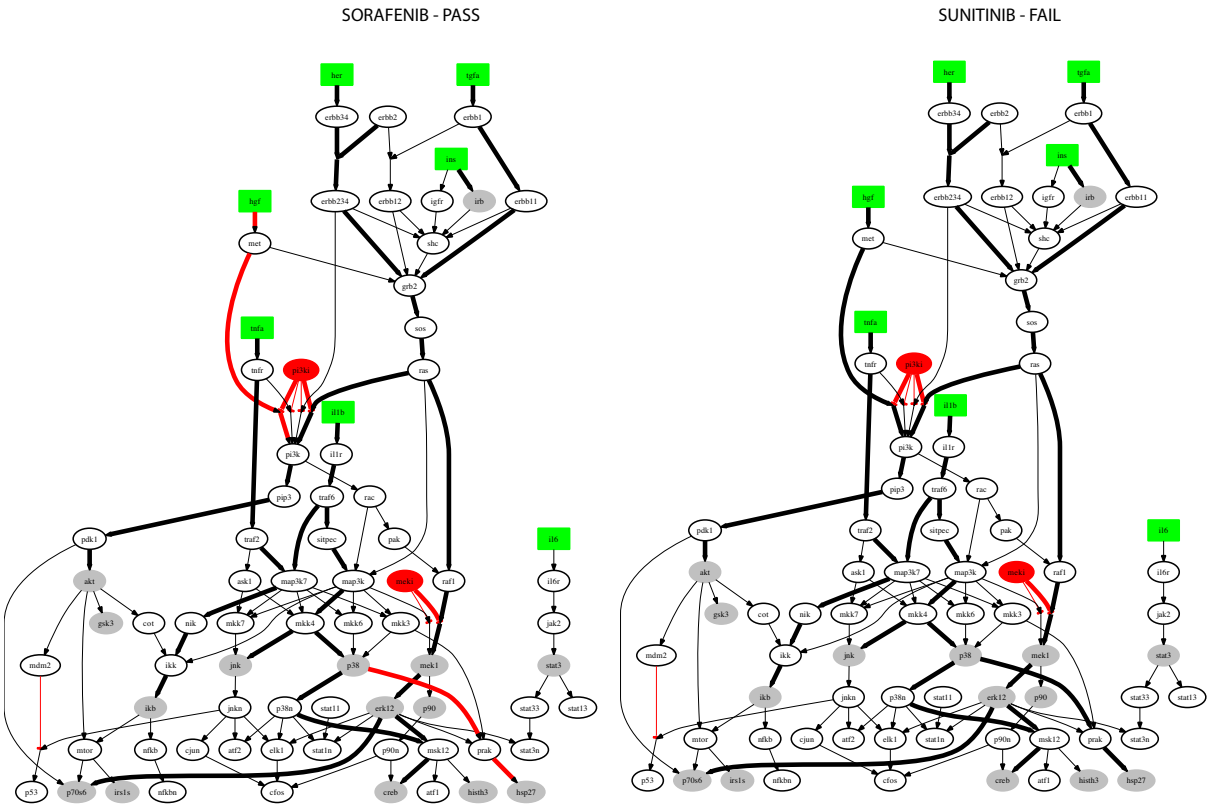
Με βάση τις αντιδράσεις που μπλοκάρονται. Χρησιμοποιώντας παρόμοια μεθοδολογία με αυτήν της προηγούμενης παραγράφου για τα φωσφοπρωτεομικά δεδομένα, εδώ εισάγουμε τις αντιδράσεις που μπλοκάρονται από το εκάστοτε φάρμακο στο SVM με σκοπό την αναγνώριση των αντιδράσεων εκείνων που είναι ενδεικτικές της κλινικής αποτελεσματικότητας των υπό εξέταση φαρμάκων. Τα αποτελέσματα φαίνονται στο σχήμα 3.40. Οι πιο ενδεικτικές αντιδράσεις είναι (i) η αντίδραση ERK12 → MSK12, (ιι) η αντίδραση RAS → PI3K, και (ιιι) η αντίδραση PRAK →



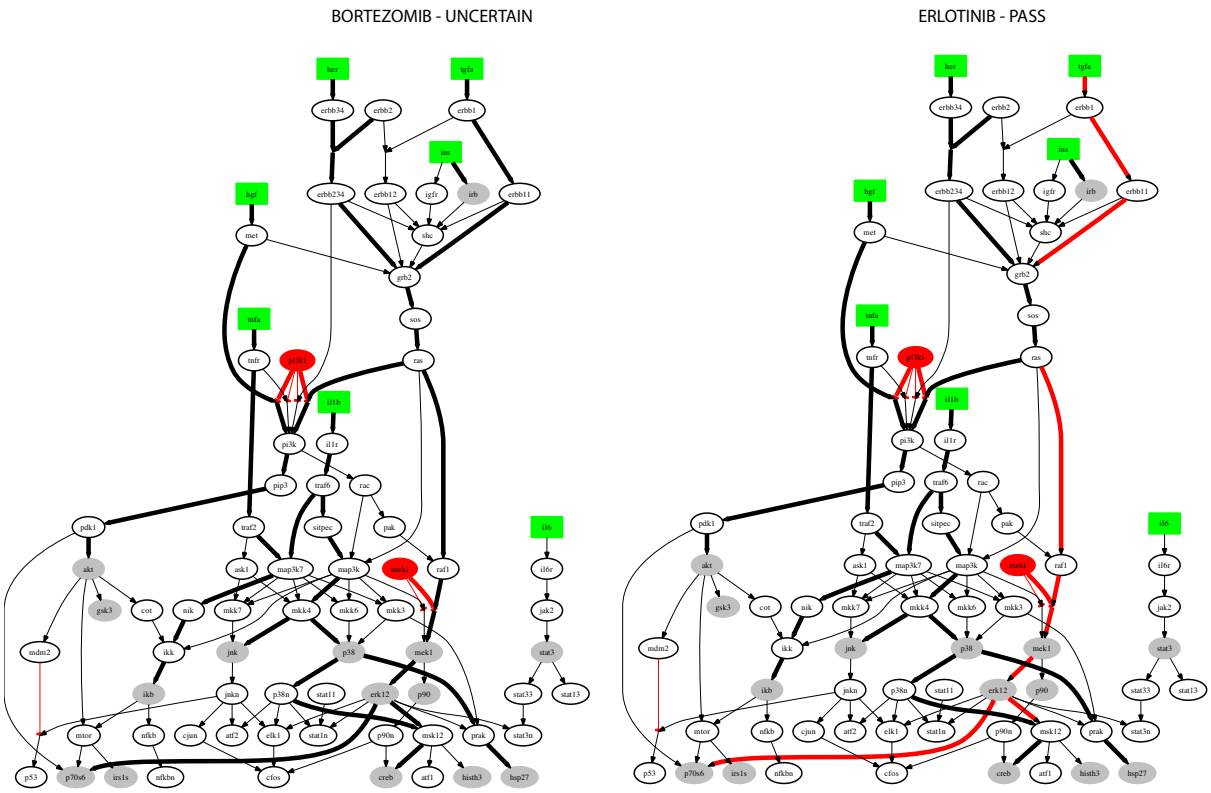
Σχήμα 3.35: Σηματοδοτικά δίκτυα για το gefitinib και το lapantinib

HSP27 (ακρίβεια 80%). Συγκεκριμένα, το SVM αναγνώρισε ότι το μπλοκάρισμα των αντιδράσεων (i) και (ii) είναι ενδεικτικό ενός φαρμάκου που απέτυχε σε κλινικές δοκιμές ενώ το μπλοκάρισμα της αντίδρασης (iii), είναι ενδεικτικό ενός φαρμάκου που πέτυχε στις κλινικές δοκιμές. (ακρίβεια 80%).

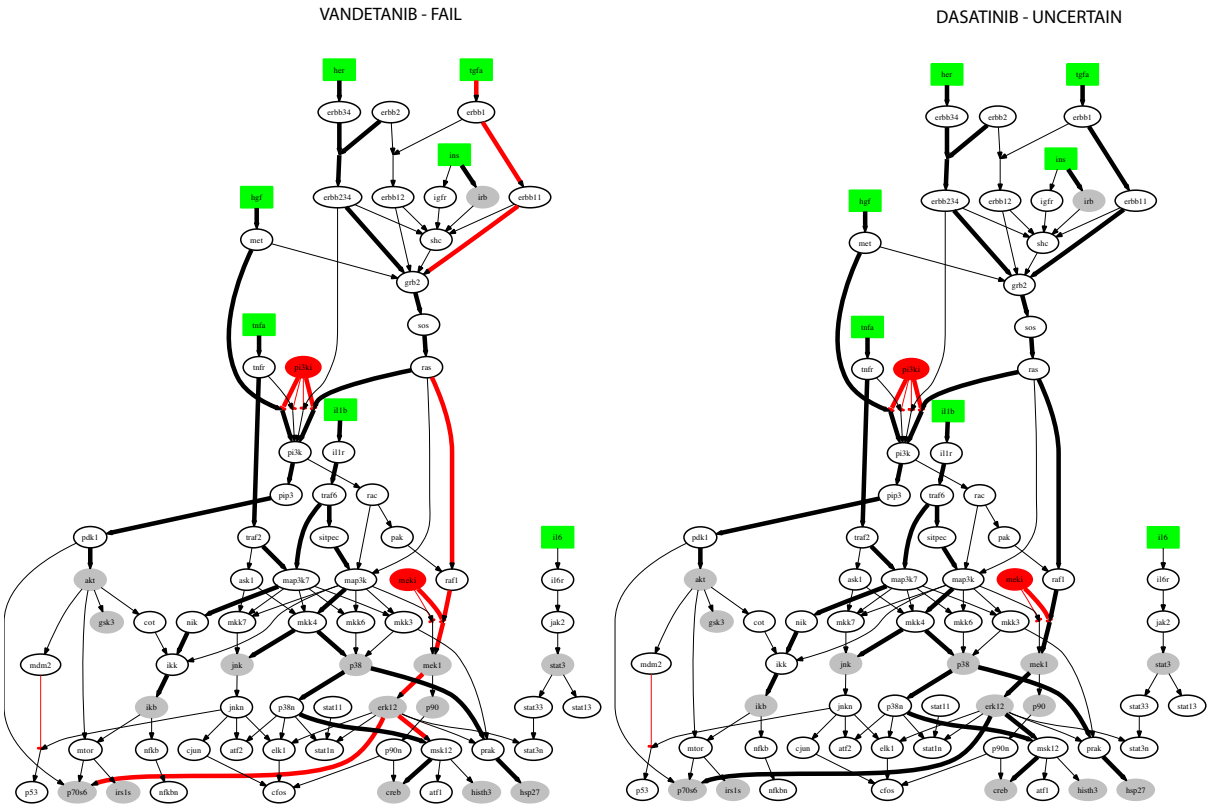
Η προτεινόμενη μεθοδολογία είναι από τις πρώτες απόπειρες για την εκμετάλλευση μαζικών φωσφοπρωτεομικών δεδομένων μέσω αλγορίθμου μηχανικής μάθησης SVM, με σκοπό την πρόβλεψη της κλινική αποτελεσματικότητας φαρμάκων.



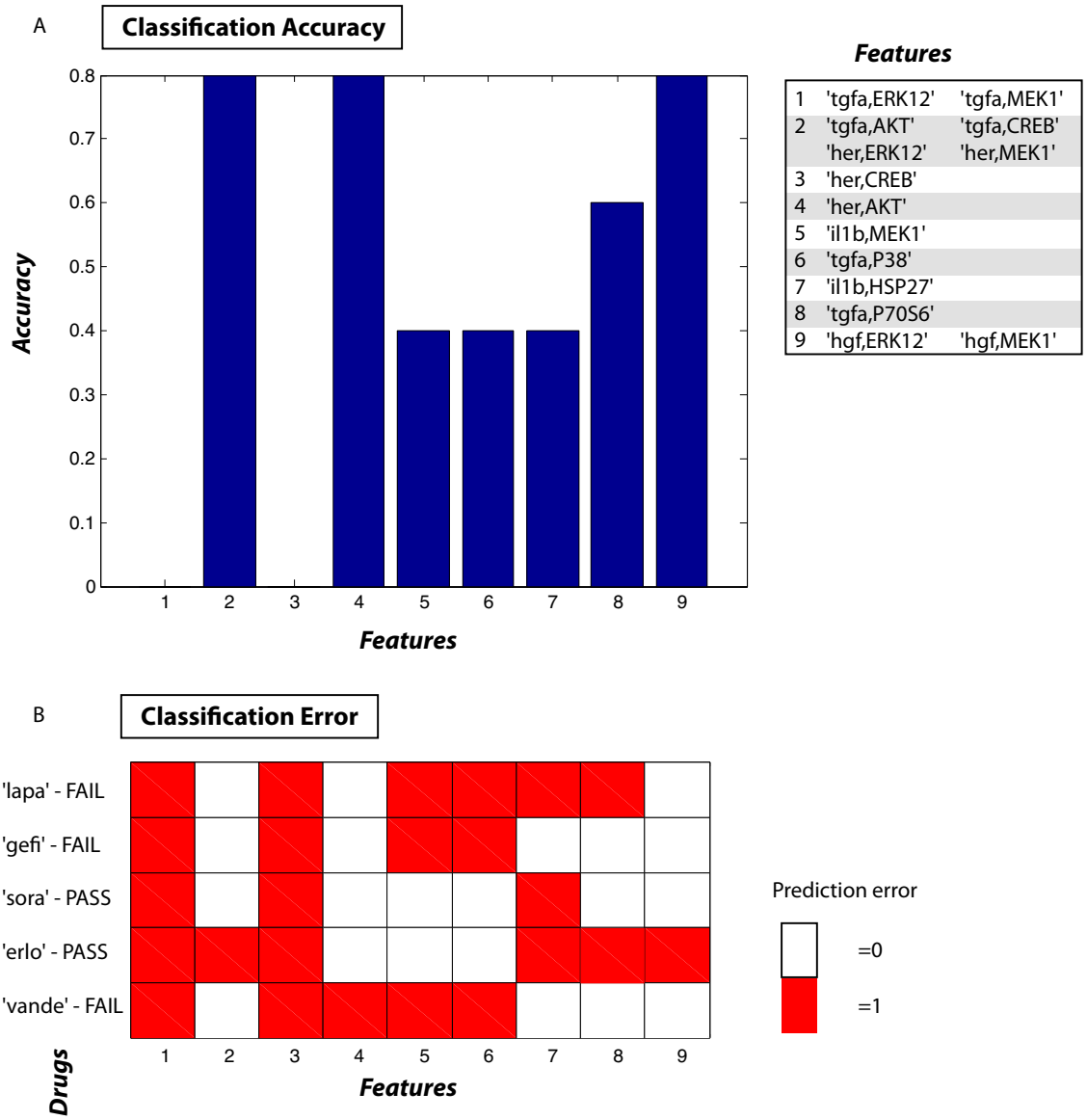
Σχήμα 3.36: Σηματοδοτικά δίκτυα για το sorafenib και το sunitinib



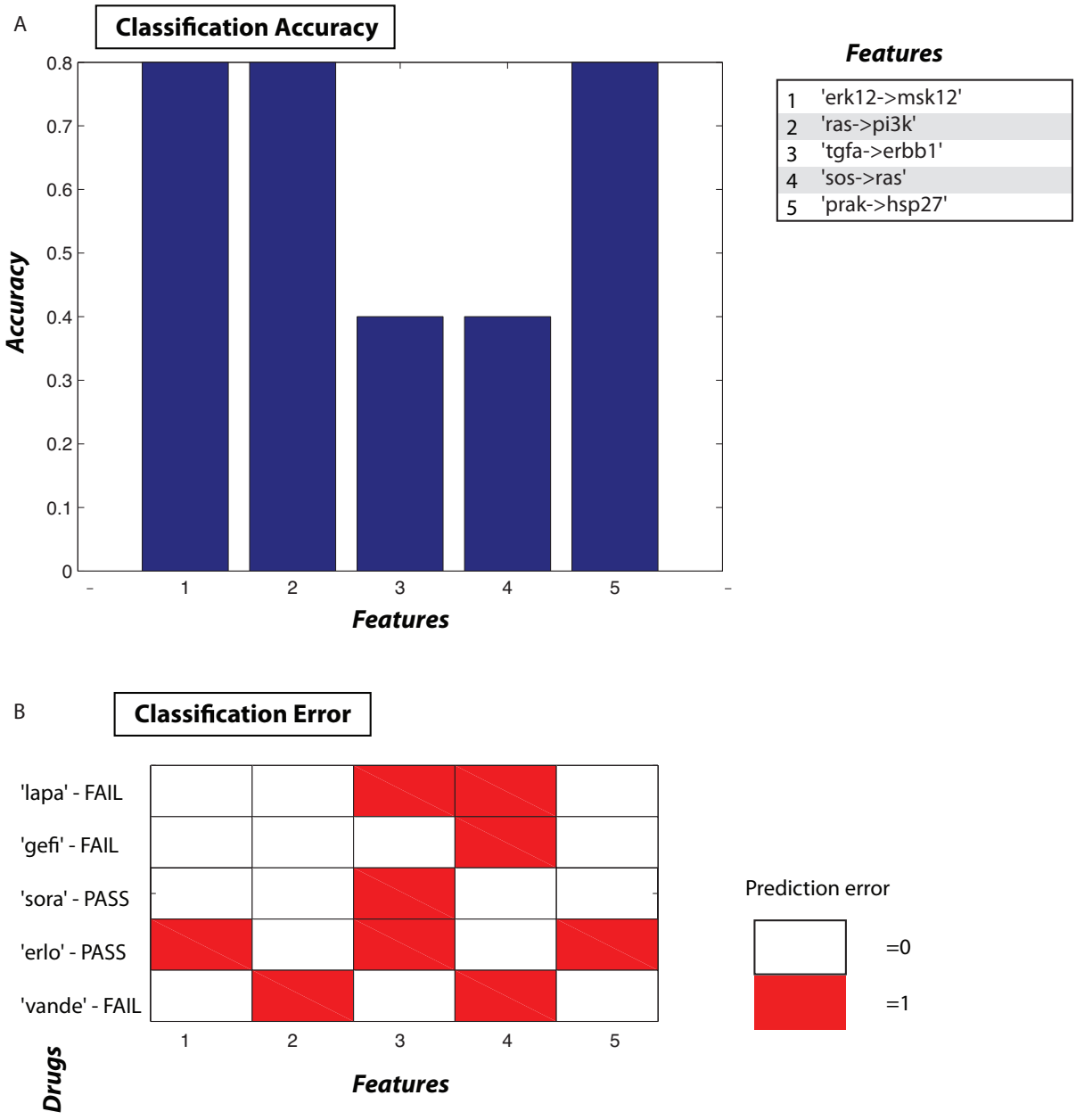
Σχήμα 3.37: Σηματοδοτικά δίκτυα για το bortezomib, και το erlotinib



Σχήμα 3.38: Σηματοδοτικά δίκτυα για το vandetanib



Σχήμα 3.39: Εξαγωγή χαρακτηριστικών - φωσφοπρωτεομικά δεδομένα



Σχήμα 3.40: Εξαγωγή χαρακτηριστικών - Αντιδράσεις που μπλοκάρονται από το εκάστοτε φάρμακο

7 Συμπεράσματα

Σε αυτή τη διδακτορική διατριβή ο συγγραφέας εισήγαγε μια νέα κατηγορία μεθόδων για την μοντελοποίηση ενδοκυτταρικών σηματοδοτικών μονοπατιών, με βάση τον Ακέραιο Γραμμικό και μη-Γραμμικό Προγραμματισμό (ILP/NLP). Η χρήση των εν λόγω μεθόδων επιτρέπει την μοντελοποίηση πολύ μεγαλύτερων δικτύων από ότι ήταν δυνατό πρώτα, και την κατασκεύη υπολογιστικών μοντέλων που αναπαριστούν πιστά τους σηματοδοτικούς μηχανισμούς της υπο εξέταση κυτταρικής σειράς.

Πιο συγκεκριμένα, αναπτύχθηκαν 3 διαφορετικοί φορμαλισμοί: **(i)** Ένας φορμαλισμός ILP που μοντελοποιεί τα δίκτυα μεταγωγής σήματος με χρήση της Boolean λογικής και ελαχιστοποιεί την ασυμφωνία με πειραματικά δεδομένα αφαιρώντας ακμές από το δίκτυο, που φαίνεται να μην είναι λειτουργικές στον υπο εξέταση κυτταρικό τύπο. **(ii)** Ένας αλγόριθμος NLP που μοντελοποιεί το σηματοδοτικό δίκτυο με χρήση λογικής fuzzy και προσδιορίζει την βέλτιστη τιμή των παραμέτρων ώστε να ελαχιστοποιηθεί η ασυμφωνία με πειραματικά δεδομένα. **(iii)** Αλγόριθμος ILP που μοντελοποιεί το δίκτυο ως προσημασμένο άκυκλο γράφο και βελτιστοποιεί την τοπολογία του ώστε να ελαχιστοποιηθεί η ασυμφωνία με πειραματικά δεδομένα.

Επιπρόσθετα, εξετάστηκαν 5 διαφορετικές εφαρμογές: **(i)** Αναγνώριση των επιδράσεων αντικαρκινικών φαρμάκων στο δίκτυο σηματοδότησης καρκινικών ηπατοκυττάρων. **(ii)** Βελτιστοποίηση εκτεταμένου δικτύου μεταγωγής σήματος σε φυσιολογικά ηπατοκύτταρα. **(iii)** Κατασκευή εκτεταμένων σηματοδοτικών μονοπατιών ώστε να περιλαμβάνουν ενδοκυτταρική και εξωκυτταρική σηματοδότηση. **(iv)** Μοντελοποίηση σηματοδοτικών δικτύων στα χονδροκύτταρα και αναγνώριση καινούριων κυττοκινών που μπορεί να ευθύνονται για την αποδόμηση του αρθρικού χόνδρου σε παθολογικές περιπτώσεις. **(v)** Αναγνώριση σηματοδοτικών μονοπατιών που σχετίζονται με την κλινική αποτελεσματικότητα φαρμάκων στον ηπατικό καρκίνο χρησιμοποιώντας φωσφοπρωτεομικά, γενομικά και κλινικά δεδομένα.

8 Εργασιές που δημοσιεύτηκαν στα πλαίσια αυτού του διδακτορικού

Συνολικά 12 εργασίες δημοσιεύτηκαν στα πλαίσια αυτού του διδακτορικού σε έγκριτα περιοδικά και συνέδρια της IEEE. Αυτές είναι οι ακόλουθες (με χρονολογική σειρά):

Σε έγκριτα περιοδικά

1. *Modeling of signaling pathways in chondrocytes based on phosphoproteomic and cytokine release data.* **Ioannis N. Melas***, Aikaterini D. Chairakaki*, Elisavet I. Chatzopoulou*, Dimitris E. Messinis, Alexander Mitsos, Zoe Dailiana, Panagoula Kollia, Leonidas G. Alexopoulos. Submitted, 2013. *Equal contributors

2. *Phosphoproteomics in drug discovery.* Melody K Morris, An Chi, **Ioannis N. Melas**, Leonidas G Alexopoulos. Drug Discovery Today, Accepted, 2013

3. *Leveraging systems biology approaches in clinical pharmacology.* **Ioannis N. Melas**, Kosmas Kretsos, Leonidas G Alexopoulos. Biopharm Drug Dispos. 2013 Aug 23. doi: 10.1002/bdd.1859

4. *Detecting and Removing Inconsistencies Between Experimental Data and Signaling Network Topologies using Integer Linear Programming on Interaction Graphs.* **Ioannis N. Melas***, Regina Samaga*, Leonidas G Alexopoulos, and Steffen Klamt. PLoS Comput Biol. 2013;9:e1003204. doi:10.1371/journal.pcbi.1003204, *Equal contributors

5. *Identification of signaling pathways related to drug efficacy in HCC via integration of phosphoproteomic, genomic and clinical data.* **Ioannis N. Melas**, Douglas A. Lauffenburger, Leonidas G. Alexopoulos. *Accepted*, 13th IEEE International Conference on BioInformatics and BioEngineering. 2013

6. *Construction of cell type-specific logic models of signaling networks using CellNetOptimizer.* M. K. Morris, **I. Melas**, J. Saez-Rodriguez. Methods in Molecular Biology:Computational Toxicology, Ed. B. Reisfeld and A. Mayeno, Humana Press. 2012

7. *Non Linear Programming (NLP) formulation for quantitative modeling of protein signal transduction pathways.* Alexander Mitsos*, **Ioannis N. Melas***, Melody K. Morris, Julio Saez-Rodriguez, Douglas A. Lauffenburger, Leonidas G. Alexopoulos. PLoS One. 2012;7(11):e50085. doi: 10.1371/journal.pone.0050085. Epub 2012 Nov 30 2012, *Equal contributors

8. *Construction of large signaling pathways using an adaptive perturbation approach with phosphoproteomic data.* **Ioannis N. Melas***, Alexander Mitsos*, Dimitris E. Messinis, Thomas S. Weiss, Julio Saez-Rodriguez, and Leonidas G. Alexopoulos, Mol. BioSyst., 10.1039/C2MB05482E, 2012. *Equal Contributors.

9. *Combined logical and data-driven models for linking signaling pathways to cellular response.* **I. Melas***, A. Mitsos*, D. Messinis, T. Weiss, L. Alexopoulos, Bmc Syst Biol 5, 107 (2011).(*equal contributors)

10. *Modeling signaling pathways in articular cartilage.* **Ioannis N. Melas**, Aikaterini D. Chairakaki, Alexander Mitsos, Zoe Dailiana, Christopher G. Provatidis, Leonidas G. Alexopoulos 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2011), Boston, MA, US

11. *Construction of Pathways and identification of drug effects in liver cancer cells via an Integer Linear Programming (ILP) formulation.* Leonidas G Alexopoulos, **Ioannis N. Melas**, Aikaterini D. Chairakaki, Julio Saez-Rodriguez, Alexander Mitsos. 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2010, Buenos Aires, Argentina

12. *Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data.* Alexander Mitsos, **Ioannis N. Melas**, Paraskevas S Siminelakis, Aikaterini D. Chairakaki, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. PloS Comp Biol. 2009. 5(12): e1000591. doi:10.1371/journal.pcbi.1000591

Συνολικά 24 εργασίες δημοσιεύτηκαν στα πλαίσια αυτού του διδακτορικού σε συνέδρια. Αυτές είναι οι ακόλουθες (με χρονολογική σειρά):

Εργασίες συνεδρείων

1. Network analysis of signaling in chondrocytes Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Zoe H. Dailiana, Christopher Provatidis, Leonidas G. Alexopoulos. European Society of Biomechanics 2013.

2. Real time evaluation of the mechanical properties of articular cartilage during collagenase induced digestion. Evangelos Zympeloudis, Elisavet I. Chatzopoulou, Ioannis N. Melas, Zoe H. Dailiana, Christopher Provatidis, Leonidas G. Alexopoulos. European Society of Biomechanics 2013.

3. A device for simultaneous compression measurements of human cartilage. Nikolaos V. Georgiou , Nikolaos D. Nikolaou, Panagiotis D. Alevras, Elisavet I. Chatzopoulou, Ioannis N. Melas, Christopher G. Provatidis, Leonidas G. Alexopoulos European Society of Biomechanics 2013.

4. A comparison of chondrocyte response in cartilage 2d cultures. Elisavet I. Chatzopoulou, Ioannis N. Melas, Christoforos G. Provatidis , Zoe Dailiana, Leonidas G. Alexopoulos European Society of Biomechanics 2013.

5. Modeling signaling pathways in chondrocytes. Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Zoe Dailiana, Christopher Provatidis, Leonidas G. Alexopoulos. Fifth International Conference on Computational Bioengineering in Leuven, Belgium. 2013. Submitted

6. A Non Linear Programming (NLP) formulation for modeling signaling transduction pathways, Ioannis N. Melas*, Alexander Mitsos*, Melody K. Morris, Julio Saez-Rodriguez, Douglas A. Lauffenburger, Leonidas G. Alexopoulos (*eq contributors). 4th Hellenic Conference on Biomedical Technology 2012, Athens, Greece.

7. Monitoring cartilage degeneration in a high throughput format via its mechanical properties, Nikolaos Nikolaou, Panagiotis Alevras, Elisavet Chatzopoulou, Ioannis Melas, Zoe Dailiana, Christoforos Provatidis, Leonidas Alexopoulos. 4th Hellenic Conference on Biomedical Technology 2012, Athens, Greece.

8. Optimization of a large scale signaling network using an Integer Linear Programming formulation. Ioannis N. Melas*, Alexander Mitsos*, Julio Saez-Rodriguez, Leonidas G

Alexopoulos, *Equal Contributors. 7th GRACM International Congress on Computational Mechanics 2011, Athens, Greece,

9. A device for multiple indentation tests of human cartilage. Nikolaos D. Nikolaou, Panagiotis D. Alevras, Ioannis N. Melas, Christopher P. Provatidis, Leonidas G Alexopoulos. 7th GRACM International Congress on Computational Mechanics 2011, Athens, Greece

10. Construction of large signaling pathways from phosphoproteomic data using an ILP computational approach, Dimitris E. Messinis, Ioannis N. Melas, Alexander Mitsos, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. 6th conference of the Hellenic Society for Computational Biology and Bioinformatics, 2011, Patra, Greece.

11. Construction of large signaling pathways from phosphoproteomic data, Dimitris E. Messinis, Ioannis N. Melas, Alexander Mitsos, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. Planet xMap Congress 2011, Vienna, Austria.

12. Construction of large signaling pathways from phosphoproteomic data, Leonidas G. Alexopoulos, Ioannis N. Melas, Julio Saez-Rodriguez, Thomas S. Weiss, Dimitris E. Messinis, Alexander Mitsos. ICSB, Heidelberg and Mannheim, 2011.

13. Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data. Leonidas G. Alexopoulos, Ioannis N. Melas, Aikaterini D. Chairakaki, Julio Saez-Rodriguez, Alexander Mitsos., BIO-IT world conference 2010, Boston, USA.

14. Extending logical models of signaling pathways to predict cytokine release. Ioannis N. Melas, Alexander Mitsos, Thomas S. Weiss, Leonidas G Alexopoulos. 10th International Conference on Systems Biology 2010, Edinburgh, Scotland.

15. A systems biology approach to modeling signaling pathways in cartilage degeneration. Ioannis N. Melas, Aikaterini D. Chairakaki, Zoe Dailiana, Panagoula Kolia, Leonidas G Alexopoulos. Gordon Research Conference on Musculoskeletal Biology & Bioengineering 2010, Andover NH, USA

16. Identifying Drug effects on HepG2 cells via pathway alterations using an Integer Linear Programming (ILP) formulation. Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Julio Saez-Rodriguez, Dimitris E. Messinis, Danai Kirli-Florou, Leonidas G. Alexopoulos. European Association for the study of the Liver monothematic conference, signaling in the liver 2010, Amsterdam, Netherlands

17. Biomechanical and Systems Biology Approach for Modelling Cartilage Degeneration. I.N. Melas, A. Chairakaki, A. Mitsos, C.P. Provatidis, Z. Dailiana, P. Kolia, L.G. Alexopoulos. 4th Conference of the Helenic Society of Biomechanics 2010, Athens, Greece

18. Linking signaling pathways to cellular behavior using proteomic data. Ioannis N. Melas, Alexander Mitsos, Thomas S. Weiss, Leonidas G Alexopoulos. Planet xMap Congress 2010, Vienna, Austria

19. Crosstalk between EGF, HGF and Insulin Signaling in Hepatocytes: A logical modeling approach. Regina Samaga, Ioannis N. Melas, Leonidas G. Alexopoulos, Steffen Klamt. Conference on Systems Biology of Mammalian Cells, SBMC 2010, Freiburg, Germany

20. Systems biology approach and high-throughput proteomic analysis identifies Toll-Like-Receptor activators as major players of cartilage degeneration. Leonidas G Alexopoulos, Aikaterini D. Chairakaki, Ioannis N. Melas, Christopher P Provatidis, Panagoula Kolia, Zoe Dailiana. Gordon Research Conference on Musculoskeletal Biology & Bioengineering 2010, Andover NH, USA

21. Systems biology approach and high throughput proteomic analysis identifies Toll-Like-Receptor acivators as major players of cartilage degeneration. Leonidas G Alexopoulos, Aikaterini D. Chairakaki, Ioannis N. Melas, Christopher P. Provatidis, Panagoula Kolia, Zoe Dailiana. OARSI conference 2010, Brussels, Belgium.

22. Drug Effects via Pathway Alterations using Integer Linear Programming Optimization on Phosphoproteomic Data. A Mitsos, IN Melas, P Siminelakis, AD Chairakaki, J Saez-Rodriguez, LG Alexopoulos. 13th Annual International Conference on Research in Computational Molecular Biology RECOMB 2009, Boston, USA

23. Drug Effects Identification using an Integer Linear Programming Optimization Formulation Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Julio Saez- Rodriguez, Leonidas G. Alexopoulos. Conference of Hellenic Society of Biomechanics and Systems Biology 2009, Ioannina, Greece

24. Systems Biology for phosphoproteomic-based Drug Targeting, Efficacy, and Safety. Aikaterini D. Chairakaki, Ioannis N. Melas, Georgios Manikis, Paraskevas S Siminelakis, Steffen Klamt, Alexander Mitsos, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. International Greek Biotechnology Forum, 2009, Athens, Greece.

Δύο βραβεία έχουν απονεμηθεί σε εργασίες που εκπονήθηκαν στα πλαίσια αυτού του διδακτορικού. Αυτά είναι:

Βραβεία

1. "Best Poster First Prize" 9th Planet xMAP congress. 2011

2. "Best Practice award" Bio-IT world conference and expo. 2010

9 Ευχαριστίες

Ο συγγραφέας θέλει να ευχαριστήσει για την βοήθειά τους, τους ακόλουθους ανθρώπους. Απο το ΕΜΠ, σχολή Μηχανολόγων Μηχανικών, τον Καθ. Χριστόφορο Προβατίδη και τον Καθ. Ιωάννη Αντωνιάδη για τις συμβουλές τους και πολλές χρήσιμες συζητήσεις. Από το πανεπιστήμιο του Aachen τον Καθ. Αλέξανδρο Μητσό για την ενεργή βοήθειά του σε όλη την διάρκεια των διδακτορικών σπουδών του υποψηφίου και την πολύ επικοδομητική μας συνεργασία. Από το European Bioinformatics Institute τον Dr. Julio Saez-Rodriguez για την φιλοξενία του κατά την επίσκεψη του υποψήφιου στο γκρούπ του το καλοκαίρι του 2010 και την ενεργή καθοδήγησή του εκείνο το διάστημα. Από το Massachusetts Institute of Technology, department of Biological Engineering, τον Prof. Douglas A. Lauffenburger για την φιλοξενία του κατά την επίσκεψη του υποψήφιου στο γκρούπ του τα καλοκαίρια του 2011 και 2012 και την ενεργή καθοδήγησή του εκείνο το διάστημα. Από το Max Planck Institute, Department for complex technical systems, τον Dr. Steffen Klamt για την φιλοξενία του κατά την επίσκεψη του υποψήφιου στο γκρούπ του το φθηνόπωρο του 2012 και την ενεργή καθοδήγησή του εκείνο το διάστημα. Τον Καθ. Σωκράτη Τσαγγάρη, τον Καθ. Νικόλαο Χρόνη, τον Καθ. Αριστοτέλη Χατζηιωάννου και τον Καθ. Φραγκίσκο Κολίση, μέλη της επταμελούς επιτροπής του υποψηφίου για την εξέταση και επικοδομητική κριτική της διατριβής του. Ολα τα μέλη της ομάδας συστημικής βιολογίας και εμβιομηχανικής του Καθ. Λεωνίδα Αλεξόπουλου στο ΕΜΠ, σχολή Μηχανολόγων Μηχανικών, από το καλοκαίρι του 2008 έως το καλοκαίρι του 2013 για την πολύ επικοδομητική συνεργασία μας, και με ειδική αναφορά στην Αικατερίνη Δ. Χαιρακάκη (αυτή τη στιγμή Υποψήφια Διδάκτορα στο Division of Immunogenetics, Center of Immunology and Transplantation, Biomedical Research Foundation of the Academy of Athens), στον Δημήτρη Μεσσήνη, στην Δανάη Κυρλή-Φλώρου, στον Θεόδωρο Σακελλαρόπουλο, στον Ευάγγελο Ζυμπελούδη και στην Ελισάβετ Χατζοπούλου για την στενή μας συνεργασία. Και τέλος τον Καθ. Λεωνίδα Αλεξόπουλο για την καθοδήγησή του επί πέντε χρόνια και που έδωσε το ερέθισμα για την βαθύτερη κατανόηση των βιολογικών συστημάτων που κινητοποίησε τον υποψήφιο σε όλη τη διάρκεια των διδακτορικών του σπουδών.

Ioannis N. Melas

PhD Candidate,
National Technical University of Athens,
Department of Mechanical Engineering,
Systems Biology and Bioengineering Group.

Address: Heroon Polytechneiou 9
15780 Zografou, Athens, Greece
Tel: +30 6932 252 170
e-mail: giannis.melas@gmail.com
web: https://sites.google.com/site/giannismelas/

Expertise

Computational biology, network modeling, modeling of signal transduction pathways, optimization of network models to experimental data

Education

03/2009 – 9/2013 PhD in systems biology
Expected Research Topic: Modeling of signal transduction networks via regular optimization formulations
Department of Mechanical Engineering,
National Technical University of Athens, Greece
Advisors: Dr. Leonidas G. Alexopoulos, Dr. Nikolaos Chronis, Dr. Ioannis Antoniadis

2003 – 2008 Diploma, Mechanical Engineering Department.
National Technical University of Athens, Greece

Research Experience

9/2012 - 10/2012 *Visiting Student*
Max Planck Institute,
Department for Dynamics of Complex Technical Systems
Advisor: Steffen Klamt

8/2011 – 11/2011 *Visiting Student*
& 5/12 - 8/12 Massachusetts Institute of Technology (MIT),
Biological Engineering Department
Advisor: Douglas Lauffenburger

7/2010 – 10/2010 *Visiting Student*
European Bioinformatics Institute (EMBL-EBI)
Systems Biomedicine Group
Advisor: Julio Saez-Rodriguez

Peer reviewed publications

[1] Modeling of signaling pathways in chondrocytes based on phosphoproteomic and cytokine release data. Ioannis N. Melas*, Aikaterini D. Chairakaki*, Elisavet I. Chatzopoulou*, Dimitris E. Messinis, Alexander Mitsos, Zoe Dailiana, Panagoula Kollia, Leonidas G. Alexopoulos. Submitted, 2013. *Equal contributors

[2] Phosphoproteomics in drug discovery. Melody K Morris, An Chi, Ioannis N. Melas, Leonidas G Alexopoulos. Drug Discovery Today, Accepted, 2013

[3] Leveraging systems biology approaches in clinical pharmacology. Ioannis N. Melas, Kosmas Kretsos, Leonidas G Alexopoulos. Biopharm Drug Dispos. 2013 Aug 23. doi: 10.1002/bdd.1859

[4] Detecting and Removing Inconsistencies Between Experimental Data and Signaling Network Topologies using Integer Linear Programming on Interaction Graphs. Ioannis N. Melas*, Regina Samaga*, Leonidas G Alexopoulos, and Steffen Klamt. PLoS Comput Biol. 2013;9:e1003204. doi:10.1371/journal.pcbi.1003204, *Equal contributors

[5] Identification of signaling pathways related to drug efficacy in HCC via integration of phosphoproteomic, genomic and clinical data. Ioannis N. Melas, Douglas A. Lauffenburger, Leonidas G. Alexopoulos. Accepted, 13th IEEE International Conference on BioInformatics and BioEngineering. 2013

[6] Construction of cell type-specific logic models of signaling networks using CellNetOptimizer. M. K. Morris, I. Melas, J. Saez-Rodriguez. Methods in Molecular Biology:Computational Toxicology, Ed. B. Reisfeld and A. Mayeno, Humana Press. 2012

[7] Non Linear Programming (NLP) formulation for quantitative modeling of protein signal transduction pathways. Alexander Mitsos*, Ioannis N. Melas*, Melody K. Morris, Julio Saez-Rodriguez, Douglas A. Lauffenburger, Leonidas G. Alexopoulos. PLoS One. 2012;7(11):e50085. doi: 10.1371/journal.pone.0050085. Epub 2012 Nov 30 2012, *Equal contributors

[8] Construction of large signaling pathways using an adaptive perturbation approach with phosphoproteomic data. Ioannis N. Melas*, Alexander Mitsos*, Dimitris E. Messinis, Thomas S. Weiss, Julio Saez-Rodriguez, and Leonidas G. Alexopoulos, Mol. BioSyst., 10.1039/C2MB05482E, 2012. *Equal Contributors.

[9] Combined logical and data-driven models for linking signaling pathways to cellular response. I. Melas*, A. Mitsos*, D. Messinis, T. Weiss, L. Alexopoulos, Bmc Syst Biol 5, 107 (2011).(*equal contributors)

[10] Modeling signaling pathways in articular cartilage. Ioannis N. Melas, Aikaterini D.

Chairakaki, Alexander Mitsos, Zoe Dailiana, Christopher G. Provatidis, Leonidas G. Alexopoulos 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2011), Boston, MA, US

[11] Construction of Pathways and identification of drug effects in liver cancer cells via an Integer Linear Programming (ILP) formulation. Leonidas G Alexopoulos, Ioannis N. Melas, Aikaterini D. Chairakaki, Julio Saez-Rodriguez, Alexander Mitsos. 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2010, Buenos Aires, Argentina

[12] Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data. Alexander Mitsos, Ioannis N. Melas, Paraskevas S Siminelakis, Aikaterini D. Chairakaki, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. PloS Comp Biol. 2009. 5(12): e1000591. doi:10.1371/journal.pcbi.1000591

Refereed abstracts/conferences – partial list

[1] Network analysis of signaling in chondrocytes Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Zoe H. Dailiana, Christopher Provatidis, Leonidas G. Alexopoulos. European Society of Biomechanics 2013.

[2] Real time evaluation of the mechanical properties of articular cartilage during collagenase induced digestion. Evangelos Zympeloudis, Elisavet I. Chatzopoulou, Ioannis N. Melas, Zoe H. Dailiana, Christopher Provatidis, Leonidas G. Alexopoulos. European Society of Biomechanics 2013.

[3] A device for simultaneous compression measurements of human cartilage. Nikolaos V. Georgiou , Nikolaos D. Nikolaou, Panagiotis D. Alevras, Elisavet I. Chatzopoulou, Ioannis N. Melas, Christopher G. Provatidis, Leonidas G. Alexopoulos European Society of Biomechanics 2013.

[4] A comparison of chondrocyte response in cartilage 2d cultures. Elisavet I. Chatzopoulou, Ioannis N. Melas, Christoforos G. Provatidis , Zoe Dailiana, Leonidas G. Alexopoulos European Society of Biomechanics 2013.

[5] Modeling signaling pathways in chondrocytes. Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Zoe Dailiana, Christopher Provatidis, Leonidas G. Alexopoulos. Fifth International Conference on Computational Bioengineering in Leuven, Belgium. 2013. Submitted

[6] A Non Linear Programming (NLP) formulation for modeling signaling transduction pathways, Ioannis N. Melas*, Alexander Mitsos*, Melody K. Morris, Julio Saez-Rodriguez, Douglas A. Lauffenburger, Leonidas G. Alexopoulos (*eq contributors). 4th Hellenic Conference on Biomedical Technology 2012, Athens, Greece.

[7] Monitoring cartilage degeneration in a high throughput format via its mechanical properties, Nikolaos Nikolaou, Panagiotis Alevras, Elisavet Chatzopoulou, Ioannis Melas, Zoe Dailiana, Christoforos Provatidis, Leonidas Alexopoulos. 4th Hellenic Conference on

Biomedical Technology 2012, Athens, Greece.

[8] Optimization of a large scale signaling network using an Integer Linear Programming formulation. Ioannis N. Melas*, Alexander Mitsos*, Julio Saez-Rodriguez, Leonidas G Alexopoulos, *Equal Contributors. 7th GRACM International Congress on Computational Mechanics 2011, Athens, Greece,

[9] A device for multiple indentation tests of human cartilage. Nikolaos D. Nikolaou, Panagiotis D. Alevras, Ioannis N. Melas, Christopher P. Provatidis, Leonidas G Alexopoulos. 7th GRACM International Congress on Computational Mechanics 2011, Athens, Greece

[10] Construction of large signaling pathways from phosphoproteomic data using an ILP computational approach, Dimitris E. Messinis, Ioannis N. Melas, Alexander Mitsos, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. 6th conference of the Hellenic Society for Computational Biology and Bioinformatics, 2011, Patra, Greece.

[11] Construction of large signaling pathways from phosphoproteomic data, Dimitris E. Messinis, Ioannis N. Melas, Alexander Mitsos, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. Planet xMap Congress 2011, Vienna, Austria.

[12] Construction of large signaling pathways from phosphoproteomic data, Leonidas G. Alexopoulos, Ioannis N. Melas, Julio Saez-Rodriguez, Thomas S. Weiss, Dimitris E. Messinis, Alexander Mitsos. ICSB, Heidelberg and Mannheim, 2011.

[13] Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data. Leonidas G. Alexopoulos, Ioannis N. Melas, Aikaterini D. Chairakaki, Julio Saez-Rodriguez, Alexander Mitsos., BIO-IT world conference 2010, Boston, USA.

[14] Extending logical models of signaling pathways to predict cytokine release. Ioannis N. Melas, Alexander Mitsos, Thomas S. Weiss, Leonidas G Alexopoulos. 10th International Conference on Systems Biology 2010, Edinburgh, Scotland.

[15] A systems biology approach to modeling signaling pathways in cartilage degeneration. Ioannis N. Melas, Aikaterini D. Chairakaki, Zoe Dailiana, Panagoula Kolia, Leonidas G Alexopoulos. Gordon Research Conference on Musculoskeletal Biology & Bioengineering 2010, Andover NH, USA

[16] Identifying Drug effects on HepG2 cells via pathway alterations using an Integer Linear Programming (ILP) formulation. Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Julio Saez-Rodriguez, Dimitris E. Messinis, Danai Kirli-Florou, Leonidas G. Alexopoulos. European Association for the study of the Liver monothematic conference, signaling in the liver 2010, Amsterdam, Netherlands

[17] Biomechanical and Systems Biology Approach for Modelling Cartilage Degeneration. I.N. Melas, A. Chairakaki, A. Mitsos, C.P. Provatidis, Z. Dailiana, P. Kolia, L.G. Alexopoulos. 4th Conference of the Helenic Society of Biomechanics 2010, Athens, Greece

[18] Linking signaling pathways to cellular behavior using proteomic data. Ioannis N. Melas, Alexander Mitsos, Thomas S. Weiss, Leonidas G Alexopoulos. Planet xMap Congress 2010, Vienna, Austria

[19] Crosstalk between EGF, HGF and Insulin Signaling in Hepatocytes: A logical modeling approach. Regina Samaga, Ioannis N. Melas, Leonidas G. Alexopoulos, Steffen Klamt. Conference on Systems Biology of Mammalian Cells, SBMC 2010, Freiburg, Germany

[20] Systems biology approach and high-throughput proteomic analysis identifies Toll-Like-Receptor activators as major players of cartilage degeneration. Leonidas G Alexopoulos, Aikaterini D. Chairakaki, Ioannis N. Melas, Christopher P Provatidis, Panagoula Kolia, Zoe Dailiana. Gordon Research Conference on Musculoskeletal Biology & Bioengineering 2010, Andover NH, USA

[21] Systems biology approach and high throughput proteomic analysis identifies Toll-Like-Receptor acivators as major players of cartilage degeneration. Leonidas G Alexopoulos, Aikaterini D. Chairakaki, Ioannis N. Melas, Christopher P. Provatidis, Panagoula Kolia, Zoe Dailiana. OARSI conference 2010, Brussels, Belgium.

[22] Drug Effects via Pathway Alterations using Integer Linear Programming Optimization on Phosphoproteomic Data. A Mitsos, IN Melas, P Siminelakis, AD Chairakaki, J Saez- Rodriguez, LG Alexopoulos. 13th Annual International Conference on Research in Computational Molecular Biology RECOMB 2009, Boston, USA

[23] Drug Effects Identification using an Integer Linear Programming Optimization Formulation Ioannis N. Melas, Aikaterini D. Chairakaki, Alexander Mitsos, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. Conference of Hellenic Society of Biomechanics and Systems Biology 2009, Ioannina, Greece

[24] Systems Biology for phosphoproteomic-based Drug Targeting, Efficacy, and Safety. Aikaterini D. Chairakaki, Ioannis N. Melas, Georgios Manikis, Paraskevas S Siminelakis, Steffen Klamt, Alexander Mitsos, Julio Saez-Rodriguez, Leonidas G. Alexopoulos. International Greek Biotechnology Forum, 2009, Athens, Greece.

Honors/Awards

2011	"Best Poster First Prize" 9th Planet xMAP congress.
2010	"Best Practice award" Bio-IT world conference and expo

IT Skills

User:	Experienced in Windows, Linux, Mac operating systems
Developer:	Experienced in programming with Matlab, C, Fortran, Bash Scripting Perl, Python, R

On Systems Biology

Developer: CellNetOptimizer, SigNetTrainer (toolbox for detecting and removing inconsistencies between experimental data and signaling network topologies. Available at: http://www.mpi-magdeburg.mpg.de/projects/cna/etcdownloads.html