

Analyse intégrative et modélisation des voies moléculaires dérégulées dans la polyarthrite rhumatoïde

Integrative analysis and modeling of molecular pathways dysregulated in rheumatoid arthritis

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, Structure et Dynamique des Systèmes Vivants
(SDSV)

Spécialité de doctorat : Biologie computationnelle

Unité de recherche : Laboratoire Européen de Recherche pour la Polyarthrite
rhumatoïde, Genhotel, Evry

Référent : Université d'Évry Val d'Essonne

Thèse présentée et soutenue à Évry,
le 03/12/2020, par

Vidisha SINGH

Composition du Jury

Bruno COLOMBO

Professor, UEVE, Paris Saclay

Président

Vassili SOUMELIS

Professor, IRSL

Rapporteur

Dagmar WALTEMATH

Professor, MIL, Greifswald, Germany

Rapporteur

Denis THIEFFRY

Professor, IBENS

Examineur

Laurence CALZONE

Research engineer, Institut curie

Examinatrice

Mohamed ELATI

Professor, Université de Lille

Examineur

Direction de la thèse

Elisabeth PETIT-TEIXEIRA

Professor, UEVE, Paris Saclay

Directrice de thèse

Anna NIARAKIS

Associate Professor, UEVE, Paris Saclay

Co-directrice de thèse

Acknowledgements

First of all, I would like to thank Dr. Dagmar Waltemath and Dr. Vassili Soumelis for accepting to be the reviewers for my PhD manuscript, also Dr. Denis Thieffry, Dr. Laurence Calzone, Dr. Bruno Colombo and Dr. Mohamed Elati for agreeing to be a part of my jury as examiners.

I would like to express my deepest gratitude to Dr. Anna Niarakis for being an extraordinary supervisor, for her invaluable guidance, support, suggestions and endless encouragement. She always gave me the freedom to pursue my own interests and provided me with insightful suggestions and support in developing independent thinking and research skills. She has been an exceptional mentor and I appreciate both our professional and personal conversations over the years. The knowledge and wisdom I have gained from her will forever guide me in education and in life.

I would also like to thank Dr. Elisabeth Petit Teixeira, my thesis and laboratory director, for giving me an opportunity to work in her lab and providing all the essential facilities for carrying out the work. I thank her for being extremely kind and supportive, for providing me constructive feedbacks in my work and for being always there for me.

I am beyond thrilled to have known and working with both my mentors. I greatly admire them for being such a positive source of dynamism and their constant motivation.

I would also like to thank Dr. Valerie Chaudru for being very kind and helpful. She was always there for me to help and discuss whenever I needed her. A big thanks to her.

I would like to thank my thesis committee members, Dr. Laurence Calzone, Dr. Bruno Colombo and Dr. Mohamed Elati for all their feedbacks and help over the years that helped significantly in the progression of my PhD.

A special thanks to Dr. George Kalliolias for all his time and discussions for the pathways. I really enjoyed working with him and thank him for all the fruitful discussion sessions. I would also like to thank Dr. Sylvain Soliman for all his help and support. I worked with him for the development of the tool CaSQ and he had been incredibly patient and helpful with all the different problems and issues we had. I appreciate all his hard work and greatly enjoyed our meeting together which were always followed by the pizza treat.

I would like to thank Dr. Marek Ostaszewski for his prompt responses, discussions and help for MINERVA, Dr. Alexander Mazein for all his suggestions for the SBGN standards and Dr.

Tomas Helikar for his help with Cell Collective. I would also like to thank Dr. Aurelien Naldi for his help regarding GINsim and BioLQM.

I would like to thank Dr. Maeva Veyssiere, one of my closest friend for all her help and talks. She had been always there for me during the three years of my PhD and continued to do so for which I will be forever grateful. I greatly enjoyed working with her and discussing both professional and personal fronts with her. I would also like to thank Melissa Saichi and Nawel Zerrouk, also my closest friends for their friendship, emotional support and motivations. I would also like to thank Quentin Miagoux for being an incredible friend and colleague. I greatly enjoy our talks and discussion and I thank him for his kindness, joyous personality and all the helps. It is hard to think my life in France without all of your support and love. I would like to thank Sara Aghamiri for her friendship, all great talks and her help in my project. I would like to thank Dereck de Mezquita for all the interesting discussions and being a great colleague and friend. I would also like to thank Dr. Smahane Chalabi, my colleague for her kindness and suggestions. I would like to thank Saran Pankaew for his preliminary work in the project. I thank Dr. Javier Perea, for all the talks and discussions. Also, thank you Sahar Aghakhani who joined just now for her PhD for being such a nice colleague.

I would like to thank my mother Kusum Singh and father Mohan Singh for their unconditional love and encouragement throughout my life. I am here because of them and I could not thank them enough for everything they did for me. Thank you my dear Pramod for being in my life, for being an incredible person, for your unconditional love and support, for always guiding and helping me. Thank you again for your consistent motivation and kindness. I would like to thank my other family members Skand Singh, Snigdha Singh and Sakshi Singh for being my strength. I would like to thank Dr. Amir Ali Feiz and his loving family for being great friends and their unending love and kindness. Thank you Ajay Sharma and Palak Rawat for your friendship and love.

I would also like to express my gratitude to Genopole and Roxane Brachet for selecting me for the Fondagen scholarship twice to attend prestigious workshops at WGC, Cambridge and EMBL, Germany. These workshops helped greatly in the progression of my PhD work and also provided me opportunity to interact with the experts in my domain. I am also thankful to Open health Institute for providing me scholarship to support publishing the RA map paper. Also thanks to disease maps and COVID-19 community from where I learnt the challenges and the important issues of the field.

Lastly, I thank you Genhotel lab again for providing me the support, acceptance and love that greatly helped me to grow both as a person and as a researcher. You are my second family and I will forever cherish the precious moments spent in this lab. I love you all.

Résumé en français

Les maladies complexes ou multifactorielles sont définies comme des maladies dont l'étiologie et la progression sont déterminées par un certain nombre de facteurs génétiques, environnementaux et épigénétiques. On peut citer à cet égard les maladies auto-immunes comme la polyarthrite rhumatoïde (PR), les maladies neurologiques comme la maladie d'Alzheimer, la maladie de Parkinson et la sclérose en plaques, les maladies respiratoires comme l'asthme, les maladies du tissu conjonctif comme la sclérodermie, les maladies rénales, et bien d'autres encore. Bien que les chercheurs dans le domaine de la biomédecine aient fait des progrès significatifs contre diverses autres maladies mortelles, ils sont encore confrontés à des défis pour comprendre les mécanismes fondamentaux sous-jacents à l'initiation et à la progression de maladies complexes. La raison en est que la méthodologie de la biologie expérimentale classique est principalement basée sur l'étude de gènes et de protéines individuels, en considérant l'organisme comme simple et linéaire. Cependant, les maladies complexes impliquent de nombreuses interactions complexes entre les différentes entités biologiques, ce qui entraîne certaines conséquences pathologiques critiques. Dans ce projet, nous avons essayé de comprendre le mécanisme sous-jacent à l'initiation d'une maladie complexe courante, la polyarthrite rhumatoïde (PR). La PR touche environ 0.5-1.0% de la population adulte dans la plupart des pays développés. La prévalence mondiale de la PR serait un peu plus faible (environ 0.24%), avec une prévalence trois fois plus élevée chez les femmes que chez les hommes et environ 0.5 à 1% chez les individus blancs selon une étude basée sur les pays occidentaux.

La PR est une maladie auto-immune inflammatoire chronique dont la pathogenèse implique une interaction complexe entre des facteurs génétiques, épigénétiques et environnementaux, entraînant une cascade de réactions immunitaires. La PR affecte principalement le cartilage et les articulations synoviales des mains, des poignets et des pieds, des épaules, des coudes, des genoux et des chevilles. La PR est caractérisée par une inflammation et une hyperplasie synoviale, la production d'auto-anticorps (facteur rhumatoïde [FR] et anticorps anti-peptides citrullinés [ACPA]), la destruction du cartilage et des os, et des caractéristiques systémiques, notamment des troubles cardiovasculaires, pulmonaires, psychologiques et squelettiques. La complexité des multiples voies de signalisation qui interagissent les unes avec les autres, entraînant l'expression de diverses molécules pro-inflammatoires comme les cytokines, chimiokines et métalloprotéinases matricielles (MMP), rend l'étude de l'étiologie de la maladie quelque peu déroutante. Les processus du système immunitaire inné

et adaptatif jouent un rôle majeur dans la propagation des signaux qui entraînent l'activation de multiples cellules immunitaires comme les macrophages, les neutrophiles, les cellules T et les cellules B, ainsi que l'activation de cellules articulaires comme les fibroblastes synoviaux. L'infiltration et l'accumulation de cellules inflammatoires comme les macrophages, les lymphocytes (lymphocytes T, lymphocytes B) et les neutrophiles entraînent une hyperplasie de la membrane synoviale et, par la suite, la formation de pannus. En plus des diverses cellules immunitaires étudiées pour l'intervention thérapeutique dans la maladie, les fibroblastes synoviaux (FS) sont apparus comme une cible thérapeutique supplémentaire. De nombreuses études se concentrent sur les FS pour mieux comprendre ses interactions complexes avec les cellules immunitaires et comment la multiplication des FS peut éventuellement induire la destruction du cartilage et l'érosion osseuse dans les articulations. En général, dans les tissus sains, les fibroblastes synoviaux fournissent à la cavité articulaire et au cartilage adjacent des protéines plasmatiques et des molécules lubrifiantes telles que l'acide hyaluronique. Les FS produisent également des protéines de matrice telles que le collagène et l'hyaluronane, une variété d'enzymes dégradant la matrice, comme les métalloprotéases de matrice (MMPS) qui contribuent au remodelage de la matrice.

Le diagnostic de la PR est principalement basé sur les manifestations cliniques et les biomarqueurs sériques tels que le FR et les ACPA, ce dernier étant plus spécifique de la maladie. La positivité des ACPA est maintenant largement utilisée pour diagnostiquer la PR dans la pratique clinique en raison de sa grande spécificité (>97%). Un examen médical typique comprend une recherche de gonflement et de sensibilité avec atrophie des muscles à proximité des articulations concernées. Les niveaux de protéine C réactive et la vitesse de sédimentation des érythrocytes sont souvent augmentés en cas de PR active, et ces réactifs en phase aiguë font partie des nouveaux critères de classification de la PR. L'héritabilité de la PR est d'environ 60%, avec des facteurs de risque génétique tels que certains allèles du gène HLA-DRB1 (HLA-DRB1*01 et HLA-DRB1*04 étant les plus significativement associés) .., des variants de gènes non-HLA, de type SNP (Single Nucleotide Polymorphism) ont également été associés à la PR, tels que PTPN22, IL23R, TRAF1, CTLA4, IRF5, STAT4, CCR6 et PADI4.

La PR est généralement associée à des preuves sérologiques d'auto-immunité systémique, comme l'indique la présence d'auto-anticorps dans le sérum et le liquide synovial. Le premier auto-anticorps dans la PR, le RF, a été décrit par Waaler en 1940 et est dirigé contre la région Fc des IgG. Les auto-antigènes, qui comprennent un large éventail de composants du cartilage, de protéines de stress, d'enzymes, de protéines nucléaires et de protéines citrullinées, ciblés par

un certain nombre d'auto-anticorps, ont été caractérisés dans la PR. Cela démontre que la PR est caractérisée par des auto-réactivités accumulées dans les cellules B et T.

La pathogenèse de la progression de la PR est médiée par des molécules qui activent différentes voies cellulaires. Ces médiateurs interagissent avec les récepteurs membranaires et déclenchent des cascades de signalisation. Après l'interaction ligand-récepteur, le stimulus en amont se propage par une série de réactions couplées de la membrane plasmique au cytoplasme, pour réguler les 10 facteurs clés qui sont responsables des différents phénotypes cellulaires. Pour la PR, les principaux stimuli en amont sont les suivants : cytokines et chimiokines, facteurs de croissance et TLR. Ils conduisent à l'activation des voies en aval qui comprennent : la voie JAK-STAT, la voie MAPK et la voie PI3K-AKT.

Les biomolécules des systèmes vivants ne fonctionnent pas individuellement mais interagissent entre elles en formant des réseaux biomoléculaires. Par exemple, une maladie est rarement la conséquence d'une anomalie dans un seul gène mais reflète les perturbations ou les dysfonctionnements des réseaux biologiques complexes qui relient les tissus et les systèmes d'organes. Afin d'organiser et d'analyser cette complexité, il est urgent de développer de nouvelles méthodologies permettant d'obtenir des informations grâce à une vision plus globale.

La biologie systémique est l'étude au niveau des systèmes, qui permet de déchiffrer la complexité de la maladie, son principe fondamental d'initiation et de progression. Elle est considérée comme une approche analytique puissante, qui considère un organisme vivant comme un réseau interactif et dynamique de protéines, de gènes et de réactions biochimiques. L'une des approches initiales communes à cet égard est le domaine de la biologie en réseau qui comprend la construction de réseaux biologiques en exploitant les données expérimentales en constante expansion. La disponibilité de grands ensembles de données en réseau et la capacité de calcul abordable ont conduit au développement d'algorithmes bioinformatiques et d'approches de biologie computationnelle pour analyser les données afin d'offrir des aperçus des dynamiques biologiques. L'étude des systèmes biologiques du point de vue des réseaux a récemment attiré beaucoup d'attention sous la forme de cartes d'interactions présentant les liens entre les protéines (gènes ou neurones), les phénotypes qui leur sont associés (maladies) et les facteurs environnementaux correspondants (médicaments). La modélisation informatique peut être utilisée pour fournir un réseau exécutable et dynamique qui peut révéler des propriétés cachées et prendre en compte les comportements émergents au niveau des systèmes grâce à des simulations et des perturbations *in silico*. Un cycle de construction de modèles, de simulations

et de validation expérimentale des prédictions peut contribuer aux approches diagnostiques et thérapeutiques actuelles de la médecine.

La construction de modèles mathématiques implique de construire la structure du modèle, de choisir des expressions mathématiques pour caractériser les relations entre ses composantes, de trouver les valeurs des paramètres et les conditions initiales, et de réaliser des simulations numériques et d'autres analyses mathématiques qui peuvent à la fois reproduire les observations et conduire à des prédictions. De nombreuses méthodes de modélisation basées sur la représentation temporelle et la valeur variable sont en cours d'élaboration pour modéliser et simuler les réseaux moléculaires et génétiques et pourraient être classées de manière générale en deux catégories : les approches quantitatives et qualitatives. Les modèles quantitatifs sont basés sur l'application de la théorie des systèmes à la cinétique chimique et ont été utilisés pour décrire les réseaux métaboliques, les voies de signalisation et la régulation des gènes. La modélisation logique qualitative est basée sur l'idée qu'une variable peut prendre un nombre discret d'états ou de valeurs (deux dans le cas des modèles booléens) et que l'état d'une variable est décidé par une combinaison logique des états d'autres variables.

Bien que la modélisation cinétique quantitative fournisse des prédictions plus précises, le manque de paramètres cinétiques entrave son utilisation dans de nombreuses situations, ce qui est abordé par d'autres approches comme la modélisation logique. Dans un modèle logique, chaque composant est associé à une variable discrète, qui est une abstraction logique (souvent booléenne, c'est-à-dire binaire) de son niveau d'activité (ou de concentration). Ce cadre repose sur son évolutivité (des modèles logiques de quelques centaines de composants ont été simulés) ainsi que sur sa nature qualitative, car les paramètres cinétiques et autres connaissances précises sur les mécanismes moléculaires ne sont pas nécessaires. Un modèle booléen fournit une représentation qualitative d'un système. Chaque variable booléenne ne peut prendre que deux valeurs possibles désignées par 1 (ON) ou 0 (OFF) correspondant aux valeurs logiques FAUX et VRAI. ON et OFF représentent l'état d'une entité biologique correspondant à la variable binaire. Cela indique si un gène est exprimé ou non, si un facteur de transcription est actif ou inactif, ou si la concentration d'une molécule est supérieure ou inférieure à un certain seuil. Dans les modèles booléens, l'état futur d'un nœud est déterminé sur la base d'une déclaration logique sur l'état actuel de ses régulateurs. Cette déclaration, appelée règle booléenne (fonction), est généralement exprimée par les opérateurs logiques *ET*, *OU* et *NON*. Les modèles booléens de réseau peuvent être projetés comme un graphe dirigé où les nœuds A, B et C correspondent aux variables booléennes reliées par des arêtes.

Pour comprendre les détails mécaniques des processus biologiques sous-jacents de la maladie complexe qu'est la PR, nous devons systématiquement rassembler et analyser les informations disponibles dans la littérature, les ensembles de données, et les bases de données. L'objectif de ce travail de doctorat est de faire la lumière sur les mécanismes moléculaires impliqués dans la pathogenèse de la PR en utilisant des approches, des outils et des techniques de pointe en biologie systémique. Le premier objectif du travail est de résumer les connaissances biologiques actuelles (expression des gènes, voies de signalisation, phénotypes cellulaires) concernant la PR en une carte détaillée d'interactions moléculaires. La construction d'une carte de la PR à la pointe de la technologie nécessite une analyse exhaustive de la littérature, une réévaluation des tentatives publiées précédemment, ainsi que des critères de conservation stricts et, surtout, des conseils d'experts afin de limiter les faux positifs en se concentrant sur la maladie et la spécificité humaine. Les cartes d'interactions moléculaires peuvent servir de base de connaissances autonomes, ou elles peuvent être utilisées comme un échafaudage pour la construction de modèles informatiques. Un aspect important est également l'utilisation de normes communautaires telles que la notation SBGN (systems biology graphical notation), qui est une représentation graphique standard destinée à favoriser le stockage, l'échange et la réutilisation efficaces des informations sur les voies de signalisation, les réseaux métaboliques et les réseaux de régulation des gènes parmi les communautés de biochimistes, de biologistes et de théoriciens. Ces cartes résument alors les connaissances actuelles sur les voies biologiques dans une représentation de description de processus, tout en tenant compte du plus grand nombre possible de détails mécanistiques. C'est la première tentative de construction d'une carte des interactions moléculaires pour la PR en utilisant presque exclusivement des données spécifiques à l'homme et à la maladie, et en utilisant la norme SBGN. Le deuxième objectif du projet est de développer une approche de modélisation informatique pour fournir un réseau dynamique exécutable qui peut révéler des propriétés cachées et prendre en compte les comportements émergents au niveau du système par le biais de simulations et de perturbations *in silico*. Parmi les différents types de cellules impliquées dans la PR, les FS jouent un rôle crucial dans la persistance des caractéristiques destructrices de la maladie. Il est démontré qu'ils expriment des cytokines immunomodulatrices, diverses molécules d'adhésion et des enzymes de modélisation de la matrice. En outre, les fibroblastes de la PR présentent des taux de prolifération élevés et un phénotype résistant à l'apoptose. Ces cellules peuvent également se comporter comme des moteurs primaires de l'inflammation, et les thérapies dirigées contre les FS dans la PR pourraient devenir une approche complémentaire aux thérapies ciblant le système immunitaire.

Notre objectif est de construire un modèle à grande échelle pour étudier le comportement des FS de la PR afin de mieux comprendre les mécanismes qui contrôlent leur phénotype agressif. Il s'agit, à notre connaissance, de la première tentative de construction d'un modèle dynamique à grande échelle pour ce type d'étude. Le troisième objectif du projet est un effort pour automatiser le passage d'une représentation statique des mécanismes de la maladie à un modèle dynamique développant un cadre pour la conversion des graphes et l'inférence de formules logiques basées sur la topologie et la sémantique encodées sur la carte moléculaire. La construction d'un modèle dynamique exécutable, soit à partir de cartes/modèles moléculaires préexistants, soit de manière autonome, est un processus qui prend du temps. Le cadre proposé pourrait considérablement faciliter le processus de génération de modèles exécutables prêts à l'emploi pour les simulations *in silico* (perturbations/déclenchements). Notre modèle est le premier modèle booléen à grande échelle à être construit de manière entièrement automatisée à partir d'une carte d'interaction moléculaire. En outre, cela ouvre également la voie pour la mise à l'échelle de la génèse automatique de modèles de grande échelle. Jusqu'à présent, le nombre de nœuds pouvant être inclus dans un modèle biologique dynamique était limité. Cela était conforme aux outils de modélisation disponibles, dont le potentiel de simulation était également limité. Dans cette thèse, nous voudrions remettre en question ce potentiel et augmenter l'échelle des modèles couramment utilisés pour étudier les réseaux biologiques, afin d'élucider les mécanismes de régulation complexes et de donner des réponses à des questions plus sophistiquées qui ne pourraient pas être traitées avec des modèles comprenant un nombre limité de facteurs.

Dans ce sens, une carte complète des interactions moléculaires pour la PR, construite avec le logiciel CellDesigner, a été publiée en 2010. Elle était basée sur des données expérimentales à haut débit combinées avec des informations provenant de la base de données KEGG (<http://www.genome.jp/kegg/pathway.html>). Vingt-huit études publiées ont été utilisées pour la construction de la carte qui comprenait des expériences réalisées dans différents types de cellules et de tissus tels que les cellules mononucléaires du sang périphérique (PBMC), les fibroblastes synoviaux, le tissu synovial et les macrophages, entre autres. Nous avons utilisé cette carte comme base et l'avons étendue pour créer une carte de la PR basée sur les connaissances les plus récentes. Nous l'avons transformé en base de connaissances interactive de pointe sur la maladie, qui s'interface avec diverses bases de données pour l'annotation du contenu et l'analyse d'enrichissement des résultats expérimentaux. Nous avons également utilisé des outils bioinformatiques tels que BioInfoMiner (<https://bioinforminer.com>) et

Cytoscape (<https://cytoscape.org/>) pour l'analyse de la carte de la PR comme un réseau biologique complexe, révélant ainsi les aspects topologiques et fonctionnels de la carte.

La carte de la PR illustre graphiquement les voies de signalisation, la régulation de l'expression des gènes, les mécanismes moléculaires et les phénotypes cellulaires impliqués dans la pathogenèse de la maladie. Cette carte implique une recherche documentaire exhaustive, l'extraction d'informations des bases de données pertinentes, ainsi qu'une mise à jour continue et des conseils d'experts du domaine. Il est important de noter que les interactions illustrées dans le diagramme représentent un modèle graphique codé selon un format standardisé, ce qui rend la carte facile à calculer. Tous les composants de la carte ont au moins deux références PubMed ajoutées manuellement, ce qui donne un total de 353 publications couvrant une période allant de 1975 à 2019.

Nous présentons donc ici une carte des interactions moléculaires à grande échelle pour la PR, qui est à notre connaissance la première carte des maladies normalisée conforme au SBGN. Tous les composants et réactions sont annotés en utilisant uniquement la PR et les études spécifiques à l'homme. Afin de limiter les faux positifs, les conseils d'experts sont intégrés et les normes SBGN ont été utilisées pour la représentation de la carte afin d'assurer sa réutilisation. La carte de la PR que nous présentons ici comprend 506 espèces, 446 réactions et 8 phénotypes. Les espèces de la carte sont classées en 303 protéines, 61 complexes, 106 gènes, 106 entités ARN, 2 ions et 7 molécules simples. La carte de la PR peut également être utilisée comme une base de connaissances interactive, en utilisant la plateforme MINERVA, et servir de modèle pour la superposition de plusieurs ensembles de données. La visualisation des données expérimentales peut en effet aider à mettre en évidence certains aspects du processus biologique affecté et aider à rendre plus évidentes les différences entre les conditions expérimentales. La visualisation des résultats de l'analyse de l'expression différentielle de trois ensembles de données d'expression de gènes dans le tissu synovial de la PR a montré un enrichissement pour tous les phénotypes cellulaires sauf l'apoptose. Cette constatation est conforme au fait que les fibroblastes, qui constituent un pourcentage important des synoviocytes de la PR, présentent un phénotype résistant à l'apoptose.

Nous avons effectué une analyse fonctionnelle et une priorisation des gènes à l'aide du logiciel BioInfoMiner. Les gènes qui sont les mieux classés dans cette analyse sont associés à de nombreux processus systémiques et sont considérés comme des nœuds dans le réseau sémantique. Suite à cette priorisation, une analyse pharmacogénomique est fournie, puisque les

hubs proposés sont considérés comme des cibles médicamenteuses putatives. Les résultats de l'analyse utilisant les termes GO et PHO ont révélé des acteurs connus de la PR, dont la plupart ont déjà été utilisés comme cibles de médicaments, ce qui démontre que la carte de la PR comprend des facteurs bien caractérisés et saisit la plupart des processus systémiques pertinents impliqués dans la maladie. L'analyse topologique peut révéler des caractéristiques structurelles sous-jacentes de la carte de la PR comme des parties non connectées du réseau ou des nœuds importants (nœuds bien connectés) qui sont autrement difficiles à percevoir dans les réseaux à grande échelle. L'analyse topologique réalisée dans le cadre de cette étude a révélé des parties connectées et non connectées du réseau. Ce résultat reflète notre connaissance fragmentée d'une part, mais aussi l'utilisation de critères stricts pour les nœuds inclus dans la carte : interactions validées expérimentalement dans au moins deux études publiées, utilisation de données d'origine strictement humaine et spécifiques à la maladie. La carte de la PR, fruit de collaborations interdisciplinaires entre cliniciens, biologistes et bio-informaticiens, est accessible sur ramap.elixir-luxembourg.org.

La construction de modèles dynamiques à grande échelle peut être un travail long et fastidieux qui nécessite non seulement la construction du graphique réglementaire mais aussi l'écriture et le réglage des formules logiques. Nous avons développé un outil, CaSQ, visant à faciliter la construction de modèles booléens à grande échelle, en tirant parti des similitudes partagées entre les cartes d'interaction moléculaire et les modèles dynamiques. Dans le cadre proposé, nous utilisons des normes de biologie systémique pour la construction de modèles (SBML-qual), de sorte que l'outil CaSQ puisse être interopérable avec d'autres outils et logiciels de modélisation. Pour l'inférence des formules logiques, nous avons basé nos hypothèses sur la topologie et la sémantique des cartes moléculaires. Plus précisément, nous avons décidé d'aborder le processus de conversion en utilisant principalement des fonctions logiques *OU* et *ET*, de sorte qu'une cible est activée si l'une des réactions qui la produit est activée, et une réaction est activée si tous les réactifs sont activés, tous les inhibiteurs sont désactivés et l'un des catalyseurs est activé. L'idée qui sous-tend cette hypothèse est que nous disposons très rarement d'informations exactes sur la nécessité de la présence de deux ou plusieurs activateurs pour une cible. Même si une synergie est définie, très souvent une activation relative peut se produire même par la présence d'un seul activateur. En outre, le nombre d'événements pour lesquels nous disposons de telles informations est nettement inférieur à celui des événements incertains et le réglage des règles à la main devrait être un processus rapide. Grâce à l'utilisation de CaSQ, comme le montre cette étude, nous pouvons maintenant obtenir des modèles booléens

à grande échelle qui peuvent être exécutés à l'aide de logiciels de modélisation populaires capables d'importer des fichiers de qualité SBML. Dans ce travail, pour comparer les performances et la précision de l'outil, nous avons comparé les nœuds communs entre les modèles inférés par CaSQ et les modèles construits manuellement, leur capacité à reproduire des scénarios biologiques en effectuant des simulations, et enfin, nous avons effectué une comparaison des états stables, lorsque cela était possible. Nous avons également effectué des simulations pour voir si les modèles inférés par CaSQ pouvaient reproduire une partie de la dynamique du système original. L'étape suivante a consisté à effectuer une analyse logique de l'état stable.

Pour cela, nous avons utilisé GINsim, un logiciel puissant de modélisation logique. Le but était de voir si, dans les états stables du modèle interférentiel CaSQ, nous pouvions récupérer les états stables du modèle construit manuellement et publié. Il faut noter que CaSQ déduit des règles booléennes préliminaires, de sorte que le modélisateur doit encore affiner le modèle et trouver les meilleures règles logiques pour reproduire les données avec précision. Bekkar et al, 2018 montrent que les modèles logiques avec une curation humaine ajoutée ont de meilleures performances que les modèles où les règles sont extraites automatiquement d'une topologie donnée. Comme le montrent les résultats, l'outil CaSQ produit des modèles qui sont largement en accord avec le modèle qu'un modélisateur humain construirait, ce qui accélère de façon impressionnante le temps de construction du modèle. Ce travail a également motivé le travail communautaire, car il a abordé les questions de réutilisation des modèles, d'utilisation des formats standard de la biologie systémique et d'interopérabilité entre différents outils ayant des fonctionnalités complémentaires. Comme démontré, notre méthode est évolutive, et les modèles SBML-qual à grande échelle produits par CaSQ peuvent être importés dans Cell Collective et conserver la mise en page et les annotations. L'objectif est de proposer un pipeline transparent pour la production de modèles booléens exécutables à partir de cartes d'interactions moléculaires qui peuvent être analysées en profondeur à l'aide de divers outils de modélisation informatique. L'outil CaSQ peut jouer le rôle d'un pont réunissant deux communautés distinctes, les conservateurs et les modélisateurs, pour produire des modèles annotés interopérables de meilleure qualité, plus précis et réutilisables.

Pour l'objectif final du doctorat, nous avons mis à notre connaissance nos efforts pour créer un modèle booléen à grande échelle pour les FS de PR, le premier modèle exécutable pour ce système. Nous utilisons la carte de la PR, une source de connaissances de haute qualité basée sur la curation humaine, et les réalisations techniques décrites dans les chapitres précédents

pour construire un modèle axé sur des résultats cellulaires spécifiques, à savoir l'apoptose, la prolifération cellulaire, la dégradation de la matrice, l'érosion osseuse et l'inflammation. Nous avons dû faire un compromis dans le choix des phénotypes, car la taille du modèle produit était déjà importante. Nous avons fait ce choix en fonction de nos principales questions biologiques, et plus particulièrement de notre intérêt pour l'étude de la résistance à l'apoptose, de la contribution à l'inflammation et des résultats liés aux dommages structurels. Cependant, le phénotype que nous aimerions inclure à l'avenir est la chimiotaxie et le recrutement cellulaire, car les FS de PR sont connus pour sécréter un certain nombre de cytokines et de chimiokines qui fonctionnent comme des appels de signalisation vers d'autres cellules qui s'infiltrent dans les articulations. Pour faire face à la complexité et aux défis des simulations à grande échelle, nous utilisons une approche modulaire pour construire notre modèle et les sous-modules qui le créent. Cette approche nous a permis d'étudier des parties plus petites des systèmes, et en même temps d'évaluer le comportement au niveau des modules et du modèle entier.

Les simulations en temps réel des modules et de l'ensemble du modèle ont révélé des incohérences et d'éventuelles interactions manquantes, notamment pour les modules d'apoptose et de dégradation de la matrice. Une observation des simulations en temps réel est que les comportements des modules individuels et du modèle entier pour les scénarios testés étaient cohérents. Cela signifie que le comportement observé dans le module a été maintenu intact dans l'ensemble du modèle. Cela pourrait être dû au fait que la majorité des simulations impliquent l'activation ou l'inactivation d'une ou de quelques entrées seulement, mais ce n'était pas un résultat de facto. Nous avons pu calculer des attracteurs pour notre système en utilisant le logiciel BoolNet de R et des heuristiques qui ont contribué à limiter la recherche de l'espace d'état. Nous avons choisi comme conditions initiales soit tous les nœuds mis à zéro, soit tous mis à un. Nous avons calculé les attracteurs pour les deux conditions, pour chaque module individuel et pour l'ensemble du modèle. Leur analyse, préliminaire à l'heure actuelle car ils sont de taille considérable, a confirmé les comportements observés lors des simulations en temps réel, et a également révélé quelques incohérences concernant le comportement de certains nœuds (par exemple AKT2 dans l'inflammation). Une observation intéressante concerne les oscillations entre les entités TP53 et MDM2 dans notre modèle. L'expression abondante de MDM2 dans les FS de PR a été démontrée. MDM2 est le principal régulateur négatif de p53 et, dans les tumeurs, elle contribue à augmenter la prolifération cellulaire. Dans la PR, elle pourrait être un facteur contribuant au phénotype hypo-apoptotique des tissus de la muqueuse grâce à sa capacité à réguler à la baisse les niveaux et les effets de p53. Les anomalies

de p53 dans la PR pourraient soutenir et accélérer l'inflammation synoviale principalement par l'intermédiaire de l'IL-6, comme l'ont montré des études utilisant des rats Lewis avec une arthrite induite par un adjuvant (AIA). Une analyse approfondie des attracteurs est nécessaire et est actuellement en cours, afin de fournir une vision plus détaillée du comportement du système. L'étape suivante consisterait à examiner attentivement l'état de tous les composants dans des conditions données afin d'évaluer dans quelle mesure le comportement du modèle est cohérent avec la biologie du système. Différentes conditions initiales pourraient également être utilisées pour tester des hypothèses spécifiques. Les probabilités phénotypiques utilisant MaBoSS ont fourni une troisième couche d'analyse des comportements observés, notamment en ce qui concerne la dépendance de l'apoptose et le rôle de la voie WNT dans l'érosion osseuse. D'autres scénarios de simulation devraient être testés afin d'obtenir des informations sur des mécanismes spécifiques.

Au cours de ma thèse, la taille des modèles obtenus, même pour les modules, a remis en question le potentiel d'analyse et de simulation de la plupart des logiciels utilisés pour les modèles logiques. L'un des problèmes que nous avons rencontrés au début était le fait que la plupart des logiciels nécessitent l'utilisation du même logiciel pour la construction et l'analyse des modèles. C'était également le cas pour Cell Collective. Notre collaboration avec le Dr Tomas Helikar de l'université du Nebraska-Lincoln, aux États-Unis, a conduit à la mise en œuvre de l'importation de la qualité SBML pour Cell Collective, ce qui a grandement facilité l'analyse. Le Dr Helikar nous a également aidé à récupérer les références de nos modèles encodés dans les identificateurs MIRIAM, dans la plate-forme Cell Collective, ce qui nous a permis de créer des modèles annotés instantanément qui peuvent être publiés sans avoir à refaire le travail de conservation. Tout au long de cette thèse, les développeurs de Cell Collective ont ajusté la plateforme pour satisfaire nos demandes toujours croissantes et ont soutenu sans réserve notre quête de modélisation à grande échelle. Nous avons également travaillé en étroite collaboration avec le Pr. Denis Thieffry (ENS Paris) et le Dr. Aurélien Naldi (ENS Paris), pour surmonter les problèmes d'interopérabilité concernant l'importation de nos fichiers SBML-qual dans GINsim. Lorsque la taille et la complexité des modèles ont remis en question la capacité de GINsim, bioLQM a été utilisé pour effectuer des réductions. Le Dr. Laurence Calzone (Institut Curie Paris), a également fourni des conseils utiles pour l'utilisation de MaBoSS et les réglages nécessaires au fonctionnement du logiciel. Enfin, nous avons eu le soutien du Dr. Sylvain Soliman (INRIA), qui, entre autres, a aidé à nommer les différents composants du modèle, facilitant ainsi l'identification des gènes, des ARN et des protéines

ayant un nom commun. Sans les efforts collectifs et le soutien de la communauté, la plupart des travaux présentés n'auraient pas été possibles. Maintenant que l'interopérabilité est réalisée et que le cadre carte-modèle est établi, nous pouvons espérer que l'analyse en aval des réseaux booléens biologiques à grande échelle sera plus facile et moins longue d'un point de vue technique, ce qui laissera plus de temps pour se concentrer sur les questions biologiques en jeu.

Pour conclure, les systèmes vivants ne peuvent être compris en étudiant uniquement leurs parties individuelles. Avec une production sans cesse croissante d'ensembles de données biologiques, accélérée par le séquençage des génomes et les techniques omiques à haut débit, l'objectif général de la recherche biomédicale sur les maladies complexes devrait être progressivement déplacé d'une analyse essentiellement au niveau moléculaire vers un niveau de biologie systémique capturant le comportement dynamique caractéristique du système. La biologie systémique est une approche holistique qui permet d'étudier la vue d'ensemble des systèmes biologiques et de leur organisation. Une tentative de définition des qualités de la biologie des systèmes a été proposée, avec trois caractéristiques principales. Premièrement, la diversité qui se réfère à la compréhension biologique que chaque interaction d'un composant apporte au système. Deuxièmement, la simplicité qui fait référence à l'approche réductionniste en décomposant le système en descriptions simples. Troisièmement, la complexité, qui renvoie à la compréhension de l'interaction complexe d'un réseau moléculaire. Ces trois qualités font de la biologie systémique un domaine interdisciplinaire par nature, qui combine l'informatique, l'informatique et les mathématiques avec la biologie.

Entre autres applications, cette approche s'est révélée être un outil analytique puissant pour comprendre les interactions dynamiques qui sont au cœur des maladies complexes.

Pour résumer, dans ma thèse :

a) J'ai utilisé mes connaissances préalables pour construire un réseau spécifique aux FS de la PR. Pour récapituler tout ce qui est connu et publié sur la maladie, une carte globale, annotée, spécifique à la PR a été construite sur la base d'une recherche bibliographique exhaustive, d'une curation manuelle et d'une validation par des experts du domaine. Cette carte présente les interactions impliquées dans la PR provenant de divers types de cellules. Grâce aux annotations étendues de chaque entité et de chaque réaction incluses dans la carte, et aux fonctionnalités avancées offertes par la plateforme MINERVA, l'utilisateur peut opter pour des interactions spécifiques aux cellules et extraire les réseaux correspondants spécifiques aux cellules. Si de

telles cartes de connaissances sont très instructives, elles pourraient être pleinement exploitées en les traduisant en objets dynamiques et en modèles mathématiques prédictifs.

b) J'ai ensuite ajouté une couche dynamique au réseau, et ceci de manière automatisée. Jusqu'à présent, la construction des modèles dynamiques était manuelle, sauf pour quelques-uns. Pendant mes études doctorales, un effort pour ajouter une couche dynamique sur cette carte descriptive a été exploré, et l'approche choisie pour étudier la carte la PR de manière dynamique en a été le formalisme logique (booléen), pour sa simplicité et son absence de paramètres cinétiques. En collaboration avec le Dr Soliman, (Lifeware INRIA, Saclay) nous avons développé CaSQ (CellDesigner as SBML-Qual), <https://lifeware.inria.fr/~soliman/post/casq/>, un outil pour l'inférence automatisée de modèles booléens à grande échelle, sans paramètres, à partir de cartes d'interactions moléculaires basées sur la topologie et la sémantique des réseaux. Il s'agit, à notre connaissance, du premier outil qui produit des modèles logiques exécutables de centaines de nœuds (jusqu'à plusieurs centaines), dans un format standard SBML pour la description des modèles, qui peuvent être simulés et analysés plus avant à l'aide d'outils de modélisation connus. Dans ce cadre, des formules logiques préliminaires pour le modèle sont déduites automatiquement selon des règles et des contraintes prédéfinies. Ces règles se sont avérées suffisamment générales pour couvrir divers scénarios de représentation des connaissances biologiques, mais n'ont pas été testées systématiquement pour évaluer la robustesse du modèle.

c) Le problème de l'analyse des modèles booléens à grande échelle constitue un défi sur le terrain. La taille accrue du modèle en termes de nombre de nœuds et la complexité due aux crosstalks de signalisation et à la présence de nombreuses boucles de rétroaction ne permettent pas une analyse dynamique simple. Pour faire face à la complexité, j'ai utilisé une approche modulaire qui m'a permis d'étudier chaque phénotype fonctionnel des FS de PR séparément et de le comparer avec le système fusionné. Pour ce faire, j'ai utilisé la carte globale et le plugin de flux de la plateforme MINERVA pour me concentrer sur les sous-réseaux pertinents. L'idée de modules qui partent d'un phénotype et incluent tous les régulateurs possibles en amont a simplifié la charge de simulation des relations entrée-sortie et a aidé à mieux comprendre le système des FS de PR.

d) Outre la modularité, qui m'a aidé à faire face à la complexité du système en utilisant la stratégie du "diviser pour mieux régner", j'ai également essayé de regrouper les entrées du modèle qui étaient nombreuses et qui contribuaient à l'explosion des calculs lorsque j'essayais

de faire une analyse dynamique. Bien que le regroupement des entrées entraîne une certaine perte d'informations (ainsi, je ne pouvais pas effectuer de simulations *in silico* pour le TNF et l'IL6 séparément car ils étaient regroupés sous le terme de cytokines), cela a rendu le calcul des probabilités phénotypiques accessible et a contribué à une meilleure compréhension du système.

e) J'ai utilisé des normes de biologie systémique pour la conservation, la notation graphique et la modélisation afin de promouvoir l'utilisation de normes communautaires, la transparence des résultats scientifiques, l'interopérabilité et la réutilisation des fichiers et du contexte scientifique, respectivement. J'ai utilisé le SBGN pour la construction de la carte RA, les identifiants MIRIAM pour la conservation, les identifiants uniques pour les gènes et les protéines, et les fichiers SBML qual pour la modélisation dynamique. Tous mes travaux sont ou seront accessibles en accès libre, ce qui favorise la science ouverte.

f) J'ai également eu la chance de participer à un travail communautaire et d'apprendre de cette expérience. Je suis membre du consortium Disease Map et j'ai contribué activement à l'initiative COVID-19 Disease Map (voir chapitre 6 - contributions à la communauté).

Orientations futures:

- a) Il est difficile d'estimer la robustesse d'un modèle lorsque les hypothèses pour l'inférence des règles logiques sont modifiées (par exemple, utiliser les opérateurs ET au lieu de OU pour les formules logiques). La robustesse du modèle pourrait être estimée comme la capacité du modèle à reproduire un scénario biologique bien établi (validé expérimentalement), basé uniquement sur les règles déduites (sans réglage manuel supplémentaire). Une étude approfondie de l'impact d'une modification des règles de traduction du modèle permettra de mieux comprendre les règles et l'effet de la topologie sur la dynamique du modèle. Cet impact pourrait être résumé en probabilités phénotypiques de résultats fonctionnels selon différents scénarios et ensembles de règles. Les résultats pourraient ensuite être systématiquement confrontés aux données expérimentales à petite échelle et aux connaissances des experts encodées dans la carte de PR afin d'identifier l'ensemble de règles de traduction le plus robuste/fiable.
- b) La présence de nombreux inputs dans les modèles inférés est le résultat combiné des règles de traduction de carte en modèle et de la structure de la carte qui conduit à la création de pseudo-intrants. Ces pseudo-intrants augmentent le coût de calcul de

l'analyse, et une stratégie qui pourrait aider à limiter leur présence/impact (c'est-à-dire en fixant leurs valeurs) dans les modèles obtenus faciliterait l'analyse en aval.

- c) L'ambition de ce travail est de mettre en place un cadre qui puisse faciliter l'identification de nouvelles cibles thérapeutiques pour la PR. L'objectif est de prédire les conditions optimales qui favoriseraient l'apoptose et minimiseraient l'érosion osseuse, la destruction du cartilage et les résultats de l'inflammation. Des tests systématiques de différentes conditions initiales pourraient en outre permettre de prédire les résultats de perturbations spécifiques, telles que des effets uniques ou combinés, simulés avec le modèle en forçant ou en supprimant systématiquement l'activité de chaque gène/protéine. L'objectif final est de mieux comprendre le mécanisme de dégradation du cartilage et de l'os, deux symptômes débilissants majeurs de la PR, et de proposer une stratégie qui pourrait aider à bloquer, voire à inverser ces résultats.
- d) Pour obtenir un modèle affiné et plus précis, les données issues d'expériences à petite échelle et les ensembles de données Omic publiées sur les fibroblastes de la PR (transcriptomique, RNAseq, RNAseq mono-cellulaire) devraient être utilisés pour augmenter la spécificité des données sur les cellules et la maladie, ainsi que pour définir des signatures phénotypiques (biomarqueurs) qui pourraient être testées par rapport à des états stables.
- e) Les approches de test unitaire et de propagation de la valeur telles que celles décrites dans Hernandez et al, 2020, pourraient également être appliquées pour vérifier le comportement du modèle au niveau local.
- f) Une autre application potentielle consisterait à identifier les nœuds du système FS de la PR qui seraient des cibles possibles pour différents traitements, non uniquement antirhumatismaux. En effectuant des simulations avec des combinaisons de perturbations, cette approche pourrait permettre, en évaluant l'impact sur les phénotypes, d'identifier les traitements thérapeutiques combinés ou de réorienter des médicaments existants.
- g) Enfin, les expériences in vitro seraient un complément idéal de ces travaux. Les prédictions du modèle pourraient être utilisées pour approfondir l'étude du mécanisme de la maladie au niveau intracellulaire, en ciblant la voie/les facteurs proposés et en mesurant l'apoptose et/ou l'inflammation. De plus, au niveau des crosstalks

cellulaires, nous pourrions évaluer l'impact des résultats fonctionnels des FS de la PR sur les cellules du système immunitaire ainsi que sur d'autres cellules voisines telles que les chondrocytes, responsables des lésions du cartilage et de l'érosion osseuse qui s'ensuit. Des expériences in vitro pourraient également être utilisées pour tester les prévisions concernant les traitements combinés.

Table of Contents

List of Figures	i
List of Tables	vi
List of Abbreviations	viii
Chapter 1. Introduction	1
1.1 Complex human diseases.....	1
1.2 Rheumatoid arthritis (RA)	1
1.3 Diagnosis of RA.....	3
1.4 Genetic factors involved in RA	4
1.5 Environmental factors.....	5
1.6 Autoimmunity and cross talks in RA.....	7
1.7 Disease stages / phases	9
1.8 Signaling pathways involved in RA	9
1.9 Current RA treatment.....	11
1.10 Systems biology for complex diseases	12
1.10.1 SBGN standards and networks representation	13
1.10.2 MIRIAM and NOTE.....	14
1.11 Computational Modelling in biology	16
1.12 Objectives	18
Chapter 2. RA map.....	21
2.1 Introduction.....	21
2.2 Construction of the RA map	22
2.2.1 Annotation and curation criteria	24
2.2.2 Evaluation of components and reactions	25
2.2.3 Compartments, structure, and layout	25
2.2.4 Experts' advice and feedback	26
2.2.5 SBGN standards and process description (PD) map validation	27

2.3 RA map	28
2.3.1 A comprehensive molecular interaction map for Rheumatoid Arthritis (RA).....	28
2.3.2 Molecular pathways covered in the RA map.....	30
2.4 Transforming RA map into a state of the art knowledge base using MINERVA.....	30
2.4.1 The RA map as a template for visualizing cell-specific overlays	34
2.4.2 Comparing overlap with respective disease databases such as Disnor, DisGeNet and Ingenuity pathway analysis (IPA).....	36
2.4.3 Visualizing Omic datasets	39
2.5 Topological analysis with Cytoscape	41
2.5.1 Functional enrichment of the whole RA network.....	46
2.5.2 Clustering and functional enrichment of the clusters	47
2.5.3 Systemic Interpretation and Pharmacogenomics Analysis using BioInfoMiner	51
2.6 Conclusion	55
Chapter 3. Automated inference of Boolean models from molecular interaction maps using CaSQ.....	57
3.1 Biological network representations and molecular interaction maps.....	57
3.2 Boolean models for dynamical studies	58
3.3 Bridging the gap between static and dynamic representation	59
3.4 CaSQ.....	60
3.5 Molecular interaction maps and logic models.....	66
3.6 Model comparison	68
3.7 In silico simulations and calculation of stable states.....	68
3.7.1 Cell Collective	68
3.7.2 GINsim.....	69
3.8 Graph reduction and model inference.....	70
3.9 CaSQ run time	71
3.10 CaSQ-inferred Boolean models versus manually built models.....	72

3.10.1 Shared nodes	72
3.10.2 In silico simulations and dynamic analysis	77
3.11 Discussion	87
Chapter 4. Inference of a modular, large-scale Boolean network for modelling the Rheumatoid Arthritis fibroblast-like synoviocytes.....	89
4.1 Introduction.....	89
4.2 Methods and data	91
4.2.1 Using prior knowledge to build an RA FLS specific network	91
4.2.2 Using CaSQ to infer executable Boolean networks from the RA FLS specific network	92
4.2.3 Real-time <i>in silico</i> simulations using the Cell Collective modelling platform.....	92
4.2.4 Attractors search using BoolNet.....	93
4.2.5 Model reduction using GINsim	93
4.2.6 Probabilistic Boolean modelling with MaBoSS	94
4.3 Results.....	94
4.3.1 Real time <i>in silico</i> simulation using the platform Cell Collective....	98
4.3.2 Calculating Attractors using BoolNet.....	109
4.3.3 Creating a more compact version of the RA FLS model with five phenotypes, and using MaBoSS to cross check observations from real time simulations	111
4.4 Discussion.....	115
Chapter 5. General discussion and future perspectives	118
5.1 Future directions	120
5.2 Economic and social impact of computational systems biology approaches in RA.....	121
Chapter 6. Contributions to the community.....	123
6.1 The Disease Maps consortium.....	123

6.2 COVID-19 Disease Map, a large-scale community effort to create graphical and executable models of SARS-CoV-2 virus-host interaction mechanisms.....	123
6.3 Apoptosis diagram	125
Bibliography.....	127
ANNEX A	150
ANNEX B	162
ANNEX C	163

List of Figures

Figure 1.1: The pathology of RA (figure taken from (Clancy & Hasthorpe, 2011)).....	3
Figure 1.2: Current rheumatoid arthritis (RA) genetic risk loci (taken from (McAllister et al., 2011)).....	5
Figure 1.3: Multistep progression towards RA disease development (figure taken from (McInnes & Schett, 2011))	6
Figure 1.4: Cytokine mediated synovial regulation and interaction between various cells of both innate and adaptive immune systems during the progression of RA (taken from (McInnes & Schett, 2007)).....	8
Figure 1.5 : Three different types of SBGN networks used to represent biological processes (taken from (Le Novère, 2015)).....	14
Figure 1.6: CellDesigner Annotation section. A) MIRIAM and NOTE section in CellDesigner B) Bio models qualifiers representing relation between a model component and the resource used to annotate it C) Different data types used to annotate the components.	15
Figure 1.7: Various modelling approaches based on time representation and variable values (taken from (Le Novère, 2015)).....	17
Figure 1.8: A simple Boolean network model: A simple regulatory network containing three nodes A, B and C is shown. The edges with sharp arrows represent activation (positive effects) while the edges with blunt arrows indicate inhibition (negative effects).	18
Figure 2.1: Workflow for the construction and use of the RA map.	21
Figure 2.2: Current Celldesigner graphical notations (http://celldesigner.org/help/images/components42.png).....	23
Figure 2.3: Example of an SBGN Process Description uses two kinds of nodes (that is, nodes representing biochemically-indistinguishable entities such as molecules). One kind of node whose glyph is a rounded rectangle represents different macromolecules, and the other whose glyph is a circle represents pools of simple chemicals (https://sbgn.github.io/examples).	23
Figure 2.4: Annotations added to the MIRIAM section of CellDesigner.....	25
Figure 2.5: Cell layout of the RA map.....	26
Figure 2.6: Gene regulatory representation according to SBGN standards.....	27
Figure 2.7: RA map references. Barplot showing the yearly distribution of peer-reviewed scientific articles and reviews included in the RA map.....	29
Figure 2.8: Snapshot of the SBGN compliant RA map. The map is colour-coded with proteins in purple, genes in green, RNAs in red and phenotypes in yellow. State transitions and catalysis reactions are displayed in black and the inhibitions are in red. Compartments are distinguished as bounding boxes. The map was built using CellDesigner, version 4.4 (16).....	29

Figure 2.9: The RA map in MINERVA platform. A) Users can type in the search box the element to look for in the map. The resulting elements are shown as pins on the map. Corresponding annotations of the searched element, like HGNC, Entrez Gene, RefSeq and Ensembl identifiers are displayed on the left panel along with the PubMed identifiers of the manually curated annotations. B) Further clicking on the pin will display additional information about interacting drugs, chemicals and microRNAs for the element.	32
Figure 2.10: Tree expansion plugin in MINERVA.	33
Figure 2.11: Stream export plugin in MINERVA.	34
Figure 2.12: Cell/tissue/fluid categories in the RA map. Percentage of map components in the seven different overlays of the RA-map based on the scientific articles used.....	35
Figure 2.13: Visualizing cell/tissue/fluid-specific parts of the RA map using dedicated overlays. Snapshot of the visualisation of the Synovial Tissue overlay.	36
Figure 2.14: Overlap of the RA map with DisGeNet and Disnor databases. Mapping of A) DisGeNet all sources list, B) Disnor list.	37
Figure 2.15: Early versus erosive RA components extracted from IPA. Mapping of components that correspond to A) early-stage RA, B) erosive stage of RA, C) proteins in the endoplasmic reticulum in early and erosive RA. Only PDIA3 is shared between the molecules that map in both states.....	39
Figure 2.16: Mapping of omic datasets from RA synovial tissue. The apoptosis and angiogenesis phenotypes appear to be inactive as no molecule leading to these cellular phenotypes is mapped.	40
Figure 2.17: The RA map as a complex network with spring embedded layout in Cytoscape (Shannon et al., 2003). One connected core and several smaller unconnected parts are shown.	41
Figure 2.18: Node degree distributions of the RA map with a fitted power law. A) Overall degree distribution, B) In degree, C) Out degree distribution.	44
Figure 2.19: RA map enrichment in RA disease terms using DAVID. Gene count corresponding to RA related disease terms in DAVID, using the RA map components as the input list.	46
Figure 2.20: Functional analysis of the RA map using DAVID. The list of RA map components was used as the input list in the functional annotation tool DAVID. Disease enrichment analysis showed enrichment in Type 2 Diabetes, multiple sclerosis, various cancer types, asthma, HIV, atherosclerosis and rheumatoid arthritis.	47
Figure 2.21: 57 Glay clusters in Cytoscape.	48
Figure 2.22: Heatmap of the 48 priority genes and their systemic interpretation using BioInfoMiner and GO terms.....	53
Figure 2.23: Heatmap of the 32 priority genes and their systemic interpretation using BioInfoMiner and HPO terms.....	55

Figure 3.1: The repertoire of CellDesigner graphical notation schemes used to illustrate CaSQ's rules. For CaSQ's conversion rules we use the notation schemes for association, transport, catalysis, state transition and also the glyphs for receptor, protein, modified protein (here we show phosphorylation as an example) and the empty set. The empty set can account for degradation or in SBGN-PD terms, can represent the creation (respectively, the disappearance) of an entity from an unspecified source (resp. sink) that we do not need or wish to express explicitly.	61
Figure 3.2: Illustration of the 1st rule. If two species of the map are only reactants in a heterodimer association, and if one of the reactants is annotated as a receptor, then the receptor is deleted from the map (its annotations are added to the product of the reaction).	62
Figure 3.3: Illustration of the 2nd rule. Compression of the complex formation, where none of the reactants is denoted as a receptor, and both reactants do not participate in any other reaction. As a result, both reactants are removed and modifiers are rewired to have the complex as a product.	62
Figure 3.4: Illustration of the 3rd rule. Removing inactive forms that do not participate in other reactions.	63
Figure 3.5: Combination of rules 2 and 3. CaSQ retains components that contribute further to the propagation of the signal.	63
Figure 3.6: Combination of the 2nd and the 4th rule. Components that are translocated across other compartments (for example transcription factors) are merged in one component that inherits all influences, provided that the original component does not participate in another reaction/ regulation.	64
Figure 3.7: A complex part of the RA map translated into an activity flow (AF) like diagram with preliminary logical rules.	65
Figure 3.8: Snapshot of the RA map built with the software CellDesigner. The annotations for every node are stored in the MIRIAM section, using the bqbiol:isDescribedBy tag. Here we see the MIRIAM annotations for the node IL6.	67
Figure 3.9: Integration and understanding of the biological knowledge with CellCollective (Taken from (Helikar et al., 2012)).	69
Figure 3.10: View of the CaSQ inferred models using the modelling platform Cell Collective. The CaSQ tool produces annotated Boolean models, including bibliographical references stored in MIRIAM. References stored in the MIRIAM section of the xml file of the molecular maps built with the software CellDesigner can be retrieved and visualised in the Cell Collective modelling platform. The original map layout is also conserved facilitating simulations. In the left panel, the selected node that corresponds to the IRAK1/IRAK4 complex is shown in blue while on the right panel the corresponding references are displayed.	70
Figure 3.11: A) Screenshot of simulations for Btk knockout of the CaSQ derived mast cell activation model using Cell Collective. B) In the graph panel, one can see that when Btk is set to zero, Erk and PLCG1 are not expressed.	78

Figure 3.12: A) Screenshot of simulations for Syk knockout of the CaSQ derived mast cell activation model using Cell Collective. B) In the graph panel, one can see that when Syk is set to zero, Erk, JNK, NFAT, NFkB, Ca2+, PKC, Elk1, PLCG1 are not expressed.	78
Figure 3.13: Screenshot of the stable states table for the manually built model for mast cell activation with GINsim.....	79
Figure 3.14: Simulations of the CaSQ inferred model using the modelling platform Cell Collective. The CaSQ inferred model for MAPK was able to reproduce known biological scenarios, either completely or partially. The results of the <i>in silico</i> simulations for the three first biological conditions described in Table 9 showed perfect agreement with the results of manually built model, as depicted in panels a, b and c. For conditions described in scenarios 4 and 5 of Table 9 the CaSQ inferred model could partially reproduce the attended behaviour (panels d and e) while simulation results for scenario 6, were inconsistent with the literature and the results of the manually built model (panels f, g and h).	87
Figure 4.1: The notion of modules in our approach involves the selection of one (or more) phenotype(s) and follows the upstream regulators all the way to the initial inputs.....	92
Figure 4.2: Diagram comprising the five functional outcomes for RA FLS. Two different data sets were used as overlays. The color purple depicts the synovial fibroblasts overlay from the extensive source annotations of the RA map while the yellow color represents the differentially expressed genes from the scRNA seq data of RA FLS (0.05 threshold p value)	95
Figure 4.3: The stream plugin in MINERVA platform allows for the selection of upstream (downstream) regulations starting from a given node. In our case, either one or multiple phenotypes were selected and the upstream direction to create the individual modules and the merged network.	96
Figure 4.4: Venn Diagram of the five modules. A core of 153 nodes is shared among all five modules and only a small number of nodes is characteristic of the corresponding module. This distribution explains also why the size of the merged model is not the sum of the sizes of the individual modules.....	97
Figure 4.5: Simulation with CellCollective. A. Oscillatory behaviour for MDM2 and TP53 entities with initial condition set to all inactive and asynchronous updating B. Same oscillatory behaviour for MDM2 and TP53 with a span of over 400 time steps.....	108
Figure 4.6: Simulation with CellCollective. A. All phenotypes except Matrix Degradation get activated and B. stay in this state for more than 400 time steps.	109
Figure 4.7: A. Pie chart showing the probability of 80% chances of apoptosis activation via FASL B. Diagram showing the trajectories for FASL and Apoptosis when FASL is kept at 80% ON.....	112
Figure 4.8: MaBoss simulation showing the absence of Bone erosion when RANKL complex kept 80% ON.....	113
Figure 4.9: MaBoss simulation when WNT complex was set as 80% ON.	114

Figure 4.10: **A.** Pie chart representing a MaBoSS simulation, showing 65% of bone erosion activation via both RANKL complex and WNT and 15% bone erosion activation via WNT alone **B.** Diagram showing the trajectories of RANKL, WNT and Bone erosion phenotype in MaBoSS simulation. 115

Figure 6.1: Screenshot of the Fairdomhub space for the COVID-19 Disease Map project. 236 people from 123 institutions around the globe are participating in this large-scale community effort to tackle the pandemic..... 124

Figure 6.2: Screenshot of the MINERVA build for the COVID-19 Disease Map diagrams.125

Figure 6.3: Apoptosis map built with Celldesigner graph editing software is structured into three compartments, namely Extracellular space containing the ligands, plasma membrane with receptor-ligand complexes and cytoplasm with all signaling and viral proteins. Green boxes represent generic proteins while peach colored boxes represent viral proteins. Red colored interactions are inhibitions while the black interactions are activations..... 126

List of Tables

Table 2.1: Overlays of RA-map and corresponding sub-categories.	35
Table 2.2: Simple Topological Parameters obtained with NetworkAnalyzer for the RA network.	43
Table 2.3: Top ten hubs of the RA map.	45
Table 2.4: 84 RA specific genes after analysis with DAVID.	47
Table 2.5: Top 5 Glay clusters enriched in RA.	49
Table 2.6: Top 5 Glay clusters enriched in RA and their corresponding Pathway enrichment and P value.	49
Table 2.7: Top ten priority genes using BioInfoMiner and GO terms.	52
Table 2.8: Top ten priority genes using BioInfoMiner and PHO terms.	54
Table 3.1: Molecular maps and corresponding manually built models used for benchmarking CaSQ.	67
Table 3.2: Size (number of components) of the CaSQ inferred model using the default and BCC options.	71
Table 3.3: The run times of CaSQ for producing executable SBML-qual files with default options.	71
Table 3.4: Comparison of CaSQ inferred Boolean models with manually built models (MM).	73
Table 3.5: Shared nodes between CaSQ and manually built model for mast cell activation. .	73
Table 3.6: Shared nodes between CaSQ and manually built model for MAPK.	74
Table 3.7: Examples of different naming of shared nodes between CaSQ inferred and manually built models for mast cell activation and MAPK.	76
Table 3.8: Logical steady-state analysis obtained with GINsim, for manually built and CaSQ inferred mast cell activation models. The nine stable states of the manually built model are given in this table (SS1 to SS9). The bottom lines show which of these stable states are correctly recovered in the CaSQ inferred model, or have 1 or 2 mismatches.	80
Table 3.9: Biological data and corresponding behaviours of the manually built and the CaSQ inferred models for MAPK.	82
Table 4.1: Size in terms of number of nodes and number of edges for the five modules and the whole model.	96
Table 4.2: <i>In silico</i> simulation of the Inflammation module and the merged model (Simulation figures are available at ANNEX figure A1).	98
Table 4.3: <i>In silico</i> simulations of the Bone erosion module and the merged model (Simulation figures are available at ANNEX figure A2).	101

Table 4.4: <i>In silico</i> simulation of Cell growth/survival/proliferation module and merged model (Simulation figures are available at ANNEX figure A3).....	103
Table 4.5: <i>In silico</i> simulation of the Apoptosis module and the merged model (Simulation figures are available at ANNEX figure A4).	105
Table 4.6: <i>In silico</i> simulation of the Matrix degradation module and the merged model (Simulation figures are available at ANNEX figure A5).....	106
Table 4.7: Attractors calculation for predefined initial conditions (all set to 0 or 1) for the five modules and the merged model (files available at ANNEX B).....	110

List of Abbreviations

ACPA: Anti-citrullinated protein

AF: Activity Flow

CTNNB1: Catenin beta 1

DAVID: Database for annotation, visualization and integrated discovery

ECM: ExtraCellular Matrix

EGF: Epidermal growth factor

ER: Entity-Relationship

FADD: Fas-associated via death domain

FAS: Fas cell surface death receptor

FGF: Fibroblast growth factor

GAD: Genetic association database

GO: Gene ontology

HGNC: HUGO gene nomenclature committee

HLA-DRB1: Major histocompatibility complex, class II, DR beta 1

IGF: Insulin-like growth factor

IL12A: Interleukin 12 alpha

IL1B: Interleukin 1 beta

IPA: Ingenuity pathway analysis

IRF5: Interferon regulatory factor 5

JAK2: Janus kinase 2

MAPKs: Mitogen-activated protein kinases

MINERVA: Molecular Interaction NEtworks VisuAlization

MIRIAM: Minimal Information Requested In the Annotation of Models

MS: Multiple Sclerosis

NF- κ B: Nuclear factor kappa-light-chain-enhancer of activated B cells

PBMC: Peripheral blood mononuclear cell

PD: Process Description

PDGF: Platelet-derived growth factor

PHO: Human phenotype ontology

PTPN22: Protein tyrosine phosphatase non-receptor type 22

RA: Rheumatoid arthritis
RF: Rheumatoid factor
RIPK1: Receptor-interacting serine/threonine kinase 1
RIPK2: Receptor-interacting serine/threonine kinase 2
SBGN: Systems Biology Graphical Notation
SBML: Systems Biology Markup Language
SOCS: Suppression of cytokine signalling
STAT3: Signal transducer and activator of transcription 3
STAT4: Signal transducer and activator of transcription 4
T2D: Type 2 diabetes
TBK: TANK-binding kinase
TLRs: Toll-like receptors
TNF: Tumor Necrosis Factor
TRAF6: TNF receptor-associated factor 6
VANTED: Visualisation and Analysis of Networks containing Experimental Data
VEGF: Vascular endothelial growth factor
WNT5A: Wnt family member 5A

Chapter 1. Introduction

1.1 Complex human diseases

Complex or multifactorial diseases are defined as diseases whose aetiology and progression is determined by a number of genetic, environmental and epigenetic factors. Some examples in this regard are autoimmune diseases like rheumatoid arthritis (RA), neurological diseases like alzheimer's disease, parkinson's disease and multiple sclerosis, respiratory diseases like asthma, connective tissue diseases like scleroderma, kidney diseases, and many more (Craig, 2008). Genetic factors dictate our susceptibility to some diseases, as well as how well we respond to different treatments. However, they only represent a part of the risk associated with complex disease phenotypes. The actual development of the disease is a combined outcome which relies heavily on environment and lifestyle factors.

Although researchers in the biomedicine area have made significant progress against various other life threatening diseases, they are still facing challenges to understand the basic underlying mechanisms of initiation and progression in complex diseases. The reason lies behind the fact that the methodology of classical experimental biology is mainly based on studying individual genes and proteins, considering the organism as simple and linear. However, complex disease involves numerous complicated interactions among the biological entities resulting in some critical pathological outcomes (D.-Y. Cho et al., 2012). In this project we tried to understand the underlying mechanism of a common complex disease, Rheumatoid Arthritis (RA). RA affects ~0.5-1.0% of the adult population in most developed countries (Scott et al., 2010). The global prevalence of RA is reported to be somewhat lower (~0.24%), with a 3 times higher prevalence in women compared with men (Cross et al., 2014) and ~0.5 to 1% in white individuals by a study based on western countries (Smolen et al., 2018).

1.2 Rheumatoid arthritis (RA)

RA is an inflammatory, chronic autoimmune disease whose pathogenesis involves a complex interplay between genetic, epigenetic and environmental factors resulting in a cascade of immune reactions ("Rheumatoid arthritis," 2018; Sparks, 2019). RA results in inflammation of the joints primarily of hands, wrists, and small joints of the feet and, in particular, the synovial membrane that covers them. RA main characteristics are synovial inflammation and hyperplasia, autoantibody production (rheumatoid factor [RF] and anti-citrullinated protein

antibody [ACPA]), cartilage and bone destruction, and systemic features, including cardiovascular, pulmonary, psychological, and skeletal disorders. The complexity of multiple signalling pathways interacting with one another resulting in the expression of various proinflammatory molecules like cytokines, chemokines and matrix metalloproteinases (MMPs) makes RA disease puzzling in order to elucidate its aetiology (McInnes & Schett, 2011) (**Fig 1.1**). Both innate and adaptive immune system processes play a major role in the propagation of signals resulting in the activation of multiple immune cells like macrophages, neutrophils, T cells and B cells along with the activation of joint cells like synovial fibroblasts. Infiltration and accumulation of inflammatory cells like macrophages and lymphocytes (T lymphocytes, B lymphocytes) and neutrophils results in the hyperplasia of the synovial membrane and subsequently results in pannus formation.

In addition to various immune cells being studied for the therapeutic intervention in RA disease, synovial fibroblasts (SFs) have emerged as an additional therapeutic target for RA. Many studies focus on SFs to gain more insights about the complex interactions with the immune cells and how the multiplication of SFs can eventually induce cartilage destruction and bone erosion in the joints (Bustamante et al., 2017; Huber et al., 2006). Generally, in healthy tissue, synovial fibroblasts provide the joint cavity and the adjacent cartilage with plasma proteins and lubricating molecules such as hyaluronic acid and lubricin. SFs also produce matrix proteins such as collagen and hyaluronan, a variety of matrix-degrading enzymes, such as MMPS contributing to matrix remodelling (Müller-Ladner et al., 2007).

Innate immune cells, like macrophages, act by releasing cytokines (such as tumor necrosis factor (TNF- α) and interleukins (IL-1, IL-6)), MMPs, cellular processes like phagocytosis and antigen presentation and in turn are activated by toll-like receptors (TLR-2/6, TLR-3, TLR-4, TLR-8) and nucleotide-binding oligomerization domain (NOD)-like receptors (NLRs), cytokines, T cells interactions and so on. Neutrophils, another innate cell type, produce and release prostaglandins and proteases contributing to synovitis (McInnes & Schett, 2011). Cytokines play a major role in the propagation of signals between cell types in the synovial membrane and synovial fluid.

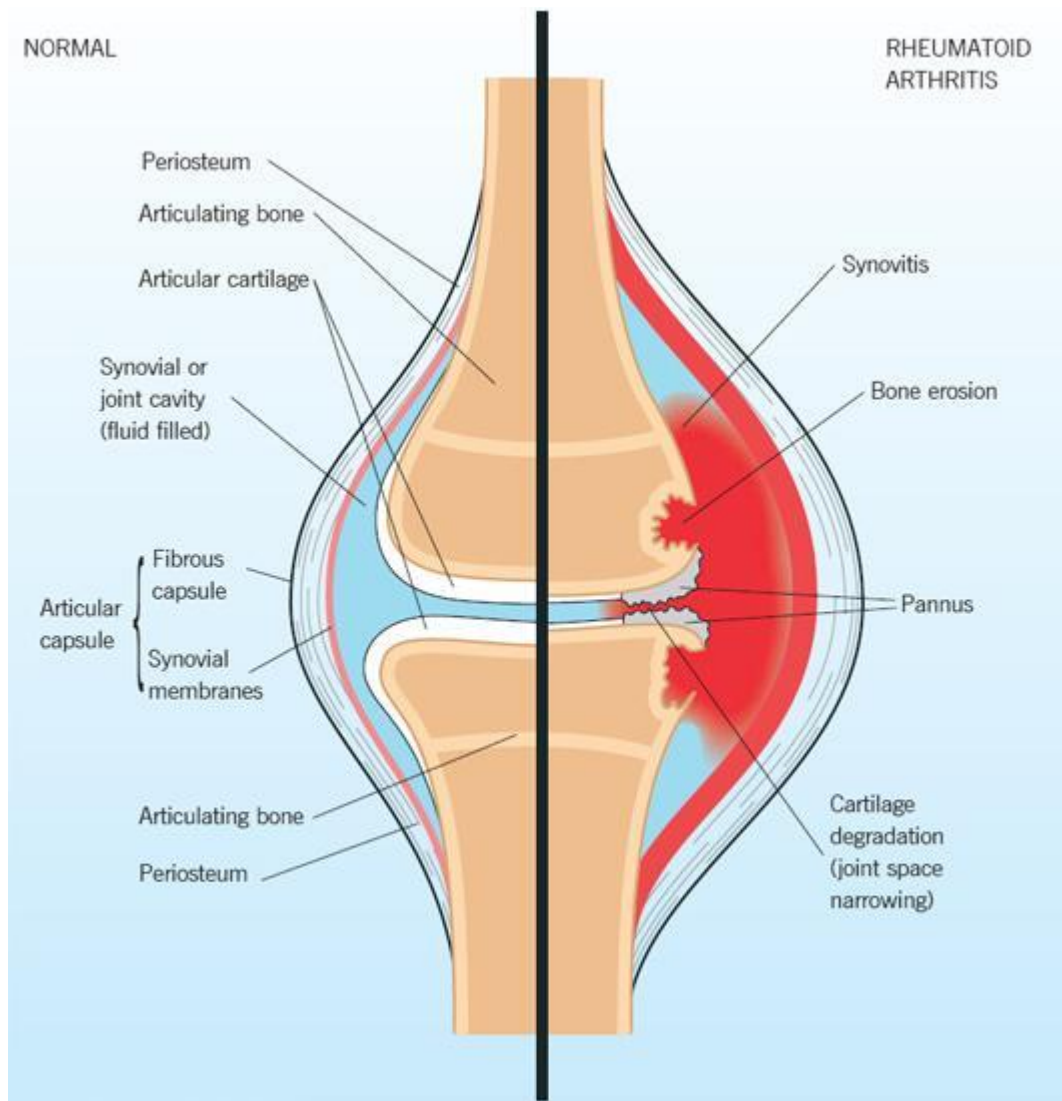


Figure 1.1: The pathology of RA (figure taken from (Clancy & Hasthorpe, 2011)).

1.3 Diagnosis of RA

The diagnosis of RA is mainly based on clinical manifestations and serum biomarkers such as RF and ACPA with the latter to be more disease-specific. The positivity of ACPA is now widely used to diagnose RA due to its high specificity (>97%) in clinical practice (Guo et al., 2018). A typical medical examination includes a search for swelling, boggiess and tenderness with atrophy of muscles near the involved joints. C-reactive protein levels and erythrocyte sedimentation rate are often increased with active RA, and these acute phase reactants are part of the new RA classification criteria (Aletaha & Smolen, 2018). The American College of Rheumatology criteria is helpful not only in the diagnosis of RA but also for following the patients during research trials in RA (Arnett et al., 1988). In 2010 the ACR and the European League Against Rheumatism (EULAR) put forward revised classification criteria emphasizing

RA characteristics that emerge early in the disease course, including ACPAs, a biomarker that predicts aggressive disease. The 2010 criteria do not include the presence of rheumatoid nodules or radiographic erosive changes, both of which are less likely in early RA (Kay & Upchurch, 2012).

1.4 Genetic factors involved in RA

The heritability of RA is about 60% with genetic factors like class II major histocompatibility antigens/human leukocyte antigens (HLA-DR) and non-HLA genes associated with its pathogenesis (Kurkó et al., 2013; MacGregor et al., 2000). Shared epitope alleles which are most significantly associated with RA are HLA-DRB1*01 and HLA-DRB1*04. The most relevant non-HLA genes with single nucleotide polymorphisms (SNPs) associated with RA are PTPN22, IL23R, TRAF1, CTLA4, IRF5, STAT4, CCR6 and PADI4 (McInnes & Schett, 2011). Many studies are being conducted supporting the genetic basis of RA through the identification of genetic susceptibility variants. Three main approaches have been used to identify these susceptibility loci: candidate gene study, genetic linkage study, and genome wide association studies (GWAS), the third being the most successful. The latest GWAS meta-analysis association study has brought the total number of confirmed RA risk loci to 101 (Okada et al., 2014) (**Fig 1.2**).

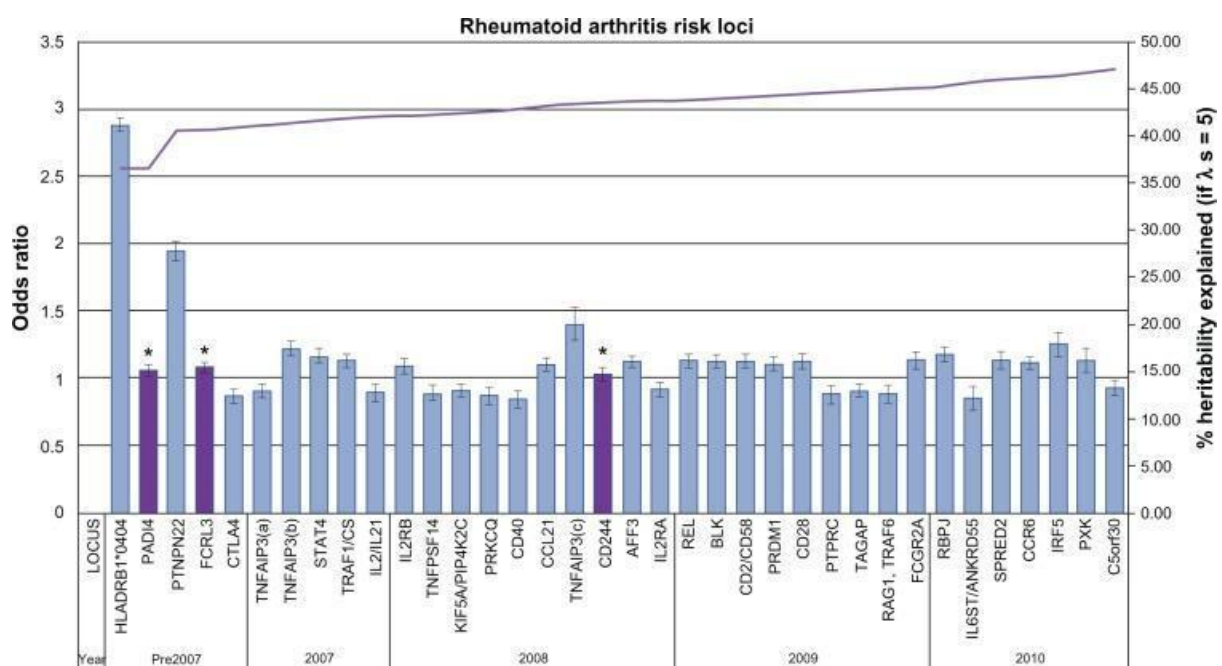


Figure 1.2: Current rheumatoid arthritis (RA) genetic risk loci (taken from (McAllister et al., 2011))

1.5 Environmental factors

Environmental factors confer great susceptibility in the pathogenesis of RA. These factors are classified as regional environmental factors, including geography, climate, and socio-cultural environmental factors, such as lifestyle, smoking and dietary habits. The most significant environmental factor is smoking which is also highly dependent on the genetic background of the person. The risk of developing RA with ACPAs is increased by 21 folds in smokers carrying two copies of the shared epitope, as compared to non-smokers without the shared epitope (Klareskog et al., 2006). Some other environmental risk factors are poor socioeconomic status and educational level, exposure to infectious agents, obesity, high dietary sodium intake, smoking, as well as occupational exposure to mineral dust, insecticides, and textiles. Multistep progression to the development of rheumatoid arthritis including the environmental factors is shown in **Fig 1.3 (McInnes & Schett, 2011)**.

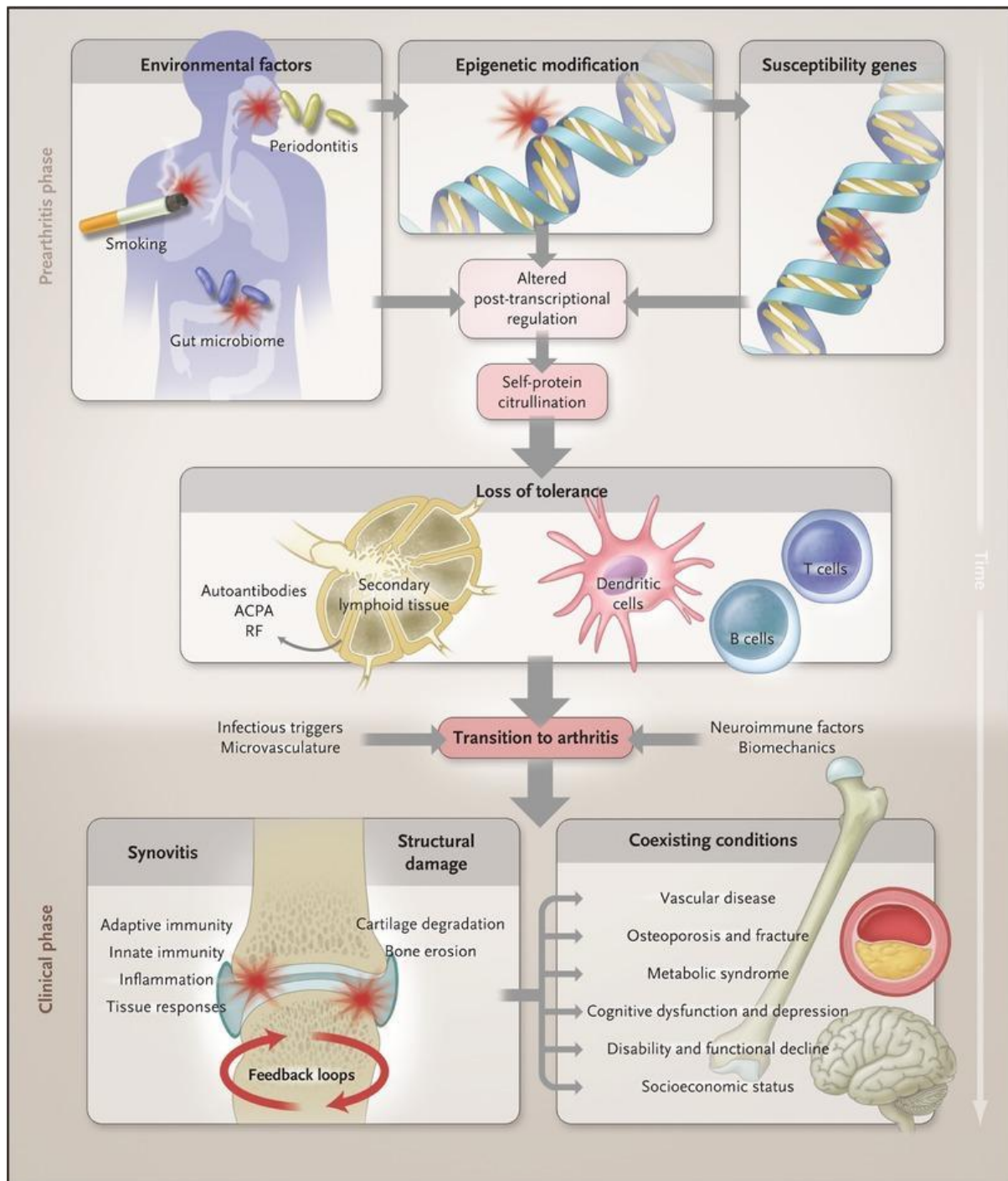


Figure 1.3: Multistep progression towards RA disease development (figure taken from (McInnes & Schett, 2011))

1.6 Autoimmunity and cross talks in RA

RA is generally associated with serological evidence of systemic autoimmunity as indicated by the presence of autoantibodies in serum and synovial fluid. The first autoantibody in RA, RF was described by Waaler in 1940 (Waalder, 2007) which is directed to the Fc region of IgG. Autoantigens which include a wide range of cartilage components, stress proteins, enzymes, nuclear proteins and citrullinated proteins, targeted by a number of autoantibodies are eventually found in RA. This demonstrates that RA is characterized by accumulated auto reactivities in both B and T cells. Another autoantibody ACPA marks the initial stage of RA development which is an abnormal response to various citrullinated proteins such as fibrin, vimentin, fibronectin, α -enolase, type II collagen, and histones. Initial ACPA levels predict the time interval to disease onset and also reflects the break of immunological tolerance thus inducing pain, bone loss and inflammation in RA (Guo et al., 2018).

Both innate and adaptive immune systems are involved in the perpetuation of inflammation, cartilage destruction and bone erosion eventually making it systemic, affecting other parts and organs of the body such as the cardiovascular system, respiratory system, eyes, mouth, kidney and so on. Various cell types of both innate and adaptive immune systems such as macrophages, dendritic cells, T-lymphocytes and B-lymphocytes participates in the diffusion of inflammatory signal via secretion of pro inflammatory mediators activating immune cells along with the activation of synovial cells of the joints. Major pro-inflammatory mediators involved in the progression of inflammation in RA are cytokines such as IL-1, IL-6, IL-17, TNFs and chemokines which vary greatly in numbers and subtypes affecting and activating different cell types (**Fig 1.4**). Cytokines regulate inflammation, autoimmunity and articular destruction in the joints of patients with rheumatoid arthritis (McInnes & Schett, 2007, 2011).

The two key pro-inflammatory cytokines in RA are IL-1 and TNF- α whose regulation is of crucial importance in the RA disease. IL-1 and TNF- α also initiates a cascade of signaling pathways further resulting in the expression of pro inflammatory mediators. In particular, TNF- α has already proved to be of particular utility as a therapeutic target. Other cytokines found to be implicated in RA are IL-12, IL-15, IL-17, IL-18, IL-23 and members of type 1 interferon family like IFN- α and IFN- β (McInnes & Schett, 2007). The essential cytokine mediators such as receptor activator of nuclear factor- κ B (RANK) ligand (RANKL) and macrophage colony-stimulating factor (M-CSF), which are expressed by synovial fibroblasts and T helper 1 (TH1)

cells are responsible for the maturation and activation of bone resorption, osteoclast maturation and activation. As shown in **Fig 1.4**, critical cytokine pathways are depicted in both innate and adaptive immune system where activation of dendritic cells (DCs), T lymphocytes, B lymphocytes and macrophages underpins the dysregulated expression of cytokines that in turn drive activation of effector cells, including neutrophils, mast cells, endothelial cells and synovial fibroblasts. Such activation further triggers disease phenotypes like angiogenesis, cartilage matrix degradation, inflammation and synovial hyperplasia among others (McInnes & Schett, 2007).

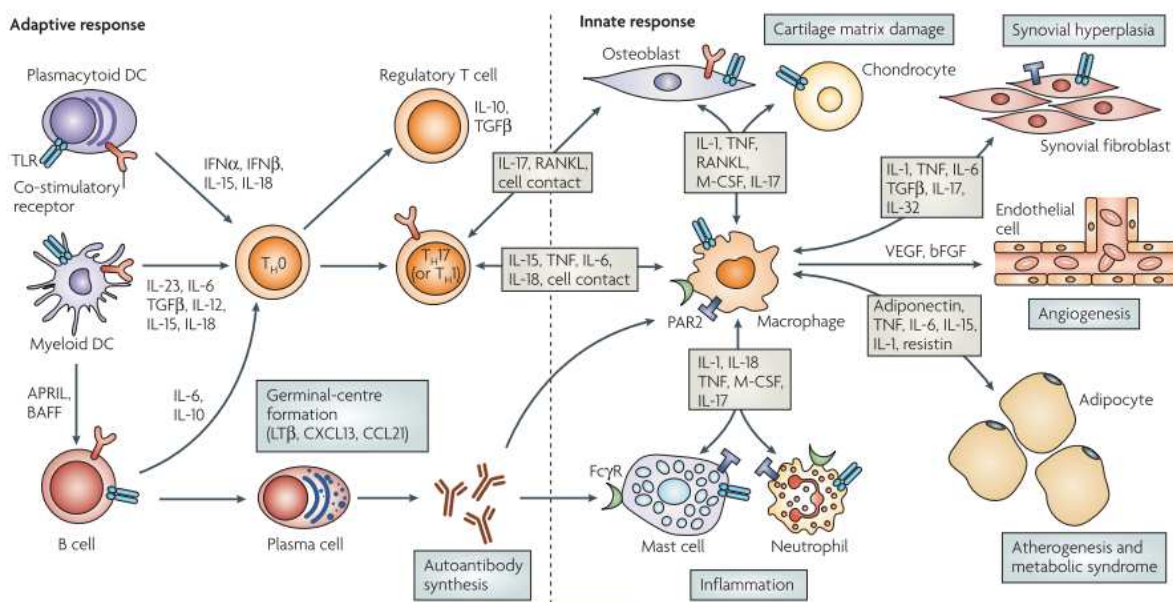


Figure 1.4: Cytokine mediated synovial regulation and interaction between various cells of both innate and adaptive immune systems during the progression of RA (taken from (McInnes & Schett, 2007))

Inflammatory chemokines including IL-8/CXCL8, ENA78/CXCL5, Mig/CXCL9, IP10/CXCL10, MCP-1/CCL2, RANTES/CCL5, MIP-1 α /CCL3 and LARC/CCL20 are mainly produced by activated synovial fibroblasts and monocytes/macrophages in the RA synovial tissue and elevated levels of these chemokines were also detected in the synovial fluid and serum of RA patients (Szekanecz et al., 2003). Chemokines not only play a role in inflammatory cell migration but are also involved in the activation of fibroblasts in the RA synovium, possibly in an autocrine or paracrine manner (Nanki et al., 2001).

1.7 Disease stages / phases

Disease onset and progression of RA in joints starts with synovitis which develops with the infiltration and accumulation of leukocytes such as T cells, B cells, neutrophils in the synovial compartment. Synovial fluid is flooded with pro-inflammatory mediators that begins to interact initiating an inflammatory cascade, characterized by the interactions of synovial cells like fibroblasts like synoviocytes (FLSs) with the cells of the innate immune system (such as dendritic cells (DCs), monocytes, macrophages and so on) along with the cells of the adaptive immune system like T-lymphocytes (cell-mediated immunity) and B-lymphocytes (humoral immunity) (Guo et al., 2018). The synovial membrane is infiltrated by immune cells more specifically monocytes/macrophages which are also central to the pathophysiology of RA. The imbalance between pro-inflammatory M1 macrophage and anti-inflammatory M2 macrophage contributes to osteoclastogenesis in RA patients, especially in ACPA-positive RA (Fukui et al., 2017) . Antigen-presenting cells such as DCs begin to accumulate in the articular cavity especially myeloid DCs which induce T cell differentiation. T cells of the adaptive immune system are activated and subsequently release a variety of inflammatory cytokines. CD4 effector T cells in RA sustain chronic synovitis and support autoantibody production. B cells are mainly studied for their antigen presentation, antibody formation and release and cytokine release (Gary S Firestein & McInnes, 2017).

Synovium is generally characterized by cells like bone marrow-derived macrophages and specialized FLSs which secrete lubricants like hyaluronic acid and lubricin for maintaining normal health and mobility of the joints. In the inflamed rheumatoid synovium, the healthy lining structure is transformed into a pannus-like structure, a hyperplastic synovial lining containing a higher number of activated FLS and macrophages that extends into the joint space, attaches to the cartilage surface (cartilage–pannus junction), and invades and degrades the cartilage matrix promoting joint destruction. The sublining layer contains proliferating blood vessels and is invaded by inflammatory cells such as lymphocytes, plasma cells, and macrophages (Bustamante et al., 2017).

1.8 Signaling pathways involved in RA

The pathogenesis of RA progression is mediated by mediators discussed below which activates different cellular pathways. These mediators interact with the membrane receptors and initiate signaling cascades. After ligand-receptor interaction, the upstream stimulus propagates through a series of coupled reactions from the plasma membrane to the cytoplasm, to regulate key

factors that are responsible for different cellular phenotypes. RA includes the following key upstream stimuli:

Cytokines and Chemokines: Cytokines and chemokines are small proteins produced by cells for regulating specific biological functions that regulate and determine the nature of immune responses while controlling immune cell trafficking. Cytokines are the general category of the messenger/ signalling molecules while chemokines, a subcategory of cytokines, play the role of chemo-attractants, calling other cells to migrate to the sites of infection/inflammation. Cytokines like TNF- α , IL6 and chemokines like CCL2, CXCL10 to list a few, are implicated in various phases of RA pathogenesis by promoting autoimmunity, initiating and maintaining chronic inflammatory synovitis and driving cartilage and bone destruction (Hwang & Kim, 2017; Kalliolias & Ivashkiv, 2016; Noack & Miossec, 2017);

Growth factors: such as epidermal growth factor (EGF), fibroblast growth factor (FGF), insulin-like growth factor (IGF), vascular endothelial growth factor (VEGF), platelet-derived growth factor (PDGF) activate intracellular signaling pathways (such as PI3K-AKT pathway) and regulate a broad range of cellular functions like cell growth, survival, cell motility and apoptosis (Malemud, 2007; Song et al., 2005);

TLRs: primarily expressed in synovial fibroblasts and infiltrating myeloid cells in human RA joints. Activation of TLR-2, TLR-4, TLR-5 results in recruitment of adaptor molecules such as MyD88, IRAK, TRAF6, and TANK-binding kinase (TBK)-1 and leads to the activation of MAPKs and NF- κ B and the increased expression of various pro-inflammatory and tissue-destructive mediators (such as TNF, IL-6, chemokines and MMPs) (M.-L. Cho et al., 2007; Elshabrawy et al., 2017; Kim et al., 2014; Xu et al., 2013; W. Zhu et al., 2011).

The activation of these upstream components leads to the activation of downstream pathways that include:

1) **the JAK-STAT pathway:** this is an effective target in RA therapy. Many cytokines, including IL-6 and TNF, which are validated therapeutic targets in RA, activate directly (for example IL-6) or indirectly (for example TNF) this pathway by phosphorylating JAK proteins. JAKs in turn phosphorylate STATs, which then dimerize and translocate to the nucleus and bind to regulatory elements of DNA modulating the expression of target genes (Ivashkiv & Hu, 2003; Schinnerling et al., 2017). Activation of JAK-STAT pathway also results in the activation of suppression of cytokine signalling (SOCS), which operates as a feedback inhibitory loop aiming to terminate excessive activation of JAK-STAT (Malemud, 2017).

2) **the NF- κ B pathway:** it is involved in inflammation, cell survival, and proliferation. Activated NF κ B is detected in immune cells (such as macrophages and lymphocytes) as well as in stromal cells (such as FLS and endothelial cells) and stimulates the transcription of arthritogenic mediators like IL-1, TNF, IL-12, RANKL, PTGS2 and IL-6 in RA synovium. TNF, IL-1, and RANKL are key upstream RA-relevant triggers of the activation of the NF- κ B pathway (Han et al., 1998).

3) **the MAPK pathway:** All the three classes of MAPKs, namely ERK, JNK and p38, are found to be expressed and activated in synovial tissue in RA (25) . A series of cytokines including among others TNF, IL-1, and IL-6 activate ERK, JNK and p38 MAPK in synovial tissue with successive induction of proinflammatory mediators such as cytokines and tissue destructive enzymes (e.g., MMP-1 and MMP-13). Negative feedbacks are required to keep in check the constitutive activation of MAPK proteins in order to control the excessive prolonged expression of pro-inflammatory genes (Clark & Dean, 2012; Schett et al., 2008).

4) **the PI3K-AKT pathway:** Growth factors like vascular endothelial growth factor (VEGF) and fibroblast growth factor (FGF) induce PI3K-AKT pathway. Activated cellular AKT regulates immune cells, and survival of synoviocytes and chondrocytes by phosphorylating several downstream signalling proteins modulating mTOR, Bad, FOXO3 and tumour protein-73 (TP-73) (Higgs, 2010; Malesud, 2013; Song et al., 2005).

1.9 Current RA treatment

RA involves yet unknown numbers of molecules and their alterations resulting in complete loss of mobility. Thus, extensive research is required into the treatment of the disease with the identification of new therapies. The common therapy includes the use of Disease-modifying anti-rheumatic drugs (DMARDs), which are immunosuppressive and immunomodulatory drugs indicated for the treatment of inflammatory arthritis including RA. DMARDs are categorized into conventional DMARDs or biologic DMARDs, both differing in their mode of action (Benjamin et al., 2020).

Conventional DMARDs include drugs that target the entire immune system whereas biologic DMARDs are monoclonal antibodies (mAbs) and soluble receptors that target protein messenger molecules or cells. Commonly used conventional DMARDs include methotrexate (MTX), hydroxychloroquine (HCQ) and while biologics DMARDs include infliximab (INX), adalimumab (ADM) and etanercept (ETC) targeting tumor necrosis factor (TNF- α or TNF- α), rituximab (RIX) targeting CD20 expressed on the surface of B-cells, abatacept (ABC)

downregulating T cells, among others (29939640). Glucocorticoids are also used in low doses when methotrexate (or another conventional synthetic DMARD) should have reached full effectiveness (Smolen et al., 2018).

Many above-mentioned drugs cause one or another adverse effects in the body which could be dermatological, gastroenterological, haematological, respiratory among many others. MTX, the most common DMARDs used to treat RA can cause liver, lung and kidney damage as well as strong immunosuppression and the most advanced therapies that target focused pathways (anti-TNF α) are extremely costly (Benjamin et al., 2020). Long-term use of glucocorticoids can also cause many harmful effects, such as skin atrophy, osteoporosis, disturbed glucose tolerance, hypertension, cataract development and a higher risk of infections.

There are also non responders who either show no adequate response from a certain drug (primary non-responders) or whose response diminishes after the production of anti drug antibodies (secondary non-responders). The use of a molecule with same mode of action but different immunogenicity often works well however, more studies are being undertaken especially regarding primary non responders (Smolen et al., 2018).

1.10 Systems biology for complex diseases

Complex disease mechanisms in humans involve numerous interconnected biological processes, signaling pathways resulting in certain gene expression and cellular phenotypes.

Biomolecules in living systems do not function individually but rather interact with one another forming biomolecular networks. For instance, a disease is rarely a consequence of an abnormality in a single gene but reflects the perturbations or malfunctions of the complex biological networks that link tissues and organ systems (Barabási et al., 2011). In order to organize and analyse this complexity, there is an urgent need to develop new methodologies to gain insights through a more holistic view. Until now human disease assessments were generally based on statistics and mutation correlation which do not provide mechanistic details of the processes (Schrodi et al., 2014).

Systems biology is the systems-level study, which deciphers the complexity of the disease, its fundamental principle of initiation and progression. It is considered to be a powerful analytical approach, viewing a living organism as an interacting and dynamical network of proteins, genes and biochemical reactions. One of the common initial approaches in this regard is the construction of biological networks by exploiting the ever-expanding experimental data. The

availability of large network datasets and the affordable computing capability has driven the development of bioinformatics algorithms and computational biology approaches to analyze data to offer biological insights. Studying the biological systems from the network perspective has attracted much attention recently (Barabási & Oltvai, 2004; Hu et al., 2016; Vidal et al., 2011) in the form of linkage maps among proteins (genes or neurons), their associated phenotypes (diseases), and the corresponding environmental factors (drugs) (Ideker & Nussinov, 2017).

1.10.1 SBGN standards and networks representation

The systems biology graphical notation (SBGN)(Le Novère et al., 2009)is a standard for the visual representation of biological/biochemical processes as networks (Le Novère et al., 2009). There are three types of SBGN levels covering different granularity of biological processes namely - Process Description (PD), Entity-Relationship (ER) and Activity Flow (AF) (Le Novère, 2015) (**Fig 1.5**).

- a) Process descriptions (PD): directed, sequential representation of the mechanistic details of the underlying process; often used in the construction of metabolic pathways and signalling networks. PD diagrams have been standardized with the SBGN process description language (Le Novère, 2015; Moodie et al., 2015)
- b) Activity flows: influence graphs representing the flow of information concisely often signalling pathways or gene regulatory networks. They have been standardized with the Systems Biology Graphical Notation (SBGN) activity flow language. Because of the qualitative nature of the information provided, activity flows are the natural representations for qualitative models and, in particular, for logic models (Le Novère, 2015; Mi et al., 2015)
- c) Entity relationships: represents entities, statements about those entities and the influence of entities on statements (Le Novère, 2015). They have been standardized using the SBGN entity relationship language (Le Novère et al., 2009), and such maps have been constructed to represent molecular events underlying, for instance, the cell cycle (Tozluoğlu et al., 2008) and apoptosis (Kohn, 1999)

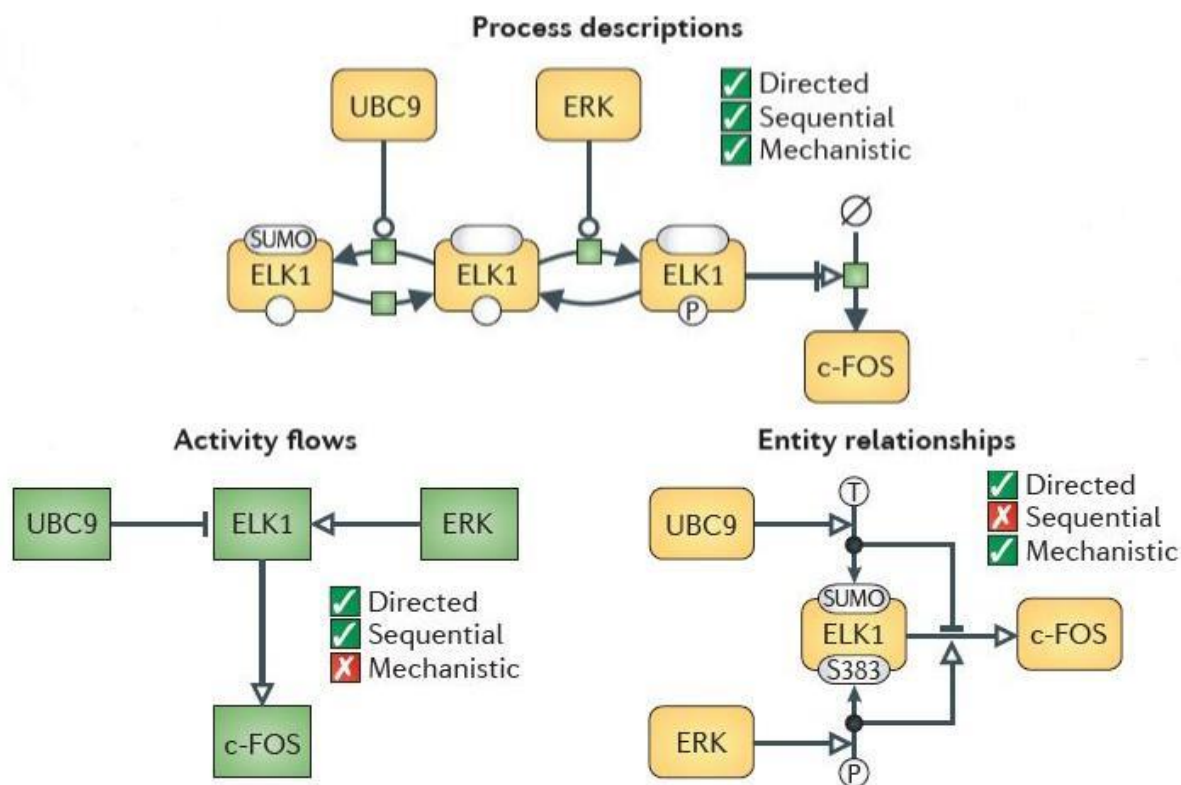


Figure 1.5 : Three different types of SBGN networks used to represent biological processes (taken from (Le Novère, 2015))

1.10.2 MIRIAM and NOTE

The Minimal Information Required In the Annotation of biochemical Models (MIRIAM) is a set of guidelines for the proper annotation and curation processes of computational models (Le Novère et al., 2005) whereas the NOTE allows you to type in additional text information for the Components (http://www.celldesigner.org/help/CDH_NotesMiriam_01.html) (**Fig 1.6**). The proposed encoding and annotating with MIRIAM of the computational models are in a machine-readable format which facilitates their exchange and that can be unambiguously parsed by software to perform simulations and analysis. MIRIAM is divided into two parts. The first is a proposed standard for model qualifiers which represents relation between a model component and the resource used to annotate it (**Fig 1.6b**), whereas the second is a proposed annotation scheme that specifies the documentation of the model by external knowledge (Le Novère et al., 2005) (**Fig 1.6c**). Here is a screenshot of MIRIAM annotation in the CellDesigner tool (Kitano et al., 2005).

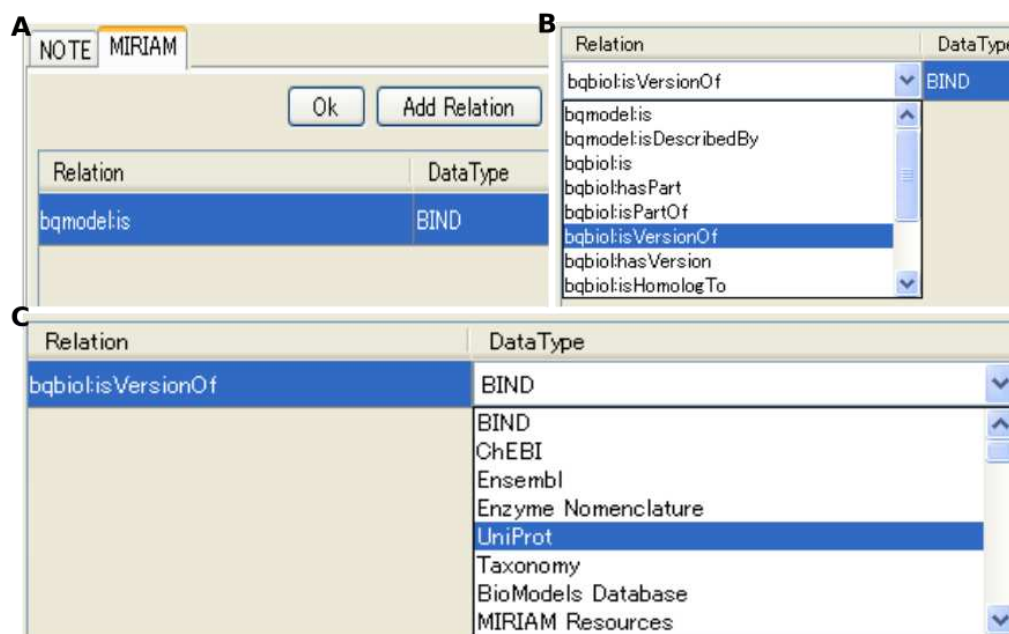


Figure 1.6: **CellDesigner Annotation section.** **A)** MIRIAM and NOTE section in CellDesigner **B)** Bio models qualifiers representing relation between a model component and the resource used to annotate it **C)** Different data types used to annotate the components.

Common objectives in disease study include the identification of disease biomarkers, molecular mechanisms, potential drug targets and disease subtypes for better diagnostics and stratification of patients. Toward this direction, initiatives have emerged, such as The Cancer Cell Map Initiative (Krogan et al., 2015), the Atlas of Cancer Signaling Networks (<http://acsncurie.fr>) concerning cancer, and the Disease Maps Project (<http://disease-maps.org>), an open, large-scale community effort that consists of a network of groups working together for developing best practices, standards and tools in order to better represent disease mechanisms. Disease maps describe disease-related signalling, metabolic and gene regulatory processes with evidence of their relationships to pathophysiological causes and outcomes (Mazein, Ostaszewski, et al., 2018). Some examples of such published disease maps are Rheumatoid arthritis (Wu et al., 2010), Alzheimer's pathway (Mizuno et al., 2012), Parkinson's (Fujita et al., 2014), Asthma (Mazein, Knowles, et al., 2018), Atherosclerosis (Parton et al., 2019) and so on. Disease maps are not only the encyclopedia of the molecular systems involved in the disease but also act as a source for the construction of mathematical models to capture the dynamics of otherwise complex systems. Disease maps provide a comprehensive template for

visualization and analysis of omics datasets, and can also be analyzed in terms of the underlying network structure. However, their static nature cannot account for the coordination of multiple biological processes, or how the regulation of several nodes due to the presence or absence of certain factors can alter the functional outcome (i.e. activation of a particular pathway following the repression of a given factor).

1.11 Computational Modelling in biology

Computational modelling can be used to provide an executable, dynamic network that can reveal hidden properties and account for emerging system-level behaviours through *in silico* simulations and perturbations. A cycle of model construction, simulations and experimental validation of model predictions can help in the current diagnostic and therapeutic approaches to medicine (Kriete & Eils, 2013).

Mathematical models construction involves building the structure of the model, choosing mathematical expressions to characterize the relationships between its components, finding parameter values and initial conditions, and performing numerical simulations and other mathematical analyses that can both reproduce observations and lead to predictions. Many modelling methods based on time representation and variable value are being developed to model and simulate molecular and gene networks (**Fig 1.7**) and could be categorized generally into Quantitative and Qualitative approaches. Quantitative models are based on the application of systems theory to chemical kinetics and have been used to describe metabolic networks (Chance et al., 1952; Joshi & Palsson, 1989; Savageau, 1970), signalling pathways (Bray et al., 1993; Goldbeter & Koshland, 1981) and gene regulation (Arkin et al., 1998; Elowitz & Leibler, 2000; von Dassow et al., 2000). Qualitative logical modelling is based on the idea that a variable can take a discrete number of states or values (two in the case of Boolean models) and that the state of a variable is decided by a logical combination of the states of other variables. The system can be updated synchronously, with the values of all variables being calculated after a transition, or asynchronously, when variables undergo transitions one at a time.

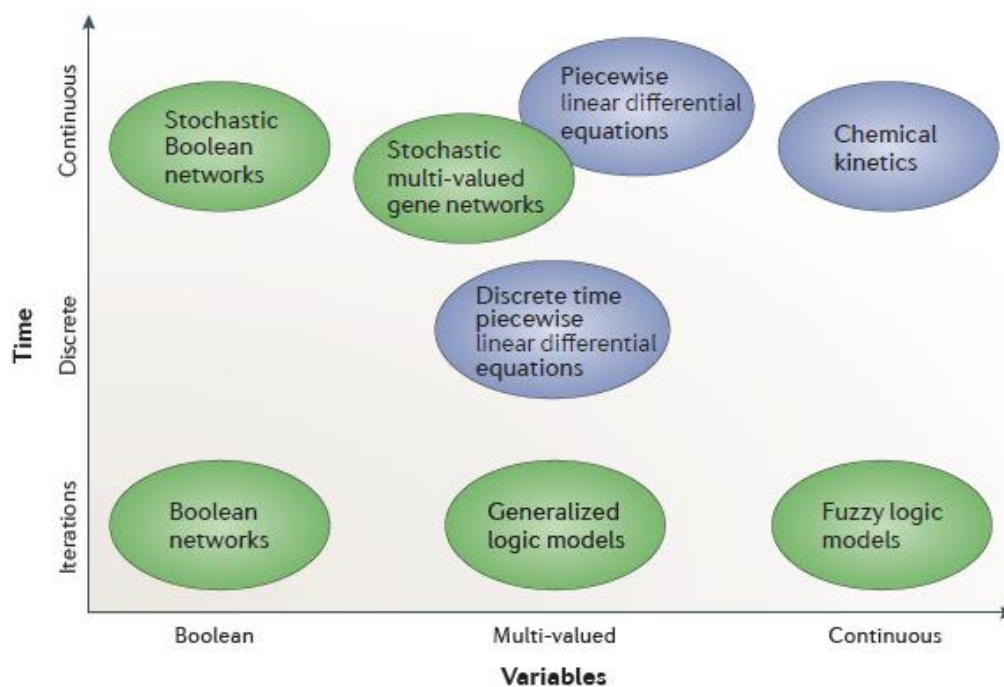


Figure 1.7: Various modelling approaches based on time representation and variable values (taken from (Le Novère, 2015)).

Despite quantitative kinetic modeling providing most precise predictions, the lack of kinetic parameters hampers its use in many situations which is approached by other approaches like logical modelling. In a logical model, each component is associated with a discrete variable, which is a logical (often Boolean i.e., binary) abstraction of its level of activity (or concentration). This framework relies on its scalability (logical models with few hundreds components have been simulated) along with its qualitative nature, as kinetic parameters and other precise knowledge about the molecular mechanisms are not required (Abou-Jaoudé et al., 2016).

A Boolean model provides a qualitative representation of a system. Each boolean variable can assume only two possible values denoted by 1 (ON) or 0 (OFF) corresponding to the logic values FALSE and TRUE. ON and OFF represents the state of a biological entity corresponding to the binary variable. This indicates if a gene is expressed or not expressed, a transcription factor is active or inactive, and a molecule's concentration is above or below a certain threshold. In Boolean models, the future state of a node is determined based on a logic statement on the current states of its regulators. This statement, called a Boolean rule (function), is usually expressed via the logic operators AND, OR, and NOT. Boolean network models can be projected like a directed graph where the nodes A, B and C correspond to the Boolean

variables connected through edges (**Fig 1.8**). Each edge has a sign implying whether the input node has a positive or negative effect on the target node (R.-S. Wang et al., 2012) . Model dynamics are conveniently represented in terms of State Transition Graphs (STG), where nodes denote states, while directed edges represent state transitions.

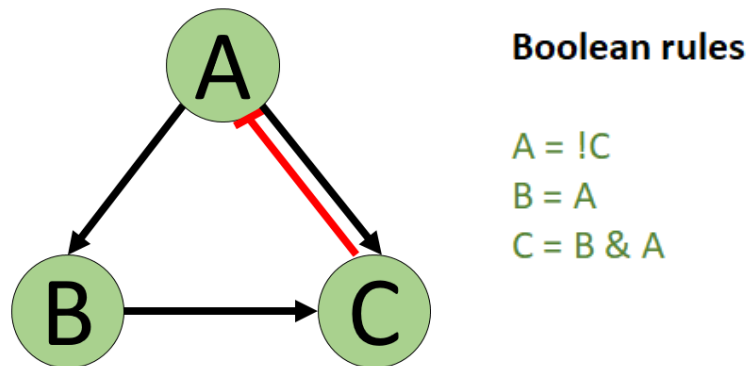


Figure 1.8: A simple Boolean network model: A simple regulatory network containing three nodes A, B and C is shown. The edges with sharp arrows represent activation (positive effects) while the edges with blunt arrows indicate inhibition (negative effects).

1.12 Objectives

RA's aetiology is still unknown due to complex interaction between biological entities including the signaling crosstalks between the cells of the immune system and the joints. To understand the mechanistic details of the underlying biological processes in this complex disease RA, we need to systematically assemble and analyse the available information from literature/datasets/databases. The aim of this PhD work is to shed light on molecular mechanisms involved in the pathogenesis of RA using state-of-the-art systems biology approaches, tools and techniques.

- **The first main objective** of the work is to summarize current biological knowledge (gene expression, signaling pathways, cellular phenotypes) concerning RA into a detailed molecular interaction map based on data available in the literature, public repositories and pathway databases. The construction of a state of the art RA map requires exhaustive literature scanning, re-evaluation of previously published attempts along with strict curation criteria and most importantly experts' advice in order to limit false positives focussing on disease and human specificity. Molecular interaction maps can serve as a stand-alone knowledge base, or they can be used as a scaffold for building

computational models. Based on information mining, human curation and expert advice, these maps summarize current knowledge about biological pathways in a process description representation, while accounting for as many mechanistic details as possible. An important aspect is also the use of community standards such as SBGN, which is a standard graphical representation intended to foster the efficient storage, exchange and reuse of information about signaling pathways, metabolic networks, and gene regulatory networks amongst communities of biochemists, biologists, and theoreticians. This is the first attempt to construct a molecular interaction map for RA using human and disease specific data, and utilizing the SBGN standard.

- **The second main objective** of the project is to develop a computational modelling approach to provide an executable, dynamic network that can reveal hidden properties and account for emerging system-level behaviours through *in silico* simulations and perturbations. Amongst the different cell types involved in RA, synovial fibroblasts' play a crucial role in driving the persistent, destructive characteristics of the disease. They are shown to express immune-modulating cytokines, various adhesion molecules, and matrix-modelling enzymes. Moreover, RA FLS display high proliferative rates and an apoptosis-resistant phenotype. These cells can also behave as primary drivers of inflammation, and RA FLS-directed therapies could become a complementary approach to immune-directed therapies (Huber et al., 2006; Turner & Filer, 2015).

We aim at building a large scale model to study the behaviour RA FLS in order to gain better insights on the mechanisms that control their aggressive phenotype (apoptosis-resistant, migration, high rate of cell proliferation etc). This is to our knowledge the first attempt to build a large-scale dynamic model for the study of RA FLS.

- **The third main objective** of the project is an effort to automatize the passage from a static representation of disease mechanisms to a dynamical model developing a framework for graph conversion and inference of logical formulae based on topology and semantics encoded on the molecular map. Construction of an executable dynamic model either from pre-existing molecular maps/ models or standalone is a time-consuming process. The proposed framework could considerably ease the process of generation of executable models ready-to-use for *in silico* simulations (perturbations/knockouts). Our model is the first large-scale Boolean model to be constructed in a fully automated way from a molecular interaction map. In addition, this also paves the way for scaling up the automatic generation of large-scale models.

Until now, there had been limitations with the number of nodes (elements-proteins/genes) that could be included in a dynamic, biological model. This was in accordance with the available modeling tools with limited simulating potential. In this thesis we would like to challenge this potential and increase the scale of the models routinely used to study biological networks, in an effort to elucidate complex regulatory mechanisms and give answers to more sophisticated questions that could not be addressed with models comprising a limited number of factors.

Chapter 2. RA map

2.1 Introduction

A comprehensive molecular interaction map for RA built with the software CellDesigner (Matsuoka et al., 2014) was published in 2010 (Wu et al., 2010). It was based on high throughput experimental data combined with information derived from the KEGG pathway database (Kanehisa, 2009) (<http://www.genome.jp/kegg/pathway.html>). Twenty-eight published studies were used for the construction of the map that included experiments performed in different cell types and tissues such as the peripheral blood mononuclear cell (PBMC), synovial fibroblasts, synovial tissues and macrophages among others. We used this RA map as a basis and extended it to create a state of the art knowledge-based map for RA. This chapter comprises three parts. In the first part, we present the process of constructing the RA map, highlighting the most critical pathways. In the second part, we transform the RA map into a state-of-the-art interactive knowledge base for the disease, which interfaces with various databases for content annotation and enrichment analysis of experimental results. In the third part, we use bioinformatics tools such as BioInfoMiner (Lhomond et al., 2018) (29311133) (<https://bioinforminer.com>) and Cytoscape (Shannon et al., 2003) (14597658) for the analysis of the RA map as a complex biological network, revealing topological and functional aspects of the map (**Fig 2.1**).

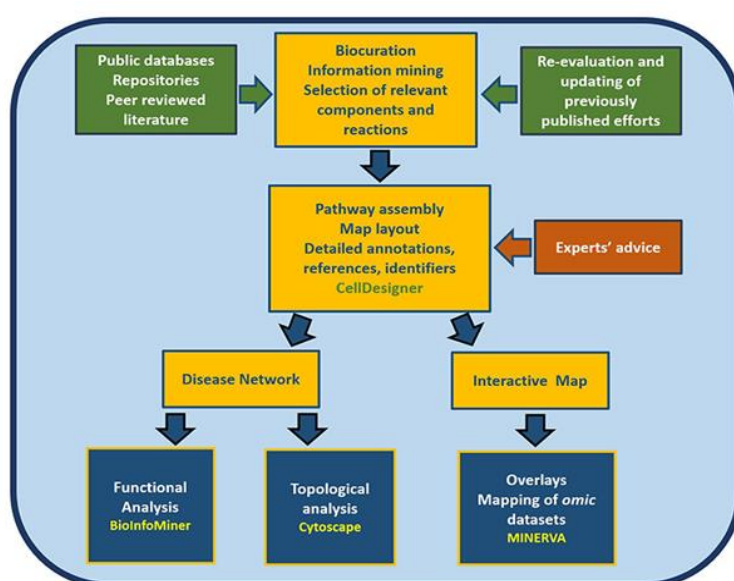


Figure 2.1: Workflow for the construction and use of the RA map.

2.2 Construction of the RA map

CellDesigner (Matsuoka et al., 2014) is a structured diagram editor for the construction of gene regulatory and biochemical reaction networks. In a CellDesigner diagram, nodes represent chemical species like proteins, genes, complexes, and other molecules and the edges denote the interaction between the nodes, which can be activation, inhibition, catalysis and state transition among many other possible interactions (Kitano et al., 2005; P. Wang et al., 2014). CellDesigner supports graphical notation and listing of the symbols based on the proposal by Kitano and adopting process description diagram (PD) compliant to the guidelines of the Systems Biology Graphical Notation (SBGN; <http://sbgn.org>) (Le Novère et al., 2009) (19668183). Current graphical notation of CellDesigner is shown in **Fig 2.2**. PD is closest to metabolic and regulatory pathways found in biological literature and textbooks offering a well-defined semantics for a superior precision in expressing biological knowledge. PD represents mechanistic and temporal dependencies of biological interactions and transformations as a graph. Its different types of nodes include entities (e.g. metabolites, proteins, genes and complexes) and processes (e.g. reactions, associations and influences). The edges describe relationships between the nodes (e.g. consumption, production, stimulation and inhibition) (Rougnny et al., 2019). A simple PD example of MAPK pathway is shown in **Fig 2.3**

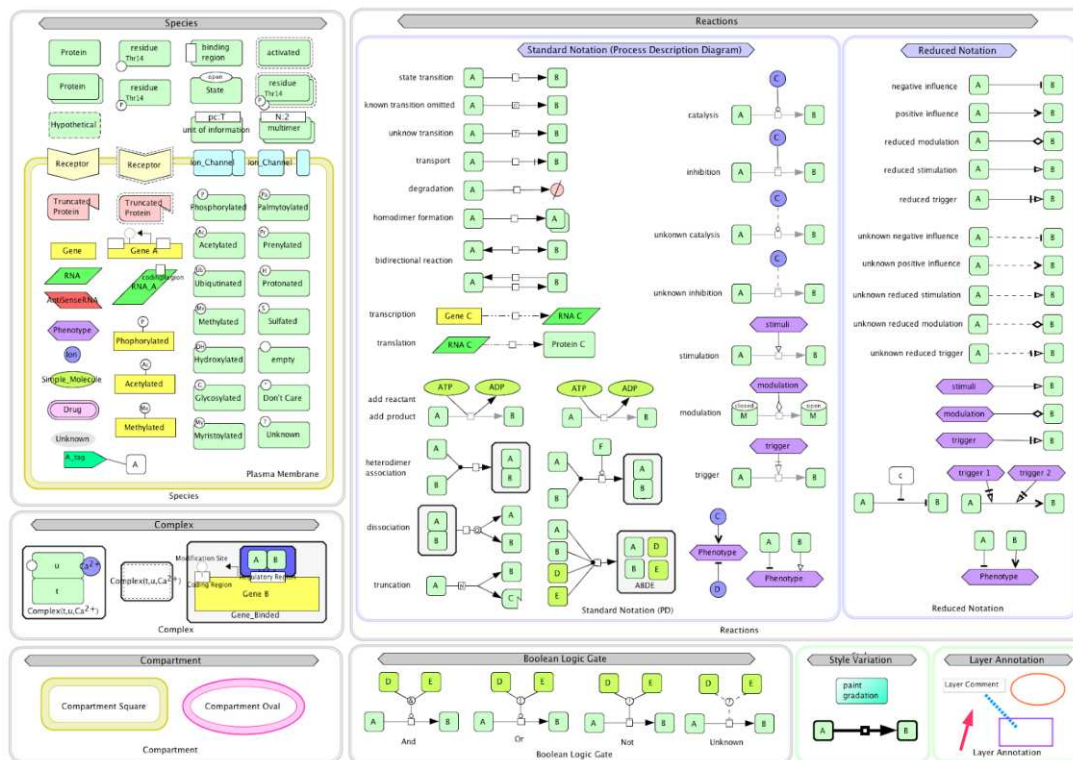


Figure 2.2: Current Celldesigner graphical notations
(<http://celldesigner.org/help/images/components42.png>).

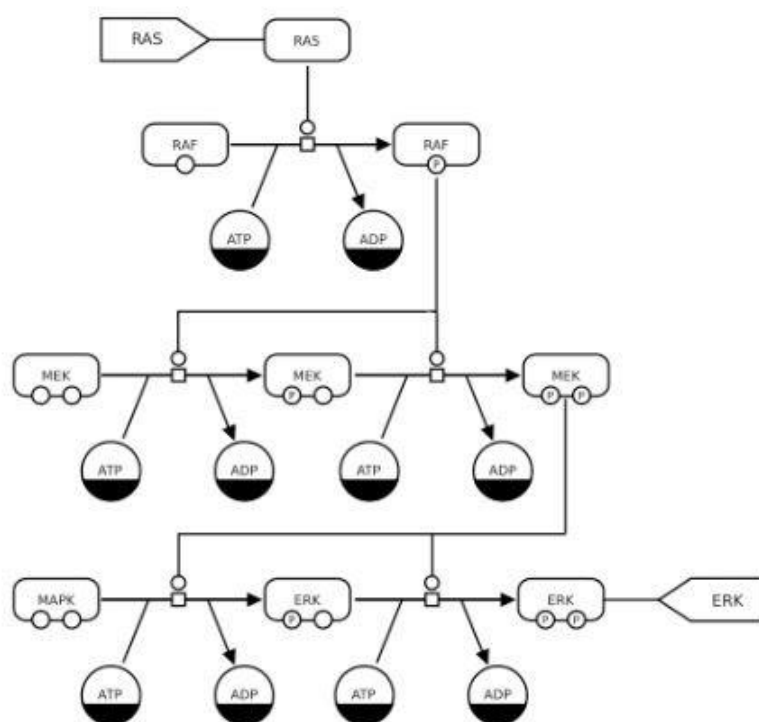


Figure 2.3: Example of an SBGN Process Description uses two kinds of nodes (that is, nodes representing biochemically-indistinguishable entities such as molecules). One kind of node

whose glyph is a rounded rectangle represents different macromolecules, and the other whose glyph is a circle represents pools of simple chemicals (<https://sbgn.github.io/examples>).

The models constructed in CellDesigner are stored in Systems Biology Markup Language (SBML, <http://sbml.org>) format (Hucka et al., 2003), a standard for representing models of biochemical and gene-regulatory networks. The new updated version CellDesigner 4.4 was used for the construction of the RA map.

2.2.1 Annotation and curation criteria

We focussed on studies based on cells, fluids, and tissues of human origin only using small-scale experiments, in an attempt to limit false positives from gene expression data used to construct the first RA map. With small-scale, we refer to experiments that are not high throughput, such as ELISA, Western blot etc. In this direction, all components and reactions of the previous RA map were carefully evaluated and categorized. Annotations concerning experimental validation with small-scale experiments were added to all the components where possible. Molecules, for which small-scale experiments were not found, were kept if they appeared in at least two high throughput studies. Molecules that failed to fulfil these criteria were removed from the map.

An exhaustive literature search was carried out for new proteins, genes and other molecules involved in the pathogenesis of RA since 2010. Relevant keywords and key phrases like ‘Pathogenesis of RA’, ‘Cytokines in the pathogenesis of RA’, ‘Therapeutic targets in RA’ among many others were used to filter the literature abstracts and studies in PubMed and Google Scholar. Along with it, peer review articles concerning RA and their bibliography were searched, and information was mined. New RA mediators were added to the RA map and referenced with at least two PubMed IDs. However, this criterion got restricted for few molecules published only very recently in 2018/2019.

Annotations were added for all the components (proteins, RNAs and genes) and reactions present in the CellDesigner XML file using the sections text NOTE and Minimal Information Requested In the Annotation of Models (MIRIAM) (Le Novère et al., 2005), which are human and machine-readable formats respectively. MIRIAM comprises the set of guidelines for the consistent annotation and curation of computational models in biology. In the MIRIAM segment, we added PubMed IDs for different cell types with the tag “bqbiol: isDescribedBy”

(Fig 2.4). In the NOTE section, we added text information about KEGG pathway identifiers used to cross-validate interactions.

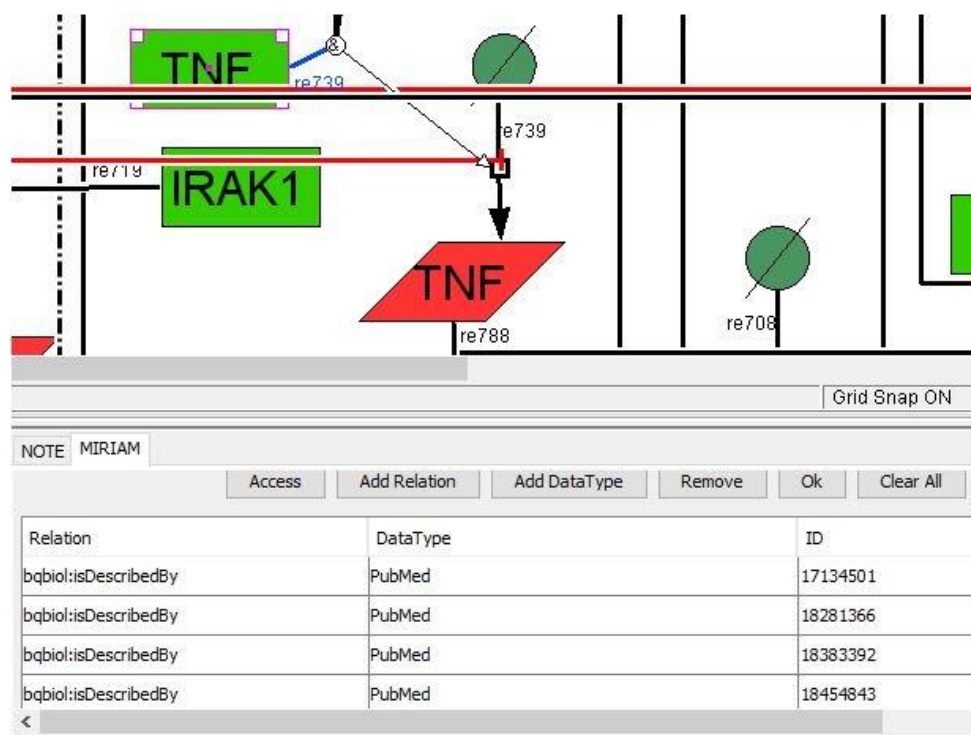


Figure 2.4: Annotations added to the MIRIAM section of CellDesigner.

2.2.2 Evaluation of components and reactions

All components and reactions of the previous RA map were carefully evaluated and categorized. Annotations concerning experimental validation with small-scale experiments were added where possible. Molecules, for which small-scale experiments were not found, were kept if they appeared in at least two high throughput studies. Molecules that failed to fulfill these criteria were removed from the map. In order to connect and add interactions among the map components, pathway databases like KEGG and Ingenuity Pathway Analysis (IPA) were used.

2.2.3 Compartments, structure, and layout

CellDesigner plugin Relay Layout Model (<http://www.celldesigner.org/plugins.html>) was used to improve the layout of the molecular map. The RA map was restructured to follow a cell layout with surrounding extracellular space, the cytoplasmic area containing organelles, proteins and

small molecules, the nucleus with gene-regulatory mechanisms, secreted molecules and cellular phenotypes (**Fig 2.5**).

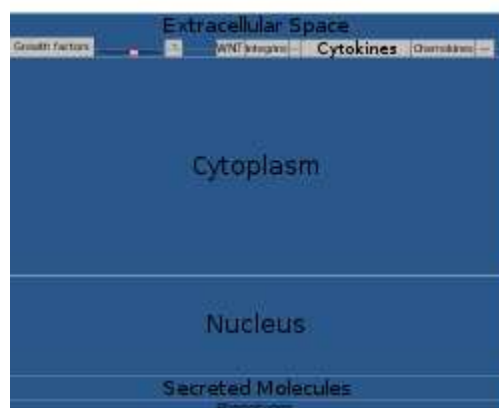


Figure 2.5: Cell layout of the RA map.

Receptors were grouped by category (growth factors, cytokines, chemokines, integrins, toll-like receptors). The RA map includes six compartments, namely extracellular space, plasma membrane, cytoplasm (including Golgi apparatus, endoplasmic reticulum, and mitochondria), nucleus, secreted molecules and phenotypes. Extracellular space includes the protein ligands outside the cell that can form a complex with the plasma membrane receptors and proteins resulting in the activation of several signalling cascades. Cytoplasm compartment includes the proteins (signaling proteins, transcription factors (TFs)) and small molecules. TFs are subsequently transported to the nucleus where they regulate gene expression. The nucleus compartment includes transcription factors transported from the cytoplasm, genes, and RNAs (miRNA and mRNA). A separate compartment contains proteins secreted out of the cell, and finally, a dedicated compartment contains cellular phenotypes relevant for RA. We used a distinct colour code for the components in the RA map: plasma membrane receptors in peach, proteins in purple, genes in green, RNAs in red and phenotypes in Yellow. Inhibition arrows/edges are represented in red colour, while for all others like state transition, catalysis, transport, reduced physical stimulation and heterodimer association are in black colour.

2.2.4 Experts' advice and feedback

Experts' curation is critical in order to reconstruct molecular and cellular interactions from the available literature. Due to the complexity of RA regarding cell types (macrophages, endothelial cells, synovial fibroblasts), mediators of inflammation (cytokines, chemokines, growth factors) and the variety of biological processes implicated in the disease, the review of

the map by RA experts was important for an accurate representation of disease hallmarks. In order to provide a systematic and comprehensive molecular map we used SBGN standards and a cell layout. We took advice from experienced scientists in both biological and computational domains to make the content comprehensive and functional for different types of users such as experimental biologists, clinicians, computational modellers, and bioinformaticians. The RA map layout, the representation of various levels of information and the validity of molecules and pathways included in the RA map, were carefully examined in this context.

2.2.5 SBGN standards and process description (PD) map validation

The systems biology graphical notation (SBGN) is a standard for the visual representation of biological/biochemical processes as networks as discussed in Chapter 1, section 1.8 (Le Novère et al., 2009). There are four types of SBGN levels covering different granularity of biological processes namely - Interaction Network (IN), Process Description (PD), Entity Relationship (ER) and Activity Flow (AF) (Le Novère, 2015). The RA map is a PD map showing the detailed biological processes implicated in RA. The RA map was systematically analysed for compliance with SBGN. All non-compliant reactions and complexes were transformed by SBGN standards. In this representation, genes act as part of an enhancer complex along with transcription factors for the activation of RNAs. Gene regulatory mechanisms in the nucleus compartment were also modified with the addition of new null nodes as a source (e.g., nucleotides) leading to the production of RNAs (**Fig 2.6**).

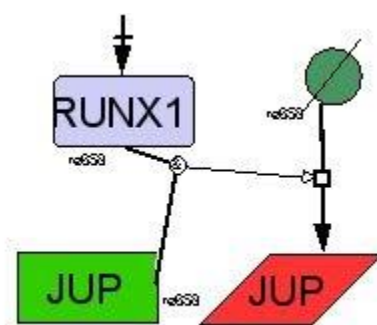


Figure 2.6: Gene regulatory representation according to SBGN standards.

VANTED (Visualisation and Analysis of Networks containing Experimental Data) (Rohn et al., 2012), is a framework for systems biology applications with functionalities ranging from network reconstruction, data visualisation, integration of various data types to network simulation using systems biology standards for visualisation and data exchange. We used

SBGN-ED (an add-on for VANTED for editing, validating and translating SBGN maps) (Czauderna et al., 2010) to validate our SBGN-PD encoding of the RA map. As this tool works with SBGN-ML file format, we utilised the CellDesigner to SBGN converter (<https://royludo.github.io/cd2sbgnml>) for converting the CellDesigner XML file into SBGN-ML format which was then imported in VANTED for further analysis.

2.3 RA map

2.3.1 A comprehensive molecular interaction map for Rheumatoid Arthritis (RA)

The RA map graphically illustrates signalling pathways, gene expression regulation, molecular mechanisms, and cellular phenotypes involved in the pathogenesis of the disease. The RA map involves exhaustive literature curation, information mining from relevant databases along with continuous updating and advice from domain experts. Importantly, the interactions shown in the diagram represent a graphical model encoded using a standardized format, making the map computationally tractable. All the components in the map have at least two manually curated PubMed references, giving overall 353 publications covering a time span from 1975 to 2019 (Fig 2.7).

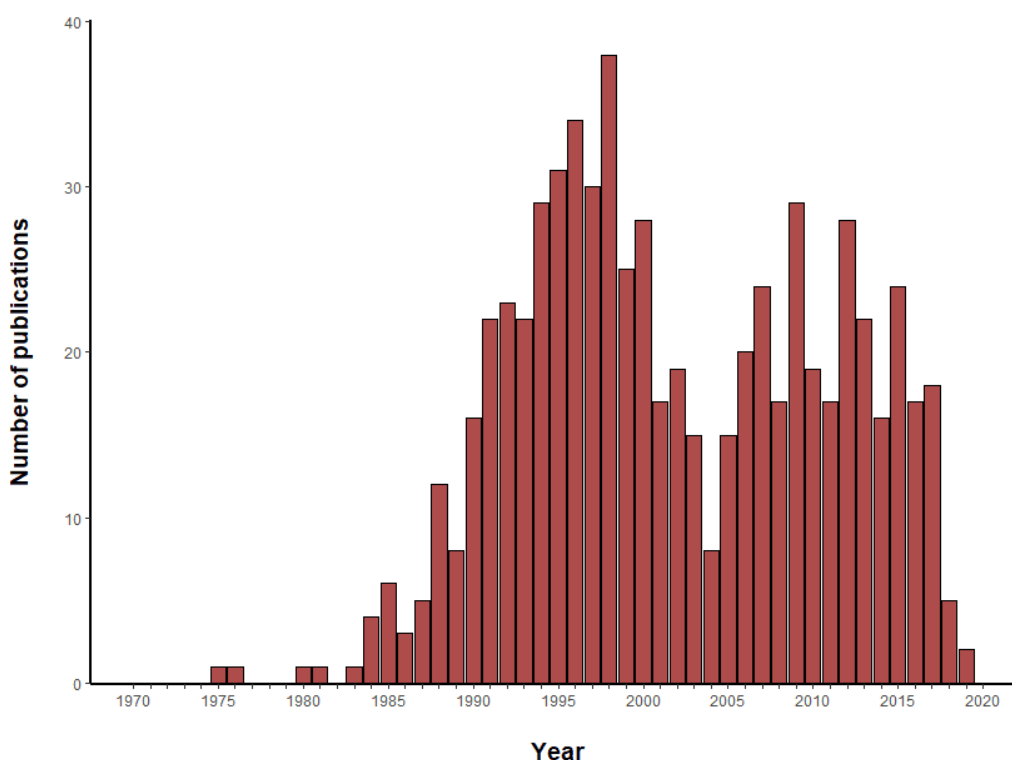


Figure 2.7: RA map references. Barplot showing the yearly distribution of peer-reviewed scientific articles and reviews included in the RA map

An overview of the RA map is shown in **Fig 2.8**. The RA map is fully SBGN compliant and features 506 species, 449 reactions and 8 cellular phenotypes. The biomolecules in the map are 303 proteins, 61 molecular complexes, 106 genes, 106 RNA entities, two ions and 7 simple molecules like cAMP, H₂O₂, PIP₃. Proteins include extracellular proteins, membrane proteins, and cytoplasmic proteins both as signalling and transcription factors. The reactions are classified as state transitions, catalyses, inhibitions, transports, heterodimer associations, dissociations, boolean AND gates and reduced physical stimulations. The RA map is organised in the form of a cell representing the flow of information from the extracellular space (ligands) to the plasma membrane (ligand-receptor complexes) and then to the cytoplasm (signalling pathways), the nucleus (gene regulation) and the secreted compartment or cellular phenotypes.

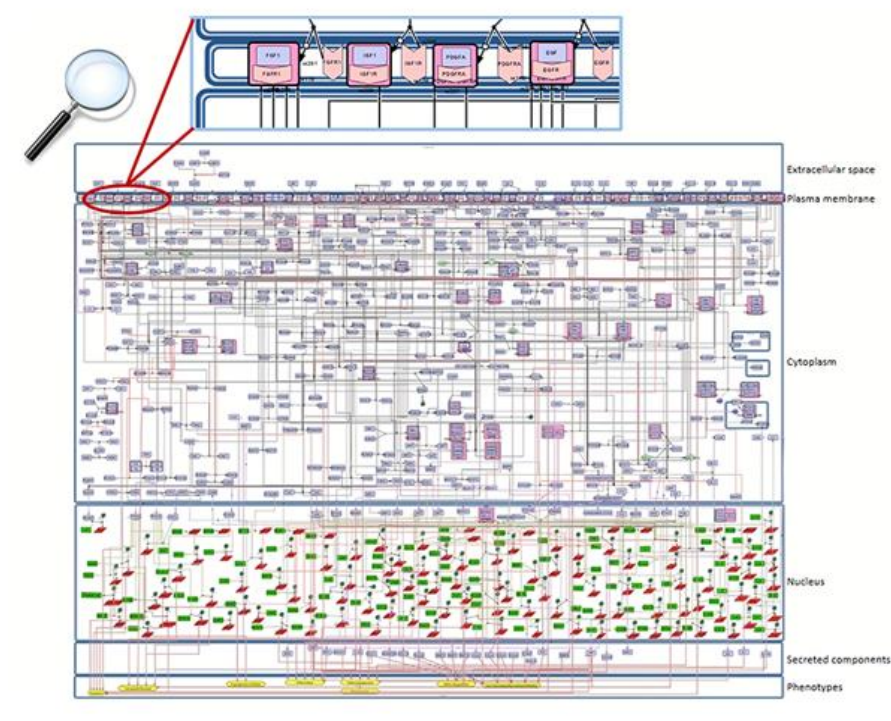


Figure 2.8: Snapshot of the SBGN compliant RA map. The map is colour-coded with proteins in purple, genes in green, RNAs in red and phenotypes in yellow. State transitions and catalysis reactions are displayed in black and the inhibitions are in red. Compartments are distinguished as bounding boxes. The map was built using CellDesigner, version 4.4 (16).

2.3.2 Molecular pathways covered in the RA map

The RA map contains the key cellular and molecular pathways that according to the published literature it has been suggested to be involved in disease pathogenesis. In signaling cascades, activation occurs as a response to an upstream stimulus. After activation, the signal propagates through a series of coupled reactions from the plasma membrane to the cytoplasm, to regulate key factors that are responsible for different cellular phenotypes. The RA map includes the key upstream stimuli of Cytokines and Chemokines: a diverse group of proteins like Tumor Necrosis Factor (TNF) and Interleukins (ILs); Growth factors like EGF, FGF, IGF, VEGF, PDGF ; TLRs. The activation of these upstream components leads to the activation of downstream pathways that mainly include the JAK-STAT pathway, the MAPK pathway, the NF- κ B pathway and the PI3K-AKT pathway (discussed in detail in Chapter 1, section 1.8)

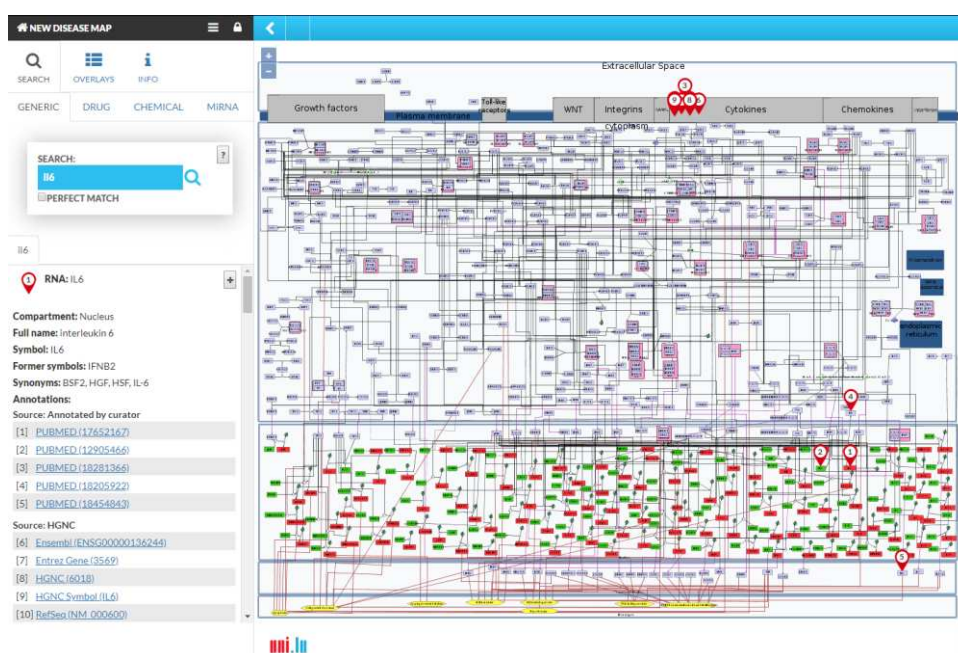
All signalling cascades lead to specific cellular outcomes grouped in eight distinct phenotypes such as apoptosis (G S Firestein et al., 1995; Korb et al., 2009; H. Li & Wan, 2013; Müller-Ladner et al., 2007), cell survival/growth/proliferation (Jacobs et al., 1995; Kramer et al., 2003; Malemud, 2007; Mongan & Jacox, 1964), angiogenesis (Elshabrawy et al., 2015; Malemud, 2007; Müller-Ladner et al., 2007), inflammation (Demoruelle et al., 2014; Malemud, 2007; McInnes & Schett, 2011; Müller-Ladner et al., 2007), matrix degradation (Ainola et al., 2005; Müller-Ladner et al., 2007; Shiozawa et al., 2011; Yasuda, 2006; Yoshihara & Yamada, 2007), cell chemotaxis/recruitment/infiltration (Goddard et al., 1984; Mellado et al., 2015), osteoclastogenesis (Müller-Ladner et al., 2007; Sato & Takayanagi, 2006) and bone erosion (Goldring, 2002; McInnes & Schett, 2011; Müller-Ladner et al., 2007; Panagopoulos & Lambrou, 2018).

2.4 Transforming RA map into a state of the art knowledge base using MINERVA

MINERVA (Molecular Interaction NEtworkRks VisuAlization) is a web service that supports curation, annotation, and visualisation of molecular interaction networks in SBGN-compliant format (Gawron et al., 2016). MINERVA provides automated content annotation and verification, along with mapping of drug targets and overlaying experimental data on the visualised networks. MINERVA integrates functionalities of pathway databases and features a user-centric view for content display. Automated annotations (HGCN) and curator's annotations for every component and reaction are displayed in the left panel, as seen in **Fig 2.9**.

The user can also visualise cell-specific data based on curated overlays or analyse patients *omic* datasets. Moreover, MINERVA provides an interface for interrogating several other databases.

The RA map is available at ramap.elixir-luxembourg.org in the form of an interactive diagram, using the platform MINERVA (**Fig 2.9**). Clicking on a biomolecule in the map, the user can choose to visualize interacting drugs, chemicals, and miRNAs. These interactions are provided by the interfaces to the DrugBank (<https://www.drugbank.ca/>), ChEMBL (<https://www.ebi.ac.uk/chembl/>), CTD (<http://ctdbase.org>) and miRTarBase (<http://mirtarbase.mbc.nctu.edu.tw>).



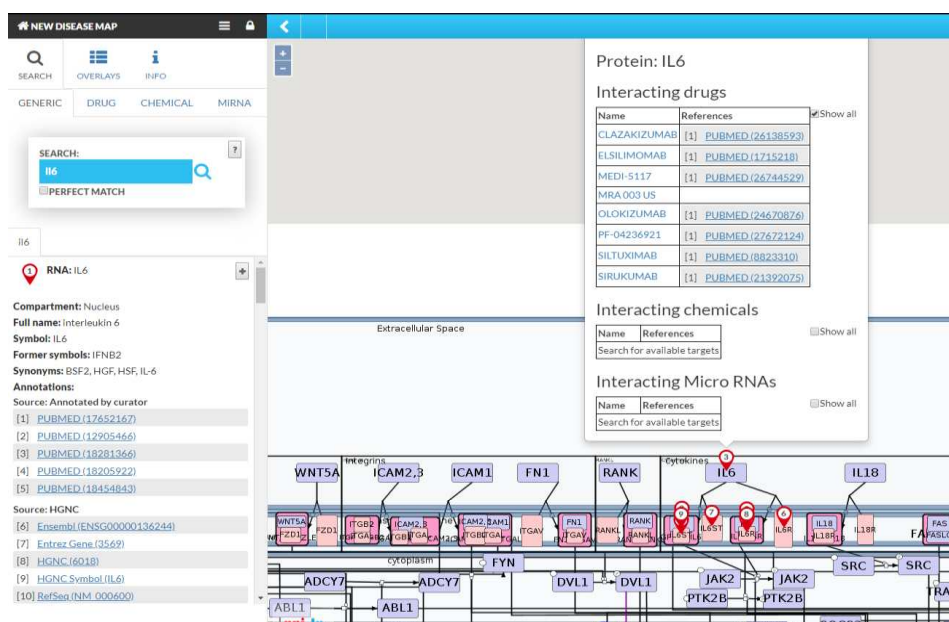


Figure 2.9: **The RA map in MINERVA platform.** **A)** Users can type in the search box the element to look for in the map. The resulting elements are shown as pins on the map. Corresponding annotations of the searched element, like HGNC, Entrez Gene, RefSeq and Ensembl identifiers are displayed on the left panel along with the PubMed identifiers of the manually curated annotations. **B)** Further clicking on the pin will display additional information about interacting drugs, chemicals and microRNAs for the element.

RA map offers custom visualization and export capabilities via MINERVA plugins (Hoksza et al., 2019). For instance, users can explore the RA map starting from a molecule of interest and easily follow its interactions, even through a dense and complex network. This functionality greatly facilitates navigating through the contents and tracking the flow of the signal from the ligand to the corresponding phenotype (see Fig 2.10).

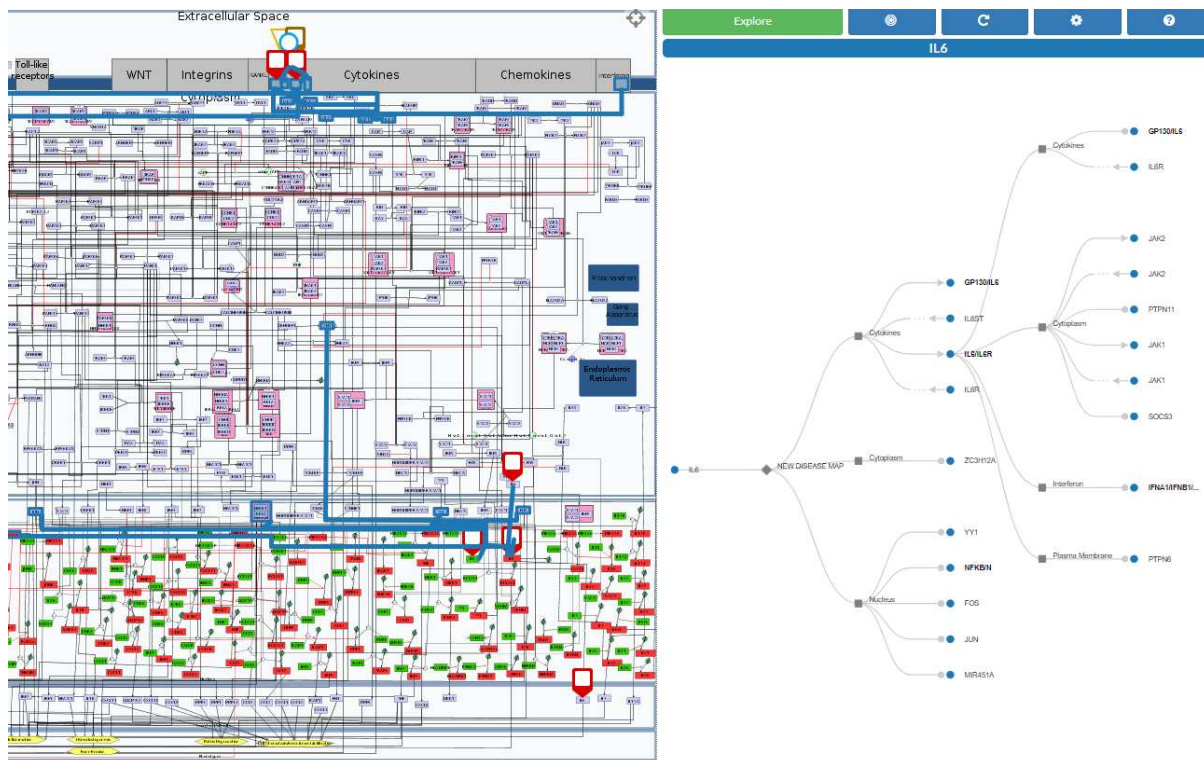


Figure 2.10: Tree expansion plugin in MINERVA.

Another feature of RA map is the stream plugin, allowing for highlighting and export of entire subnetworks in the map in one click. This is especially important for visualizing the ensemble of signaling pathways converging on the same disease-related phenotype (see **Fig 2.11**).

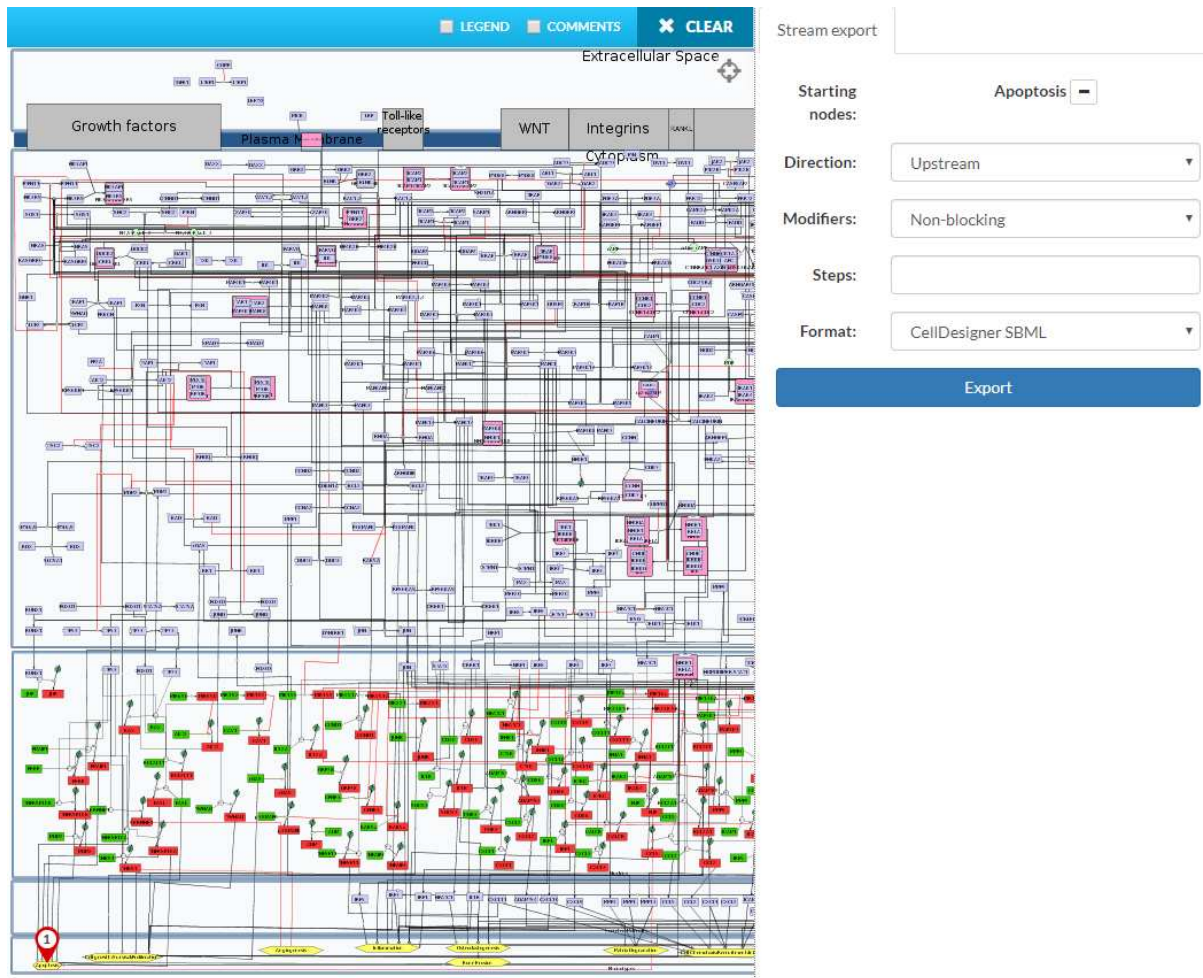


Figure 2.11: Stream export plugin in MINERVA.

2.4.1 The RA map as a template for visualizing cell-specific overlays

The RA map contains information from various sources serving as a generic blueprint for disease mechanisms. However, due to extensive annotation and reference, the user can opt for visualising cell-specific nodes and interactions. The RA map is a global map, integrating data and information from various sources. As a result, it has reactions and components that come from various cell or tissue types. We have grouped the sources into seven distinct groups namely Synovial fibroblasts, Synovial tissue, Peripheral blood mononuclear cells, Blood, Synovial fluid, Chondrocytes and Macrophages. **Fig 2.12** summarizes percentages of every group in the RA map. **Table 2.1** lists the main overlays and subcategories included in every overlay. In the RA map, the user can select to visualise one of the corresponding overlays, as it can be seen in **Fig 2.13** for synovial tissue.

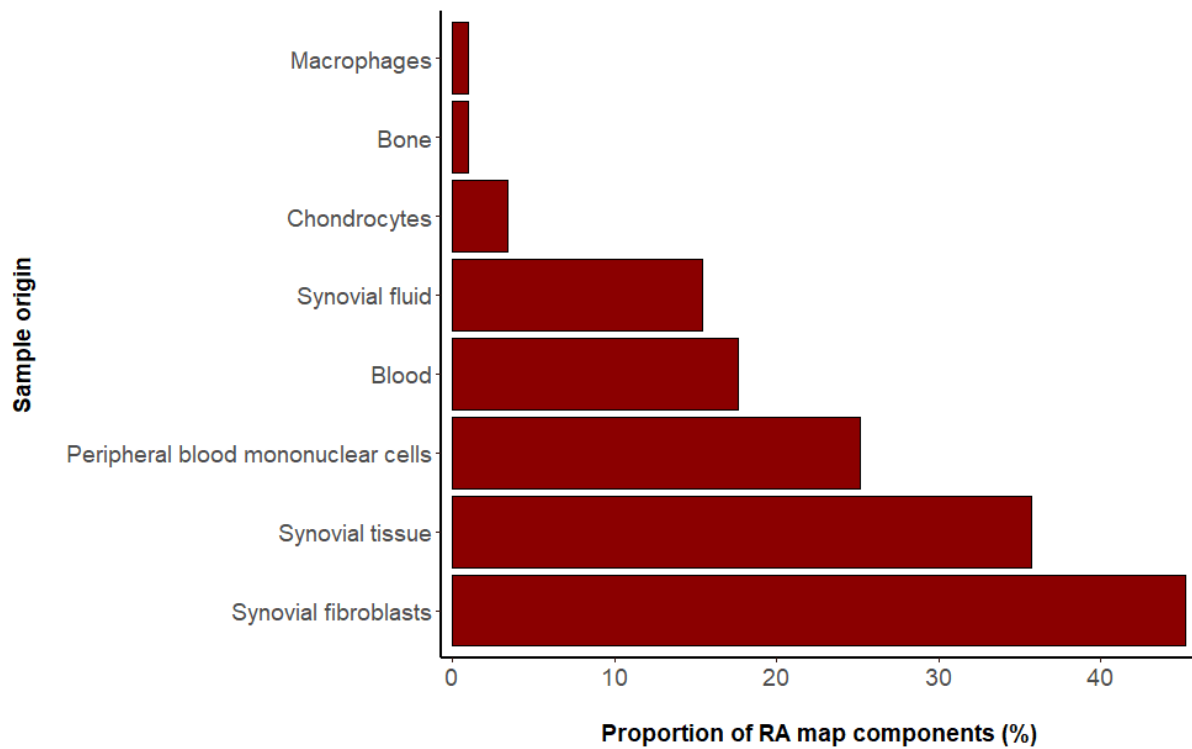


Figure 2.12: **Cell/tissue/fluid categories in the RA map.** Percentage of map components in the seven different overlays of the RA-map based on the scientific articles used.

Table 2.1: Overlays of RA-map and corresponding sub-categories.

Overlays	Subcategories
Synovial fibroblasts	RA synovial fibroblasts (RASf), RA fibroblast-like synoviocytes (RA-FLS), Synovial fibroblasts (SFs)
Blood	RA peripheral whole blood, RA peripheral blood, RA peripheral blood T cells, RA peripheral blood neutrophils, RA peripheral blood monocytes, RA whole blood, RA whole blood B cells, RA whole blood T cells, RA peripheral blood leukocytes, RA peripheral blood B cells, RA peripheral blood CD14+ monocytes, RA plasma, RA serum
Synovial tissue	RA synovial tissue, RA synovial membrane, RA synovium, RA synovial tissue monocytes, RA pannus tissue, RA endothelial cells
Peripheral blood mononuclear cells	Peripheral blood mononuclear cells (PBMCs), RA peripheral blood polymorphonuclear (PMN)
Synovial fluid	RA synovial fluid mononuclear cells, RA synovial fluid neutrophils, RA synovial fluid, RA synovial fluid T cells, RA synovial fluid dendritic cells, RA synovial fluid monocytes

Chondrocytes	RA chondrocytes, RASFs stimulated chondrocytes
Bone	Bone
Macrophages	CD68+ macrophage-like synoviocytes, macrophages, RA peripheral blood macrophages, RA synovial fluid macrophages



Figure 2.13: **Visualizing cell/tissue/fluid-specific parts of the RA map using dedicated overlays.** Snapshot of the visualisation of the Synovial Tissue overlay.

2.4.2 Comparing overlap with respective disease databases such as Disnor, DisGeNet and Ingenuity pathway analysis (IPA)

In order to assess the coverage of our map, we created lists of RA-specific biomolecules using two disease databases, Disnor (Lo Surdo et al., 2018) and DisGeNet (Piñero et al., 2017). Disnor is an open resource of disease networks that includes disease-associated genes annotated in the DisGeNET resource, gene-disease association (GDA) data to assemble inferred disease pathways. SIGNOR database is used to infer causal interactions related to disease genes with the highest possible coverage. DisGeNET involves collections of genes and variants involved in human diseases from expert curated repositories, GWAS catalogues, animal models and the scientific literature.

Concerning DisGeNet, we extracted two different lists, the first consisting of biomolecules of all sources, and the second one that is curated and comprising biomolecules from human samples.

209 out of 1847 biomolecules relevant to RA of the DisGeNet integrated list mapped to the RA map (11,31%) and 36 out of the 173 biomolecules of the human, curated list (20,8%). Concerning Disnor list, 72 biomolecules out of 238 in total mapped to the RA map (30, 25%) (Fig 2.14).

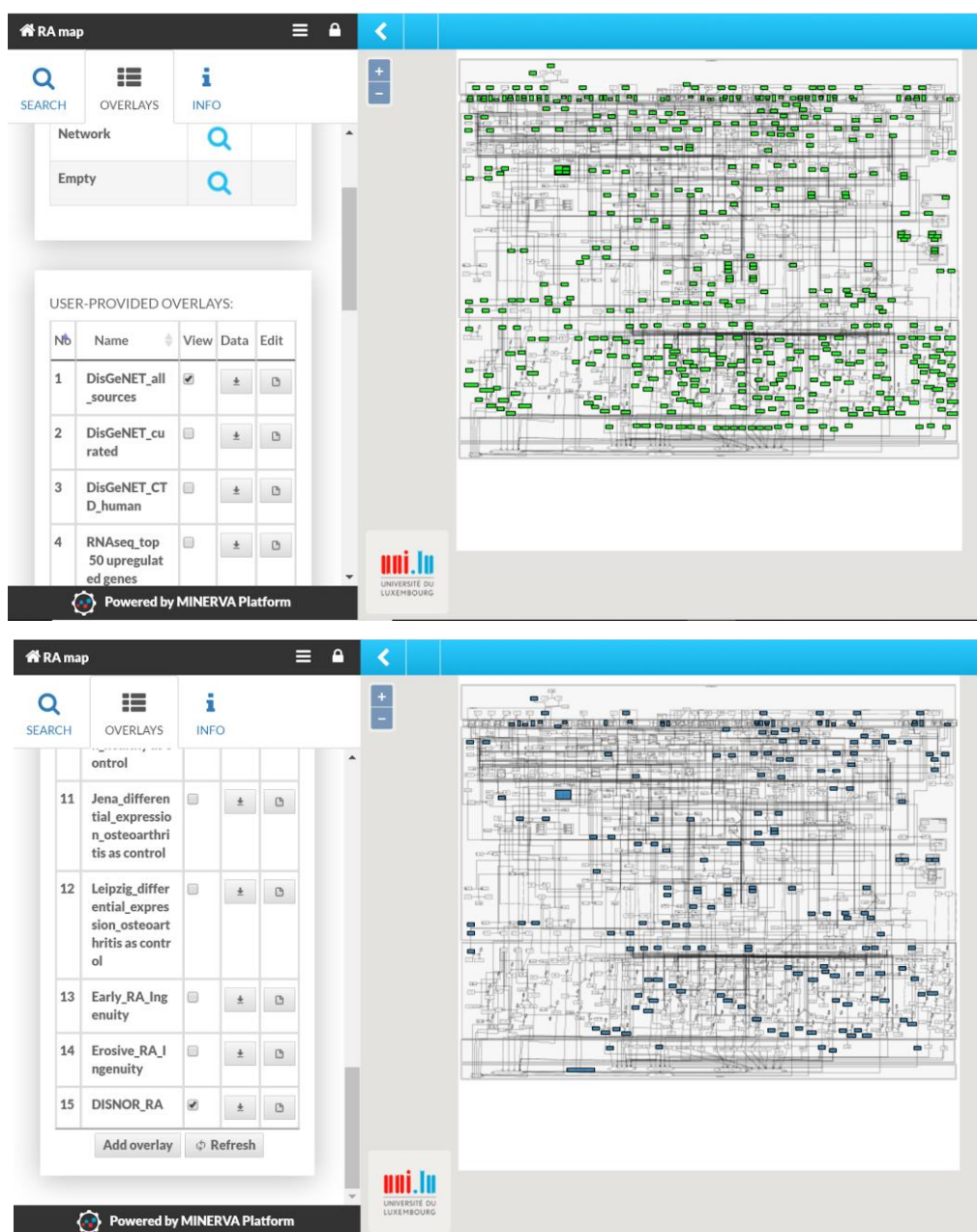
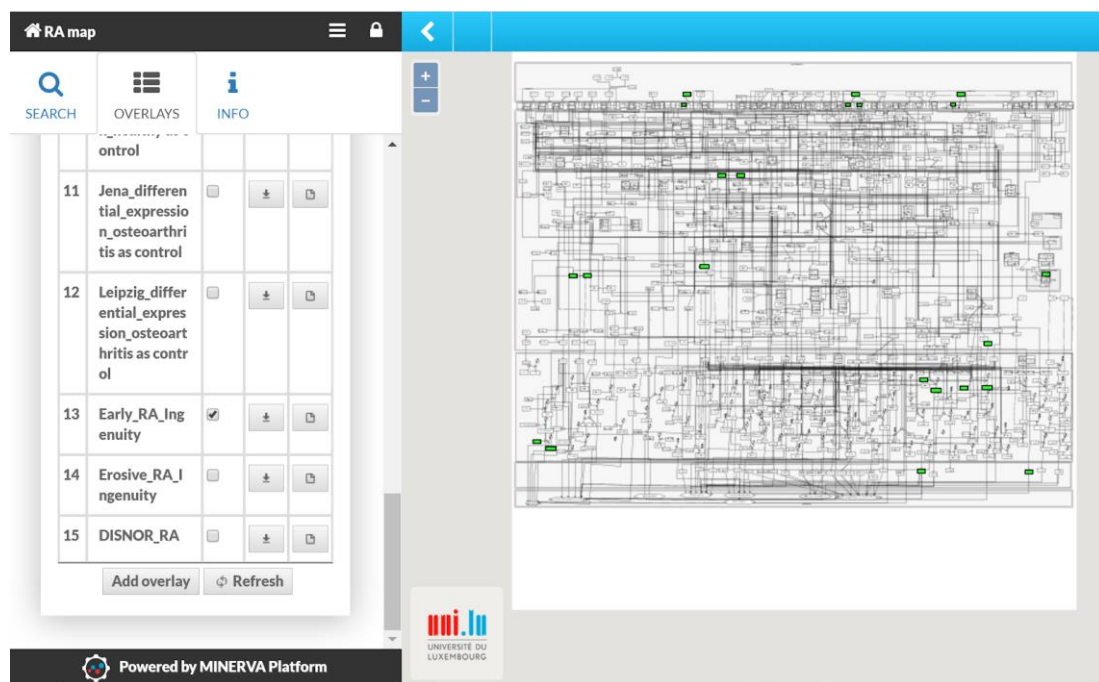


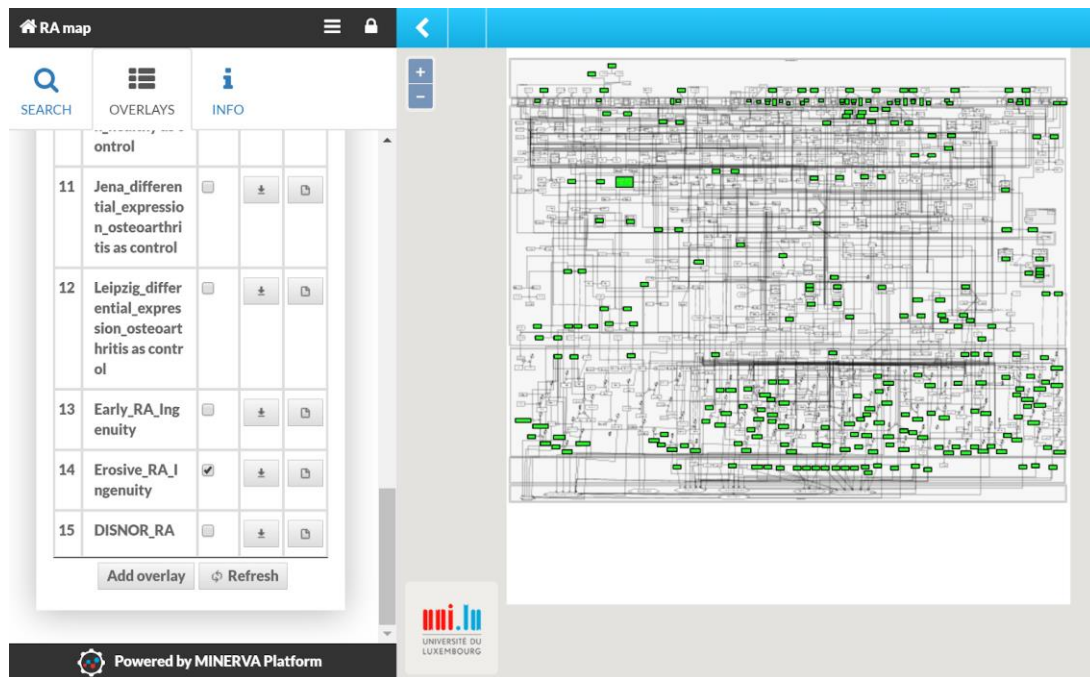
Figure 2.14: Overlap of the RA map with DisGeNet and Disnor databases. Mapping of A) DisGeNet all sources list, B) Disnor list.

Next, we extracted from IPA (Krämer et al., 2014) components corresponding to early and erosive RA. We used the RA map for visualising these two distinct RA phases. Results show a mapping of a few elements that correspond to early phase RA (7 proteins out of 99 used in total), while for the erosive stage of RA, we see a broader coverage (101 proteins mapped out of 1295 used in total) (**Fig 2.15A and B**). Not surprisingly, in early RA the few pathways that are affected implicate IL6, TNF, MAP4K4, and MDM2. What is also interesting is that in early RA the only protein in the endoplasmic reticulum that is present is PDIA3 while in erosive RA the whole complex is present (**Fig 2.14C**).

A



B



C

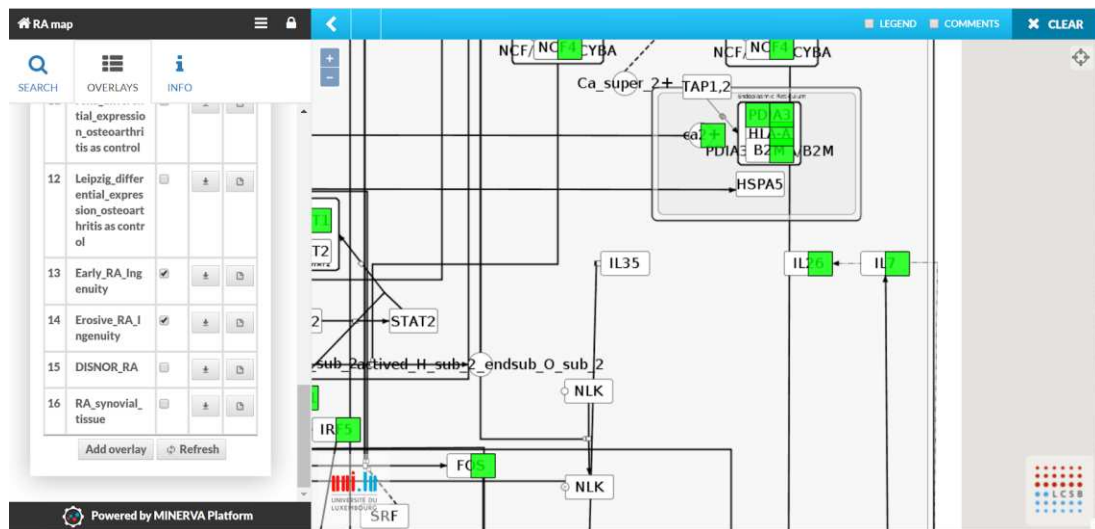


Figure 2.15: Early versus erosive RA components extracted from IPA. Mapping of components that correspond to A) early-stage RA, B) erosive stage of RA, C) proteins in the endoplasmic reticulum in early and erosive RA. Only PDIA3 is shared between the molecules that map in both states.

2.4.3 Visualizing Omic datasets

We used publicly available datasets for visualisation with the RA map. Our goal was to evaluate the coverage of the map using high throughput expression data and to compare the differentially

expressed pathways or regions in different datasets. For this purpose, we used the datasets from transcriptomic data of synovial tissue (Woetzel et al., 2014). In this study, three multicenter genome-wide transcriptomic datasets (from three clinical groups located in Jena, Berlin, and Leipzig and hence the names of the three datasets) from 79 patients/donors were used to infer rule-based classifiers to discriminate RA, OA, and healthy controls. We performed differential expression analysis between Berlin, Leipzig and Jena datasets with osteoarthritis as control. As it can be seen in **Fig 2.16**, 122 molecules were mapped to the RA map, with most pathways highlighted, as molecules that lead to most phenotypes were present (secreted component compartment). Interestingly, we found enrichment for almost all cellular phenotypes except for apoptosis and angiogenesis. Molecules leading to six out of eight phenotypes were expressed, while molecules linked to the two mentioned phenotypes were absent. For the absence of apoptosis enrichment, one can think of this result as coherent with the fact that synoviocytes and especially fibroblasts in RA are apoptosis-resistant (Baier et al., 2003).

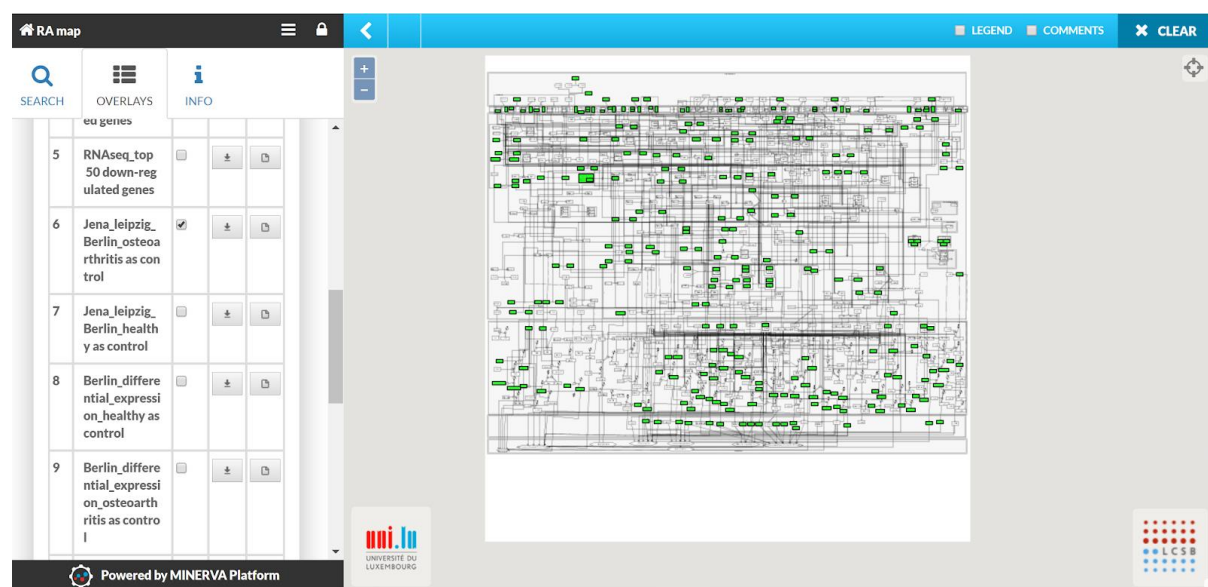


Figure 2.16: **Mapping of omic datasets from RA synovial tissue.** The apoptosis and angiogenesis phenotypes appear to be inactive as no molecule leading to these cellular phenotypes is mapped.

2.5 Topological analysis with Cytoscape

Cytoscape (Shannon et al., 2003) is an open-source software platform for visualising molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles, and other data types.

RA map XML file was imported in the Cytoscape version 3.5.0 and analyzed further with the help of different plugins which are easily downloadable from the plugin/app manager (<https://apps.cytoscape.org/>). The RA network comprises 1227 nodes and 1471 interactions.

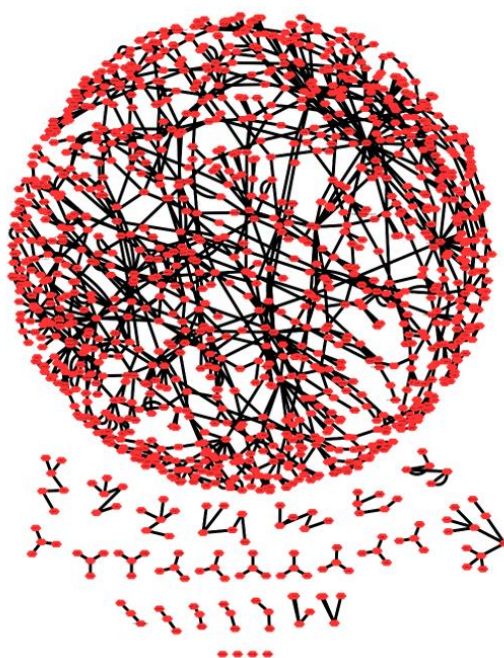


Figure 2.17: The RA map as a complex network with spring embedded layout in Cytoscape (Shannon et al., 2003). One connected core and several smaller unconnected parts are shown.

The analysis using NetworkAnalyzer, a built-in tool of Cytoscape, revealed that the RA network consists of 30 connected components. These connected components correspond to the connected subgraphs, i.e., parts of the graph in which any node can be reached from any other node by a path, with a core subgraph of 1106 nodes and 1379 reactions and 29 smaller ones (**Fig 2.17**). The nodes of a network can be characterized by the number of adjacent nodes (nodes directly connected to them). This node property is called the node degree. In directed networks such as signalling networks where the reactions are oriented (i.e from the ECM to the nucleus), we can distinguish two types of node degree: the in-degree, meaning the number of directed edges that have the node as target, and the out-degree that is the number of directed edges that

have the node as source. Node degree is an individual characteristic for each node, but one can define a degree distribution to quantify the diversity of the whole network.

The degree distribution describes the fraction of nodes that have degree k and is obtained by counting the number of nodes $N(k)$ that have $k = 1, 2, 3...$ edges and dividing it by the total number of nodes N . The majority of biological networks display scale-free properties (Barabási, 2009), which means that they contain a few central nodes that are highly connected (hubs) and several other loosely connected peripheral nodes. These networks follow a well-defined functional form $P(k) = ax^{-b}$ called a power law with the degree exponent b being usually in the range $2 < b < 3$ (Albert & Barabási, 2002). This function indicates that there is a high diversity of node degrees which is why these networks are described as 'scale-free'.

First, we performed the analysis considering the network as undirected in order to obtain the overall degree distribution (in and out) and then as directed to obtain the In degree and Out degree distributions. All node degree distributions follow a power-law, showing that the RA network is indeed a scale-free network, as seen in **Fig 2.18**.

When analysed as undirected, the RA network gives the topological parameters seen in **Table 2.2**. Each node has an average of 2.299 neighbours (nodes to which it is connected). The network density (number of edges divided by the number of pairs of nodes) is 0.002 meaning that it is a sparse network. The clustering coefficient is a ratio N / M , where N is the number of edges between the neighbors of n , and M is the maximum number of edges that could possibly exist between the neighbors of n . The clustering coefficient of a node is always a number between 0 and 1. The network clustering coefficient is the average of the clustering coefficients for all nodes in the network. Here, nodes with less than two neighbours are assumed to have a clustering coefficient of 0.

The clustering coefficient for the RA network equals to zero suggesting that there are no particular links between the neighbours of a node, giving space to more star-like shapes (node with several edges connected to it) than cliques (node that its neighbours have also edges in common) in the network. Networks whose topologies resemble a star have a centralization close to 1, whereas decentralized networks are characterized by having a centralization close to 0. The network heterogeneity reflects the tendency of a network to contain hub nodes.

Hubs allow to reach different parts of the network faster than passing through peripheral nodes but make scale-free networks sensitive to hub removal, as this would mean the destruction of

the network's structure. We used the degree distribution to obtain the hubs of the RA network and in **Table 2.3** we display the top ten hubs.

Table 2.2: Simple Topological Parameters obtained with NetworkAnalyzer for the RA network.

Topological parameters	Corresponding values
Clustering coefficient	0.0
Connected component	30
Network diameter	24
Network radius	1
Network centralisation	0.021
Shortest paths	1222614 (81%)
Characteristic path length	10.099
Average number of neighbours	2.299
Number of nodes	1225
Network density	0.002
Network heterogeneity	0.802
Isolated nodes	4
Number of self-loops	0
Multi-edge node pairs	63
Analysis time (sec)	0.5

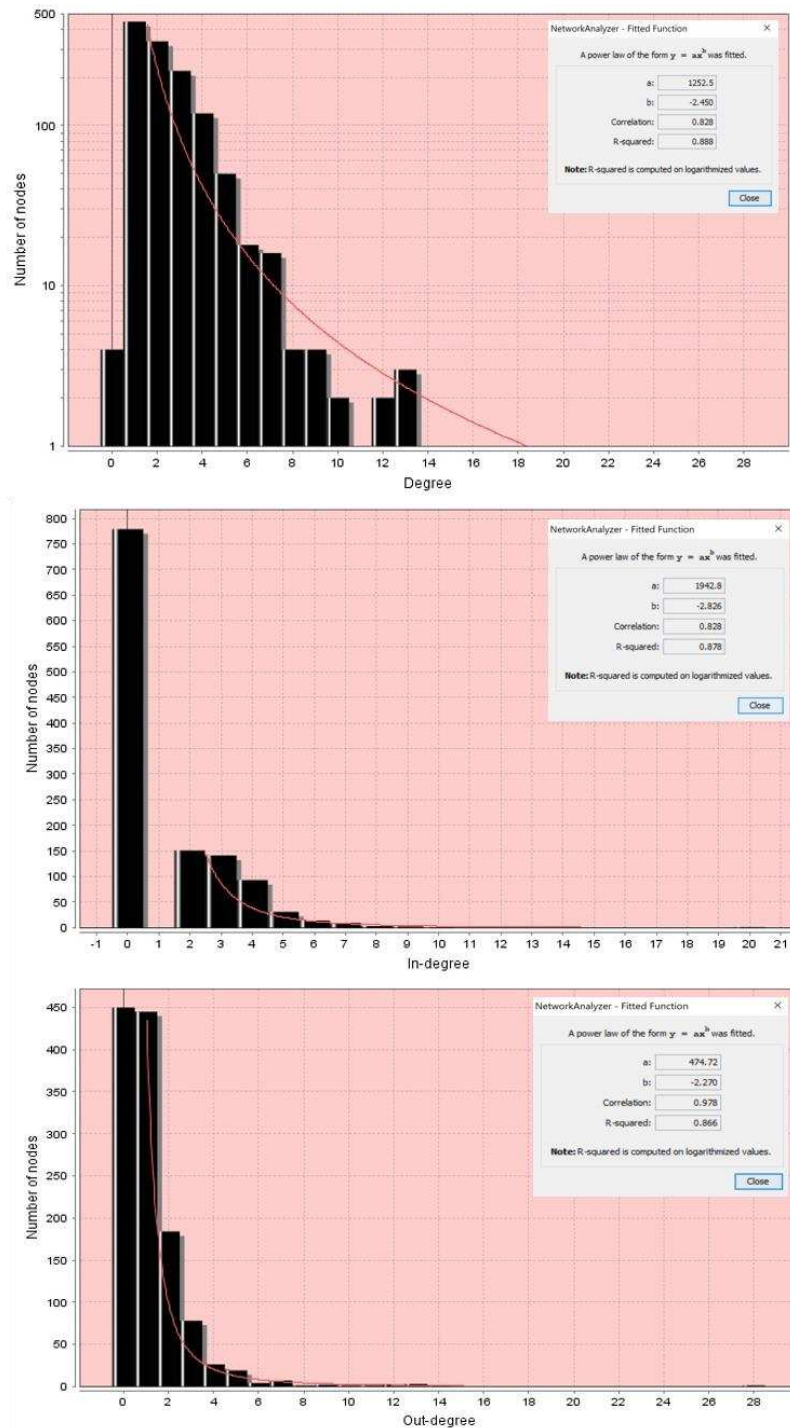


Figure 2.18: **Node degree distributions of the RA map with a fitted power law.** A) Overall degree distribution, B) In degree, C) Out degree distribution.

In addition, the number of multi-edge node pairs indicates how often neighboring nodes are linked by more than one edge. The network diameter of the RA network that corresponds to the maximum length of shortest paths between two nodes, is 24 suggesting that the signal starting from ligand-receptor complexes in the membrane reaches most of the network within 24 steps. The characteristic path length of the network that corresponds to the expected distance

between two connected nodes is approximately 10, meaning that the response to a signal and its propagation can occur relatively rapidly.

Table 2.3: Top ten hubs of the RA map.

RA map nodes	Node Degree	Role	Reference
NFKB	28	Implicated in RA and inflammation	(Liu et al., 2017) (Noort et al., 2015) (Simmonds & Foxwell, 2008) (Makarov, 2001)
Inflammation	14	A major characteristic of RA	(Demoruelle et al., 2014; Malemud, 2007; McInnes & Schett, 2011; Müller-Ladner et al., 2007)
AKT	13	Regulates apoptosis in RA	(García et al., 2010; Mountz et al., 2001)
Cell chemotaxis/recruitment/infiltration	13	Implicated in RA	(Goddard et al., 1984; Mellado et al., 2015)
JUN	13	Implicated in RA	(Hannemann et al., 2017; Han et al., 2001)
MAPK1	12	Implicated in RA	(Namba et al., 2017; Thalhamer et al., 2008)
RAC1,2	12	Implicated in RA	(Bartok et al., 2014; Gary S Firestein, 2010)
Cell growth/ Survival	11	Major characteristic of RA	(Jacobs et al., 1995; Kramer et al., 2003; Malemud, 2007; Mongan & Jacox, 1964)
Osteoclastogenesis	10	Results in bone damage in RA	(Müller-Ladner et al., 2007; Sato & Takayanagi, 2006)
TP53	9	Involved in the apoptosis pathway implicated in RA	(Seemayer et al., 2003; Tak et al., 2000)

2.5.1 Functional enrichment of the whole RA network

We performed a functional analysis of the RA map content using the DAVID annotation tool (Dennis et al., 2003). DAVID is a web based Database for Annotation, Visualization, and Integrated Discovery. DAVID provides a comprehensive set of functionalities included in four main modules: Annotation Tool, GoCharts, KeggCharts, and DomainCharts. Annotation tool provides functional annotation of gene list; GOCharts shows the distribution of genes among functional categories using Gene Ontology Consortium (GO) vocabulary; KeggCharts display the distribution of genes among KEGG biochemical pathways; DomainCharts shows the distribution genes among PFAM protein domains (Dennis et al., 2003; Punta et al., 2012).

However, the overall disease enrichment was low concerning RA, even after regrouping of all RA related terms (**Fig. 2.19**). The results of DAVID using GAD (Genetic association database) (Becker et al., 2004) disease are shown in **Fig. 2.20**. From the 340 unique gene ids recognized by DAVID, only 84 genes are annotated as RA specific, as shown in **Table 2.4**.

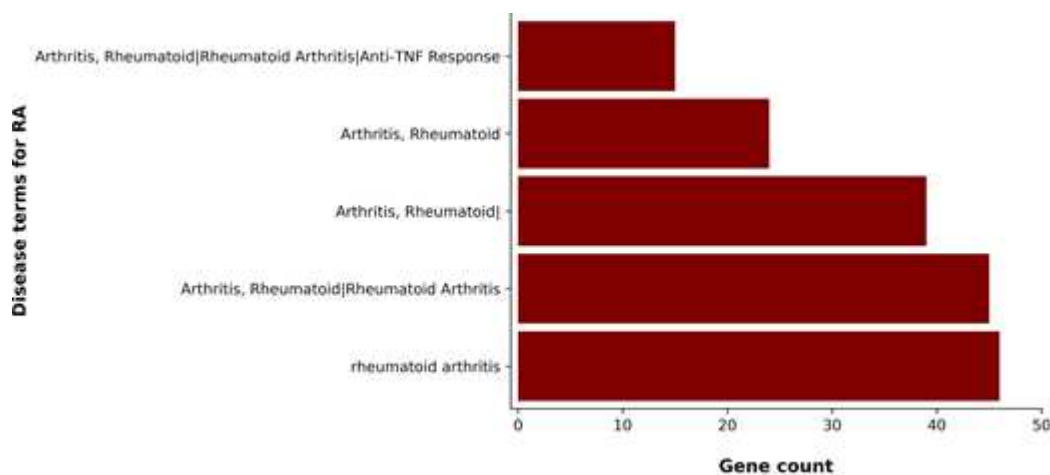


Figure 2.19: **RA map enrichment in RA disease terms using DAVID.** Gene count corresponding to RA related disease terms in DAVID, using the RA map components as the input list.

However, 74 genes were identified as relevant to bone mineral density which can be related to RA, as bone erosion is one of the main characteristics of the disease. It is worth noting that the GAD disease database used for annotation in DAVID has not been updated since 2009.

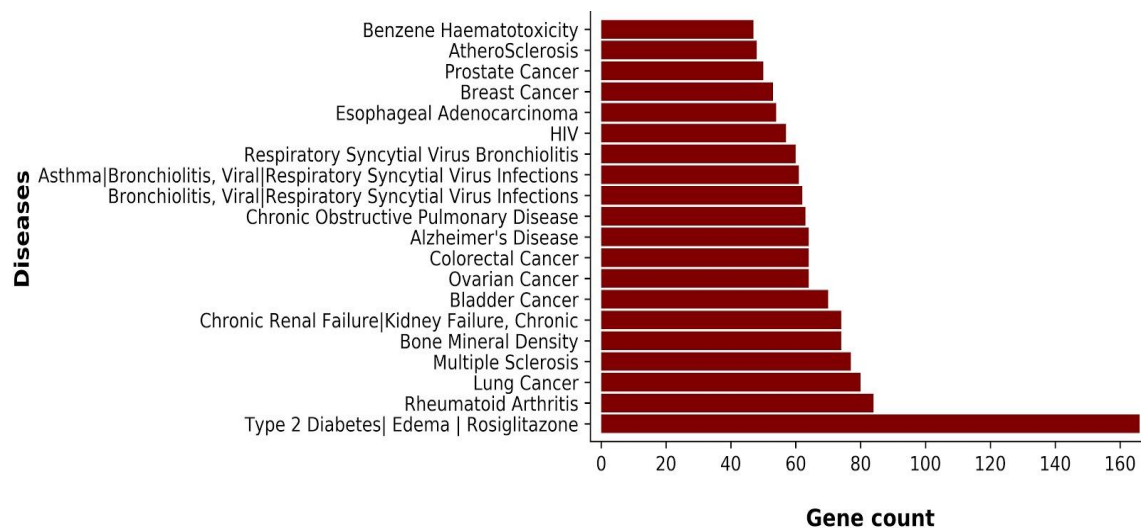


Figure 2.20: **Functional analysis of the RA map using DAVID.** The list of RA map components was used as the input list in the functional annotation tool DAVID. Disease enrichment analysis showed enrichment in Type 2 Diabetes, multiple sclerosis, various cancer types, asthma, HIV, atherosclerosis and rheumatoid arthritis.

Table 2.4: 84 RA specific genes after analysis with DAVID.

Disease	Gene count	Genes
RA (all RA- related disease terms)	84	TRAF1, MICB, TNF, HLA-DRB1, PTGS2, MMP9, IL18, TLR2, NFKBIA, PTPN22, NFKB1, TLR4, MMP3, CCL5, MMP1, TGFB1, TNFRSF1A, STAT4, TNFRSF1B, NOD2, TAP2, ITGAV, TAP1, SERPINE1, IL1B, RUNX1, TRAF5, PTPN6, PTPRC, IL6, IL1RN, PRKCH, IL26, IL6R, CD40, MMP13, TAB2, TYK2, CD86, TNFSF11, CD19, IRF5, CD80, VEGFA, TNFAIP3 CD14, TRAF1, IL1R1, IL6ST, TLR5, MAPKAPK2, IL17A, IL17F, TRAF6, IL1A, MAP2K6, CD28, IRAK1, ICAM1, SOCS3, MAP2K3, TP53, TAB1, IL21, MIR146A, RPS6KA5, MAPK14, RIPK1, KLRK1, CXCL8, IRAK4, MAP3K1, RELA, ICAM3, IKBKE, IFNB1, MDM2, AP3K14, CSF2, ETS1, IRF8, MAP3K7, TIRAP, MYD88

2.5.2 Clustering and functional enrichment of the clusters

As for the construction of the map we used an approach from local to global (built interaction by interaction and not by adding ready-made pathways), we wanted to see if the automated clustering would produce clusters that correspond to biological pathways relevant to RA

allowing us to pinpoint pathways that are RA specific and not shared with other autoimmune diseases. We employed the Glay community clustering algorithm (Morris et al., 2011; Su et al., 2010) in order to cluster the RA network. The community clustering algorithm is an implementation of the Girvan-Newman fast greedy algorithm as implemented by the Glay Cytoscape plugin. The algorithm begins by simplifying the network to give it a community-like structure by removing duplicate edges and self-loops. Then, it identifies clusters by iteratively removing edges from the network and then checking to see which nodes are still connected. We decided to use the Glay community algorithm because this algorithm operates exclusively on connectivity, a principle that is very close to the biological concept for pathways. Glay algorithm produced 57 clusters: 25 medium-sized clusters (9 to 75 nodes), 28 smaller clusters (3 to 8 nodes) and four single nodes (**Fig 2.21**). A total of 25 clusters with node size ≥ 9 were analysed individually using the functional annotation tool - DAVID (Huang et al., 2007) for their enrichment in diseases and pathways (**ANNEX B**). Functional annotation with DAVID revealed explicit enrichment of 17 clusters in RA out of 25 clusters. For 8 clusters, RA appeared in the top ten terms of the disease enrichment list. Seventeen were found to be enriched in Type 2 diabetes (T2D), 7 in Multiple Sclerosis (MS) and 8 in Asthma. Type 1 diabetes also appeared in three clusters already enriched in T2D. In **Tables 2.5** and **2.6**, we can see the top 5 Glay clusters enriched in RA and their functional enrichment in pathways with corresponding p-values.

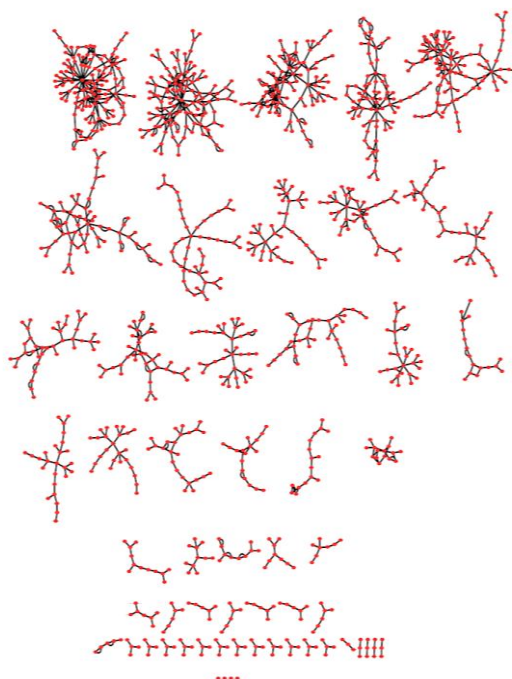


Figure 2.21: 57 Glay clusters in Cytoscape.

Table 2.5: Top 5 Glay clusters enriched in RA.

Clusters	Cluster's name (hub)	Number of elements in the cluster	P value for enrichment in RA
1	MYD88	40	3.27E-12
2	MAPK14	11	2.75E-04
3	MIR451a	10	1.24E-02
4	STAT3	6	1.08E-03
5	MIR146a	3	1.83E-04

Table 2.6: Top 5 Glay clusters enriched in RA and their corresponding Pathway enrichment and P value

Clusters	Pathway enrichment	P value
1	hsa04620:Toll-like receptor signaling pathway	5,71E-19
	hsa04064:NF-kappa B signaling pathway	4,43E-18
	hsa05145:Toxoplasmosis	2,05E-11
	hsa05140:Leishmaniasis	2,45E-11
	hsa05152:Tuberculosis	4,89E-11
	hsa05133:Pertussis	2,06E-09
	hsa05142:Chagas disease (American trypanosomiasis)	2,09E-08
	hsa04010:MAPK signaling pathway	6,13E-07
	hsa04350:TGF-beta signaling pathway	5,39E-06
	hsa05144:Malaria	1,51E-05
	hsa05168:Herpes simplex infection	1,71E-05
	hsa05134:Legionellosis	2,23E-05
	hsa05169:Epstein-Barr virus infection	3,34E-05
	hsa04380:Osteoclast differentiation	4,71E-05
	hsa05162:Measles	5,07E-05
	hsa05323:Rheumatoid arthritis	3,17E-02
2	hsa04620:Toll-like receptor signaling pathway	3,39E-03
	hsa05162:Measles	5,29E-03
	hsa05152:Tuberculosis	9,22E-03

	hsa04621:NOD-like receptor signaling pathway	4,79E-02
	hsa05140:Leishmaniasis	6,04E-02
	hsa05133:Pertussis	6,37E-02
	hsa04064:NF-kappa B signaling pathway	7,35E-02
	hsa05142:Chagas disease (American trypanosomiasis)	8,74E-02
	hsa05146:Amoebiasis	8,90E-02
	hsa04668:TNF signaling pathway	8,98E-02
	hsa05145:Toxoplasmosis	9,22E-02
3	hsa04010:MAPK signaling pathway	1,27E-10
	hsa04722:Neurotrophin signaling pathway	3,49E-07
	hsa04668:TNF signaling pathway	4,06E-04
	hsa04912:GnRH signaling pathway	7,27E-03
	hsa04380:Osteoclast differentiation	1,47E-02
	hsa04261:Adrenergic signaling in cardiomyocytes	1,62E-02
	hsa05161:Hepatitis B	1,78E-02
	hsa04921:Oxytocin signaling pathway	1,89E-02
	hsa05202:Transcriptional misregulation in cancer	2,32E-02
4	hsa05321:Inflammatory bowel disease (IBD)	5,05E-04
	hsa04920:Adipocytokine signaling pathway	4,01E-02
	hsa04917:Prolactin signaling pathway	4,07E-02
	hsa04931:Insulin resistance	6,13E-02
	hsa05160:Hepatitis C	7,51E-02
	hsa04630:Jak STAT signaling pathway	8,17E-02
5	hsa05140:Leishmaniasis	2,05E-02
	hsa05133:Pertussis	2,17E-02
	hsa04064:NF-kappa B signaling pathway	2,51E-02
	hsa05142:Chagas disease (American trypanosomiasis)	3,00E-02
	hsa04620:Toll-like receptor signaling pathway	3,06E-02
	hsa05145:Toxoplasmosis	3,17E-02
	hsa04722:Neurotrophin signaling pathway	3,46E-02
	hsa05169:Epstein-Barr virus infection	3,52E-02
	hsa05162:Measles	3,83E-02

From **Tables 2.5** and **2.6** regarding the enrichment of the top 5 clusters of the RA map, we can see that the clusters include molecules implicated in RA and anti TNF-response, Toll like receptor signalling, NF-kappa B signalling pathway, TGF-beta signalling pathway, MAPK

signalling, JAK-STAT signalling, cascades that are well characterised for their involvement in RA pathogenesis.

2.5.3 Systemic Interpretation and Pharmacogenomics Analysis using BioInfoMiner

We also used the BioInfoMiner web application (<https://bioinforminer.com>) to perform functional analysis and prioritize genes in the RA map. The application performs biological interpretation of gene sets, which comprises detection and prioritization of systemic processes and pathways, as well as prioritization of genes based on their mapping to those processes. We performed two sets of analyses using gene ontology (GO) and human phenotype ontology (PHO) terms. The first analysis using GO gave a list of 48 genes and enrichment of terms like Inflammatory response, Regulation of cytokine production, Activation of MAPK activity, all relevant to pathways included in the RA map. In **Table 2.7** we can see the top 10 priority genes. In **Fig. 2.22** is shown the signature using GO, consisting of the ranked systemic processes (y axis) and prioritized genes (x axis). The first most prioritized gene was TNF, a very common target for many anti-inflammatory drugs, including anti-rheumatic, as anti-TNF agents were the first molecular targeting drugs developed for the treatment of RA. Interestingly, the second most prioritized gene (**Table 2.7**) was TLR4 that was found as the target of 8 drugs, one of which was Methotrexate, the most common disease-modifying anti-rheumatic drug (DMARD). Concerning IL-1B, it is the target of Rilonacept that was first proposed as a treatment for RA, however after the phase I it was shown that IL-1B blockade was not very beneficial for the patients of RA (McDermott, 2009)

Concerning FADD, a study showed that transfection of FADD gene by adenoviral vector into cultured RA synoviocytes induced up-regulation of FADD expression and apoptosis. In addition, local injection of FADD adenovirus (Ad-FADD) eliminated synoviocytes in vivo by induction of apoptosis of proliferating human rheumatoid synovium engrafted in severe combined immunodeficiency mouse, which is the most suitable animal model of RA for the evaluation of treatment strategy in vivo. The study also showed that chondrocytes were not affected (Kobayashi et al., 2000). JAK1 and JAK2 kinases have been implicated in the pathogenesis of RA and various drugs have been proposed that target them. Baricitinib, an oral selective inhibitor of JAK1 and JAK2, offers an effective treatment for RA in a wide range of patients (E. H. S. Choy et al., 2019). Tofacitinib, another drug that targets JAK kinases is indicated for the treatment of moderate to severe active rheumatoid arthritis (RA) in adult

patients who have responded inadequately to, or who are intolerant of, one or more DMARDs (S. Dhillon, 2017)

Wnt5a is a member of the non-canonical family of Wnts that modulates a wide range of cell processes, including differentiation, migration, and inflammation. Wnt5a has been implicated as a possible contributor to arthritis and it is upregulated in synovial fibroblasts from RA patients (MacLauchlan et al., 2017) Numerous studies link TRAF6 with RA. Elevated expression of tumor necrosis factor receptor-associated factor 6 (TRAF6) in RA synovium correlated significantly with the severity of synovitis and the number of infiltrated inflammatory cells (L.-J. Zhu et al., 2017)

TRAF6 blockade significantly suppressed the IL-1 β -stimulated migration and invasion of human RA-FLSs. These results support a role for TRAF6 in the pathogenesis of RA, and suggest that the TRAF6 blockade may be a potential strategy in the management of RA (L.-J. Zhu et al., 2017)

MYD88 is a TLR protein adaptor that might contribute in PKD1 associated proinflammatory responses in RA (Yi et al., 2015). RIPK1 plays a critical role in mediating deleterious responses downstream of TNFR1. RIPK1 inhibitors have progressed successfully past human phase I clinical studies.

RIPK1 rRIPK1 is a key mediator of apoptotic and necrotic cell death as well as inflammatory pathways. receptor-interacting serine/threonine-protein kinase inhibition attenuates experimental autoimmune arthritis via suppression of osteoclastogenesis (Jhun et al., 2019)

Table 2.7: Top ten priority genes using BioInfoMiner and GO terms

Rank	Gene Symbol	Definition
1	TNF	tumor necrosis factor
2	TLR4	toll like receptor 4
3	RIPK2	receptor interacting serine/threonine kinase 2

4	IL1B	interleukin 1 beta
5	RIPK1	receptor interacting serine/threonine kinase 1
6	FADD	Fas associated via death domain
7	JAK2	Janus kinase 2
8	WNT5A	Wnt family member 5A
9	TRAF6	TNF receptor associated factor 6
10	MYD88	MYD88, innate immune signal transduction adaptor

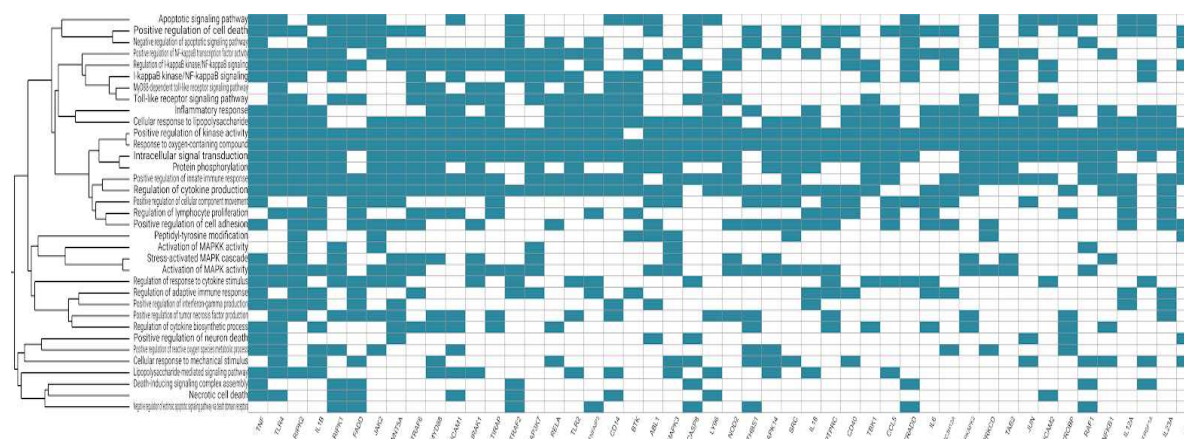


Figure 2.22: Heatmap of the 48 priority genes and their systemic interpretation using BioInfoMiner and GO terms.

Analysis using the PHO gave 32 priority genes out of which top 10 genes are shown in **Table 2.8** and enrichment in terms containing Arthralgia, Skin nodule, Abnormality of the immune system among others (**fig 2.23**). Overall, the systemic functional analysis with BioInfoMiner further confirmed the validity of the model at the semantic level, complementary to the mechanistic one.

Table 2.8: Top ten priority genes using BioInfoMiner and PHO terms.

Rank	Gene Symbol	Definition
1	IL12A	interleukin 12A
2	FAS	Fas cell surface death receptor
3	NRAS	NRAS proto-oncogene, GTPase
4	STAT3	signal transducer and activator of transcription 3
5	PTPN22	protein tyrosine phosphatase, non-receptor type 22
6	HLA-DRB1	major histocompatibility complex, class II, DR beta 1
7	JAK2	Janus kinase 2
8	IRF5	interferon regulatory factor 5
9	STAT4	signal transducer and activator of transcription 4
10	CTNNB1	catenin beta 1

All top ten priority genes are drug targets except for IRF5. Various studies in the last decade have placed IRF5 as a central regulator of the inflammatory response. IRF5 contributes to the pathogenesis of many inflammatory and autoimmune diseases, such as rheumatoid arthritis, inflammatory bowel disease and systemic lupus erythematosus and represents a potential therapeutic target. However, despite a significant interest from the pharmaceutical industry, inhibitors that interfere with the IRF5 pathway still remain elusive. (Almuttaqi & Udalova, 2019) (Cieřla et al., 2019)

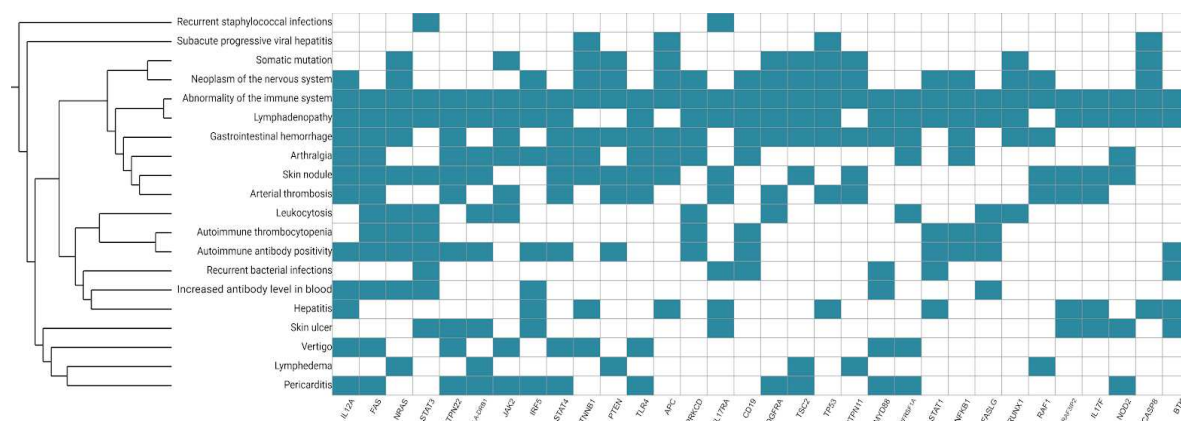


Figure 2.23: Heatmap of the 32 priority genes and their systemic interpretation using BioInfoMiner and HPO terms.

2.6 Conclusion

We present here a state-of-the-art, large-scale molecular interaction map for RA (RA map), which is to our knowledge the first SBGN-compliant Process Description disease map. All the components and reactions are annotated using only RA and human-specific studies. In an attempt to limit false positives experts' advice is incorporated and SBGN standards were used for representation of the map to assure reusability. The RA map we present here includes information from 353 peer-reviewed publications featuring 506 species, 446 reactions and 8 phenotypes. The species in the map are classified to 303 proteins, 61 complexes, 106 genes, 106 RNA entities, 2 ions and 7 simple molecules.

The RA map can also be used as an interactive knowledge base, using the platform MINERVA and serve as a template for overlaying multiple datasets. Visualization of experimental data could help highlight aspects of the affected biological process and make differences between experimental conditions more evident. Visualizing the results of differential expression analysis of three datasets of gene expression of RA synovial tissues showed enrichment in all cellular phenotypes but not in apoptosis. This finding is in line with the fact that fibroblasts, which constitute a large percentage of the RA synoviocytes, exhibit an apoptosis-resistant phenotype.

We performed functional analysis and gene prioritization using BioInfoMiner. The genes that rank higher in this analysis are associated with many systemic processes and are considered as hubs in the semantic network. Along with prioritization, a pharmacogenomic analysis is provided since the hubs proposed are considered as putative drug targets. The results of the

analyses using GO and PHO terms revealed known RA players, most of which have been already used as drug targets demonstrating that the RA map comprises well-characterized factors and captures most of the relevant systemic processes implicated in the disease.

Topological analysis can reveal underlying structural features of the RA map like unconnected parts of the network, or important hubs (well-connected nodes) which are otherwise hard to perceive in large-scale networks. The topological analysis performed in this study revealed connected and unconnected parts of the network. This result reflects our fragmented knowledge on the one hand, but also the use of stringent criteria for the nodes included in the map: experimentally validated interactions in at least two published studies, use of data of strictly human origin and disease-specific.

The RA map, the fruit of interdisciplinary collaborations between clinicians, biologists and bioinformaticians is accessible at ramap.elixir-luxembourg.org.

Chapter 3. Automated inference of Boolean models from molecular interaction maps using CaSQ

3.1 Biological network representations and molecular interaction maps

Biological processes can be represented in the form of interaction networks/graphs where components (genes, proteins, other molecules) are represented as ‘nodes’ and the interactions between components are represented as ‘edges’. Network interactions can be directed or undirected, depending on the biological information available that allows the characterization of the interaction (inhibition or activation) and also the source and the target node. Representing the complexity of biological regulatory systems using networks enables the analysis of their topology, identifying distinct clusters that may correspond to specific biological processes (‘modules’) and nodes with a high degree of connectivity (‘hubs’). These network features exercise a significant influence on the propagation of biological information (i.e. signal, regulation) (Barabási & Oltvai, 2004; Ideker & Nussinov, 2017; B. Zhang et al., 2014).

The Systems Biology Graphical Notation (SBGN) scheme uses three different languages for network representation (Le Novère, 2015) (discussed in detail in chapter 1, section 1.10.1). First, the activity flow (AF) diagram is an interaction network, which includes influence direction and mode of regulation, such as activation and inhibition. Second, the entity-relationship (ER) representation that includes mechanistic details, the direction of influences but no sequential information and third, the process description diagram (PD) which is the most detailed of all, including specifics of the direction of influences, mechanism of action and the order of events. The SBGN-PD notation scheme is based on ideas first introduced to the field by Hiroaki Kitano and co-workers (2003).

Molecular interaction maps act as an encyclopedia, describing biological mechanisms concisely and effectively. Various molecular maps describing different biological processes (Caron et al., 2010; Fujita et al., 2014; Grieco et al., 2013; Jagannadham et al., 2016; Kuperstein et al., 2015; Mazein, Knowles, et al., 2018; Niarakis et al., 2014; Ogishima et al., 2016; Singh et al., 2018, 2020; Tripathi et al., 2015) have been published, and initiatives have emerged, such as the Disease Maps Project (<http://disease-maps.org>), demonstrating the utility and need of this type of representation of biological knowledge (Mazein, Knowles, et al., 2018) (Ostaszewski et al., 2019).

Molecular interaction maps can serve as a stand-alone knowledge base for any biological process or disease, or they can be used as a scaffold for building computational models to capture the dynamics of the system. Based on information mining, human curation and expert advice, these maps summarize current knowledge about biological pathways in a process description representation, while accounting for as many mechanistic details as possible. They provide a comprehensive template for visualization and analysis of omics datasets, and can also be analyzed in terms of the underlying network structure. However, their static nature cannot account for the coordination of multiple biological processes, or how the regulation of several nodes due to the presence or absence of certain factors can alter the functional outcome (i.e. activation of a particular pathway following the repression of a given factor). These regulations that fine-tune the molecular interactions are of great importance as dysregulation or disruption can lead to disease (D.-Y. Cho et al., 2012; Furlong, 2013).

3.2 Boolean models for dynamical studies

Systems Biology approaches and especially computational modelling can be used to provide an executable, dynamic network that can reveal hidden properties and account for emerging system-level behaviours through *in silico* simulations and perturbations (Azeloglu & Iyengar, 2015; Tomás Helikar et al., 2008). Each interaction is described using mathematical formalism and the obtained machine-readable model can be used to test novel hypotheses and predict new features of the system of interest. Boolean models are well suited for addressing the lack of kinetic data and handling the large size of the biological pathways described in molecular interaction maps. These models are parameter-free; nevertheless, their simplistic nature can provide a powerful tool for dynamic analysis (Abou-Jaoudé et al., 2016; Furlong, 2013). In Boolean formalism, the simplest form of logical models, nodes represent regulatory components (proteins, enzymes, complexes, transcription factors, genes, to name a few) and arcs represent their interactions. Each regulatory component is associated with a Boolean variable (taking the values 0 or 1) denoting either its qualitative concentration (0 for absent or 1 for present) or its level of activity (0 for inactive or 1 for active). The future state of each node depends on the state of its upstream regulators and is defined by a Boolean function. The function is expressed in the form of a rule using the logical operators AND, OR and NOT. The updating of the rules can be in a synchronous, deterministic mode where all nodes are updated at the same time (Glass & Kauffman, 1973; S. A. Kauffman, 1969) or in an asynchronous

mode, where only one node can be updated every time (Thomas, 1973, 1978; Thomas et al., 1976).

3.3 Bridging the gap between static and dynamic representation

The construction of a molecular interaction map and a dynamic model are two tasks that can serve different purposes and are usually performed independently. On the one hand, it is a question of creating a knowledge base in the form of a comprehensive molecular map, and on the other of defining the underlying mechanism that links the system components and captures its dynamic behaviour. Nevertheless, these two constructs share much information, including the mode of influence (e.g. activation or inhibition) and the topology of the network. Molecular maps can be built using a structured diagram editor for drawing gene-regulatory and biochemical networks, such as CellDesigner (Kitano et al., 2005). Networks in CellDesigner are drawn as process description diagrams (PD) and are stored using the Systems Biology Markup Language (SBML), a standard for representing models of biochemical and gene-regulatory networks (Hucka et al., 2003).

The idea of obtaining executable models from a network topology is not new. In the study by (Büchel et al., 2013), researchers proposed a pipeline for the automatic generation of models using KEGG pathways as a resource. They succeed in producing SBML and Systems Biology Marked Up Language-qualitative (SBML-qual) files but these constructs can be seen as model scaffolds as they require further parameterization to become executable. In (Mendoza & Xenarios, 2006), a Standardized QUALitative Dynamical system (SQUAD) is obtained directly from an input network that is already a regulatory network and not a molecular interaction map.

Furthermore, the aim is to obtain a continuous system corresponding to it, implying a small-scale network (about 20–30 nodes). Regarding Biolayout, now Graphia (Livigni et al., 2018), researchers use the modified Edinburgh Pathway Notation scheme (mEPN) to create SBML-like maps that they interpret directly as Petri nets. This approach imposes that all ‘logics’ are conjunctive. There is no direct negation, no disjunction, whereas the only firing rule in a Petri net is that all input places should be filled in order for the reaction to fire.

However, molecular maps contain much more precise information (e.g. inhibitions) that cannot be expressed directly within this framework. Moreover, Petri nets are by nature quantitative, requiring several tokens to be assigned to each place, and having the consumption of some tokens by each rule. The rxncon language ([Romers and Krantz 2017](#)) also tackles the idea that there are standard features between maps as knowledge-bases and executable Boolean models.

However, their approach is quite different from ours in that they bridge this gap through an intermediate language based on Boolean bipartite graphs. One of the most important consequences is that the logical rules (contingencies in rxncon) are already part of the input (the map being, in a way, already a model).

Finally, the <http://pd2af.org/> initiative (Vogt et al., 2013) proposes to translate an SBGN-PD graph, similar to a CellDesigner map, into an SBGN-AF graph, similar to the structure of a Boolean model, but does not go further as to propose an executable model. We will detail in the discussion some specific rules for which we have made similar or opposite choices concerning the graph transformation. However, one should note that our method adds the layer of inferring logical rules for the obtained model based on the original topology and annotations, making possible immediate simulations and analyses using the corresponding tools [e.g. GINSim (Chaouiya et al., 2012) and Cell Collective (Tomáš Helikar et al., 2012)].

In this work, we present CaSQ (CellDesigner as SBML-qual), a tool for automated inference of large-scale, parameter-free Boolean models, from molecular interaction maps with preliminary logic rules based on network topology and semantics. CaSQ is, to the best of our knowledge, the first tool that produces executable molecular networks of hundreds of nodes (at least up to eight hundred), in the SBML-qual format that can be further simulated and analyzed using popular modelling tools.

3.4 CaSQ

CaSQ is a tool that can convert a molecular interaction map built with CellDesigner (Funahashi et al., 2003) to an executable Boolean model. The tool is developed in Python and uses as source the xml file produced by CellDesigner (SBML plus CellDesigner-specific annotations) to infer preliminary Boolean rules based solely on network topology and semantic annotations (e.g. certain arcs are noted as catalysis, inhibition, etc.). The aim is to convert a Process Description (PD) representation, i.e. a reaction model, into a complete logical model. The resulting structure is closer to an AF diagram, though not in a strict SBGN-PD to SBGN-AF notion. Moreover, logical rules that make the model executable are also obtained. For illustrating the rules of the conversion, we use the repertoire of notation schemes in CellDesigner (**Fig. 3.1**).

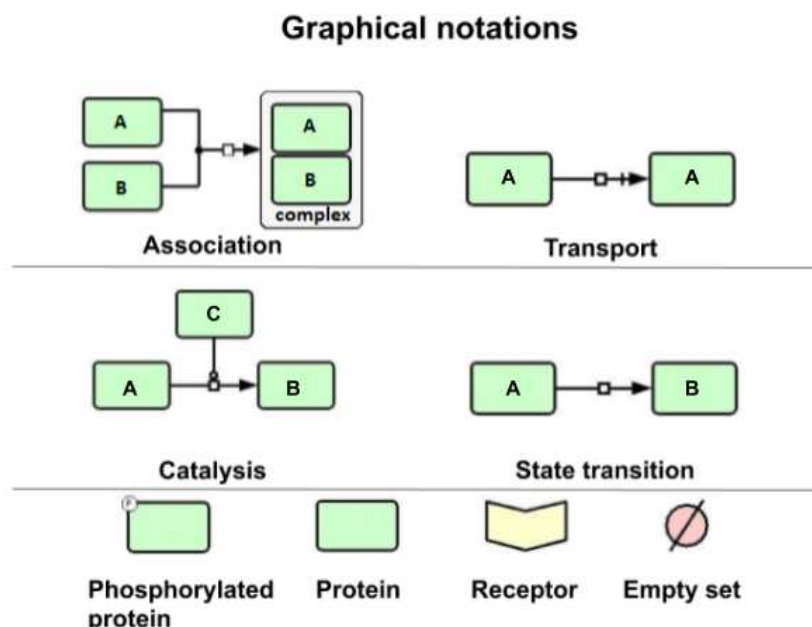


Figure 3.1: The repertoire of CellDesigner graphical notation schemes used to illustrate CaSQ's rules. For CaSQ's conversion rules we use the notation schemes for association, transport, catalysis, state transition and also the glyphs for receptor, protein, modified protein (here we show phosphorylation as an example) and the empty set. The empty set can account for degradation or in SBGN-PD terms, can represent the creation (respectively, the disappearance) of an entity from an unspecified source (resp. sink) that we do not need or wish to express explicitly.

The conversion of the graph to an executable model is a four-step process:

Step 1: First, the map is reduced through a pass of graph rewriting rules. These rules are executed in order and in a single pass, so the rewriting is terminating and confluent. The reasoning behind this reduction is that a single qualitative species of the logical model often represents by its state (active/inactive) several species of the original map. Therefore, those species might need to be merged into a single component or some inactive forms to be completely discarded to avoid redundancy in the logical model. The rules are the following:

Rule 1: If two species of the map are only reactants in a single reaction, i.e. do not take part in any other reaction, if that reaction is annotated as heterodimer association, and if one of the reactants is annotated as a receptor, then the receptor is deleted from the map (its annotations are added to the product of the reaction) (**Fig 3.2**);

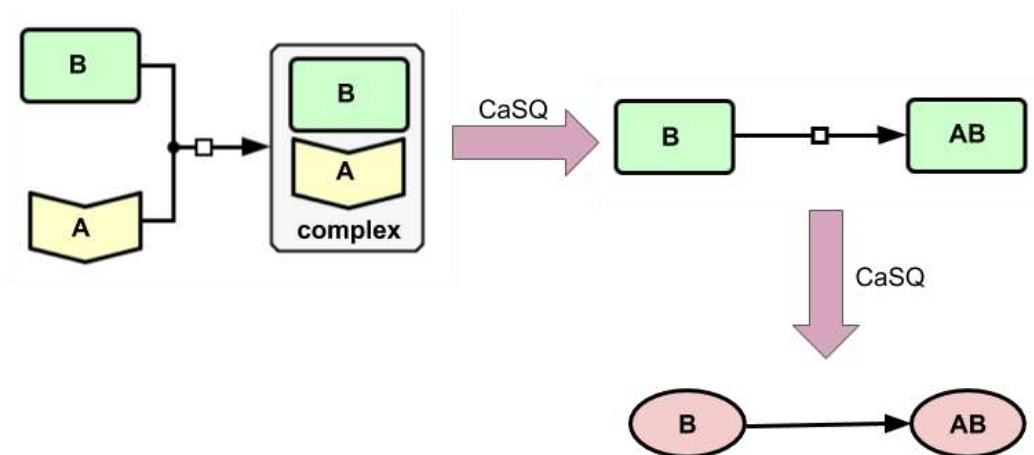


Figure 3.2: **Illustration of the 1st rule.** If two species of the map are only reactants in a heterodimer association, and if one of the reactants is annotated as a receptor, then the receptor is deleted from the map (its annotations are added to the product of the reaction).

Rule 2: If two species of the map take part in a reaction annotated as heterodimer association, if none of them are annotated as receptor, and if both do not take active part (i.e. reactant or modifier) in any other reaction, then both are merged into the complex, product of the reaction (their annotations are added to the product, and the reactions that had them as product are rewired to have the complex as product) (**Fig. 3.3**);

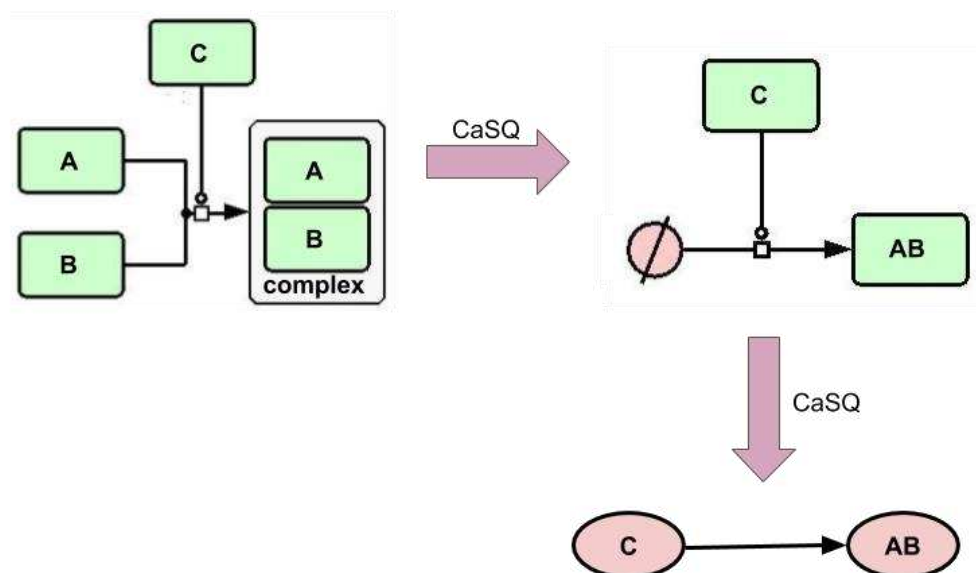


Figure 3.3: **Illustration of the 2nd rule.** Compression of the complex formation, where none of the reactants is denoted as a receptor, and both reactants do not participate in any other reaction. As a result, both reactants are removed and modifiers are rewired to have the complex as a product.

Rule 3: If one species only appears in a single reaction, if it appears there as a reactant if that reaction has a single product, and if both the reactant and the product have the same name, then the reactant is deleted (its annotations are merged into those of the product) (**Fig 3.4**);

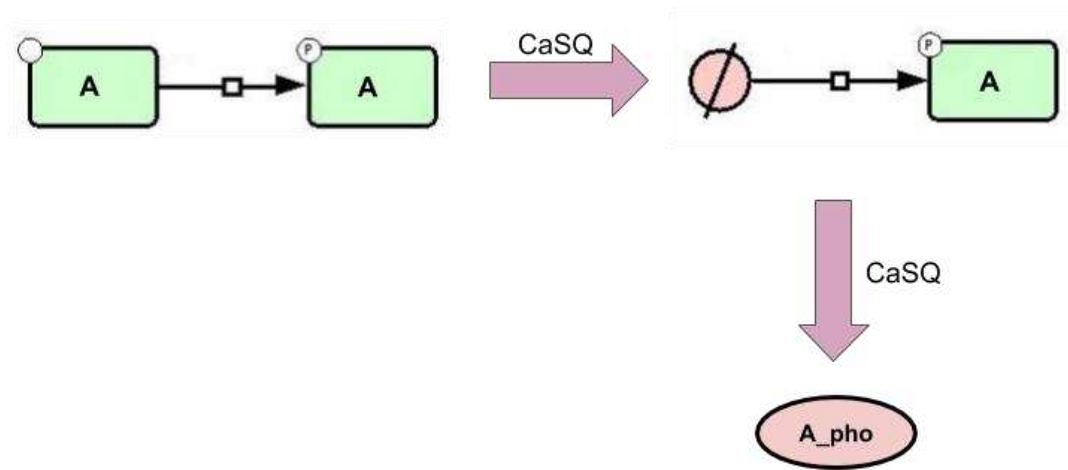


Figure 3.4: **Illustration of the 3rd rule.** Removing inactive forms that do not participate in other reactions.

Rules 2 and 3 can be combined resulting in greater graph compression, as illustrated in **Figure 3.5**.

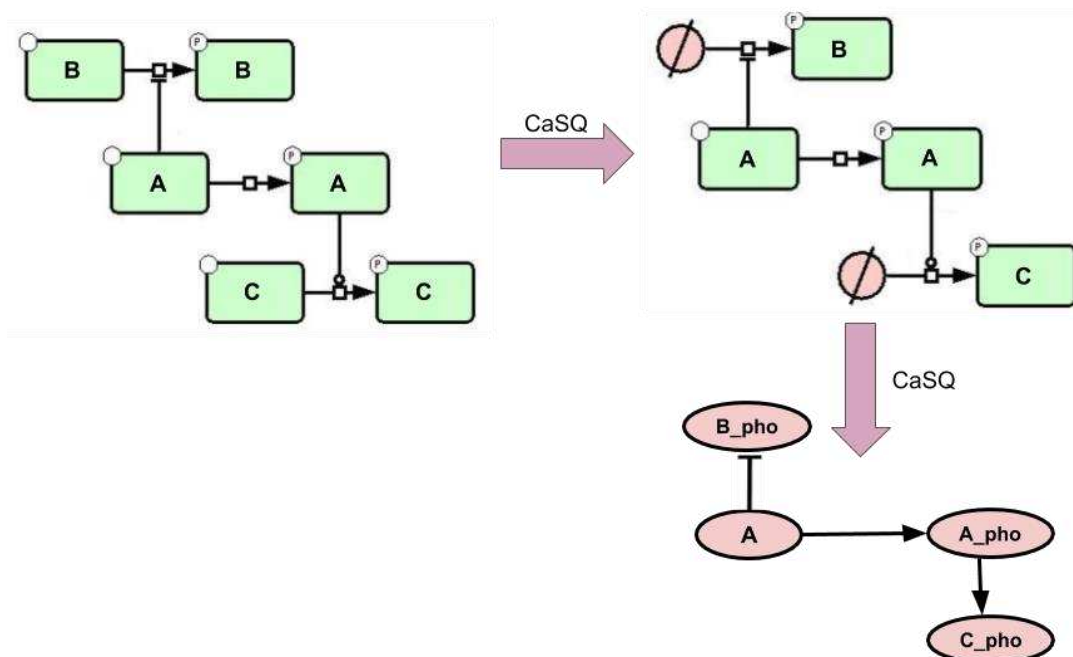


Figure 3.5: Combination of rules 2 and 3. CaSQ retains components that contribute further to the propagation of the signal.

Rule 4: If one species only appears as a reactant in a single reaction (but maybe appearing as product in another reaction) that has a single product and is annotated as transport, and if both the reactant and the product have the same name, then the reactant is merged into the product (its annotations are merged into those of the product, and the reactions producing it are rewired to the product) (**Fig 3.6**).

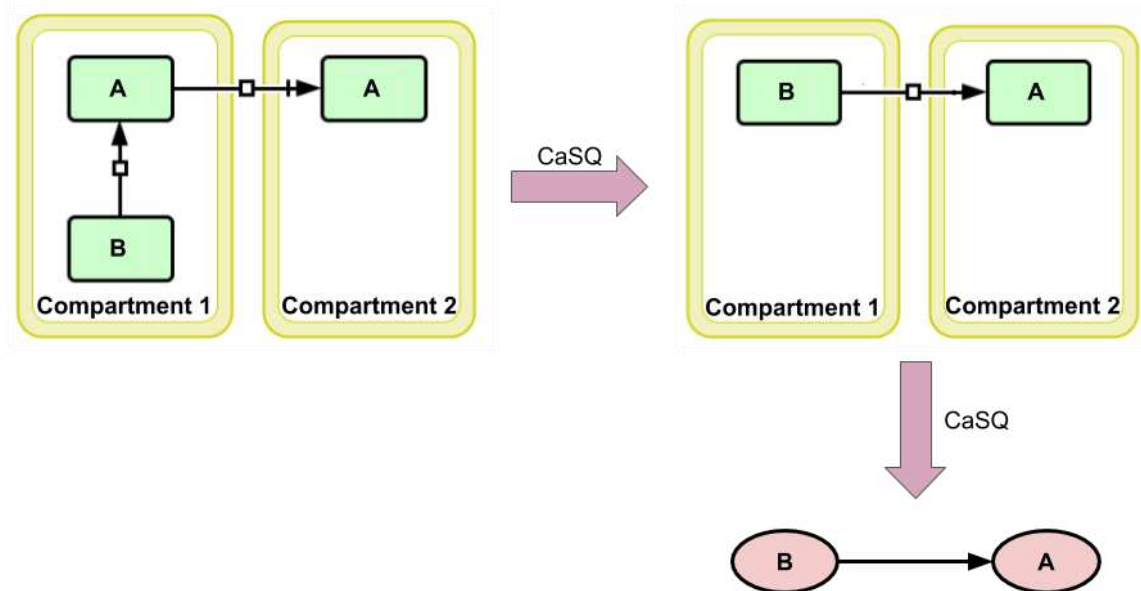


Figure 3.6: **Combination of the 2nd and the 4th rule.** Components that are translocated across other compartments (for example transcription factors) are merged in one component that inherits all influences, provided that the original component does not participate in another reaction/ regulation.

The rationale of using the name to identify the same components in different states (gene, RNA, protein, transported/phosphorylated/ methylated protein, etc.) is that we need to identify when species can be merged/discarded, to keep only what contributes further to signal propagation. However, relying on the active annotation (dotted circle) in CellDesigner maps proved to be insufficient: not all map curators use this notation, and it is not SBGN compliant.

Step 2: The topology of the model is then computed as a simple form of PD to AF conversion, with one qualitative species corresponding to each species in the reduced map obtained from Step 1. This species inherits the original map layout, using SBML3 Layout package, and MIRIAM annotations (e.g. PubMed IDs as bqbiol: isDescribedBy). The annotations have been

associated with each regulated component rather than each regulation, mostly because tools supporting the latter are quite rare. All reactants and modifiers of a reaction exert a positive influence on all the products of that reaction, whereas all inhibitors exert a negative influence. Compared to the formal abstraction of influence graphs from reaction graphs (Rizk et al., 2011), note that, the mutual inhibition between reactants is purposely ignored as in Step 1 we already condense active and inactive forms of the same species.

Step 3: The logical rules of the model are computed. For each species, its logical rule is defined as the (i) disjunction (OR), for all reactions producing it, of (ii) the disjunction (OR) for all positive modifiers of a reaction being on and (iii) the conjunction (AND) of all products of that reaction being activated and all inhibitors being inactive. Therefore, a target is on if one of the reactions producing it is on, a reaction is on if all reactants are on, all inhibitors are off and one of the catalysts is on (**Fig. 3.7**).

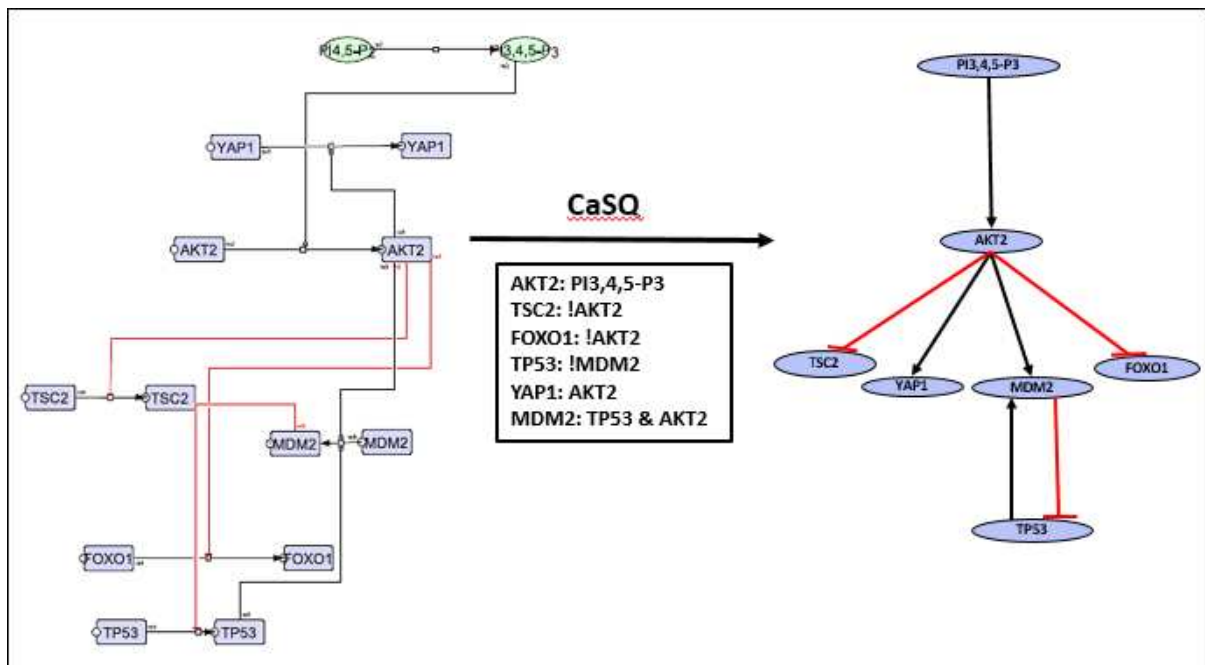


Figure 3.7: A complex part of the RA map translated into an activity flow (AF) like diagram with preliminary logical rules.

Step 4: Model refinement is performed through the optional removal of unconnected components. From our experience, keeping only the biggest connected component is what makes the most sense from a modelling perspective. However, it is possible to specify a

‘minimum size’ and keep all connected components above that size. Names of the qualitative species are also made more precise by adding the original type/modifications of the species (e.g. RNA, phosphorylated) and if there are still homonyms the original compartment is added too. More precisely, the name of the node in the model is, therefore, the name of the species in the map to which is added (separated by an underscore character ‘_’), its type as given in the map (RNA, Gene, etc.) unless that type is ‘PROTEIN’ and to which is added modifications given by the map (phosphorylation, methylation, etc.). If after that step, several species from the model are found to have the same name, the compartment is added too (once again, separated by an underscore) (**Table 3.5 and Table 3.6**).

CaSQ generates two output files; the proper logical model encoded in SBML-qual, a format that is compatible for further analysis with modelling tools such as GINsim (Chaouiya et al., 2012; Tomáš Helikar et al., 2012) or Cell Collective (Tomáš Helikar et al., 2012), and a CSV file that contains information about the names, the logic formulae and the CellDesigner alias. The second file is mostly for automated treatment.

The SBML-qual file can also be restricted to include only its biggest connected component (BCC), or only connected components above a given size threshold. This allows the modeller to obtain a more meaningful logical model even if the original map did contain several unconnected clusters corresponding to isolated pieces of information.

3.5 Molecular interaction maps and logic models

For testing the applicability of CaSQ, we used various molecular interaction maps that differ in size, complexity and use of SBGN notation, as shown in **Table 3.1**. Namely, we used one molecular interaction map comprising 125 nodes describing mast cell activation (Niarakis et al., 2014), one map comprising 232 nodes for MAPK activation (Grieco et al., 2013), one for cholecystokinin signaling with 530 nodes (Tripathi et al., 2015) and finally two large-scale molecular maps, one for rheumatoid arthritis (RA)-the only SBGN compliant (Singh et al., 2018, 2020) comprising 779 nodes, detailed annotations and references in the MIRIAM and text annotation section of the CellDesigner file (Funahashi et al., 2003) (**Fig. 3.8**) and the Alzheimer’s pathway map with 1361 nodes (Ogishima et al., 2016) (Ogishima et al., 2016). The mast cell activation and the MAPK maps were published along with their corresponding manually built logical models.

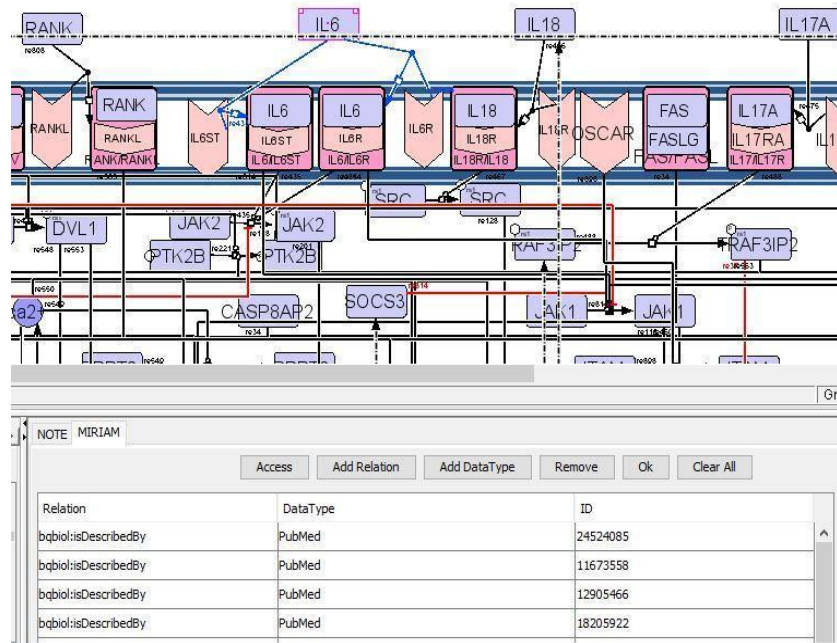


Figure 3.8: **Snapshot of the RA map built with the software CellDesigner.** The annotations for every node are stored in the MIRIAM section, using the bqbiol:isDescribedBy tag. Here we see the MIRIAM annotations for the node IL6.

Table 3.1: Molecular maps and corresponding manually built models used for benchmarking CaSQ.

Name	SBGN standard	Corresponding model	Reference
Mast cell activation	-	✓	(Niarakis et al., 2014)
Mitogen-activated protein kinase (MAPK)	-	✓	(Grieco et al., 2013)
	✓	-	(Wu et al., 2010)
Rheumatoid arthritis (RA)			(Singh et al., 2018) (Singh <i>et al.</i> , 2020)
Cholecystokinin	-	-	(Tripathi et al., 2015)
Alzheimer	-	-	(Ogishima et al., 2016)

3.6 Model comparison

For evaluating the performance of the tool, we compared size and shared nodes between manually built models that corresponded to the interaction maps (for mast cell and MAPK), with the CaSQ-inferred Boolean models. While size reduction is not the primary goal of the tool, it remains a measure of comparison between the process description static diagram of the original map and the regulatory graph that the tool produces after the conversion rules.

Conversion from a process description to an AF diagram implies a more compact network. The comparison allows us to check if such compression was achieved. We also performed simulations to see if the CaSQ-inferred models were able to reproduce known biological scenarios, and finally, we compared steady states, where feasible, between the inferred and the manually built models.

3.7 In silico simulations and calculation of stable states

3.7.1 Cell Collective

For the simulations of the CaSQ-derived models, we used CellCollective (Helikar et al., 2012). The Cell Collective is a web-based modeling platform for the collaborative construction of large-scale models of various biological processes where scientists from across the globe simulate/analyze them in real time (**Fig. 3.9**). Models in Cell Collective can be created either de novo or they can be imported using the SBML-qual standard.

Models in the Cell Collective are based on a qualitative, rule-based mathematical framework. In this framework, each species can assume either an active or inactive state. Which state a species assumes at any given time point depends on a set of rules that take into account the activation state of all immediate upstream regulators.

In the case of real-time simulations, % ON of a species represents its moving average activity, and is calculated as the fraction of the active/inactive states over a sliding window (Tomáš Helikar et al., 2012). CellCollective provides a platform with three different panels, simulation control, activity network and simulation graph. The simulation control panel allows the user to choose the activity level of external (0 to 100%) and internal components (ON and OFF) along with other simulation parameters like updating scheme (synchronous/asynchronous), simulation speed and initial state. The activity network panel provides visualization of changes in the activity of the network nodes in real time and the simulation graph follows the evolution of the dynamic behaviour between external and internal components of the model

Cell Collective SBML-qual import supports network layout, as well as model annotations. References stored in the MIRIAM section of the xml file of CellDesigner can be retrieved and visualized in the platform (**Fig. 3.10**).

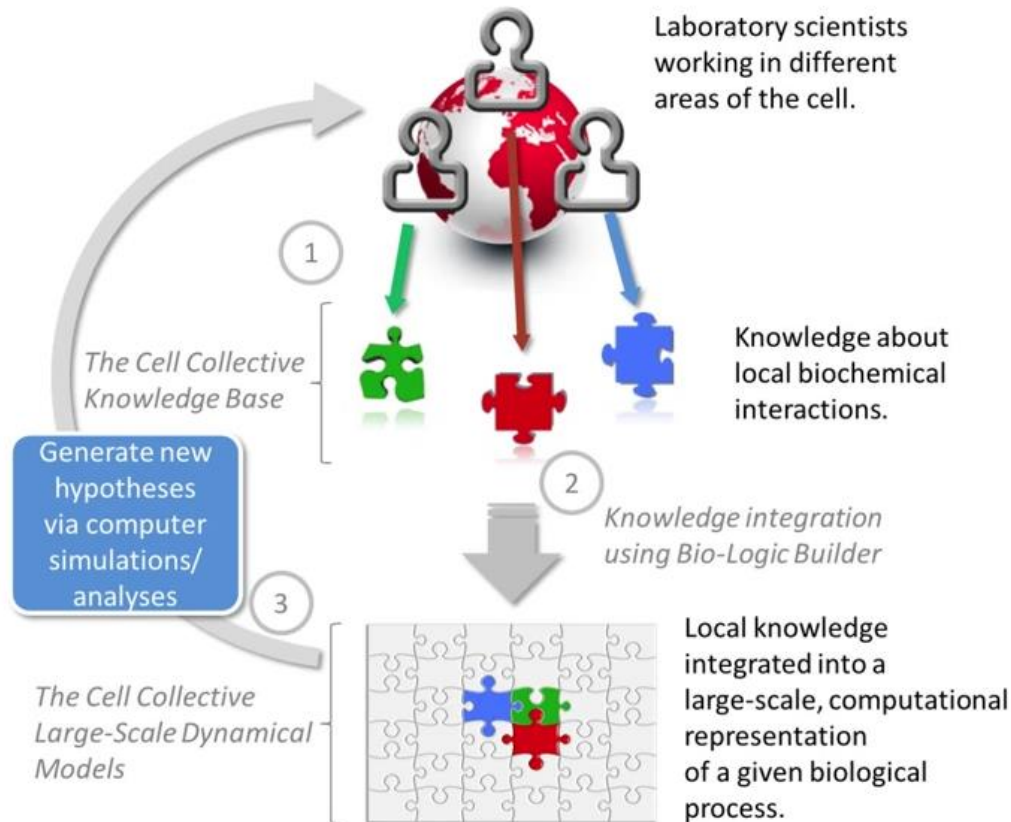


Figure 3.9: Integration and understanding of the biological knowledge with CellCollective (Taken from (Helikar et al., 2012)).

3.7.2 GINsim

For the computation of stable states, we used GINsim (Chaouiya et al., 2012), powerful software for constructing and analyzing logical models. The GINsim software supports the definition, the simulation and the analysis of regulatory graphs, based on the (multi-valued) logical formalism. It is based on the definitions of: (i) logical regulatory graphs, describing regulatory interactions between genes and (or via) their products, and (ii) state transition graphs, which represent the qualitative dynamical behaviour associated with a given regulatory graph, for given initial states.

GINsim can import SBML-qual files; however, it needs a pre-processing step to display the name and not the species IDs. Imported models retain their formulae, as well as the layout but are currently stripped from annotations during pre-processing.

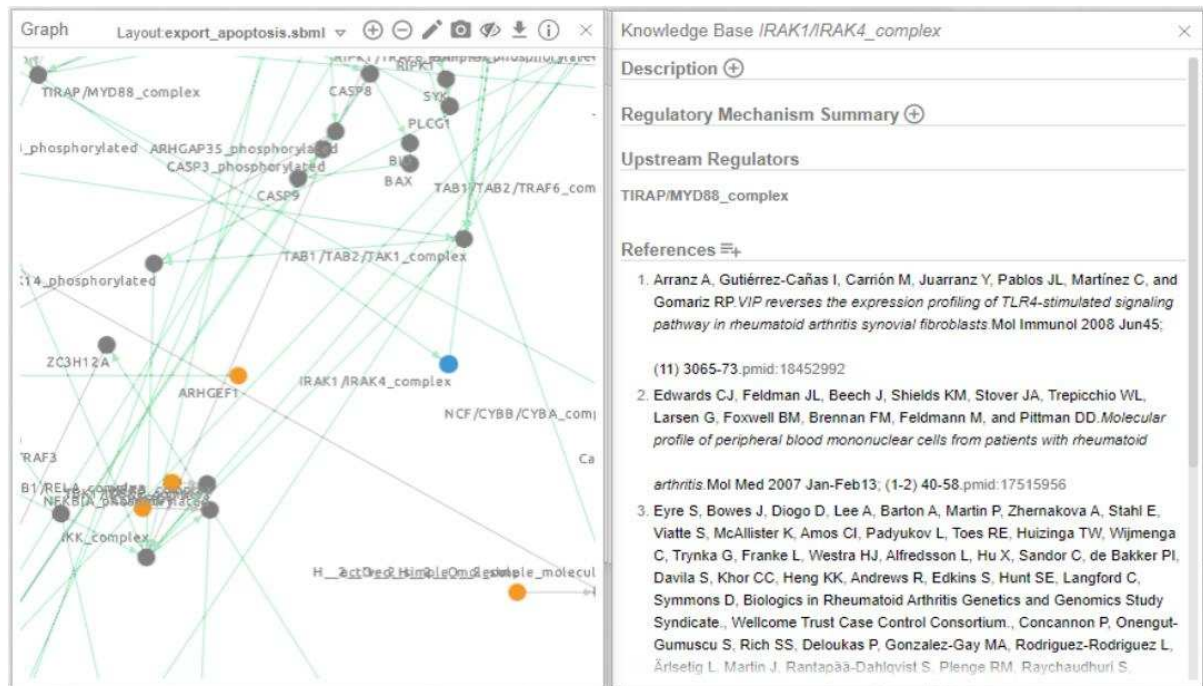


Figure 3.10: **View of the CaSQ inferred models using the modelling platform Cell Collective. The CaSQ tool produces annotated Boolean models, including bibliographical references stored in MIRIAM.** References stored in the MIRIAM section of the xml file of the molecular maps built with the software CellDesigner can be retrieved and visualised in the Cell Collective modelling platform. The original map layout is also conserved facilitating simulations. In the left panel, the selected node that corresponds to the IRAK1/IRAK4 complex is shown in blue while on the right panel the corresponding references are displayed.

3.8 Graph reduction and model inference

We first tested the tool with different molecular maps of various sizes, complexities and use of standards to see if CaSQ was able to produce corresponding executable models. We performed the analysis with CaSQ first by default and then using the BCC option. While a model should be connected to be useful, a map can include unconnected parts as the objective of a map is to represent all current knowledge for the studied biological process and this knowledge is more likely to be fragmented. The purpose of using CaSQ with default and BCC options was also to evaluate the graph reduction capacities of the tool. The size was defined by the number of nodes

included in the map (number of species in the CellDesigner files), and the number of components included in the published, manually built or CaSQ-inferred models.

CaSQ was able to handle small-, medium- and large-scale maps (ranging from 125 to 1361 nodes) with or without SBGN standards, and produce executable models smaller in size, offering a graph reduction of 21–45%. Using the BCC option that allows keeping the biggest connected component, the resulting models are slightly smaller. The size of the produced model—in terms of the number of components included—using BCC option is highly dependable on the connectivity of the initial map (**Table 3.2**).

Table 3.2: Size (number of components) of the CaSQ inferred model using the default and BCC options.

Map name	Map size	SBGN standard	CaSQ inferred model			
			Size	Graph reduction percentage	Size of BCC	Graph reduction percentage
Mast cell activation map	125	No	80	36%	73	42%
Mitogen-activated protein kinase (MAPK)	232	No	182	21%	181	22%
Cholecystokinin	530	No	404	24%	383	28%
Rheumatoid arthritis (RA)	779	Yes	431	45%	391	50%
Alzheimer's	1361	No	1169	14%	762	44%

3.9 CaSQ run time

The analysis was performed on a Dell working station with Windows 7, 64-bit Operating System, Installed memory (RAM): 64.0 GB and Processor: Intel (R) Xeon (R) CPU E5-1650 v4 @ 3.60 GHz 3.60 GHz. and the run times of CaSQ can be seen in **Table 3.3**.

Table 3.3: The run times of CaSQ for producing executable SBML-qual files with default options.

Map	Run time (seconds)
Mast cell activation map	1,42
MAPK	1,10
Cholecystokinin	1,71
RA	2,29
Alzheimer's	5,24

3.10 CaSQ-inferred Boolean models versus manually built models

3.10.1 Shared nodes

To evaluate the tool's ability to produce preliminary Boolean rules, we compared the CaSQ-inferred models with the manually built models (MM) published with the respective maps. First, we compared the size and graph reduction percentage (**Table 3.4**). For the size, we compared the shared nodes between the two models. The automated comparison gives the number of identical node names while the manual comparison accounts for differences in node names that derive from the fact that the manually built models do not correspond 100% to the maps. A modeller may choose to merge two nodes (i.e. receptor–ligand), change the name of one node (i.e. use capitals or add underscores for a complex), entirely skip it or add a node that does not exist in the initial map, making it difficult to evaluate in a fully automated way the correspondence between the manually built and the CaSQ-derived models. Manual comparison by visual inspection after the automated comparison revealed many cases where the node names were slightly different but corresponded to the exact protein or gene (**Table 3.5 and 3.6**). For example regarding the mast cell activation models, the manual model has RAS but the CaSQ model has H-RAS. Other cases concern grouping of instances, i.e. FYN in the manually built model corresponds to more instances in the CaSQ one, as the latter includes FYN with different modifications (phosphorylated, palmitoylated) (**Table 3.7**).

Table 3.4: Comparison of CaSQ inferred Boolean models with manually built models (MM).

Maps Name	Map Size (nodes)	SBGN standard	Manually built models (MM)		CaSQ inferred model (Size of the biggest connected component)		
			Size	Graph reduction percentage	Size	Graph reduction percentage	Shared Nodes between CaSQ and MM
Mast cell activation	125	No	47	62%	73	42%	64%
Mitogen-activated protein kinase (MAPK)	232	No	53	77%	181	22%	79%

Table 3.5: Shared nodes between CaSQ and manually built model for mast cell activation.

Manual model	CaSQ produced model
Bcr	Bcr
Btk	Btk, Btk_phosphorylated_Cytosol
C_Jun	c-JUN_phosphorylated_nucleus
c_Fos	Fos_phosphorylated
Ca	Ca2+_ion_Cytosol
Cbp_PAG	Cbp/PAG_palmytoylated
cCbl	cCbl
DAG	DAG_simple_molecule
Elk_1	Elk1_phosphorylated
ERK	Erk(MAPK1/3)_phosphorylated
FYN	Fyn_phosphorylated_palmytoylated, Fyn_phosphorylated_phosphorylated_palmytoylated
Gab2	Gab2
IP3	IP3_simple_molecule_Cytosol
JNK	JNK_phosphorylated
LAT	LAT1_phosphorylated_phosphorylated_phosphorylated_phosphorylated_palmytoylated
LAT2	LAT2_phosphorylated_palmytoylated

Lyn	Lyn_phosphorylated_palmytoylated
MEKK1	MEKK1(MAP3K1)_phosphorylated
NFAT	NFAT_Cytosol, NFAT_nucleus
NF_kB	NF-kappa_B_Cytosol, NF-kappa_B_nucleus
PIP2	PIP2_simple_molecule
PKC	PKC_space_Theta
PLCG1	PLCG1_phosphorylated
Raf	RAF1
Rac1	Rac1
RAS	H-RAS
RasGAP_Dok1	Dok1/RasGAP_complex
SHIP_1	SHIP-1_phosphorylated
SLP_76	SLP-76
Syk	Syk_Cytosol

Table 3.6: Shared nodes between CaSQ and manually built model for MAPK.

Manual model	CaSQ produced model
AKT	AKT_phosphorylated
Apoptosis	apoptosis_unknown
ATF2	ATF2_phosphorylated
ATM	ATM
BCL2	BCL2, BCL2_gene
CREB	CREB1_phosphorylated
DUSP1	DUSP1, DUSP1_don't care, DUSP1_gene
ELK1	ELK1_phosphorylated
ERK	ERK, ERK_phosphorylated_phosphorylated_Cytoplasm, ERK_phosphorylated_phosphorylated_Nucleus, ERK_phosphorylated_phosphorylated_Mitochondria
FOS	FOS, FOS_phosphorylated, FOS_gene
FOXO3	FOXO3_phosphorylated
FRS2	FRS2_phosphorylated

GAB1	GAB1
GADD45	GADD45_gene
GRB2	GRB2
Growth_Arrest	growth arrest_unknown
JNK	JNK, JNK_phosphorylated_phosphorylated_Cytoplasm, JNK_phosphorylated_phosphorylated_Nucleus
JUN	JUN_phosphorylated
MAP3K1_3	MAP3K1, MAP3K2_3
MAX	MAX_phosphorylated
MDM2	MDM2, MDM2_gene
MEK1_2	MEK, MEK_phosphorylated_phosphorylated
MSK	MSK_phosphorylated
MYC	MYC, MYC_gene
p14	p14
p21	P21, p21_phosphorylated
p38	p38_phosphorylated_phosphorylated_Cytoplasm, p38_phosphorylated_phosphorylated_Nucleus
p53	TP53, TP53_phosphorylated
p70	p70_phosphorylated
PDK1	PDK1_unknown
PI3K	PI3K_phosphorylated
PKC	PKC
PPP2CA	PPP2CA_unknown
Proliferation	proliferation_unknown
PTEN	PTEN, PTEN_gene
RAF	RAF1, RAF1_phosphorylated_Cytoplasm, RAF1_phosphorylated_Mitochondria
RAS	RAS
RSK	RSK_phosphorylated

SMAD	Smad2, Smad3, Smad4, Smad2/Smad4_complex, Smad3/Smad4_complex
SOS	SOS
SPRY	SPRY_phosphorylated
TAOK	TAOK_phosphorylated

For the MAPK model, an example is p53 in the manual model that corresponds to TP53 and TP53 phosphorylated in the CaSQ counterpart, or SMAD in the manually built that corresponds to a grouping of different SMAD proteins. An additional problem that made the comparison difficult was the fact that the researchers made different decisions concerning their map and model building.

For instance, the receptor tyrosine kinase (RTK) component in the MAPK map represents several different receptors (e.g. EGFR, FGFR, VEGFR, etc.) while in the model they use explicitly the different receptors. The two models used for CaSQ's benchmarking are medium-sized models (47–53 nodes). CaSQ models are twofold to fourfold bigger because they are inferred automatically from the corresponding maps (**Table 3.4**).

The CaSQ-inferred model for mast cell activation comprises 73 nodes while the manually built, 47 nodes. The authors of the manually built extracted information from the molecular map, but they also used proteomic data from bone marrow mononuclear cells (BMMCs) reported in (Bounab et al., 2013) that focused on the SLP-76 protein and its partners. Node comparison revealed that 30 of these nodes are shared between the CaSQ inferred and the manually built models (**Table 3.5**).

Table 3.7: Examples of different naming of shared nodes between CaSQ inferred and manually built models for mast cell activation and MAPK.

Manually built model for mast cell activation	CaSQ inferred model
RAS	H-RAS
FYN	Fyn_phosphorylated_palmytoylated, Fyn_phosphorylated_phosphorylated_pal mytoylated
Manually built model for MAPK	CaSQ inferred model

p53	TP53, TP53 phosphorylated
SMAD	Smad2, Smad3, Smad4, Smad2/Smad4_complex, Smad3/Smad4_complex

3.10.2 In silico simulations and dynamic analysis

Next, we simulated CaSQ-inferred models to see if they were capable of capturing the system's dynamics even though they were not identical with their manually built counterparts.

3.10.2.1 Comparison of the CaSQ-inferred model and the manually built model for mast cell activation.

One important difference, besides size and logical formulae, is also the fact that the mast cell activation model contains one multivariate variable while CaSQ -inferred models are strictly Boolean. Despite the differences, CaSQ mast cell model was able to reproduce the Btk (**Fig. 3.11**) and Syk (**Fig. 3.12**) knockout experiments described in the publication (Niarakis et al., 2014).

In **Figure 3.11 and 3.12**, we see simulation examples of the CaSQ-inferred model for mast cell activation in Cell Collective. In the case of Btk knockout, a decrease in cytokine release and degranulation, as well as a decrease of PLCG1 and ERK levels have been observed (Kajita et al., 2010; Setoguchi et al., 1998). The simulation of Btk knockout using Cell Collective platform resulted in PLCG1 and ERK set to zero, a result that is directly comparable with the simulation described in (Niarakis et al., 2014) (**Fig. 3.11**).

In Syk knockout experiments, cytokine release and degranulation are both abolished (Gilfillan & Tkaczyk, 2006). We performed an in silico simulation of Syk knockout, with Lyn and PIP2 present at the initial state in Cell Collective as described in (Niarakis et al., 2014) (**Fig. 3.12**). In this condition, the CaSQ-inferred model reaches a state where ERK, JNK, Elk-1, NF-kB, NFAT, PKC, PLCG1, Ca²⁺ are all set to zero, in agreement with the simulation described in (Niarakis et al., 2014).

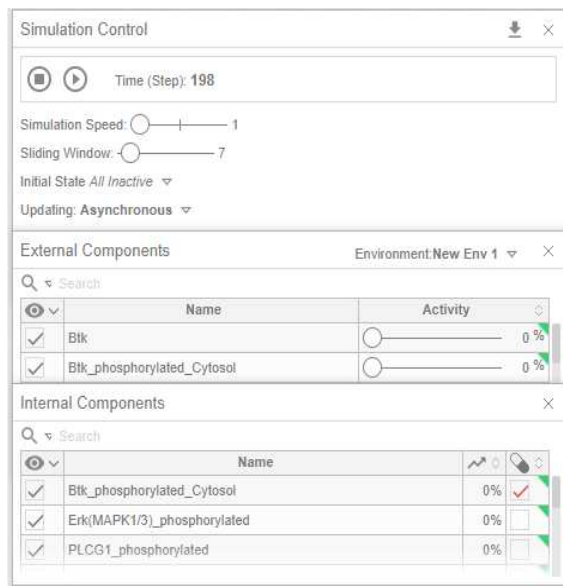
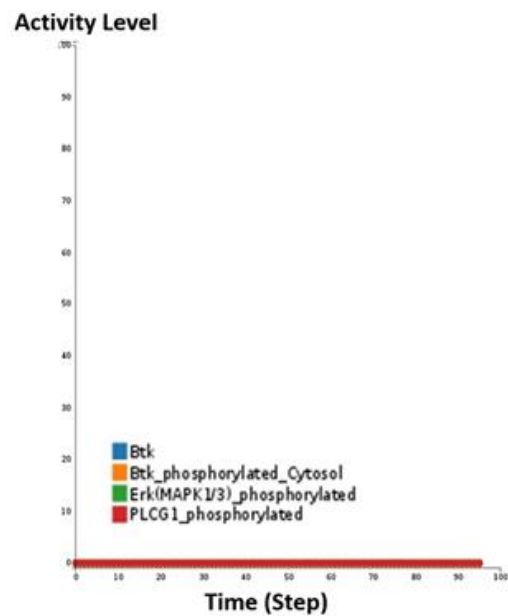
A**B**

Figure 3.11: **A)** Screenshot of simulations for Btk knockout of the CaSQ derived mast cell activation model using Cell Collective. **B)** In the graph panel, one can see that when Btk is set to zero, Erk and PLCG1 are not expressed.

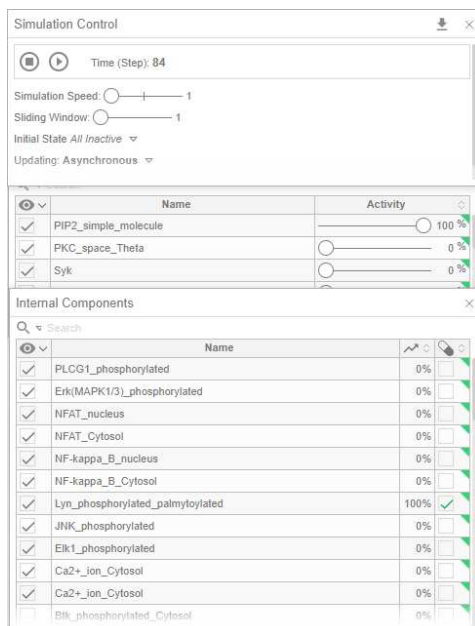
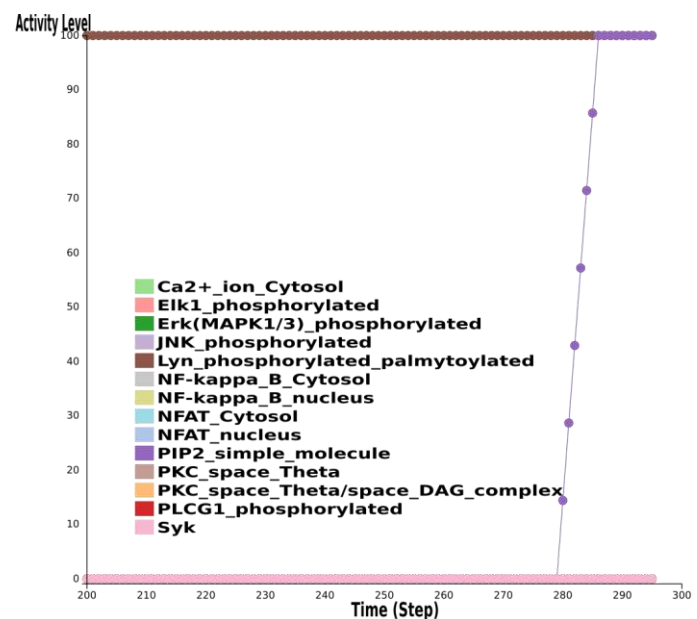
A**B**

Figure 3.12: **A)** Screenshot of simulations for Syk knockout of the CaSQ derived mast cell activation model using Cell Collective. **B)** In the graph panel, one can see that when Syk is set to zero, Erk, JNK, NFAT, NFkB, Ca2+, PKC, Elk1, PLCG1 are not expressed.

3.10.2.2 Logical steady-state analysis for the mast cell activation models

We computed all the stable states of both the CaSQ-inferred model and the manually built one for mast cell activation using bioLQM java toolkit included in GINSim (<http://colomoto.org/biolqm/>). We obtained 18 stable states for the manually built model (**Fig 3.13**) and 524.288 for the CaSQ-inferred one. The difference in the number of stable states lies in the fact that the automatically inferred model is a close representation of the system as described in a molecular map and thus significantly bigger in size, including especially a much higher number of inputs. The manual counterpart is smaller in size and also of reduced complexity as several inputs are grouped and thus, the computation of stable states leads to considerably fewer solutions.

As shown in **Table 3.5**, 30 components can be matched together between these two models. We then projected the identified stable states on these 30 components, which reduced the lists to nine stable states for the manually built model and 43.392 for the CaSQ-inferred one. Indeed, some of the original stable states only differ in the unmatched components and are thus projected on the same state. We found that three of the nine stable states of the manually built model are precisely reproduced in the CaSQ-inferred model. If we accept a single difference between the states, we can recover four additional stable states, whereas the last two stable states can be recovered with two differences (**Table 3.8**).

[illegible]

Figure 3.13: Screenshot of the stable states table for the manually built model for mast cell activation with GINsim.

Table 3.8: Logical steady-state analysis obtained with GINsim, for manually built and CaSQ inferred mast cell activation models. The nine stable states of the manually built model are given in this table (SS1 to SS9). The bottom lines show which of these stable states are correctly recovered in the CaSQ inferred model, or have 1 or 2 mismatches.

Component	SS1	SS2	SS3	SS4	SS5	SS6	SS7	SS8	SS9
Bcr	0	0	0	0	0	0	1	1	1
Btk	0	0	0	0	0	0	1	1	1
Ca	0	0	0	0	0	0	1	1	1
cCb1	0	0	0	0	0	0	0	1	1
Elk1	0	0	0	0	0	0	0	1	1
ERK	0	0	0	0	1	1	1	1	1
Gab2	0	1	0	1	0	0	0	0	0
JNK	0	0	0	0	0	0	0	1	1
LAT	0	0	0	0	0	0	1	1	1
LAT2	0	0	0	0	0	0	0	1	1
Lyn	1	1	1	1	0	0	0	0	0
NFAT	0	0	0	0	0	0	0	0	0
NFkB	0	0	0	0	0	0	0	1	1
PIP2	0	0	0	0	0	0	1	1	1
PKC	0	0	0	0	0	0	1	1	1
PLCG1	0	0	0	0	0	0	1	1	1
Rac1	0	0	0	0	1	1	1	1	1
RAS	0	0	0	0	0	0	1	1	1
RasGAP-Dok1	0	0	0	0	0	0	0	1	1
SHIP1	0	0	0	0	0	0	0	1	1
SLP76	0	0	1	1	0	1	0	1	1
Syk	0	0	0	0	0	0	0	1	1
C_Jun	0	0	0	0	0	0	1	1	1
c_Fos	0	0	0	0	0	0	0	1	1

Cbp_PAG	0	0	0	0	0	0	1	1	1
DAG	0	0	0	0	0	0	0	1	1
FYN	0	0	0	0	0	0	1	0	1
IP3	0	0	0	0	0	0	1	0	1
MEKK1	0	0	0	0	0	0	1	1	1
Raf	0	0	0	0	0	0	1	1	1
Exact match		x		x			x		
One difference	x		x					x	x
Two differences					x	x			

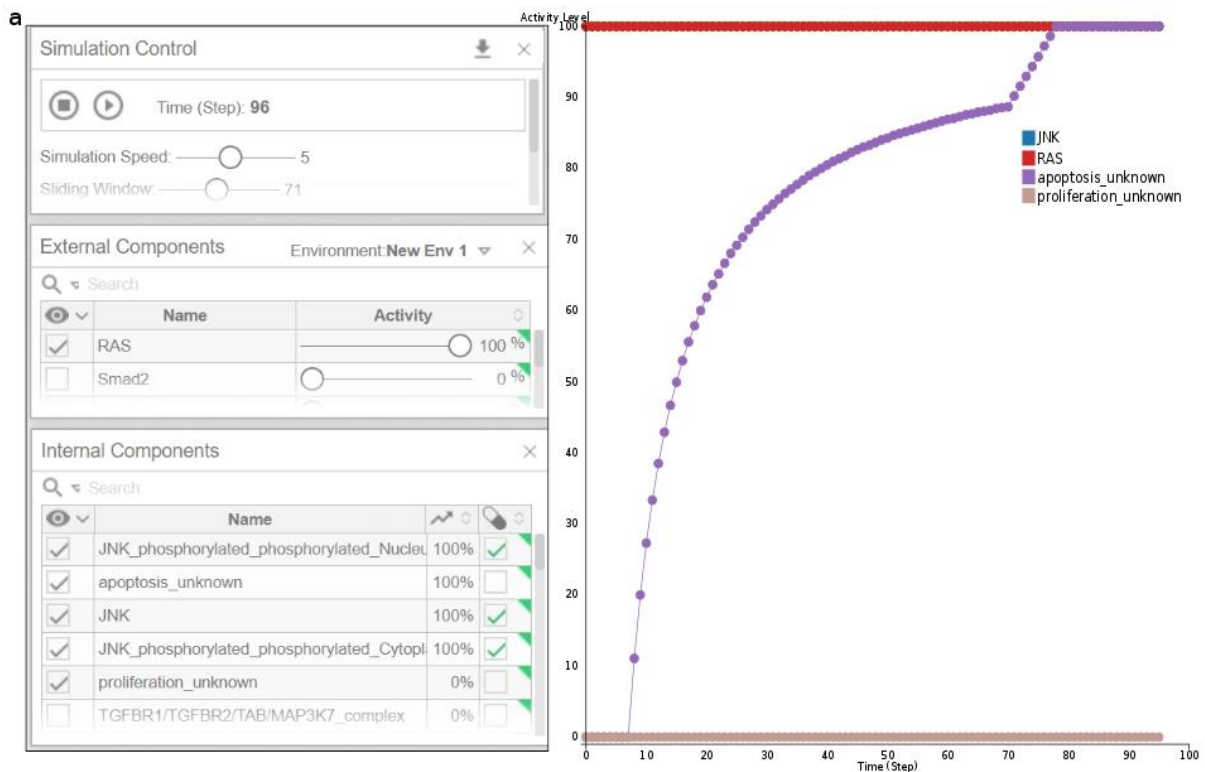
3.10.2.3 Comparison of the CaSQ-inferred model and the manually built model for MAPK

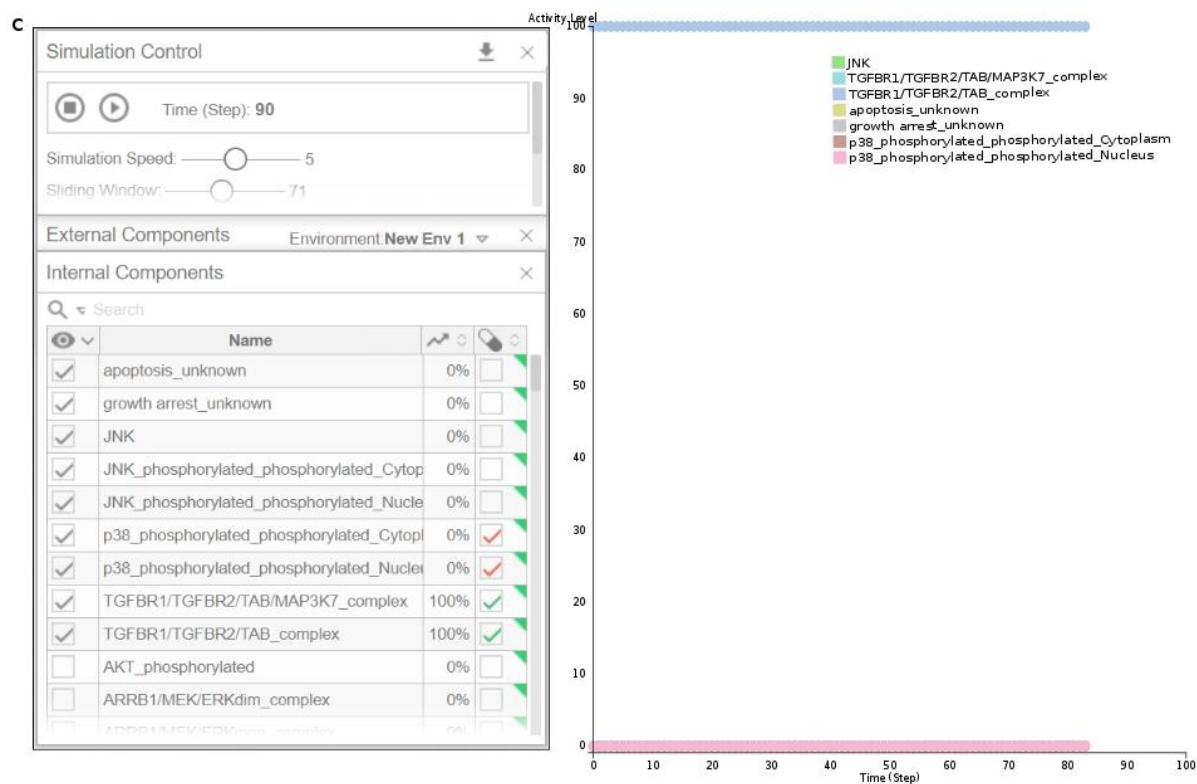
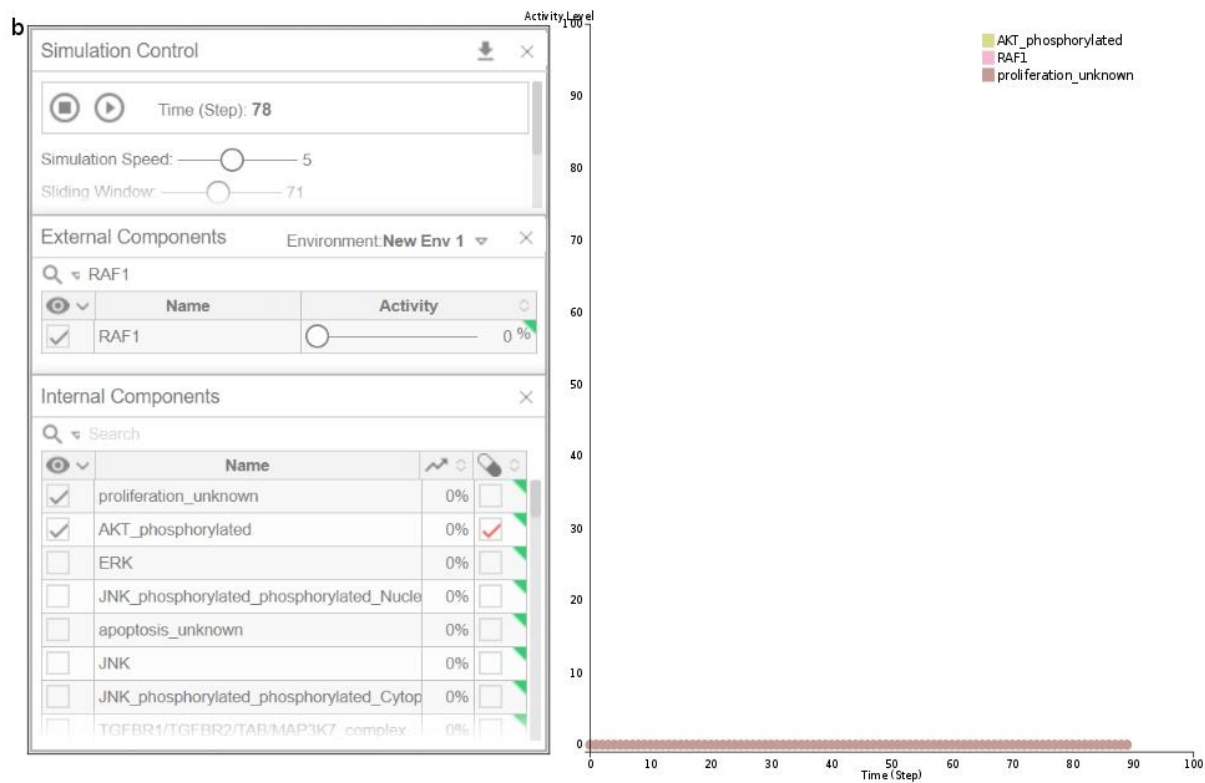
Concerning the MAPK manually built model, the authors produced a model that did not follow strictly the corresponding map (the model contained several merged inputs and merged outputs). As stated above, the RTK component in the map represents several different receptors like EGFR, FGFR and VEGFR that the researchers decided to include in the model explicitly. Besides, to cope with simulations of their model, they used the model reduction option in GINsim (Grieco et al., 2013) to produce different smaller sub-versions of the original model, each dedicated to a subset of simulations. In **Table 3.9**, we have regrouped biological scenarios modelled successfully with the MAPK manual model and the corresponding behaviour of the CaSQ counterpart. For the simulations of the CaSQ model, we used the platform Cell Collective as before (**Fig 3.14**). These reduced versions of the original MAPK model (52 components) ranged from 16 to 18 components. The CaSQ-inferred model for MAPK is inferred directly from the MAPK map and is thus significantly bigger in size and different in structure. However, comparison of the model's behaviour regarding its efficacy in capturing the systems dynamics, showed that the CaSQ model was able to reproduce partially or completely known biological scenarios. The size of the CaSQ-inferred MAPK model (181 nodes) made the calculation of stable states a non-realistic endeavour. Moreover, the fact that the manually built counterpart had to undergo multiple reductions for the dynamic analysis would not have made the comparison straightforward.

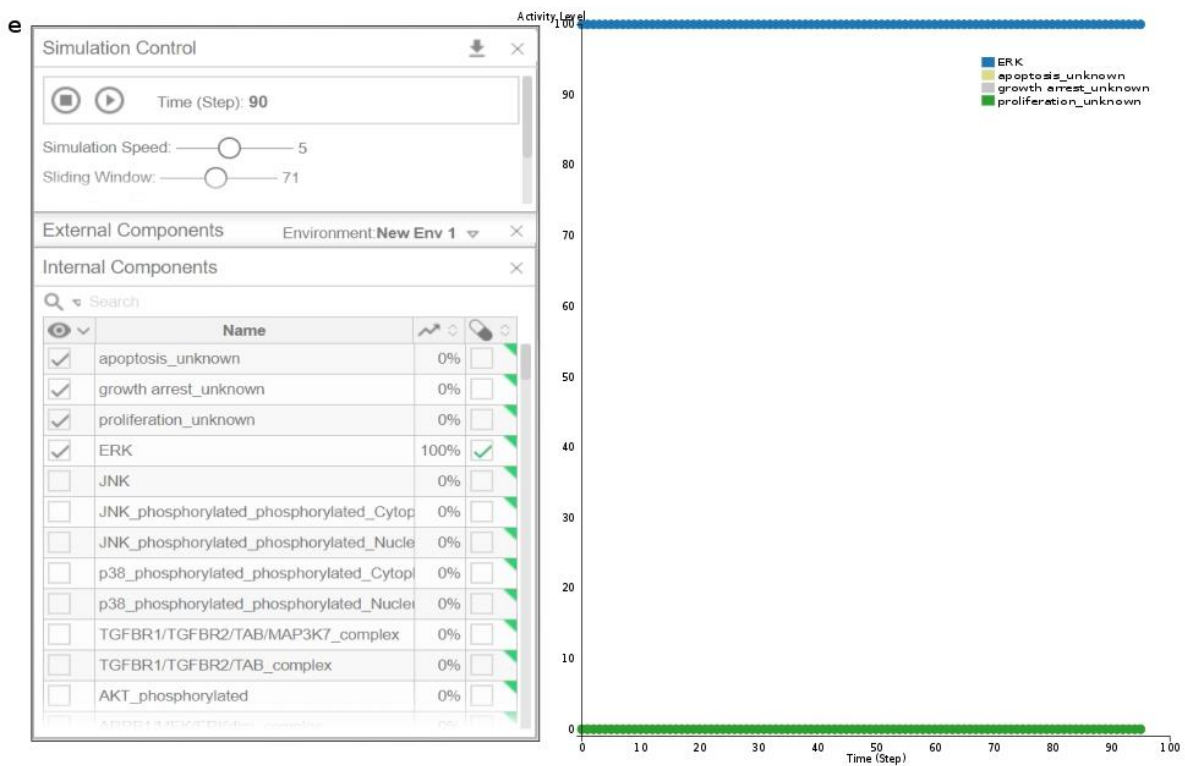
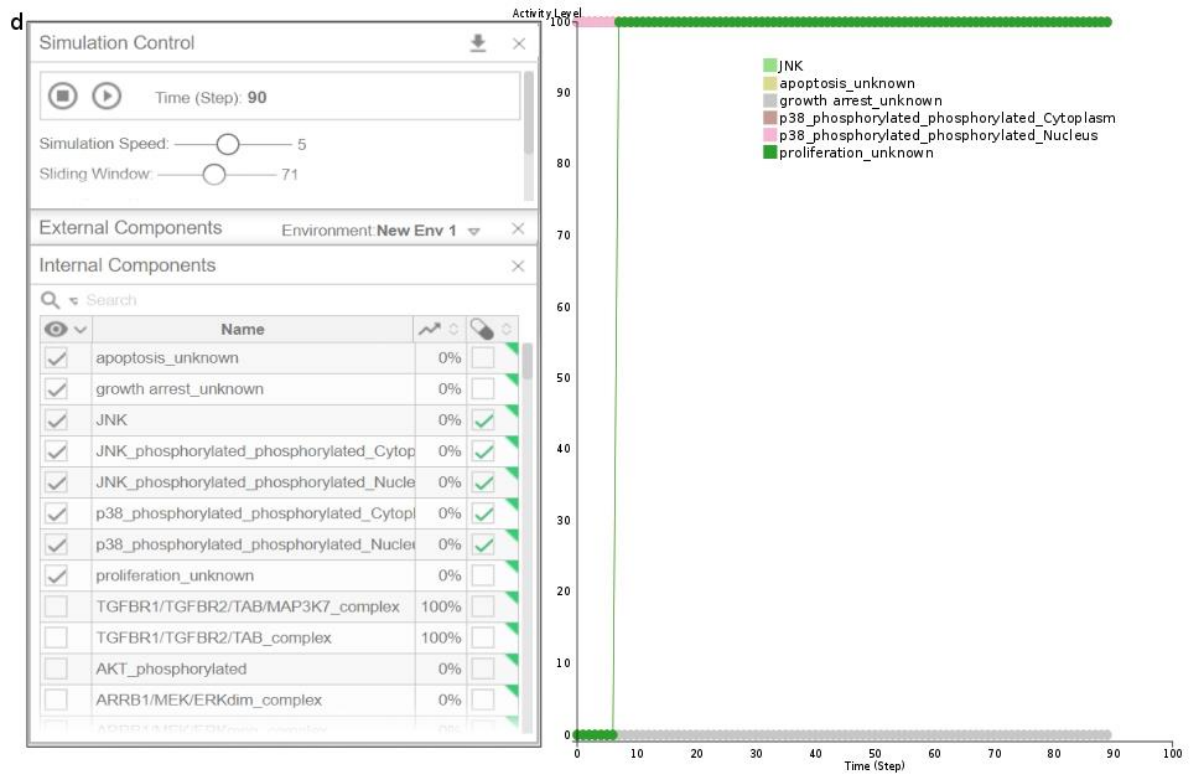
Table 3.9: Biological data and corresponding behaviours of the manually built and the CaSQ inferred models for MAPK.

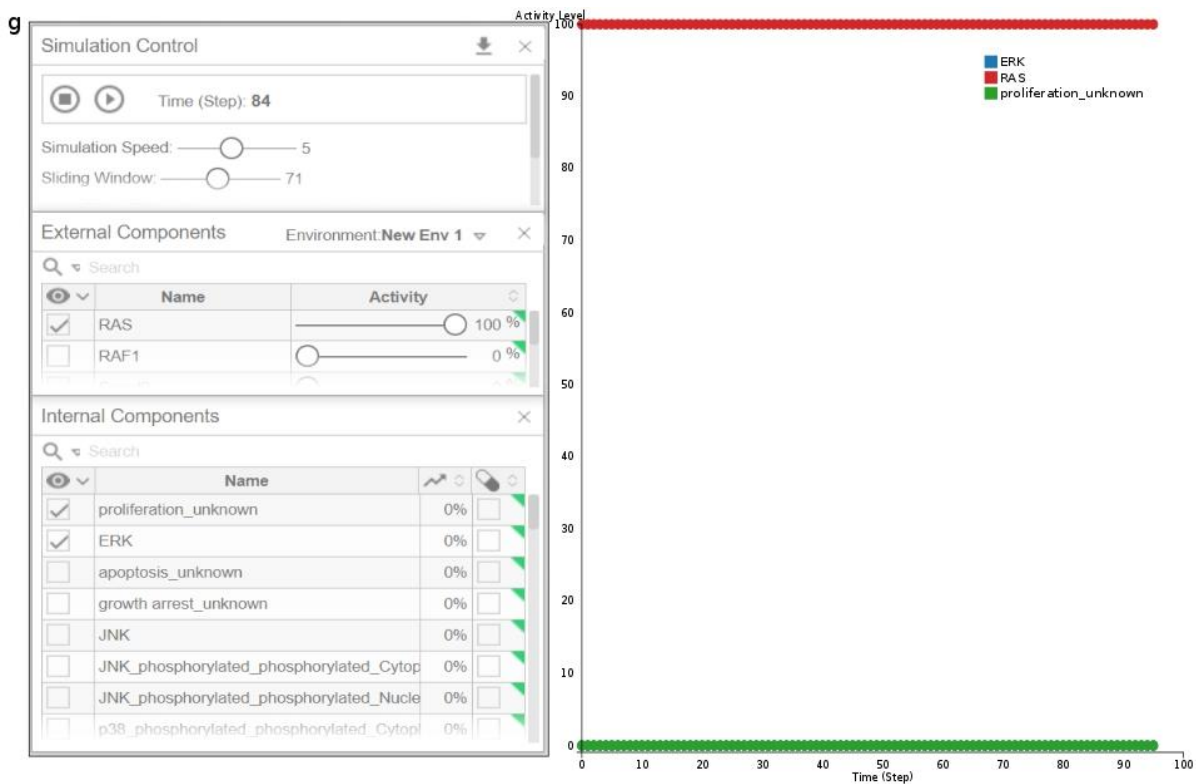
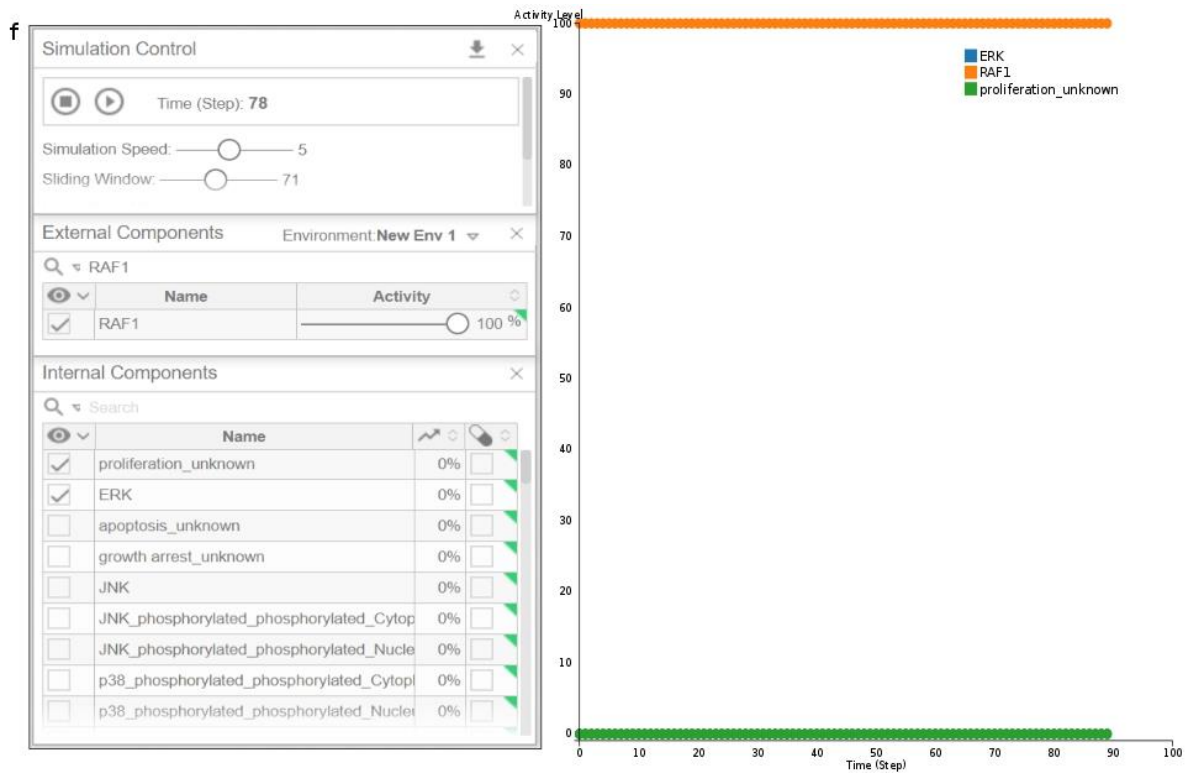
Biological data	MM built MAPK model	CaSQ inferred MAPK model	Agreement of the models behaviours
1.JNK might reduce RAS-dependent tumour formation by inhibiting proliferation and promoting apoptosis (Kennedy & Davis, 2003)	<i>When JNK is always ON and RAS is always ON then proliferation is OFF and apoptosis is ON</i>	<i>When JNK is always ON and RAS is always ON then proliferation is OFF and apoptosis is ON (Figure 3.13a)</i>	Agreement
2.HSP90 inhibitor disrupts EGFR, RAF and AKT leading to successful cancer treatment (Sharp & Workman, 2006)	<i>Concomitant RAF, EGFR, AKT deletions block proliferation</i>	<i>There is no EGFR present in the model, RAF and AKT deletions lead to proliferation being OFF (Figure 3.13b)</i>	Agreement
3.P38 and JNK play important roles in stress responses such as cell cycle arrest and apoptosis (Kyriakis & Avruch, 2001; Takekawa et al., 2011)	<i>When p38/JNK are OFF (KOs) and TGF and DNA damage are ON then there is no growth arrest or apoptosis</i>	<i>There is no DNA damage present in the model, p38/JNK constitutively OFF and TGF stimuli ON, then Growth arrest is OFF and Apoptosis is OFF (Figure 3.13c)</i>	Agreement
4.P38 and JNK, especially in the absence of mitogenic stimuli, have been shown to induce apoptotic cell death (Kyriakis & Avruch, 2001; Takekawa et al., 2011)	<i>When P38/JNK are constitutively ON then Growth arrest is ON, Apoptosis is ON and proliferation is OFF</i>	<i>When p38/ JNK are constitutively ON then Growth arrest is OFF, Apoptosis is ON, and proliferation ON (Figure 3.13d)</i>	Partial agreement

<p>5. ERK increases transcription of the cyclin genes and facilitates the formation of active Cyk/CDK complexes, leading to cell proliferation (Schramek, 2002)</p>	<p><i>When ERK is always ON then Apoptosis and Growth arrest are OFF, and proliferation is ON</i></p>	<p><i>When ERK is constitutively ON then Apoptosis and Growth arrest are OFF, and proliferation is OFF (Figure 3.13e)</i></p>	<p>Partial agreement</p>
<p>6. RAF or RAS overexpression can lead to constitutive activation of ERK (A. S. Dhillon et al., 2007)</p>	<p><i>When either RAS or RAF are constitutively active then ERK is ON and proliferation is ON</i></p>	<p><i>When either RAF or RAS or both of them are constitutively active, then ERK is OFF and proliferation is OFF (Figure 3.13f, g, h)</i></p>	<p>Disagreement</p>









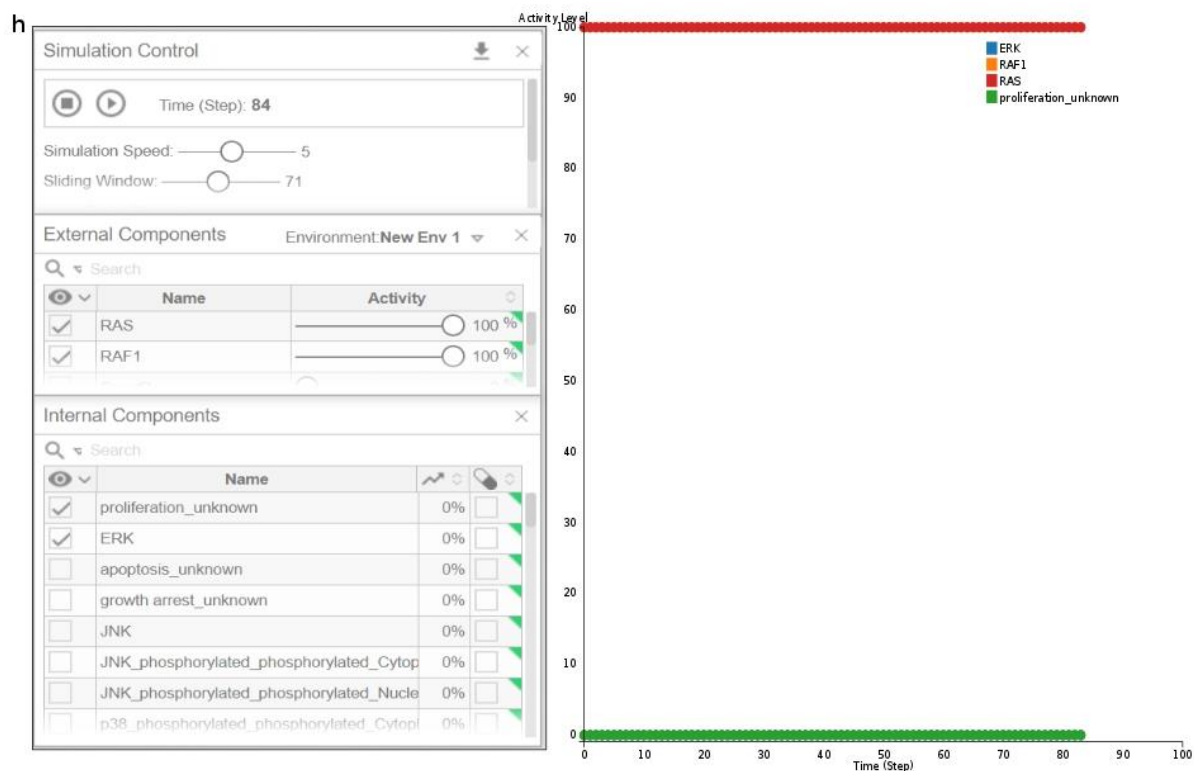


Figure 3.14: Simulations of the CaSQ inferred model using the modelling platform Cell Collective. The CaSQ inferred model for MAPK was able to reproduce known biological scenarios, either completely or partially. The results of the *in silico* simulations for the three first biological conditions described in Table 9 showed perfect agreement with the results of manually built model, as depicted in panels a, b and c. For conditions described in scenarios 4 and 5 of Table 9 the CaSQ inferred model could partially reproduce the attended behaviour (panels d and e) while simulation results for scenario 6, were inconsistent with the literature and the results of the manually built model (panels f, g and h).

3.11 Discussion

Building large-scale dynamic models can be tedious and time-consuming work that requires not only the construction of the regulatory graph but also the writing and tuning of the logical formulae. CaSQ is a tool aiming to ease the construction of large-scale Boolean models, taking advantage of the similarities shared between molecular interaction maps and dynamic models. In the framework proposed, we utilize systems biology standards for model construction (SBML-qual), so that the CaSQ tool can be interoperable with other tools and modelling software. For the inference of the logical formulae, we based our assumptions on topology and semantics of the molecular maps. More precisely, we decided to approach the conversion

process using mostly OR gates over AND, so a target is on if one of the reactions producing it is on, a reaction is on if all reactants are on, all inhibitors are off and one of the catalysts is on. The idea behind this assumption is that, very rarely, we have exact information about the need for the presence of two or more activators for one target. Even if synergy is defined, very often a relative activation can happen even by the presence of one activator. Moreover, the number of events for which we do have such information is significantly lower than the uncertain ones and tuning the rules by hand should be a quick process. With the use of CaSQ, as demonstrated in this study, we can now obtain large-scale Boolean models that can be executed using popular modelling software that can import SBML-qual files.

In this work, for comparing the tool's performance and accuracy, we compared the common nodes between the CaSQ inferred and the manually built models, their ability to reproduce biological scenarios performing simulations, and finally, we performed a comparison of stable states, where possible. We also performed simulations to see if the CaSQ-inferred models could reproduce some of the dynamics of the original system. The next step was to perform logical steady-state analysis. For this purpose, we used GINsim, powerful software for logical modelling. The goal was to see if within the stable states of the CaSQ-inferred model, we could retrieve the stable states of the published manually built model.

We should note that CaSQ infers preliminary Boolean rules, so the modeller still needs to fine-tune the model and find the best logical rules to reproduce data accurately. (Bekkar et al., 2018) show that logical models with added human curation perform better than models where rules are extracted automatically from a given topology. As demonstrated in the results, the CaSQ tool produces models that are largely in agreement with the model a human modeller would build, accelerating the time of model construction impressively.

This work was also a motivation for community work, as it addressed issues of model reusability, use of Systems Biology standard formats and interoperability between different tools that have complementary functionalities. As demonstrated, our method is scalable, and the large-scale SBML-qual models produced by CaSQ can be imported in Cell Collective and retain layout and annotations. The goal is to propose a seamless pipeline for producing executable Boolean models starting from molecular interaction maps which can be analyzed in depth using various tools for computational modelling. CaSQ tool can play the role of a bridge bringing together two distinct communities, curators and modellers to produce interoperable, annotated models of better quality, accuracy and reusability.

Chapter 4. Inference of a modular, large-scale Boolean network for modelling the Rheumatoid Arthritis fibroblast-like synoviocytes

4.1 Introduction

Rheumatoid arthritis fibroblast-like synoviocytes (RA FLS) are critical effectors in the pathogenesis of RA. They are involved in the secretion of cytokines and proteolytic enzymes, and they have also exhibited a high rate of self-proliferation. Several attempts have been made to target the activities of RA FLS with cell signalling inhibitors in order to decrease the severity of the disease. One example is the imatinib, inhibitors of PDGF receptor tyrosine kinase, which is responsible for the activation of the PI3K/AKT pathway that activates proliferation (Terabe et al., 2009). The major roles and contributions of RA FLS in RA can be classified into four main categories (Bartok & Firestein, 2010; Turner & Filer, 2015).

The first category concerns RA FLS and apoptosis. The formation of the pannus structure is caused by the irregular proliferation of RA FLS. Many studies have suggested that the expansion of the cell population results mainly from the inhibition of the pro-apoptotic pathways. Numerous oncogenes such as c-fos, ras, raf, sis, myb and myc were detected in elevated levels in RA FLS. Somatic mutation of p53 and PTEN have also been identified in the RA FLS (Smith et al., 2010).

The second category is about RA FLS and innate immune response. As mentioned before, RA FLS play a central role in the pathogenesis of the RA by activating the innate immune response. This ability contributes significantly in all stages of the RA pathogenesis such as initiation, propagation and maintenance of chronic inflammation. The immune response is maintained via the secretion of soluble molecules such as pro-inflammatory cytokines, in response to environmental stimuli and interactions with other cells. Identified pro-inflammatory molecules such as IL6, IL1 and TNF act as activators of pro-inflammatory responses through the recruitment of immune cells and promote the production of inflammatory mediators and bone and matrix-degrading enzymes. Secretion of these molecules act as a positive feedback loop and eventually trigger the activation of RA FLS into expressing the responses again and again (Bartok & Firestein, 2010).

The third category involves RA FLS and cell recruitment. A result of RA FLS activation by pro-inflammatory cytokines is the recruitment of immune cells. Secreted chemokines such as

CXCL1-3 and CXCL8 (or IL8) are responsible for the prolonged presence of immune cells in the synovium, by preventing both emigration outside the joint environment and also cell deaths (McGettrick et al., 2010).

Lastly, the fourth category links RA FLS and bone & matrix degradation. The final stages of the disease involve the degradation of both bone and cartilage. These RA characteristic features are the result of chronic inflammation within the joint area. During chronic and sustained inflammation, RASFs secrete two groups of soluble molecules: a) receptor activator of nuclear factor kappa-B ligand (RANKL), a molecule that promotes osteoblasts differentiation to osteoclasts, cells that are responsible for bone degradation and bone resorption (Boyle et al., 2003) and b) matrix metalloproteinases (MMPs), a group of matrix proteases responsible for the degradation and breakdown of numerous extracellular matrix components such as collagen, leading to the degradation of cartilage (Burrage et al., 2006). The synergistic activity of these factors (RANKL and MMPs) leads to the gradual degradation of bone and cartilage in the joint area, leading to stiffness, pain and eventually disability of movement.

Over the last two decades, mathematical modelling approaches have been utilized to understand the pathogenesis of RA. The ordinary differential equation (ODE) models are used for describing the evolution of RA, as well as the dynamics of drugs used for the RA treatment (Macfarlane et al., 2019).

In this work, we build a large-scale dynamic model of RA FLS to study apoptosis, cell proliferation and growth, osteoclastogenesis and bone erosion, matrix degradation and cartilage destruction and inflammation outcomes. The RA FLS network was created selecting fibroblast-relevant sub-parts of the state-of-the-art RA map (Singh et al., 2020), and the Boolean rules were added using the tool CaSQ that infers logical formulae based on the topology and the semantics of the network (Aghamiri et al., 2020).

In an effort to cope with complexity, we also employ a “divide and conquer” strategy by creating separate executable sub-modules that comprise each only one phenotype. We present also reduced model versions that facilitate downstream analysis.

Systematic testing of different initial conditions could further lead to predictions regarding the outcomes of specific perturbations, such as single or combined effects, simulated with the model by forcing or suppressing the activity of various factors of interest systematically. The goal is to gain a better understanding of the mechanisms that drive inflammation, resistance to

apoptosis, high proliferation rate, and cartilage and bone degradation, and their coordination, in an effort to delineate and gain control of these outcomes.

4.2 Methods and data

4.2.1 Using prior knowledge to build an RA FLS specific network

To recapitulate all that is known and published in various sources about the disease, we produced a global, comprehensive and fully annotated RA-specific map based on exhaustive literature mining, human curation and validation from domain experts (see Chapter 2 - RA map)(Singh et al., 2020). This map features interactions implicated in RA coming from various cell types. Thanks to the extensive annotations of each entity and each reaction included in the map, and the advanced functionalities offered by the MINERVA platform (Gawron et al., 2016; Hoksza et al., 2019), the user can opt for cell-specific interactions and extract the corresponding cell-specific networks. Using the synovial fibroblasts overlay and the MINERVA upstream plugin (Hoksza et al., 2019) a fibroblast specific network based on five functional outcomes was constructed (including as cellular phenotypes: osteoclastogenesis and bone erosion, matrix degradation, inflammation, proliferation and apoptosis). The cell specificity of our network was also assessed using as overlay a gene list from the meta analysis of single cell RNA-seq (Zerrouk et al., 2020) between RA and osteoarthritis (OA) patients. In this study, the authors used the gene expression omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>) dataset GSE109449 that includes freshly isolated synovial fibroblasts in 2 patients with RA and 2 patients with OA (Mizoguchi et al., 2018). The data was filtered keeping only high abundance genes with an expression higher than one and biotype as ‘protein coding genes’ eliminating low abundance, non coding RNA and pseudogenes. After performing quality control, differentially expressed genes were identified in RA samples using OA samples as reference using p value threshold equal to 0.05 and a fold change higher than 1.5 in absolute value. As mentioned above, to construct the network, we used the stream plugin. We selected as starting nodes the five functional outcomes, direction upstream, non blocking modifiers and SBML Cell Designer format. Besides obtaining the network with the five phenotypes, we repeated the procedure in order to obtain 5 individual networks corresponding each to one specific phenotype. All networks (5 phenotypes, and 5 individual subnetworks) were subsequently used with the map to model framework to obtain executable Boolean networks (Fig 4.1).

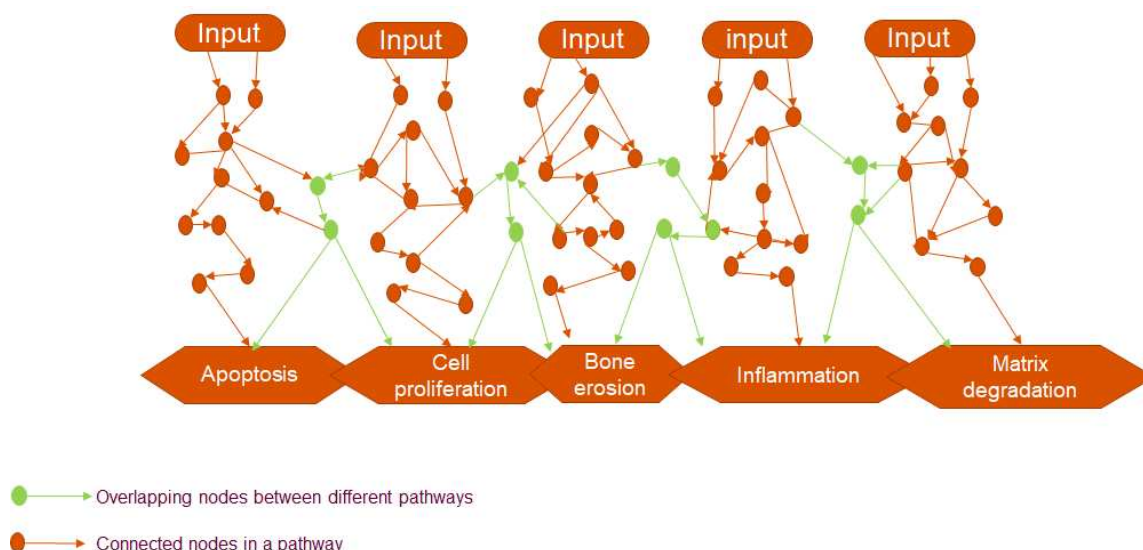


Figure 4.1: The notion of modules in our approach involves the selection of one (or more) phenotype(s) and follows the upstream regulators all the way to the initial inputs.

4.2.2 Using CaSQ to infer executable Boolean networks from the RA FLS specific network

The dynamic layer on the network was added using CaSQ (CellDesigner as SBML-Qual) (chapter 3), a tool for automated inference of large-scale, parameter-free Boolean models, from molecular interaction maps based on network topology and semantics (Aghamiri et al., 2020). The tool uses as input the XML file obtained from the RA map in MINERVA and produces an executable model in SBML-qual standard format, which can be further simulated and analysed using popular modelling tools.

We used the CellDesigner SBML files for the five individual networks and the merged network obtained in the previous step to create executable Boolean networks in SBML-qual format that were used for the dynamical analysis.

4.2.3 Real-time *in silico* simulations using the Cell Collective modelling platform

The SBML-qual file produced by CaSQ was imported in Cell Collective for further analysis. Cell Collective is a web-based platform that supports the analysis of large scale, Boolean models (see Chapter 3). In general, the activity levels of the models' individual constituents is

measured as %ON. Depending on the context of the biological process being simulated, this measure corresponds, for example, to concentration or the fraction of biological species being active at any given time. In both the real-time simulations and dynamical analysis, %ON is used as a semi-quantitative way to measure the dynamics of the modelled biological processes (Tomáš Helikar et al., 2012).

4.2.4 Attractors search using BoolNet

Identification and analysis of attractors in Boolean Networks is a crucial task. Attractors can be simple or complex stable cycles of states. A stable cycle with length equal to 1 corresponds to a stable state (fixed point). Attractors represent states where the system gets trapped and cannot escape without external intervention. They carry substantial biological implications and often can be linked to phenotypes as they can represent environments where the cell will commit/choose a certain destiny (Bornholdt, 2008; S. Kauffman et al., 2003; F. Li et al., 2004). Using BoolNet, synchronous attractors can be identified by exhaustive search of all 2^n states (for n genes). For asynchronous attractors a heuristic search starting from several predefined or randomly chosen states can be employed. For synchronous and probabilistic networks, potential attractor states and probabilities of reaching certain states can also be calculated using Markov chain simulations (Shmulevich et al., 2002).

We used BoolNet to calculate attractors for every module and the merged model, using the asynchronous updating scheme for two different initial conditions: when all node values are set to zero and when all node values are equal to 1.

4.2.5 Model reduction using GINsim

The CaSQ produced SBML-qual file was imported in GINsim (Chaouiya et al., 2012) after a slight modification to ensure interoperability (one needs to run the following command to ensure the software reads the names and not the IDs in the regulatory graph: `java -jar GINsim.jar -lm input.sbml -m sanitize output.sbml`). We used the reduction functionality of GINsim for producing more compact versions of the original model. The reduction of regulatory graphs allows to extract a "simplified" regulatory graph where a set of components are hidden. The logical rules that are associated with the targets of the hidden components account for the effects of their regulators. The reduced models preserve dynamical properties of the original model, including stable states and more complex attractors.

We removed 79 nodes from the primal model resulting in 248 nodes and 533 interactions. The smaller version was used for producing MaBoSS compatible files that were used in the downstream analysis.

4.2.6 Probabilistic Boolean modelling with MaBoSS

MaBoSS is a C++ software for simulating continuous/discrete-time Markov processes, applied to Boolean networks (Stoll et al., 2012, 2017) . MaBoSS uses a specific language for associating transition rates to each node. Given some initial conditions, MaBoSS applies Monte-Carlo kinetic algorithm to the network to produce time trajectories and thus can associate probabilities to the asymptotic solutions.

For the MaBoSS simulations we performed another step to reduce complexity. We decided to group the inputs of the merged model to ease the burden of simulations. For this step we grouped 31 inputs into 12 groups and removed 2 regulators (COMP and LEFTY2 regulating TGFB1 and SFRP5 regulating WNT) because they are specific to these mentioned components and could not be generalized for the whole class of the mediators.

4.3 Results

Using the RA map and the MINERVA stream plugin (Hoksza et al., 2019) we obtained a graph comprising the five relevant phenotypes for the RA FLS, namely osteoclastogenesis and bone erosion, matrix degradation (cartilage destruction), inflammation, proliferation and apoptosis). This network comprises 270 unique identifiers/ biological entities. To ensure that this network was fibroblast specific we used the overlays coming from the extensive annotations of the sources used to create the RA map to highlight fibroblast specific pathways. On a second step, we also used as overlay differentially expressed genes coming from transcriptomic data of RA FLS (**Fig 4.2**)(see Materials and methods for more details). The fibroblast overlay mapped 160 unique identifiers while the RNA seq list with 1447 genes gave back an exact match for only 25 entities. Among them, 12 were shared with the fibroblast overlay from annotated sources. In total, 173 biological entities of the network were mapped as fibroblast specific giving a coverage of about 64 %.

 - Differentially expressed genes single cell SF 0.05,  - Synovial fibroblasts overlay

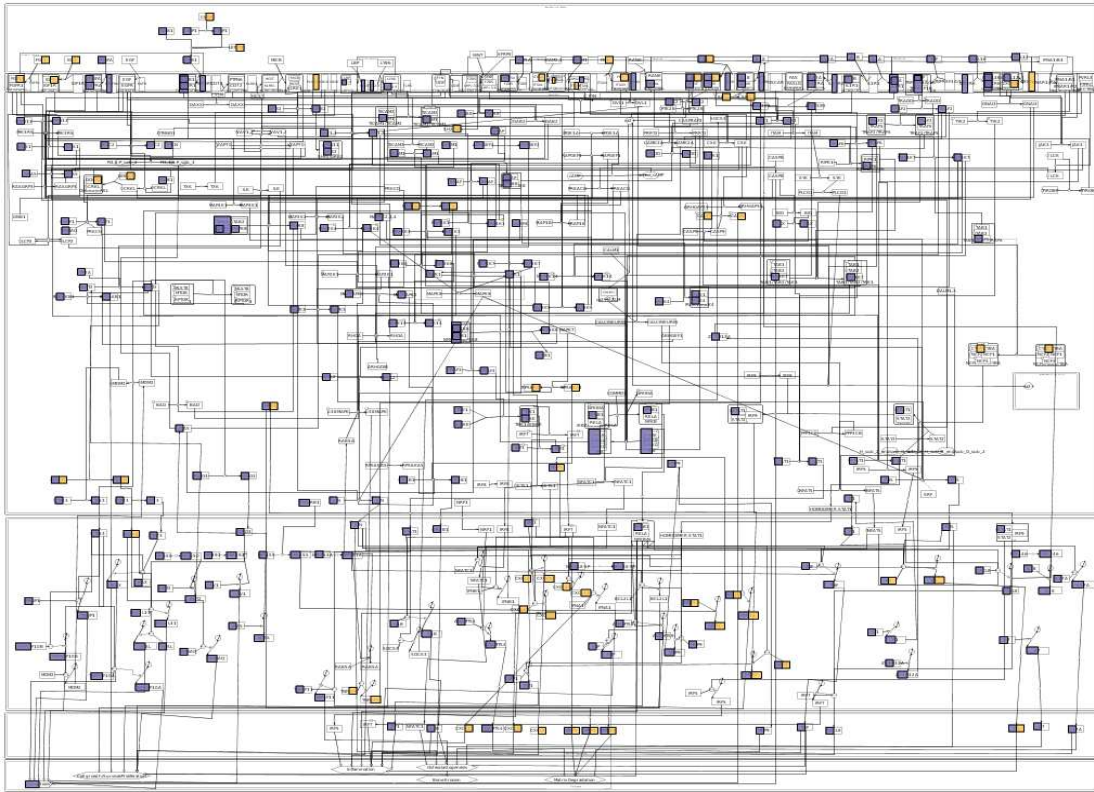


Figure 4.2: Diagram comprising the five functional outcomes for RA FLS. Two different data sets were used as overlays. The color purple depicts the synovial fibroblasts overlay from the extensive source annotations of the RA map while the yellow color represents the differentially expressed genes from the scRNA seq data of RA FLS (0.05 threshold p value)

Following our framework of map to model translation using the MINERVA platform, the upstream plugin (**Fig 4.3**) and CaSQ (as described in the Methods section) we obtained five module- models, each comprising signals that lead to a unique phenotype such as osteoclastogenesis and bone erosion, matrix degradation (cartilage destruction), inflammation, proliferation and apoptosis and one merged model combining all five phenotypes.

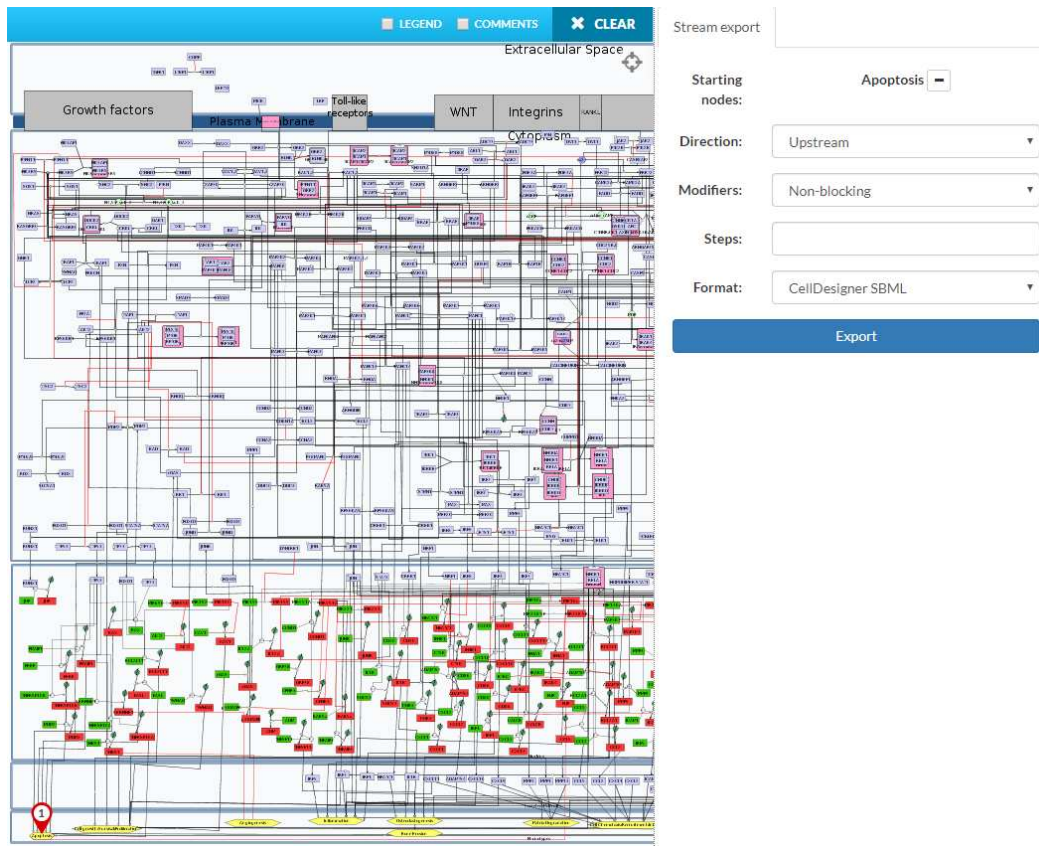


Figure 4.3: The stream plugin in MINERVA platform allows for the selection of upstream (downstream) regulations starting from a given node. In our case, either one or multiple phenotypes were selected and the upstream direction to create the individual modules and the merged network.

In **Table 4.1** we have listed the number of nodes and edges for every executable module and for the whole model.

Table 4.1: Size in terms of number of nodes and number of edges for the five modules and the whole model.

Model/ Module	Number of nodes	Number of edges
Inflammation	224	359
Osteoclastogenesis and bone erosion	211	332

Cell proliferation/ cell growth/survival	216	335
Matrix degradation	189	271
Apoptosis	184	256
Model with five phenotypes	287	465

One straightforward observation that comes from the table is that the merged model as whole is not the sum of the sub parts, meaning that the size of the merged model is not the sum of the size of all modules. Indeed, as seen in the Venn diagram of the **Fig 4.4**, there is a core of 153 nodes shared by all modules, some nodes that are shared between some of the modules and at the end a few characteristic nodes for every module.

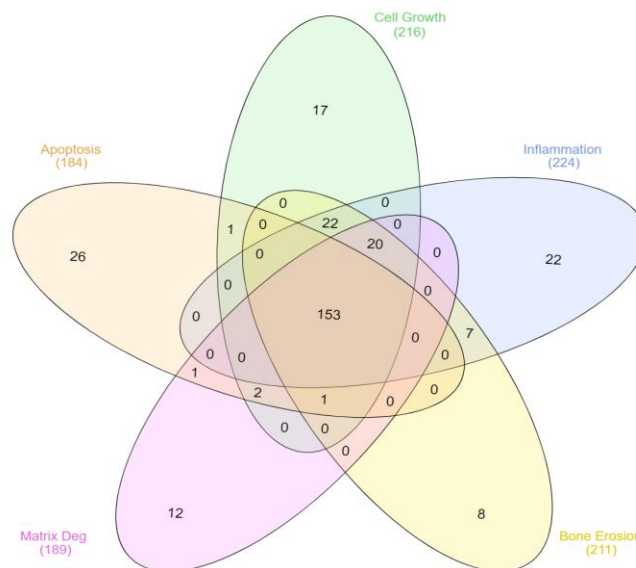


Figure 4.4: Venn Diagram of the five modules. A core of 153 nodes is shared among all five modules and only a small number of nodes is characteristic of the corresponding module. This distribution explains also why the size of the merged model is not the sum of the sizes of the individual modules.

4.3.1 Real time *in silico* simulation using the platform Cell Collective

The first part of our analysis consisted of a number of real time simulations that were used to evaluate the models' performance in regards to established biological scenarios. We formulated prior knowledge coming from small scale experiments as observations that we then tested using the Cell Collective platform. The simulations were performed using the asynchronous updating scheme. All executable files are available at ANNEX B.

4.3.1.1 *In silico* simulations for the Inflammation module and the five phenotype model

Joint inflammation is the primary characteristic of RA. During joint inflammation in RA, FLS proliferate to form the pannus, which invades and destroys the cartilage. Major pathways involved in the perpetuation of inflammation are TNF, IL-6, IL-1 and IL-17 (E. H. Choy & Panayi, 2001; Noack & Miossec, 2017). TNF and IL-17 pathway activates inflammation via NFkB pathway. As seen in **Table 4.2**, simulations of the module and the whole model showed that TNF was able to activate inflammation with the additional activation of IKBA/NFkB/RELA complex, which is required for the activation of the NFkB pathway (Matsuno et al., 2002). IL-6 activation is a sufficient signal for activating inflammation as no additional inputs are needed for the propagation of the signal until the phenotype activation. Regarding IL-17, similarly to TNF, the signal is not sufficient to activate inflammation. When NFkB pathway is activated through the additional activation of IKBA/NFkB/RELA, we observe a fluctuation activity between 40% to 80 %, both for the module and for the whole model.

Table 4.2: *In silico* simulation of the Inflammation module and the merged model (Simulation figures are available at ANNEX figure A1).

No	Initial conditions of inputs	Biological behaviour	Module behaviour	Model behaviour	References	Agreement
1	No condition		Inflammation OFF	Inflammation OFF		
2	TNF OFF	Anti-TNF- α monoclonal	Inflammation OFF	Inflammation OFF	PMID: 11934972	Yes

		antibody results in decrease in synovial inflammation				
3	TNF ON	Synovial inflammation was significantly exacerbated by TNF- α administration	Inflammation OFF	Inflammation OFF	PMID: 11934972	Activation of NF κ B pathway is also needed to activate Inflammation. However, TNF alone cannot activate NF κ B, as the latter needs the activation of the IKBA/NF κ B/RELA complex.
4	TNF ON, IKBA/NF κ B/RELA complex ON	Synovial inflammation was significantly exacerbated by TNF- α administration.	Inflammation ON	Inflammation ON	PMID: 11934972	Yes
4	IL6 OFF	Interleukin 6 knock-out mice are resistant to antigen-induced experimental arthritis	Inflammation OFF	Inflammation OFF	PMID: 10623431	Yes
5	IL6 ON	Interleukin 6 knock-out mice are resistant to antigen-induced	Inflammation ON	Inflammation ON	PMID: 10623431	Yes

		experimental arthritis				
6	IL17A OFF	Anti IL17A antibody reduces joint inflammation in collagen induced arthritis (CIA)	Inflammation OFF (fig 7e)	Inflammation OFF	PMID: 14872510 PMID: 14634133	Yes
7	IL17A ON	Anti IL17A antibody reduces joint inflammation in collagen induced arthritis (CIA)	Inflammation OFF (fig 7e)	Inflammation OFF	PMID: 14872510 PMID: 14634133	No IKK complex that activates NFkB requires more upstream regulators to be active.
8	IL17A ON, IKBA/NF KB/RELA complex ON	Anti IL17A antibody reduces joint inflammation in collagen induced arthritis (CIA)	Activity fluctuates between 40% to 80 % (fig 7f)	Activity fluctuates between 40% to 80%	PMID: 14872510 PMID: 14634133	Yes

4.3.1.2 *In silico* simulations for Bone erosion module and the five phenotype model.

Bone erosion is among the main characteristics of RA associated with disease severity. Synovitis along with the production of proinflammatory mediators like WNT (wingless-related MMTV integration site) and receptor activator of nuclear factor κ B ligand (RANKL) result in the differentiation of bone-resorbing osteoclasts, thereby stimulating local bone resorption (Cici et al., 2019). SFRP5 is the main regulator of the WNT pathway and when it is active, it inactivates bone erosion via the JNK pathway. *In silico* simulations using both the module and

the merged model as shown in **Table 4.3** showed that we could reproduce WNT and RANKL established biological behaviour (Kwon et al., 2014; MacLauchlan et al., 2017; Pettit et al., 2001). In our modelling system, SFRP5, the regulator of the WNT pathway seems to play a decisive role in the signal propagation.

Table 4.3: *In silico* simulations of the Bone erosion module and the merged model (Simulation figures are available at ANNEX figure A2).

No	Initial conditions of inputs	Biological behaviour	Module behaviour	Model behaviour	References	Agreement
1	No condition		Bone erosion ON	Bone erosion ON		Inactivation of External component SFRP5, which is the regulator of WNT pathway leads to Bone erosion activation via NFATC1
2	SFRP5 ON	Secreted frizzled-related protein 5 suppresses inflammatory response in rheumatoid arthritis fibroblast-like synoviocytes through down-regulation of c-Jun N-terminal kinase	Bone erosion OFF	Bone erosion OFF	PMID: 24764263	Yes
3	RANKL ON	RANKL expressed on synovial	Bone erosion ON	Bone erosion ON	PMID: 10693864	Yes

		fibroblasts is involved in rheumatoid bone destruction by inducing osteoclastogenesis				
4	RANKL OFF	RANKL knockout mice are protected from bone erosion in a serum transfer model of arthritis	Bone erosion ON	Bone erosion ON	PMID: 11696430	No -In the absence of SFRP5, WNT pathway gets activated and leads to bone erosion activation
5	RANKL OFF, SFRP5 ON	RANKL knockout mice are protected from bone erosion in a serum transfer model of arthritis	Bone erosion OFF	Bone erosion OFF	PMID: 11696430	Yes
7	WNT5A OFF	Wnt5a cKO mice were resistant to arthritis development	Bone erosion ON	Bone erosion ON	PMID: 28724439	No -In the absence of SFRP5, WNT pathway gets activated and leads to bone erosion activation. SFRP5 inhibition is added in the map only to canonical WNT and not to WNT5a

4.3.1.3 *In silico* simulations for the Cell growth/survival/proliferation module and the five phenotype model.

Aberrant proliferation of resident synovial fibroblasts in RA contributes to pannus formation and joint destruction. Synovial fibroblasts are the key cell types in the growth of pathological synovial tissue in RA and inhibition of their proliferation could be possible antirheumatic therapies (Sandler et al., 2006). Various growth factors like PDGFA, TGFB1 and cytokines like TNF regulate the proliferation of synovial fibroblasts. Platelet-derived growth factor (PDGF) is an important mitogen for fibroblasts including synovial fibroblasts. In both the module and the whole model, it is able to activate cell growth when kept active (ON). However, when inactive, cell growth still remains ON due to activation by other pathways. Synergetic effects of PDGFA and TGFB1 for the activation of MMP3, IL6 and TNF have also been reported (Rosengren et al., 2010) but our module and model were not able to reproduce them hinting at missing interactions in the model. The AKT role is also highlighted by various studies and the results of the simulations show a partial coherence (especially for the AKT KO). In **Table 4.4** we present a summary of the simulation results and the corresponding bibliographic references along with comments regarding the behaviour of the module and the whole model.

Table 4.4: *In silico* simulation of Cell growth/survival/proliferation module and merged model (Simulation figures are available at ANNEX figure A3).

No	Initial conditions of inputs	Biological behaviour	Module behaviour	Model behaviour	References	Agreement
1	No condition		Cell growth/survival/proliferation ON	Cell growth/survival/proliferation ON		
2	PDGFA ON	PDGF stimulate proliferation of synovial fibroblasts	Cell growth/survival/proliferation ON (Fig 4a)	Cell growth/survival/proliferation ON	PMID: 16806061 PMID: 16634808	Yes
3	PDGFA OFF	PDGF stimulates proliferation of synovial fibroblasts.	AKT2 ON and Cell growth/survival/	Cell growth/survival/proliferation ON	PMID: 16806061 PMID: 16634808	No

			proliferation ON (Fig 4b)			
4	PDGFA ON, AKT2 KO	PDGF stimulates the growth of synovial fibroblast via interference with the Akt signaling pathway	Cell growth/ survival/ proliferation ON (Fig 4c)	Cell growth/ survival/ proliferation ON	PMID: 19568828	Yes
5	PDGFA OFF, AKT2 OFF, FOXO1 OFF	JNK- dependent downregulation of FoxO1 is required to promote the survival of fibroblast- like synoviocytes in rheumatoid arthritis	Cell growth/ survival/ proliferation ON (Fig 4d)	Cell growth/ survival/ proliferation ON	PMID: 24812285	No - Inactivation of External component SFRP5, which is actually the regulator of WNT pathway leads to the cell proliferation activation via CREB1
6	TGFB1 ON, PDGF1 ON, TNF α ON	TGF- β and PDGF synergistically augmented TNF α - or IL1 β - induced matrix metalloproteinase 3 (MMP3), IL6 secretion by FLS.	MMP3, IL6, not present in the module	MMP3, IL6 OFF	PMID: 20380722	No

4.3.1.4 *In silico* simulation of the Apoptosis module and merged model.

The impaired apoptosis process of synovial fibroblasts is responsible for synovial hyperplasia and joint destruction. Major pathways regulating apoptosis include extrinsic FASL and TNF and intrinsic BCL2 family. When kept ON, both FASL and TNF can activate apoptosis, a result

that was also reproduced by both the apoptosis module and the merged five phenotype model (**Table 4.5**). AKT is another regulator of apoptosis (anti-apoptotic agent) which, when kept ON, protects RA FLS against the apoptosis induced by Fas through inhibition of Bid cleavage. However, the apoptosis module and the merged model failed to reproduce this biological scenario as apoptosis is probably activated by other pathways that remain active.

Table 4.5: *In silico* simulation of the Apoptosis module and the merged model (Simulation figures are available at ANNEX figure A4).

No	Initial conditions of inputs	Biological behaviour	Module behaviour	Model behaviour	Reference	Agreement
1	No condition		Apoptosis ON	Apoptosis ON		Inactivation of main upstream regulators such as AKT2 and microRNAs activates downstream elements TP53, FOXO1, FOXO3, MDM2 and CAV1 respectively activates Apoptosis
2	FAS/FAS L ON	Fas receptor induces apoptosis of synovial bone and cartilage progenitor populations and promotes bone loss in antigen-induced arthritis	Apoptosis ON	Apoptosis ON	PMID: 30383451	Yes
3	TNF ON, NFKB OFF	TNF α can induce apoptosis in RA-FLS, when NF- κ B was inhibited	Apoptosis ON	Apoptosis ON	PMID: 10817564	Yes
4	AKT ON	In RA FLS, phosphorylation of AKT	Apoptosis ON	Apoptosis ON	PMID: 20187936	No

		protects against Fas-induced apoptosis through inhibition of Bid cleavage.				
--	--	--	--	--	--	--

4.3.1.5 *In silico* simulation of the Matrix degradation module and the merged model.

Cartilage destruction is one of the major characteristics of RA where activated synovial fibroblasts (SF) have been shown to attach to the cartilage surface and result in the release of matrix-degrading enzymes. These proteolytic enzymes secreted locally contribute to the degradation of the articular cartilage at the pannus–cartilage interface and mediate the invasion of synovial cells into the cartilage. Out of many classes of proteolytic enzymes involved, matrix metalloproteinases (MMPs) are generally believed to be pivotal in cartilage destruction (van der Laan et al., 2003). We tested two MMPs namely MMP1 and MMP9, both of which were able to reproduce their literature established biological behaviour in RA synovial fibroblasts in both matrix degradation module and merged model (**Table 4.6**). The role of CCL5 in cartilage destruction in RA synovial fibroblasts has been described; however, CCL5 was not found to be present in both the matrix degradation module and merged model. CCL5 is indeed present in the RA map linked to Cell Chemotaxis phenotype suggesting a possible missing interaction with matrix degradation phenotype.

Table 4.6: *In silico* simulation of the Matrix degradation module and the merged model (Simulation figures are available at ANNEX figure A5).

No	Initial conditions of inputs	Biological behaviour	Module behaviour	Model behaviour	References	Comments
1	No condition		Matrix degradation OFF	Matrix degradation OFF		
2	MMP1 OFF	Inhibition of MMP1 alone results in a significant reduction	Matrix degradation OFF	Matrix degradation OFF	PMID: 15146414	Yes

		of cartilage invasion by RASFs.				
3	MMP9 ON	MMP-9 stimulates RA synovial fibroblast-mediated inflammation and degradation of cartilage	Matrix degradation ON	Matrix degradation ON	PMID: 24982240	Yes
4		CCL5 Induces Collagen Degradation by Activating MMP-1 and MMP-13 Expression in Human Rheumatoid Arthritis Synovial Fibroblasts			PMID: 29093715	CCL5 is present on the RA map but not in this module, suggesting missing interactions

4.3.1.6 *In silico* simulation of the merged model

We also performed additional simulations in CellCollective with the merged model and here we present one example. With initial condition set to all inactive and asynchronous updating, we observe an oscillatory behaviour for MDM2 and TP53 components as seen in **Fig 4.5A**. The fluctuations do not change for over 400 time steps, as shown in **Fig 4.5B**.

A



B



Figure 4.5: Simulation with CellCollective. A. Oscillatory behaviour for MDM2 and TP53 entities with initial condition set to all inactive and asynchronous updating **B.** Same oscillatory behaviour for MDM2 and TP53 with a span of over 400 time steps

Under the same conditions, we also studied the behaviour of the model regarding the five phenotypes. As seen from **Fig 4.6A**, all phenotypes except Matrix Degradation will get activated and stay in this state for more than 400 time steps. This result suggests missing key interactions that control the regulation of apoptosis, as RA FLS are supposed to be apoptosis resistant. At the same time, apoptosis phenotype and cell survival phenotypes should exhibit a switch-like behaviour as from a biology point of view they cannot happen simultaneously.

A



B



Figure 4.6: Simulation with CellCollective. A. All phenotypes except Matrix Degradation get activated and **B.** stay in this state for more than 400 time steps.

4.3.2 Calculating Attractors using BoolNet

The size and complexity of the module –models make an exhaustive search in asynchronous mode prohibited. However, by using heuristics and starting from predefined initial conditions, attractors can be identified. For all five modules, we searched for attractors using the same two initial conditions: when all nodes are inactive (equivalent to Boolean zero) and when all nodes are active (equivalent to Boolean one). In **Table 4.7** we summarize the findings for the five modules and the merged model with five phenotypes. As seen from **Table 4.7**, all modules

besides apoptosis reach one stable state when initial conditions are set to zero or one, while apoptosis and the whole model reach complex/loose attractors with numerous states, for both initial conditions.

Table 4.7: Attractors calculation for predefined initial conditions (all set to 0 or 1) for the five modules and the merged model (files available at ANNEX B).

Model/ Module	Attractors with asynchronous update when initial conditions = 0	Attractors with asynchronous update when initial conditions = 1
Apoptosis	Attractor 1 is a complex/loose attractor consisting of 248 state(s) and 966 transition(s)	Attractor 1 is a complex/loose attractor consisting of 248 state(s) and 966 transition(s)
Matrix Degradation	Attractor 1 is a simple attractor consisting of 1 state(s)	Attractor 1 is a simple attractor consisting of 1 state(s)
Bone erosion and cartilage destruction	Attractor 1 is a simple attractor consisting of 1 state(s)	Attractor 1 is a simple attractor consisting of 1 state(s)
Cell proliferation/ cell growth/survival	Attractor 1 is a simple attractor consisting of 1 state(s)	Attractor 1 is a simple attractor consisting of 1 state(s)
Inflammation	Attractor 1 is a simple attractor consisting of 1 state(s)	Attractor 1 is a simple attractor consisting of 1 state(s)
Model with five phenotypes	Attractor 1 is a complex/loose attractor consisting of 248 state(s) and 962 transition(s)	Attractor 1 is a complex/loose attractor consisting of 496 state(s) and 2163 transition(s)

Analysis of the attractors helped identify coherent behaviour of the model but also to spot inconsistencies and verify some of the scenarios tested with the real time simulations. Some examples are discussed in this paragraph. For the module of Inflammation, when all interleukins and TNF are set to zero, in the stable state reached by the system, the inflammation phenotype will also be equal to zero. Respectively, for the same module, with initial conditions set to 1, in the stable state reached IL6, TNF, IL1, IL17, IL18, IL1B and IL1A will be active and so will the inflammation phenotype. However, AKT2 is zero when inflammation is ON and ON when inflammation is OFF. This finding is not in accordance with bibliographic references where AKT2 suppression has an anti-inflammatory impact (Du et al., 2019; Malemud, 2013). Regarding Bone erosion and Osteoclastogenesis module, we observe that in the two attractors reached for both initial conditions, the phenotype of Bone erosion and Osteoclastogenesis will stay ON, independently of the state of RANK, confirming the results of simulations with Cell Collective. Indeed, in the attractor reached when the initial conditions

are set to zero, SFRP5 is zero, WNT is ON (in the absence of the inhibitor and due to the rule that describes its regulation) and NFATC1 will also be equal to 1 (ON), activating the Bone erosion and Osteoclastogenesis phenotype. Regarding Matrix degradation, in the attractor reached when all is set to zero, the phenotype is OFF, confirming previous observations, while when all is set to 1, the phenotype is switched to ON, but MMP1 and MMP9 are eventually OFF. The attractor reached when using the module of Cell growth and survival, with initial conditions set to zero, showed that the phenotype of Cell proliferation would be eventually turned on, as also shown in the real time simulations with Cell Collective. We could also verify that when the attractor AKT2 was ON, as well as WNT (in the absence of the regulator SFRP5) and led to the activation of the phenotype via CREB1 - also ON in the attractor configuration. As far as apoptosis module is concerned, for both initial conditions the module would reach a complex/ loose attractor. For initial conditions set to zero, Apoptosis phenotype was always ON, despite FAS_FASL complex being OFF. For initial conditions set to one, Apoptosis will also be always ON. As commented before, the results point to missing interactions in the apoptosis module and / or the possible need to review the inferred rules of regulation, in order to reproduce a behaviour compliant to the biological knowledge (RA FLS apoptosis-resistant, not coherent to have the phenotype always ON). In both identified complex attractors, the activity of MDM2 entities (MDM2, MDM2_rna and MDM2_phosphorylated,) and TP53 entities (TP53_phosphorylated in nucleus and cytoplasm) is fluctuating between 1 and 0, contributing to the oscillatory behaviour.

Lastly, the attractors identified for the merged model, for both initial conditions, are two complex/ loose attractors. Similarly to the apoptosis module, MDM2 entities and TP53 entities exhibit oscillatory behaviour confirming the behaviour we had observed with Cell Collective previously.

4.3.3 Creating a more compact version of the RA FLS model with five phenotypes, and using MaBoSS to cross check observations from real time simulations

In order to reduce complexity, we created a more compact version of the whole model. First, we grouped the model's inputs into the following categories: Chemokines, Cytokines, Interferon, Growth factors, Toll like receptors, FASL, OSCAR, RANKL, WNT, ICAM/integrins/ fibronectin, Plexin. Second, we imported the model in GINsim/ bioLQM (see

Chapter 3) and created a reduction by removing intermediate nodes that do not participate in feedback loops or functional circuits. The result was a smaller version of 203 nodes and 418 edges. This version was used to calculate phenotypic probabilities using the software MaBoSS.

When performing real time simulations with Cell Collective, we observed that Apoptosis phenotype could be activated independently of FAS/FASL activation. Using MaBoSS we performed an analysis with initial conditions for the FASL complex: 80% of chances to be activated. As seen in **Fig 4.7**, there are 80% of chances to have both FASL and the apoptosis phenotype activated at the same time. However, as seen from the pie chart, apoptosis has also 20% chances to be activated independently of FASL, a result that is in accordance with our previous observations.

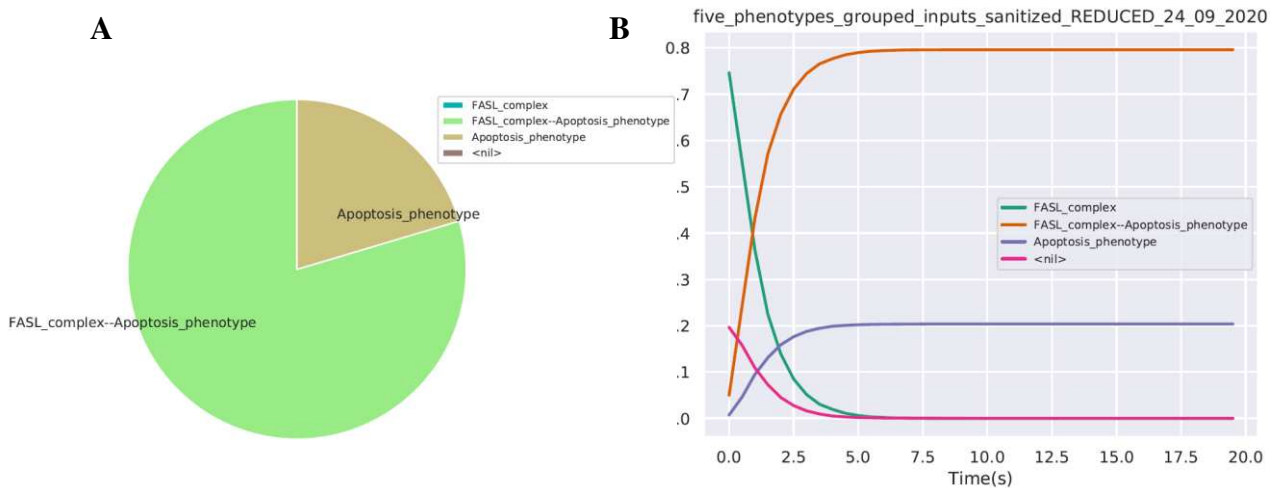


Figure 4.7: **A.** Pie chart showing the probability of 80% chances of apoptosis activation via FASL **B.** Diagram showing the trajectories for FASL and Apoptosis when FASL is kept at 80% ON.

Another scenario we wanted to check with MaBoSS, was the RANK, WNT and Bone Erosion simulations, that showed a dependency of the phenotype to the WNT pathway, regardless of the RANK state. In the simulation performed with CellCollective (Table 3) activated RANK was able to activate Bone erosion. For the MaBoSS simulations we used a grouped input version of the model that failed to reproduce this result (**Fig 4.8**). The discrepancy is due to the fact that in the ungrouped input model, WNT pathway gets activated in the absence of the external regulator SFRP5, and eventually activates Bone erosion along with RANK. However, in the grouped inputs version, we merged WNT and WNT5a into a single group, and the

regulator was removed due to its specificity for canonical WNT pathway. With the grouped inputs version we observe that when WNT is 80% ON, then Bone erosion will also get activated (80%) (**Fig 4.9**). When WNT and RANKL complex are 65% ON, there are 65% chances of activation of Bone erosion via RANKL and WNT complex together, 15% chances of Bone erosion activation via WNT alone. There exist no trajectory for Bone erosion activation via RANKL complex alone (**Fig 4.10**).

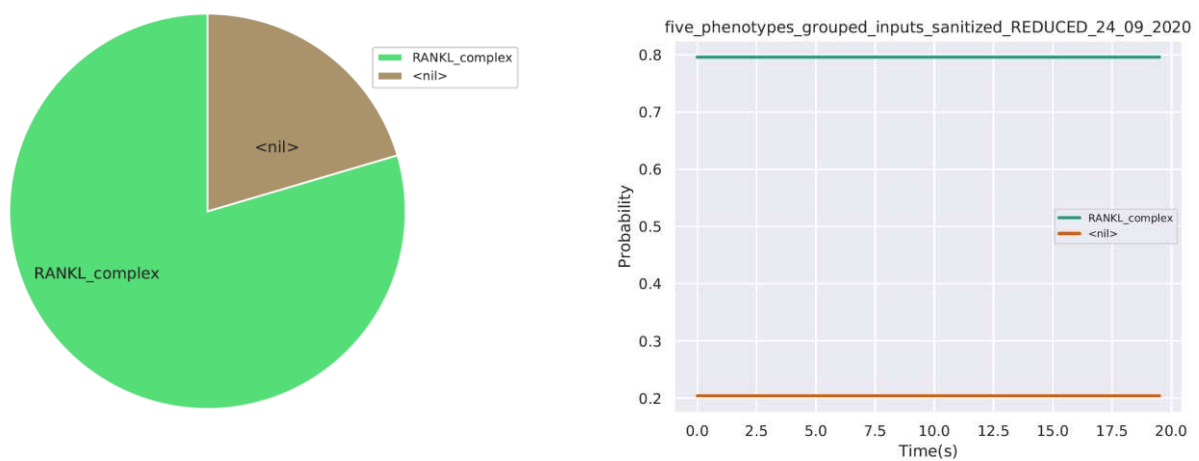


Figure 4.8: MaBoss simulation showing the absence of Bone erosion when RANKL complex kept 80% ON.

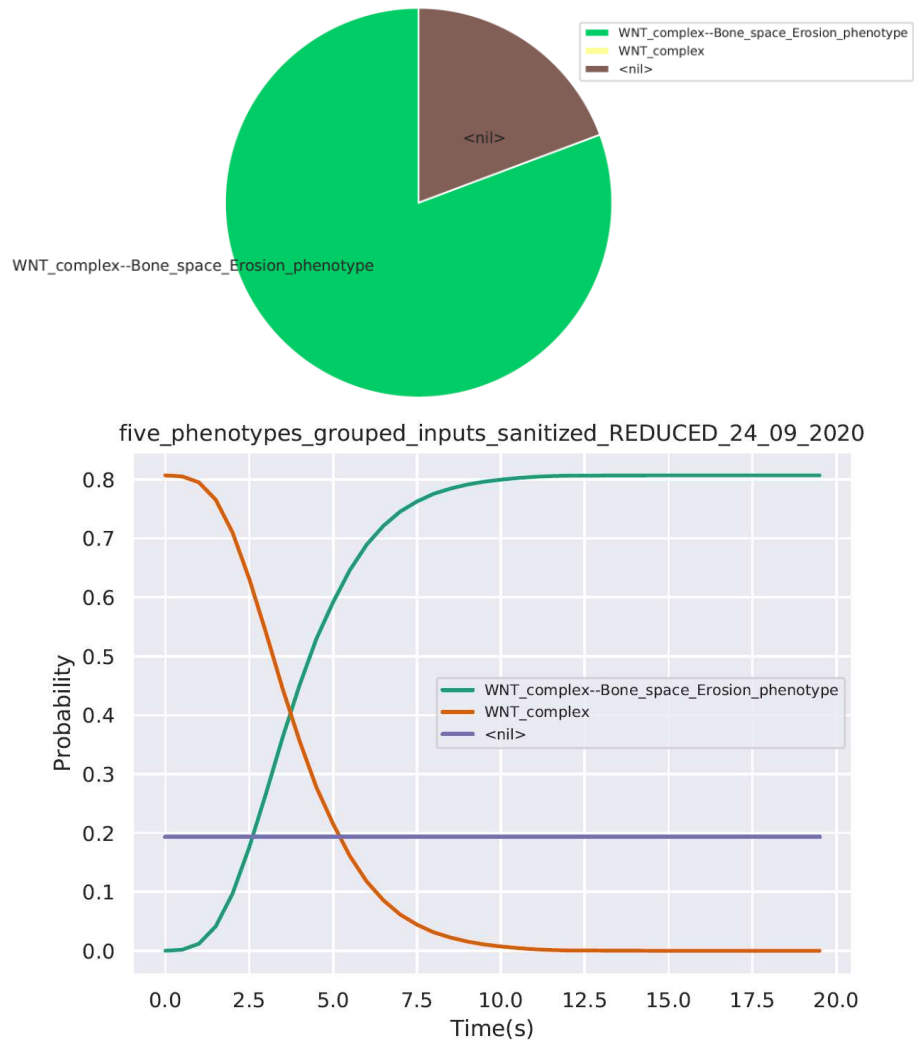
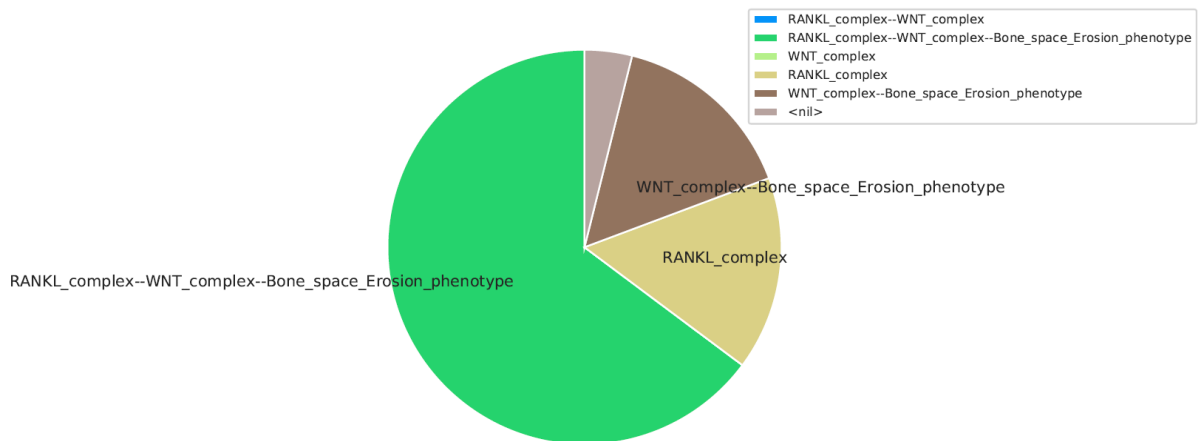


Figure 4.9: MaBoss simulation when WNT complex was set as 80% ON.



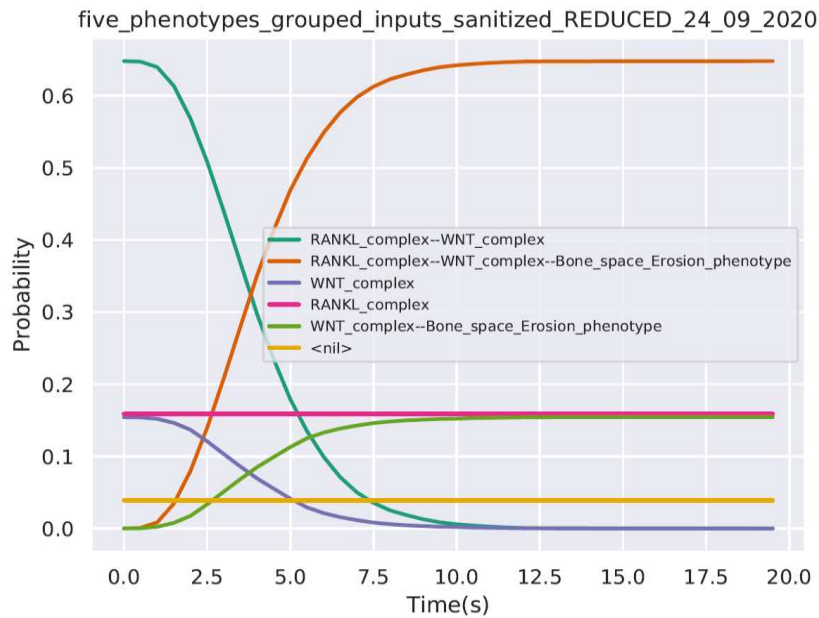


Figure 4.10: **A.** Pie chart representing a MaBoSS simulation, showing 65% of bone erosion activation via both RANKL complex and WNT and 15% bone erosion activation via WNT alone **B.** Diagram showing the trajectories of RANKL, WNT and Bone erosion phenotype in MaBoSS simulation.

4.4 Discussion

In this chapter we present our efforts to create a large-scale Boolean model for RA FLS, the first executable model for this system, to our knowledge. We make use of the RA map, a high quality source of knowledge based on human curation, and technical achievements described in previous chapters to construct a model focusing on specific cellular outcomes, namely apoptosis, cell proliferation, matrix degradation, bone erosion and inflammation. We had to make a compromise when choosing the phenotypes, as the size of the produced model was already large. We made the choice based on our main biological questions, and specifically our interest in studying apoptosis resistance, contribution to inflammation and outcomes related to the structural damage. However, the phenotype that we would like to include in the future is the chemotaxis and cell recruitment, as RA FLS are known to secrete a number of cytokines and chemokines that work as signalling calls to other cells that infiltrate the joints.

To cope with complexity and the challenges of large scale simulations, we use a modularized approach to construct our model and the sub-modules that create it. This approach allowed us to study smaller parts of the systems, and at the same time assess the behaviour on module and whole model level.

Real time simulations of the modules and the whole model revealed inconsistencies and possible missing interactions, especially for the apoptosis and the matrix degradation modules. One observation from the real time simulations is that the behaviours of the individual modules and the whole model for the scenarios tested were coherent. This means that the behaviour observed in the module was kept intact in the whole model. This might be due to the fact that the majority of the simulations involved the activation or inactivation of one or only a few inputs, however it was not a *de facto* result.

We were able to calculate attractors for our system using the R software BoolNet and heuristics that helped limit the search of the state space. We chose as initial conditions either all nodes set to zero or all set to one. We calculated the attractors for both conditions, for every individual module and the whole model. Their analysis, preliminary at the present time as they are of considerable size, confirmed behaviours observed during the real time simulations, and also revealed some inconsistencies regarding the behaviour of some nodes (ie AKT2 in inflammation). One interesting observation concerned the oscillations between TP53 and MDM2 entities in our model. The abundant expression of MDM2 in RA FLS has been demonstrated (Taranto et al., 2005). MDM2 is the major negative regulator of p53, and in tumors contributes to increased cell proliferation. In RA FLS it could be a contributing factor to the hypoapoptotic phenotype of lining tissue through its capacity to downregulate p53 levels and effects. p53 abnormalities in RA could sustain and accelerate synovial inflammation mainly through IL-6, as shown in studies using Lewis rats with adjuvant-induced arthritis (AIA) (T. Zhang et al., 2016).

A thorough, in-depth analysis of the attractors is needed and is currently ongoing, to provide a more detailed view of the system's behaviour. The next step would be to carefully examine the state of all components under given conditions to assess to which extent the model's behaviour is coherent with the biology of the system. Different initial conditions could also be used to test specific hypotheses.

Phenotypic probabilities using MaBoSS provided a third layer of analysis to observed behaviours, especially to the apoptosis -FASL dependency and the role of WNT pathway in Bone erosion. More simulation scenarios should be tested to gain insights on specific mechanisms.

I would like to mention that during my thesis the size of the obtained models, even for the modules, challenged the analysis and simulation potential of most software used for logical

models. One of the issues we came across at first was the fact that most software would require the use of the same software for model construction and analysis. This was also the case for Cell Collective. Our collaboration with Dr. Tomas Helikar, led to the implementation of the SBML qual import for Cell Collective that greatly facilitated the analysis. Dr. Helikar provided also support in retrieving the references of our models encoded in MIRIAM identifiers, in the Cell Collective platform, that allowed us to create instantly annotated models that can be published without having the burden of redoing the curation work. Throughout this thesis, Cell Collective developers adjusted the platform to satisfy our ever growing demands and supported wholeheartedly our quest for large-scale modelling.

We also worked closely with Professor Denis Thieffry and Dr. Aurelien Naldi, to overcome interoperability issues regarding the import of our SBML qual files to GINsim. When the size and the complexity of the models challenged the capacity of GINsim, bioLQM was used to perform reductions.

Dr. Laurence Calzone also provided helpful advice for the use of MaBoSS and the settings required to make the software work.

Lastly, I would like to mention that we have the support of Dr. Sylvain Soliman, who among many other things, helped with the namings of the different model components easing the task of identifying genes, RNAs, and proteins sharing common names.

Without collective efforts and the support of the community most of the work presented in this Chapter would not have been made possible.

Now that interoperability is achieved and the map-to-model framework is established, downstream analysis of large-scale biological Boolean Networks will hopefully be easier and less time consuming from a technical aspect, leaving more time to focus on the biological questions at stake.

Chapter 5. General discussion and future perspectives

Living systems cannot be understood by studying only their individual parts. With an ever-growing production of biological datasets, accelerated by genome sequencing and high-throughput omics techniques the general focus of biomedical research on complex diseases should be gradually shifted from a primarily steady state analysis at the molecular level to a systems biology level capturing the characteristic dynamic behavior of the system.

Systems biology is a holistic approach to study the big picture of biological systems and their organization. An attempt to define qualities of systems biology has been proposed (Breitling, 2010). Three main characteristics have been stated. First, diversity which refers to the biological understanding that each interaction of a component brings to the system. Second is simplicity which refers to the reductionist approach by breaking down the system into simple descriptions. Third is the complexity which refers to the understanding of complex interplay of a molecular network. These three qualities make systems biology an interdisciplinary field by nature, which combines computer science, informatics and mathematics with biology. Among other applications, it has proven to be a powerful analytical tool to understand the dynamic interactions which is the core of complex diseases.

Summarizing, in my thesis:

- a) I used prior knowledge to build an RA FLS specific network. To recapitulate all that is known and published (in various sources) about the disease, a global, fully comprehensive and annotated RA-specific map has been built based on exhaustive literature mining, human curation and validation from domain experts (Singh et al., 2018, 2020). This map features interactions implicated in RA coming from various cell types. Thanks to the extensive annotations of each entity and each reaction included in the map, and the advanced functionalities offered by the MINERVA platform (Gawron et al., 2016; Ostaszewski et al., 2019), the user can opt for cell-specific interactions and extract the corresponding cell-specific networks.. If such maps of knowledge are very informative (Kuperstein et al., 2015), they could be fully exploited by translating them into dynamical objects and mathematical predictive models.
- b) I worked to add a dynamic layer to the network in an automated fashion. Up to now the building of dynamical models were manual except for a few. During my doctoral studies an effort to add a dynamic layer on this descriptive map has been explored, and

the approach chosen to study the RA map dynamically is the logical formalism (Boolean), for its simplicity and lack of kinetic parameters. Together with Dr. Soliman, (Lifeware INRIA, Saclay) we have developed CaSQ (CellDesigner as SBML-Qual), <https://lifeware.inria.fr/~soliman/post/casq/>), a tool for automated inference of large-scale, parameter-free Boolean models, from molecular interaction maps based on network topology and semantics (Aghamiri et al., 2020). This is, to the best of our knowledge, the first tool that produces executable logical models of hundreds of nodes (up to several hundreds), in SBML-qual standard format for model description, which can be further simulated and analyzed using popular modelling tools (Chaouiya et al., 2012; Tomáš Helikar et al., 2012). In this framework, preliminary logical formulae for the model are inferred automatically according to predefined rules and constraints. These rules have proven to be general enough to cover various scenarios of biological knowledge representation but have not been tested systematically to assess the model's robustness.

- c) The problem of analyzing large-scale Boolean models constitutes a challenge in the field. The increased model size in terms of number of nodes and the complexity due to the signalling crosstalks and the presence of numerous feedback loops, do not allow for a straightforward dynamic analysis. To cope with complexity I used a modularized approach that helped me study each functional phenotype of RA FLS separately and compare it with the merged system. To do so, I used the global RA map and the stream plugin of the MINERVA platform to focus on relevant subnetworks. The idea of modules that start from one phenotype and include all possible upstream regulators simplified the simulation burden of input output relationships and helped better understand the system of RA FLS.
- d) Besides modularity, that helped me cope with the complexity of the system using the divide and conquer strategy, I also tried to group the inputs of the model that were numerous and contributed to computational explosion when trying to perform dynamical analysis. While grouping the inputs lead to a certain loss of information (ie could not perform *in silico* simulations for TNF and IL6 separately as they were grouped as cytokines), it made the computation of phenotypic probabilities accessible and helped towards a deeper understanding of the system.
- e) I used systems biology standards for curation, graphical notation and modelling in an effort to promote the use of community standards, the transparency in scientific results,

and the interoperability and reusability of the files and the scientific context, respectively. I used SBGN for the RA map construction, MIRIAM identifiers for curation, unique identifiers for genes and proteins, and SBML qual files for dynamical modelling. All my work is or will be accessible in an open access way, promoting open science.

- f) I also had the chance to do community work and learn from this experience. I am a member of the Disease Map consortium and I contributed actively to the COVID-19 Disease Map initiative (see Chapter 6 - contributions to the community).

5.1 Future directions

- a) It is difficult to estimate how robust a model is when the assumptions for the inference of logical rules are changed (for example use AND gates instead of OR for the logical formulas). The model's robustness could be estimated as the ability of the model to reproduce a well-established (experimentally validated) biological scenario, based only on the inferred rules (without additional manual tuning). An extensive study of the impact of changing the translation rules of the model will allow for a better understanding of the rules and the topology effect on the model's dynamics. This impact could be summarized in phenotypic probabilities of functional outcomes under different scenarios and rulesets. The results could then be systematically confronted with small scale experimental data and experts' knowledge encoded in the RA map in order to identify the most robust/reliable set of translation rules.
- b) The presence of numerous inputs in the inferred models is a combined result of the map to model translation rules and the map structure that leads to the creation of pseudo - inputs. These pseudo-inputs increase the computation cost of the analysis, and a strategy that could help limit their presence / impact (ie: by fixing their values) in the obtained models would facilitate the downstream analysis.
- c) The ambition of this work is to set a framework that can facilitate the identification of novel therapeutic targets for RA. The aim is to predict the optimal conditions that would favor apoptosis and minimize the bone erosion, cartilage destruction and inflammation outcomes. Systematic testing of different initial conditions could further lead to predictions regarding the outcomes of specific perturbations, such as single or combined effects, simulated with the model by forcing or suppressing the activity of each gene/protein systematically. The ultimate goal is to gain a better understanding of

the mechanism behind cartilage and bone degradation, two major debilitating symptoms of RA, and propose a strategy that could help block or even reverse these outcomes.

- d) To obtain a refined and better trained model, data from small scale experiments and also published Omic datasets of RA fibroblasts (transcriptomic, RNAseq, RNAseq single cell) should be used to increase cell and disease specificity, as well as to define phenotypic signatures (biomarkers) that could be tested against steady states.
- e) Unit testing and value propagation approaches as the ones described in Hernandez et al, 2020 (Hernandez et al., 2020), could be also applied to verify the model's behaviour locally.
- f) Another potential application would be to identify druggable points of the RA FLS system that are targeted not only by anti-rheumatic drugs, and perform simulations with combined perturbations. This approach could help identify putative drug repurposing or combined therapeutic treatments and assess the impact on the phenotypes.
- g) Lastly, *in vitro* experiments would be an ideal complement of this work. The model's predictions could be used to further investigate the disease mechanism intracellularly, by targeting the proposed pathway/factors and measuring apoptosis and/or inflammation and at the level of cellular crosstalks, by evaluating the impact of RA FLS functional outcomes to cells of the immune system and other neighboring cells such as chondrocytes, responsible for cartilage damage and subsequent bone erosion. *In vitro* experiments could also be used to test predictions regarding combined treatments.

5.2 Economic and social impact of computational systems biology approaches in RA

- **Short term benefits:** A robust computational model can reduce the need for animal models, and help elucidate human specific pathways that are not obligatory common with their mice counterparts. Induced arthritis to mice models cannot always be considered as a representative system for studying the human response to treatment. In our approach, the focus is given on RA specific human data.
- **Medium term benefits:** *In silico* simulations and predictions can significantly reduce the time of experimental design and can also be used for simulation of complex scenarios that are not easily reproduced in the lab. Besides hypothesis testing and

mechanistic insights, the model can be used for predictions of novel therapeutic targets proposed for in vitro validation. Using computational power, hundreds of data can be analyzed and numerous scenarios can be tested accelerating impressively the time needed for the production of testable hypotheses.

- **Long term benefits:** RA is associated with a heavy societal burden that stems from patients' disability and health condition and the economic costs which come with it. RA tends to have a progressive nature, with symptoms that worsen over time. Taking into account the fact that the disease onset is around the third and fourth decade of a person's life, the disease has a long-term impact on functioning and productivity, which translates to a considerable social and economic cost. The higher the disability associated with RA, the higher the health cost is. Moreover, RA can severely restrict a person's ability to work and RA patients are often obliged to reduce work time or make changes in employment to accommodate their disability. This situation causes them to lose income over the course of a lifetime. Propose new therapeutic targets that could benefit the non-responders to classic biologic treatments, or targets that could stop debilitating symptoms such as bone erosion and cartilage destruction early on, could have a significant impact on the individual's quality of life and ability to work and be independent, but also to the society by lowering the cost of prolonged health care.

Chapter 6. Contributions to the community

6.1 The Disease Maps consortium

During my thesis I had the opportunity to participate in the Disease Maps project. My supervisor, Dr. Anna Niarakis is a co-leader of the project and responsible for the Rheumatoid Arthritis Disease map. I had the chance to know first hand the challenges and the important issues of the field, to participate in community meetings, present and discuss my work and receive valuable feedback and help when needed. The interactive RA map is hosted in the MINERVA platform and supported by ELIXIR and the Luxembourg Centre for Systems Biomedicine.

6.2 COVID-19 Disease Map, a large-scale community effort to create graphical and executable models of SARS-CoV-2 virus-host interaction mechanisms

The COVID-19 Disease Map (C19DMap) is a large-scale community effort aiming at integrating available information in the form of standardized pathway representations (**Fig 6.1**). At the same time, an ecosystem of bioinformatic approaches and pipelines is developing rapidly to ensure seamless downstream analysis, including omic data integration and computational modelling (Ostaszewski et al., 2020).

As for today, the collection of maps in the repository covers molecular processes involved in SARS-CoV-2 entry, replication, and host-pathogen interactions, as well as immune response, host cell recovery and repair mechanisms. The COVID-19 map will possess a hierarchical structure of interconnected functional modules (**Fig 6.2**).

The map combines diagrams of COVID-19 mechanisms with the corresponding executable models providing a platform for clinicians, virologists, and immunologists to collaborate with data scientists and computational biologists maximizing interdisciplinarity and complementarity of skills. Computational modelling approaches are employed as they allow

for *in silico* experimentation that could lead to the formulation of testable hypotheses and finally, to predictions concerning novel therapeutic targets and candidates of drug repurposing.

This unprecedented effort uses a variety of complementary tools and approaches to create shared content and to enable a seamless downstream analysis, focusing on interoperability, reproducibility and applicability of the methods and tools. Among others, the systemic approach developed and applied in this thesis for Rheumatoid Arthritis is also being applied to the COVID19. Working with the community we help to set the basis for a large-scale systems biology framework that could be proven useful not just in the context of the C19DMap but also in similar challenges that could benefit from such holistic approaches.

More on the C19DMap project and community can be found here:

<https://fairdomhub.org/projects/190>

<https://covid.pages.uni.lu/>

The screenshot shows the Fairdomhub interface for the COVID-19 Disease Map project. At the top is a dark navigation bar with the FAIRDOM HUB logo, a 'Browse' dropdown, a 'Help' dropdown, a search bar with the text 'Search here...', and a 'Search' button with a settings gear icon. Below this is a breadcrumb trail: 'Home / Projects Index / COVID-19 Disease Map'. The main heading is 'COVID-19 Disease Map' with a small icon of a virus. A descriptive paragraph follows: 'Here we share resources and best practices to develop a disease map for COVID-19. The project is progressing as a broad community-driven effort. We establish a knowledge repository on virus-host interaction mechanisms specific to the SARS-CoV-2. The COVID-19 Disease Map is an assembly of many interaction diagrams established based on literature evidence.' Below this, project details are listed: 'Programme: Disease Maps', 'FAIRDOM PALs: No PALs for this Project', 'SEEK ID: https://fairdomhub.org/projects/190', 'Project created: 27th Mar 2020', 'Public web page: http://doi.org/10.17881/covid19-disease-map', and 'Organisms: Severe acute respiratory syndrome coronavirus 2, Homo sapiens'. A 'Related items' section at the bottom shows a list of categories: 'People (236)', 'Institutions (123)', 'Data files (1)', 'Models (25)', 'Publications (87)', and 'Documents (7)'. The 'People (236)' category is highlighted with a blue background.

Figure 6.1: Screenshot of the Fairdomhub space for the COVID-19 Disease Map project. 236 people from 123 institutions around the globe are participating in this large-scale community effort to tackle the pandemic.

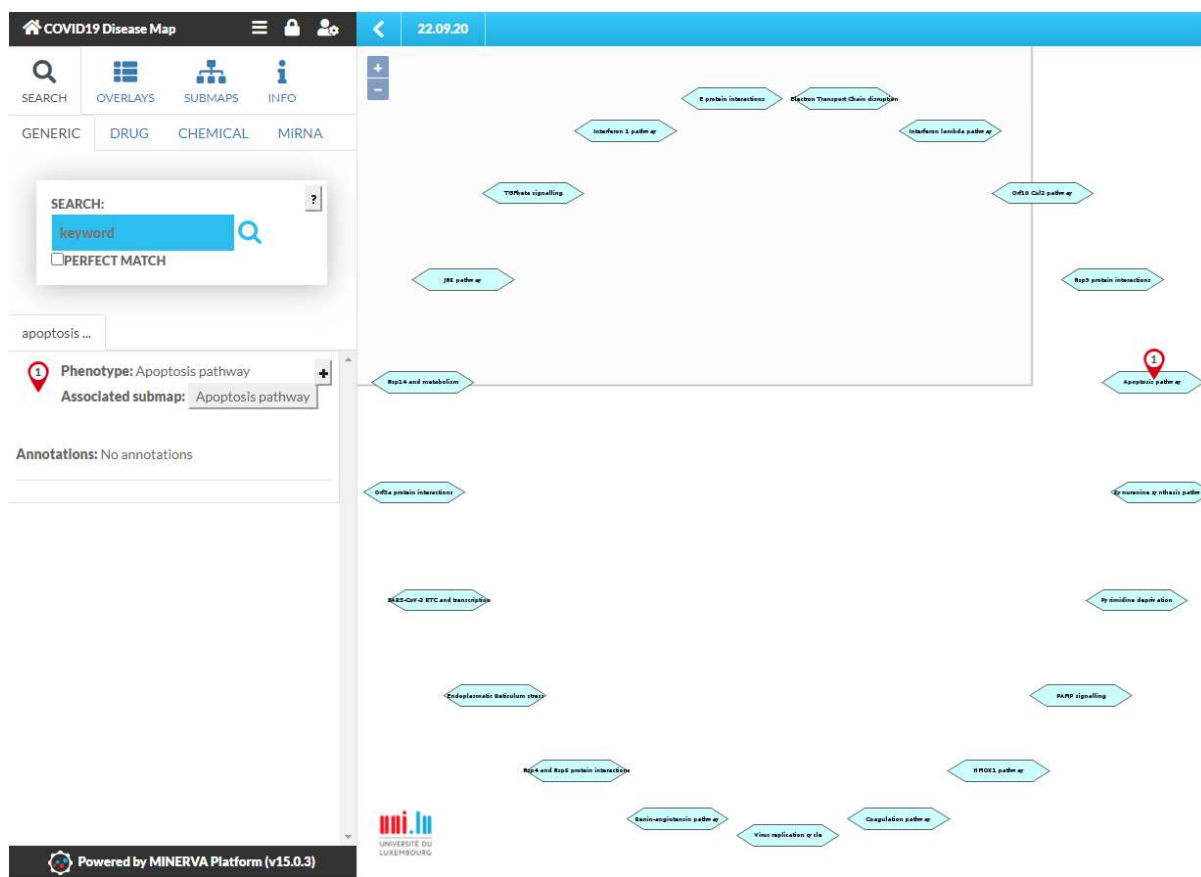


Figure 6.2: Screenshot of the MINERVA build for the COVID-19 Disease Map diagrams.

6.3 Apoptosis diagram

My contribution to the community effort is the diagram of apoptosis, which I created using the software CellDesigner. I curated manually 10 scientific articles, used SBGN rules and annotated the diagram using MIRIAM identifiers (**Fig 6.3**). The apoptosis map consists of 25 reactions, 37 chemical species and 32 proteins.

I also contributed to the development of the latest CaSQ version (0.7.8) identifying new scenarios present in the COVID-19 Disease Map graph collection that needed to be addressed by the tool in order to provide reliable executable models. CaSQ can now produce additional SIF files, so that the Activity Flow like structure of the inferred Boolean models can be used independently from the simulations part.

The CaSQ pipeline from map to the model presented in this thesis has been adopted by the community and SBML qual files of executable modules have been created and are now being analyzed.

I participated actively in weekly teleconferences, exchanged on best practices and curation criteria and had the opportunity to discuss and learn from top scientists in the field of computational biology and expand my scientific horizons.

For my contributions I am the 7th author (out of 140 co-authors) in the first community full length manuscript that is under preparation.

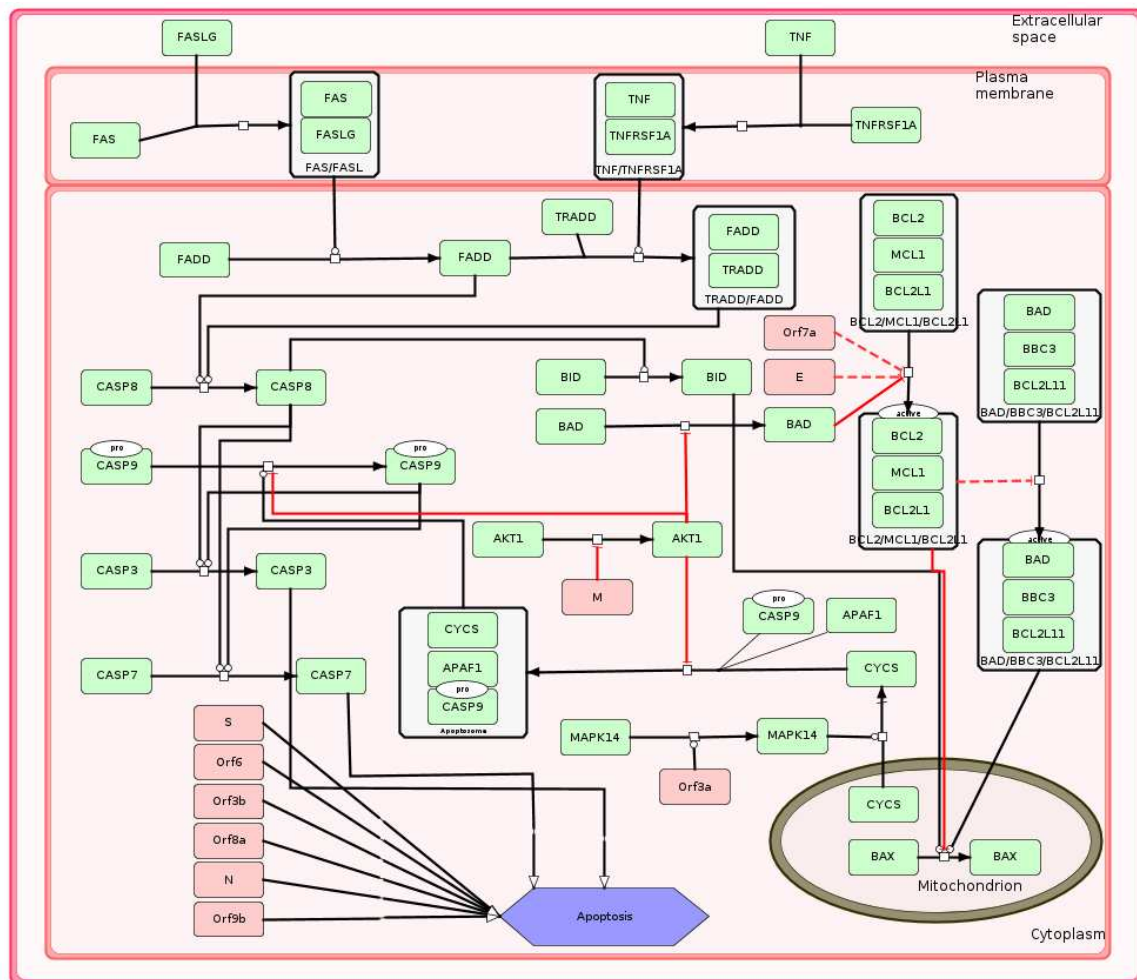


Figure 6.3: Apoptosis map built with Celldesigner graph editing software is structured into three compartments, namely Extracellular space containing the ligands, plasma membrane with receptor-ligand complexes and cytoplasm with all signaling and viral proteins. Green boxes represent generic proteins while peach colored boxes represent viral proteins. Red colored interactions are inhibitions while the black interactions are activations.

Bibliography

- Abou-Jaoudé, W., Traynard, P., Monteiro, P. T., Saez-Rodriguez, J., Helikar, T., Thieffry, D., & Chaouiya, C. (2016). Logical modeling and dynamical analysis of cellular networks. *Frontiers in Genetics*, 7, 94. <https://doi.org/10.3389/fgene.2016.00094>
- Aghamiri, S. S., Singh, V., Naldi, A., Helikar, T., Soliman, S., & Niarakis, A. (2020). Automated inference of Boolean models from molecular interaction maps using CaSQ. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btaa484>
- Ainola, M. M., Mandelin, J. A., Liljeström, M. P., Li, T. F., Hukkanen, M. V. J., & Konttinen, Y. T. (2005). Pannus invasion and cartilage degradation in rheumatoid arthritis: involvement of MMP-3 and interleukin-1beta. *Clinical and Experimental Rheumatology*, 23(5), 644–650.
- Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1), 47–97. <https://doi.org/10.1103/RevModPhys.74.47>
- Aletaha, D., & Smolen, J. S. (2018). Diagnosis and management of rheumatoid arthritis: A review. *The Journal of the American Medical Association*, 320(13), 1360–1372. <https://doi.org/10.1001/jama.2018.13103>
- Almuttaqi, H., & Udalova, I. A. (2019). Advances and challenges in targeting IRF5, a key regulator of inflammation. *The FEBS Journal*, 286(9), 1624–1637. <https://doi.org/10.1111/febs.14654>
- Alunno, A., Carubbi, F., Giacomelli, R., & Gerli, R. (2017). Cytokines in the pathogenesis of rheumatoid arthritis: new players and therapeutic targets. *BMC Rheumatology*, 1, 3. <https://doi.org/10.1186/s41927-017-0001-8>
- Arkin, A., Ross, J., & McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics*, 149(4), 1633–1648.
- Arnett, F. C., Edworthy, S. M., Bloch, D. A., McShane, D. J., Fries, J. F., Cooper, N. S., Healey, L. A., Kaplan, S. R., Liang, M. H., & Luthra, H. S. (1988). The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis and Rheumatism*, 31(3), 315–324. <https://doi.org/10.1002/art.1780310302>
- Azeloglu, E. U., & Iyengar, R. (2015). Good practices for building dynamical models in

- systems biology. *Science Signaling*, 8(371), fs8. <https://doi.org/10.1126/scisignal.aab0880>
- Baier, A., Meineckel, I., Gay, S., & Pap, T. (2003). Apoptosis in rheumatoid arthritis. *Current Opinion in Rheumatology*, 15(3), 274–279.
- Barabási, A.-L. (2009). Scale-free networks: a decade and beyond. *Science*, 325(5939), 412–413. <https://doi.org/10.1126/science.1173299>
- Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews. Genetics*, 12(1), 56–68. <https://doi.org/10.1038/nrg2918>
- Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews. Genetics*, 5(2), 101–113. <https://doi.org/10.1038/nrg1272>
- Bartok, B., & Firestein, G. S. (2010). Fibroblast-like synoviocytes: key effector cells in rheumatoid arthritis. *Immunological Reviews*, 233(1), 233–255. <https://doi.org/10.1111/j.0105-2896.2009.00859.x>
- Bartok, B., Hammaker, D., & Firestein, G. S. (2014). Phosphoinositide 3-kinase δ regulates migration and invasion of synoviocytes in rheumatoid arthritis. *Journal of Immunology*, 192(5), 2063–2070. <https://doi.org/10.4049/jimmunol.1300950>
- Becker, K. G., Barnes, K. C., Bright, T. J., & Wang, S. A. (2004). The genetic association database. *Nature Genetics*, 36(5), 431–432. <https://doi.org/10.1038/ng0504-431>
- Bekkar, A., Estreicher, A., Niknejad, A., Casals-Casas, C., Bridge, A., Xenarios, I., Dorier, J., & Crespo, I. (2018). Expert curation for building network-based dynamical models: a case study on atherosclerotic plaque formation. *Database: The Journal of Biological Databases and Curation*, 2018. <https://doi.org/10.1093/database/bay031>
- Benjamin, O., Bansal, P., Goyal, A., & Lappin, S. L. (2020). Disease Modifying Anti-Rheumatic Drugs (DMARD). In *StatPearls*. StatPearls Publishing.
- Bornholdt, S. (2008). Boolean network models of cellular regulation: prospects and limitations. *Journal of the Royal Society, Interface*, 5 Suppl 1, S85-94. <https://doi.org/10.1098/rsif.2008.0132.focus>
- Bounab, Y., Hesse, A.-M., Iannascoli, B., Grieco, L., Couté, Y., Niarakis, A., Roncagalli, R., Lie, E., Lam, K.-P., Demangel, C., Thieffry, D., Garin, J., Malissen, B., & Daëron, M.

- (2013). Proteomic analysis of the SH2 domain-containing leukocyte protein of 76 kDa (SLP76) interactome in resting and activated primary mast cells [corrected]. *Molecular & Cellular Proteomics*, 12(10), 2874–2889. <https://doi.org/10.1074/mcp.M112.025908>
- Boyle, W. J., Simonet, W. S., & Lacey, D. L. (2003). Osteoclast differentiation and activation. *Nature*, 423(6937), 337–342. <https://doi.org/10.1038/nature01658>
- Bray, D., Bourret, R. B., & Simon, M. I. (1993). Computer simulation of the phosphorylation cascade controlling bacterial chemotaxis. *Molecular Biology of the Cell*, 4(5), 469–482.
- Breitling, R. (2010). What is systems biology? *Frontiers in Physiology*, 1, 9. <https://doi.org/10.3389/fphys.2010.00009>
- Büchel, F., Rodriguez, N., Swainston, N., Wrzodek, C., Czauderna, T., Keller, R., Mittag, F., Schubert, M., Glont, M., Golebiewski, M., van Iersel, M., Keating, S., Rall, M., Wybrow, M., Hermjakob, H., Hucka, M., Kell, D. B., Müller, W., Mendes, P., ... Le Novère, N. (2013). Path2Models: large-scale generation of computational models from biochemical pathway maps. *BMC Systems Biology*, 7, 116. <https://doi.org/10.1186/1752-0509-7-116>
- Burrage, P. S., Mix, K. S., & Brinckerhoff, C. E. (2006). Matrix metalloproteinases: role in arthritis. *Frontiers in Bioscience*, 11, 529–543. <https://doi.org/10.2741/1817>
- Bustamante, M. F., Garcia-Carbonell, R., Whisenant, K. D., & Guma, M. (2017). Fibroblast-like synoviocyte metabolism in the pathogenesis of rheumatoid arthritis. *Arthritis Research & Therapy*, 19(1), 110. <https://doi.org/10.1186/s13075-017-1303-3>
- Caron, E., Ghosh, S., Matsuoka, Y., Ashton-Beaucage, D., Therrien, M., Lemieux, S., Perreault, C., Roux, P. P., & Kitano, H. (2010). A comprehensive map of the mTOR signaling network. *Molecular Systems Biology*, 6, 453. <https://doi.org/10.1038/msb.2010.108>
- Chance, B., Greenstein, D. S., Higgins, J., & Yang, C. C. (1952). The mechanism of catalase action. II. Electric analog computer studies. *Archives of Biochemistry and Biophysics*, 37(2), 322–339.
- Chaouiya, C., Naldi, A., & Thieffry, D. (2012). Logical modelling of gene regulatory networks with GINSim. *Methods in Molecular Biology*, 804, 463–479. https://doi.org/10.1007/978-1-61779-361-5_23
- Choy, E. H., & Panayi, G. S. (2001). Cytokine pathways and joint inflammation in rheumatoid arthritis. *The New England Journal of Medicine*, 344(12), 907–916.

<https://doi.org/10.1056/NEJM200103223441207>

- Cho, D.-Y., Kim, Y.-A., & Przytycka, T. M. (2012). Chapter 5: Network biology approach to complex diseases. *PLoS Computational Biology*, 8(12), e1002820. <https://doi.org/10.1371/journal.pcbi.1002820>
- Cho, M.-L., Ju, J.-H., Kim, H.-R., Oh, H.-J., Kang, C.-M., Jhun, J.-Y., Lee, S.-Y., Park, M.-K., Min, J.-K., Park, S.-H., Lee, S.-H., & Kim, H.-Y. (2007). Toll-like receptor 2 ligand mediates the upregulation of angiogenic factor, vascular endothelial growth factor and interleukin-8/CXCL8 in human rheumatoid synovial fibroblasts. *Immunology Letters*, 108(2), 121–128. <https://doi.org/10.1016/j.imlet.2006.11.005>
- Choy, E. H. S., Miceli-Richard, C., González-Gay, M. A., Sinigaglia, L., Schlichting, D. E., Meszaros, G., de la Torre, I., & Schulze-Koops, H. (2019). The effect of JAK1/JAK2 inhibition in rheumatoid arthritis: efficacy and safety of baricitinib. *Clinical and Experimental Rheumatology*, 37(4), 694–704.
- Cici, D., Corrado, A., Rotondo, C., & Cantatore, F. P. (2019). Wnt signaling and biological therapy in rheumatoid arthritis and spondyloarthritis. *International Journal of Molecular Sciences*, 20(22). <https://doi.org/10.3390/ijms20225552>
- Cieśla, M., Kolarz, B., Majdan, M., & Darmochwał-Kolarz, D. (2019). IRF5 promoter methylation as a new potential marker of rheumatoid arthritis. *Polish Archives of Internal Medicine*, 129(6), 370–376. <https://doi.org/10.20452/pamw.14863>
- Clancy, J., & Hasthorpe, H. (2011, November 21). *Pathophysiology of rheumatoid arthritis: nature or nurture?* Primary Health Care. <http://10.7748/phc2011.11.21.9.29.c8797>
- Clark, A. R., & Dean, J. L. (2012). The p38 MAPK pathway in rheumatoid arthritis: A sideways look. *The Open Rheumatology Journal*, 6, 209–219. <https://doi.org/10.2174/1874312901206010209>
- Craig, J. (2008). Complex Diseases: Research and Applications. *Scitable, from Nature Education*. <https://www.nature.com/scitable/topicpage/complex-diseases-research-and-applications-748/>
- Cross, M., Smith, E., Hoy, D., Carmona, L., Wolfe, F., Vos, T., Williams, B., Gabriel, S., Lassere, M., Johns, N., Buchbinder, R., Woolf, A., & March, L. (2014). The global burden of rheumatoid arthritis: estimates from the global burden of disease 2010 study. *Annals of the Rheumatic Diseases*, 73(7), 1316–1322. <https://doi.org/10.1136/annrheumdis-2013->

- Czauderna, T., Klukas, C., & Schreiber, F. (2010). Editing, validating and translating of SBGN maps. *Bioinformatics*, 26(18), 2340–2341. <https://doi.org/10.1093/bioinformatics/btq407>
- Demoruelle, M. K., Deane, K. D., & Holers, V. M. (2014). When and where does inflammation begin in rheumatoid arthritis? *Current Opinion in Rheumatology*, 26(1), 64–71. <https://doi.org/10.1097/BOR.0000000000000017>
- Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(5), P3. <https://doi.org/10.1186/gb-2003-4-9-r60>
- Dhillon, A. S., Hagan, S., Rath, O., & Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene*, 26(22), 3279–3290. <https://doi.org/10.1038/sj.onc.1210421>
- Dhillon, S. (2017). Tofacitinib: A review in rheumatoid arthritis. *Drugs*, 77(18), 1987–2001. <https://doi.org/10.1007/s40265-017-0835-9>
- Du, H., Zhang, X., Zeng, Y., Huang, X., Chen, H., Wang, S., Wu, J., Li, Q., Zhu, W., Li, H., Liu, T., Yu, Q., Wu, Y., & Jie, L. (2019). A Novel Phytochemical, DIM, Inhibits Proliferation, Migration, Invasion and TNF- α Induced Inflammatory Cytokine Production of Synovial Fibroblasts From Rheumatoid Arthritis Patients by Targeting MAPK and AKT/mTOR Signal Pathway. *Frontiers in Immunology*, 10, 1620. <https://doi.org/10.3389/fimmu.2019.01620>
- Elowitz, M. B., & Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767), 335–338. <https://doi.org/10.1038/35002125>
- Elshabrawy, H. A., Chen, Z., Volin, M. V., Ravella, S., Virupannavar, S., & Shahrara, S. (2015). The pathogenic role of angiogenesis in rheumatoid arthritis. *Angiogenesis*, 18(4), 433–448. <https://doi.org/10.1007/s10456-015-9477-2>
- Elshabrawy, H. A., Essani, A. E., Szekanecz, Z., Fox, D. A., & Shahrara, S. (2017). TLRs, future potential therapeutic targets for RA. *Autoimmunity Reviews*, 16(2), 103–113. <https://doi.org/10.1016/j.autrev.2016.12.003>
- Firestein, Gary S. (2010). 'Rac'-ing upstream to treat rheumatoid arthritis. *Arthritis Research & Therapy*, 12(1), 109. <https://doi.org/10.1186/ar2924>
- Firestein, Gary S., & McInnes, I. B. (2017). Immunopathogenesis of rheumatoid arthritis.

- Immunity*, 46(2), 183–196. <https://doi.org/10.1016/j.immuni.2017.02.006>
- Firestein, G S, Yeo, M., & Zvaifler, N. J. (1995). Apoptosis in rheumatoid arthritis synovium. *The Journal of Clinical Investigation*, 96(3), 1631–1638. <https://doi.org/10.1172/JCI118202>
- Fujita, K. A., Ostaszewski, M., Matsuoka, Y., Ghosh, S., Glaab, E., Trefois, C., Crespo, I., Perumal, T. M., Jurkowski, W., Antony, P. M. A., Diederich, N., Buttini, M., Kodama, A., Satagopam, V. P., Eifes, S., Del Sol, A., Schneider, R., Kitano, H., & Balling, R. (2014). Integrating pathways of Parkinson’s disease in a molecular interaction map. *Molecular Neurobiology*, 49(1), 88–102. <https://doi.org/10.1007/s12035-013-8489-4>
- Fukui, S., Iwamoto, N., Takatani, A., Igawa, T., Shimizu, T., Umeda, M., Nishino, A., Horai, Y., Hirai, Y., Koga, T., Kawashiri, S.-Y., Tamai, M., Ichinose, K., Nakamura, H., Origuchi, T., Masuyama, R., Kosai, K., Yanagihara, K., & Kawakami, A. (2017). M1 and M2 monocytes in rheumatoid arthritis: A contribution of imbalance of M1/M2 monocytes to osteoclastogenesis. *Frontiers in Immunology*, 8, 1958. <https://doi.org/10.3389/fimmu.2017.01958>
- Furlong, L. I. (2013). Human diseases through the lens of network biology. *Trends in Genetics*, 29(3), 150–159. <https://doi.org/10.1016/j.tig.2012.11.004>
- García, S., Liz, M., Gómez-Reino, J. J., & Conde, C. (2010). Akt activity protects rheumatoid synovial fibroblasts from Fas-induced apoptosis by inhibition of Bid cleavage. *Arthritis Research & Therapy*, 12(1), R33. <https://doi.org/10.1186/ar2941>
- Gawron, P., Ostaszewski, M., Satagopam, V., Gebel, S., Mazein, A., Kuzma, M., Zorzan, S., McGee, F., Otjacques, B., Balling, R., & Schneider, R. (2016). MINERVA-a platform for visualization and curation of molecular interaction networks. *NPJ Systems Biology and Applications*, 2, 16020. <https://doi.org/10.1038/npjsba.2016.20>
- Gilfillan, A. M., & Tkaczyk, C. (2006). Integrated signalling pathways for mast-cell activation. *Nature Reviews. Immunology*, 6(3), 218–230. <https://doi.org/10.1038/nri1782>
- Glass, L., & Kauffman, S. A. (1973). The logical analysis of continuous, non-linear biochemical control networks. *Journal of Theoretical Biology*, 39(1), 103–129. [https://doi.org/10.1016/0022-5193\(73\)90208-7](https://doi.org/10.1016/0022-5193(73)90208-7)
- Goddard, D. H., Kirk, A. P., Kirwan, J. R., Johnson, G. D., & Holborow, E. J. (1984). Impaired polymorphonuclear leucocyte chemotaxis in rheumatoid arthritis. *Annals of the*

- Rheumatic Diseases*, 43(2), 151–156. <https://doi.org/10.1136/ard.43.2.151>
- Goldbeter, A., & Koshland, D. E. (1981). An amplified sensitivity arising from covalent modification in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 78(11), 6840–6844. <https://doi.org/10.1073/pnas.78.11.6840>
- Goldring, S. R. (2002). Pathogenesis of bone erosions in rheumatoid arthritis. *Current Opinion in Rheumatology*, 14(4), 406–410.
- Grieco, L., Calzone, L., Bernard-Pierrot, I., Radvanyi, F., Kahn-Perlès, B., & Thieffry, D. (2013). Integrative modelling of the influence of MAPK network on cancer cell fate decision. *PLoS Computational Biology*, 9(10), e1003286. <https://doi.org/10.1371/journal.pcbi.1003286>
- Guo, Q., Wang, Y., Xu, D., Nossent, J., Pavlos, N. J., & Xu, J. (2018). Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies. *Bone Research*, 6, 15. <https://doi.org/10.1038/s41413-018-0016-9>
- Han, Z., Boyle, D. L., Chang, L., Bennett, B., Karin, M., Yang, L., Manning, A. M., & Firestein, G. S. (2001). c-Jun N-terminal kinase is required for metalloproteinase expression and joint destruction in inflammatory arthritis. *The Journal of Clinical Investigation*, 108(1), 73–81. <https://doi.org/10.1172/JCI12466>
- Han, Z., Boyle, D. L., Manning, A. M., & Firestein, G. S. (1998). AP-1 and NF-kappaB regulation in rheumatoid arthritis and murine collagen-induced arthritis. *Autoimmunity*, 28(4), 197–208. <https://doi.org/10.3109/08916939808995367>
- Hannemann, N., Jordan, J., Paul, S., Reid, S., Baenkler, H.-W., Sonnewald, S., Bäuerle, T., Vera, J., Schett, G., & Bozec, A. (2017). The AP-1 Transcription Factor c-Jun Promotes Arthritis by Regulating Cyclooxygenase-2 and Arginase-1 Expression in Macrophages. *Journal of Immunology*, 198(9), 3605–3614. <https://doi.org/10.4049/jimmunol.1601330>
- Helikar, Tomás, Konvalina, J., Heidel, J., & Rogers, J. A. (2008). Emergent decision-making in biological signal transduction networks. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6), 1913–1918. <https://doi.org/10.1073/pnas.0705088105>
- Helikar, Tomáš, Kowal, B., McClenathan, S., Bruckner, M., Rowley, T., Madrahimov, A., Wicks, B., Shrestha, M., Limbu, K., & Rogers, J. A. (2012). The Cell Collective: toward

- an open and collaborative approach to systems biology. *BMC Systems Biology*, 6, 96. <https://doi.org/10.1186/1752-0509-6-96>
- Hernandez, C., Thomas-Chollier, M., Naldi, A., & Thieffry, D. (2020). Computational verification of large logical models—application to the prediction of T cell response to checkpoint inhibitors. *Frontiers in Physiology*, 11. <https://doi.org/10.3389/fphys.2020.558606>
- Higgs, R. (2010). Rheumatoid arthritis: Synergistic effects of growth factors drive an RA phenotype in fibroblast-like synoviocytes. *Nature Reviews. Rheumatology*, 6(7), 383. <https://doi.org/10.1038/nrrheum.2010.92>
- Hoksza, D., Gawron, P., Ostaszewski, M., Smula, E., & Schneider, R. (2019). MINERVA API and plugins: opening molecular network analysis and visualization to the community. *Bioinformatics*, 35(21), 4496–4498. <https://doi.org/10.1093/bioinformatics/btz286>
- Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C., & Lempicki, R. A. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9), R183. <https://doi.org/10.1186/gb-2007-8-9-r183>
- Huber, L. C., Distler, O., Tarner, I., Gay, R. E., Gay, S., & Pap, T. (2006). Synovial fibroblasts: key players in rheumatoid arthritis. *Rheumatology (Oxford, England)*, 45(6), 669–675. <https://doi.org/10.1093/rheumatology/kel065>
- Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J. H., ... SBML Forum. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4), 524–531. <https://doi.org/10.1093/bioinformatics/btg015>
- Hu, J. X., Thomas, C. E., & Brunak, S. (2016). Network biology concepts in complex disease comorbidities. *Nature Reviews. Genetics*, 17(10), 615–629. <https://doi.org/10.1038/nrg.2016.87>
- Hwang, D., & Kim, W.-U. (2017). Rheumatoid arthritis: Modelling cytokine signalling networks. *Nature Reviews. Rheumatology*, 13(1), 5–6. <https://doi.org/10.1038/nrrheum.2016.194>

- Ideker, T., & Nussinov, R. (2017). Network approaches and applications in biology. *PLoS Computational Biology*, 13(10), e1005771. <https://doi.org/10.1371/journal.pcbi.1005771>
- Ivashkiv, L. B., & Hu, X. (2003). The JAK/STAT pathway in rheumatoid arthritis: pathogenic or protective? *Arthritis and Rheumatism*, 48(8), 2092–2096. <https://doi.org/10.1002/art.11095>
- Jacobs, R. A., Perrett, D., Axon, J. M., Herbert, K. E., & Scott, D. L. (1995). Rheumatoid synovial cell proliferation, transformation and fibronectin secretion in culture. *Clinical and Experimental Rheumatology*, 13(6), 717–723.
- Jagannadham, J., Jaiswal, H. K., Agrawal, S., & Rawal, K. (2016). Comprehensive map of molecules implicated in obesity. *Plos One*, 11(2), e0146759. <https://doi.org/10.1371/journal.pone.0146759>
- Jhun, J., Lee, S. H., Kim, S.-Y., Ryu, J., Kwon, J. Y., Na, H. S., Jung, K., Moon, S.-J., Cho, M.-L., & Min, J.-K. (2019). RIPK1 inhibition attenuates experimental autoimmune arthritis via suppression of osteoclastogenesis. *Journal of Translational Medicine*, 17(1), 84. <https://doi.org/10.1186/s12967-019-1809-3>
- Joshi, A., & Palsson, B. O. (1989). Metabolic dynamics in the human red cell. Part I--A comprehensive kinetic model. *Journal of Theoretical Biology*, 141(4), 515–528. [https://doi.org/10.1016/s0022-5193\(89\)80233-4](https://doi.org/10.1016/s0022-5193(89)80233-4)
- Kajita, M., Hogan, C., Harris, A. R., Dupre-Crochet, S., Itasaki, N., Kawakami, K., Charras, G., Tada, M., & Fujita, Y. (2010). Interaction with surrounding normal epithelial cells influences signalling pathways and behaviour of Src-transformed cells. *Journal of Cell Science*, 123(Pt 2), 171–180. <https://doi.org/10.1242/jcs.057976>
- Kalliolas, G. D., & Ivashkiv, L. B. (2016). TNF biology, pathogenic mechanisms and emerging therapeutic strategies. *Nature Reviews. Rheumatology*, 12(1), 49–62. <https://doi.org/10.1038/nrrheum.2015.169>
- Kanehisa, M. (2009). Representation and analysis of molecular networks involving diseases and drugs. *Genome Informatics. International Conference on Genome Informatics*, 23(1), 212–213.
- Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3), 437–467. [https://doi.org/10.1016/0022-5193\(69\)90015-0](https://doi.org/10.1016/0022-5193(69)90015-0)

- Kauffman, S., Peterson, C., Samuelsson, B., & Troein, C. (2003). Random Boolean network models and the yeast transcriptional network. *Proceedings of the National Academy of Sciences of the United States of America*, 100(25), 14796–14799. <https://doi.org/10.1073/pnas.2036429100>
- Kay, J., & Upchurch, K. S. (2012). ACR/EULAR 2010 rheumatoid arthritis classification criteria. *Rheumatology (Oxford, England)*, 51 Suppl 6, vi5-9. <https://doi.org/10.1093/rheumatology/kes279>
- Kennedy, N. J., & Davis, R. J. (2003). Role of JNK in tumor development. *Cell Cycle*, 2(3), 199–201.
- Kim, S.-J., Chen, Z., Chamberlain, N. D., Essani, A. B., Volin, M. V., Amin, M. A., Volkov, S., Gravallesse, E. M., Arami, S., Swedler, W., Lane, N. E., Mehta, A., Sweiss, N., & Shahrara, S. (2014). Ligation of TLR5 promotes myeloid cell infiltration and differentiation into mature osteoclasts in rheumatoid arthritis and experimental arthritis. *Journal of Immunology*, 193(8), 3902–3913. <https://doi.org/10.4049/jimmunol.1302998>
- Kitano, H., Funahashi, A., Matsuoka, Y., & Oda, K. (2005). Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology*, 23(8), 961–966. <https://doi.org/10.1038/nbt1111>
- Klareskog, L., Padyukov, L., Rönnelid, J., & Alfredsson, L. (2006). Genes, environment and immunity in the development of rheumatoid arthritis. *Current Opinion in Immunology*, 18(6), 650–655. <https://doi.org/10.1016/j.coi.2006.06.004>
- Kobayashi, T., Okamoto, K., Kobata, T., Hasunuma, T., Kato, T., Hamada, H., & Nishioka, K. (2000). Novel gene therapy for rheumatoid arthritis by FADD gene transfer: induction of apoptosis of rheumatoid synoviocytes but not chondrocytes. *Gene Therapy*, 7(6), 527–533. <https://doi.org/10.1038/sj.gt.3301127>
- Kohn, K. W. (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular Biology of the Cell*, 10(8), 2703–2734. <https://doi.org/10.1091/mbc.10.8.2703>
- Korb, A., Pavenstädt, H., & Pap, T. (2009). Cell death in rheumatoid arthritis. *Apoptosis: An International Journal on Programmed Cell Death*, 14(4), 447–454. <https://doi.org/10.1007/s10495-009-0317-y>
- Krämer, A., Green, J., Pollard, J., & Tugendreich, S. (2014). Causal analysis approaches in

- Ingenuity Pathway Analysis. *Bioinformatics*, 30(4), 523–530. <https://doi.org/10.1093/bioinformatics/btt703>
- Kramer, I., Wibulswas, A., Croft, D., & Genot, E. (2003). Rheumatoid arthritis: targeting the proliferative fibroblasts. *Progress in Cell Cycle Research*, 5, 59–70.
- Kriete, A., & Eils, R. (2013). *Computational Systems Biology: From Molecular Mechanisms to Disease*.
- Krogan, N. J., Lippman, S., Agard, D. A., Ashworth, A., & Ideker, T. (2015). The cancer cell map initiative: defining the hallmark networks of cancer. *Molecular Cell*, 58(4), 690–698. <https://doi.org/10.1016/j.molcel.2015.05.008>
- Kuperstein, I., Bonnet, E., Nguyen, H. A., Cohen, D., Viara, E., Grieco, L., Fourquet, S., Calzone, L., Russo, C., Kondratova, M., Dutreix, M., Barillot, E., & Zinovyev, A. (2015). Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, 4, e160. <https://doi.org/10.1038/oncsis.2015.19>
- Kurkó, J., Besenyei, T., Laki, J., Glant, T. T., Mikecz, K., & Szekanecz, Z. (2013). Genetics of rheumatoid arthritis - a comprehensive review. *Clinical Reviews in Allergy & Immunology*, 45(2), 170–179. <https://doi.org/10.1007/s12016-012-8346-7>
- Kwon, Y.-J., Lee, S.-W., Park, Y.-B., Lee, S.-K., & Park, M.-C. (2014). Secreted frizzled-related protein 5 suppresses inflammatory response in rheumatoid arthritis fibroblast-like synoviocytes through down-regulation of c-Jun N-terminal kinase. *Rheumatology (Oxford, England)*, 53(9), 1704–1711. <https://doi.org/10.1093/rheumatology/keu167>
- Kyriakis, J. M., & Avruch, J. (2001). Mammalian mitogen-activated protein kinase signal transduction pathways activated by stress and inflammation. *Physiological Reviews*, 81(2), 807–869. <https://doi.org/10.1152/physrev.2001.81.2.807>
- Le Novère, N. (2015). Quantitative and logic modelling of molecular and gene networks. *Nature Reviews. Genetics*, 16(3), 146–158. <https://doi.org/10.1038/nrg3885>
- Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., Crampin, E. J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J. L., Spence, H. D., & Wanner, B. L. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*, 23(12), 1509–1515. <https://doi.org/10.1038/nbt1156>

- Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S. E., Calzone, L., ... Kitano, H. (2009). The systems biology graphical notation. *Nature Biotechnology*, 27(8), 735–741. <https://doi.org/10.1038/nbt.1558>
- Lhomond, S., Avril, T., Dejeans, N., Voutetakis, K., Doultzinos, D., McMahon, M., Pineau, R., Obacz, J., Papadodima, O., Jouan, F., Bourien, H., Logotheti, M., Jégou, G., Pallares-Lupon, N., Schmit, K., Le Reste, P.-J., Etcheverry, A., Mosser, J., Barroso, K., ... Chevet, E. (2018). Dual IRE1 RNase functions dictate glioblastoma development. *EMBO Molecular Medicine*, 10(3). <https://doi.org/10.15252/emmm.201707929>
- Liu, T., Zhang, L., Joo, D., & Sun, S.-C. (2017). NF- κ B signaling in inflammation. *Signal Transduction and Targeted Therapy*, 2. <https://doi.org/10.1038/sigtrans.2017.23>
- Li, F., Long, T., Lu, Y., Ouyang, Q., & Tang, C. (2004). The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14), 4781–4786. <https://doi.org/10.1073/pnas.0305937101>
- Li, H., & Wan, A. (2013). Apoptosis of rheumatoid arthritis fibroblast-like synoviocytes: possible roles of nitric oxide and the thioredoxin 1. *Mediators of Inflammation*, 2013, 953462. <https://doi.org/10.1155/2013/953462>
- Livigni, A., O'Hara, L., Polak, M. E., Angus, T., Wright, D. W., Smith, L. B., & Freeman, T. C. (2018). A graphical and computational modeling platform for biological pathways. *Nature Protocols*, 13(4), 705–722. <https://doi.org/10.1038/nprot.2017.144>
- Lo Surdo, P., Calderone, A., Iannuccelli, M., Licata, L., Peluso, D., Castagnoli, L., Cesareni, G., & Perfetto, L. (2018). DISNOR: a disease network open resource. *Nucleic Acids Research*, 46(D1), D527–D534. <https://doi.org/10.1093/nar/gkx876>
- Macfarlane, F. R., Chaplain, M. A. J., & Eftimie, R. (2019). Quantitative predictive modelling approaches to understanding rheumatoid arthritis: A brief review. *Cells*, 9(1). <https://doi.org/10.3390/cells9010074>
- MacGregor, A. J., Snieder, H., Rigby, A. S., Koskenvuo, M., Kaprio, J., Aho, K., & Silman, A. J. (2000). Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis and Rheumatism*, 43(1), 30–37. [https://doi.org/10.1002/1529-0131\(200001\)43:1<30::AID-ANR5>3.0.CO;2-B](https://doi.org/10.1002/1529-0131(200001)43:1<30::AID-ANR5>3.0.CO;2-B)

- MacLauchlan, S., Zuriaga, M. A., Fuster, J. J., Cuda, C. M., Jonason, J., Behzadi, F., Duffen, J. P., Haines, G. K., Aprahamian, T., Perlman, H., & Walsh, K. (2017). Genetic deficiency of Wnt5a diminishes disease severity in a murine model of rheumatoid arthritis. *Arthritis Research & Therapy*, 19(1), 166. <https://doi.org/10.1186/s13075-017-1375-0>
- Makarov, S. S. (2001). NF-kappa B in rheumatoid arthritis: a pivotal regulator of inflammation, hyperplasia, and tissue destruction. *Arthritis Research*, 3(4), 200–206. <https://doi.org/10.1186/ar300>
- Malemud, C. J. (2007). Growth hormone, VEGF and FGF: involvement in rheumatoid arthritis. *Clinica Chimica Acta*, 375(1–2), 10–19. <https://doi.org/10.1016/j.cca.2006.06.033>
- Malemud, C. J. (2013). Intracellular signaling pathways in rheumatoid arthritis. *Journal of Clinical & Cellular Immunology*, 4, 160. <https://doi.org/10.4172/2155-9899.1000160>
- Malemud, C. J. (2017). Negative regulators of JAK/STAT signaling in rheumatoid arthritis and osteoarthritis. *International Journal of Molecular Sciences*, 18(3). <https://doi.org/10.3390/ijms18030484>
- Matsuno, H., Yudoh, K., Katayama, R., Nakazawa, F., Uzuki, M., Sawai, T., Yonezawa, T., Saeki, Y., Panayi, G. S., Pitzalis, C., & Kimura, T. (2002). The role of TNF-alpha in the pathogenesis of inflammation and joint destruction in rheumatoid arthritis (RA): a study using a human RA/SCID mouse chimera. *Rheumatology (Oxford, England)*, 41(3), 329–337. <https://doi.org/10.1093/rheumatology/41.3.329>
- Matsuoka, Y., Funahashi, A., Ghosh, S., & Kitano, H. (2014). Modeling and simulation using CellDesigner. *Methods in Molecular Biology*, 1164, 121–145. https://doi.org/10.1007/978-1-4939-0805-9_11
- Mazein, A., Knowles, R. G., Adcock, I., Chung, K. F., Wheelock, C. E., Maitland-van der Zee, A. H., Sterk, P. J., Auffray, C., & AsthmaMap Project Team. (2018). AsthmaMap: An expert-driven computational representation of disease mechanisms. *Clinical and Experimental Allergy*, 48(8), 916–918. <https://doi.org/10.1111/cea.13211>
- Mazein, A., Ostaszewski, M., Kuperstein, I., Watterson, S., Le Novère, N., Lefaudeux, D., De Meulder, B., Pellet, J., Balaur, I., Saqi, M., Nogueira, M. M., He, F., Parton, A., Lemonnier, N., Gawron, P., Gebel, S., Hainaut, P., Ollert, M., Dogrusoz, U., ... Auffray, C. (2018). Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *NPJ Systems Biology and Applications*, 4, 21.

<https://doi.org/10.1038/s41540-018-0059-y>

- McAllister, K., Eyre, S., & Orozco, G. (2011). Genetics of rheumatoid arthritis: GWAS and beyond. *Open Access Rheumatology: Research and Reviews*, 3, 31–46. <https://doi.org/10.2147/OARRR.S14725>
- McDermott, M. F. (2009). Rilonacept in the treatment of chronic inflammatory disorders. *Drugs of Today*, 45(6), 423–430. <https://doi.org/10.1358/dot.2009.45.6.1378935>
- McGettrick, H. M., Buckley, C. D., Filer, A., Rainger, G. E., & Nash, G. B. (2010). Stromal cells differentially regulate neutrophil and lymphocyte recruitment through the endothelium. *Immunology*, 131(3), 357–370. <https://doi.org/10.1111/j.1365-2567.2010.03307.x>
- McInnes, I. B., & Schett, G. (2007). Cytokines in the pathogenesis of rheumatoid arthritis. *Nature Reviews. Immunology*, 7(6), 429–442. <https://doi.org/10.1038/nri2094>
- McInnes, I. B., & Schett, G. (2011). The pathogenesis of rheumatoid arthritis. *The New England Journal of Medicine*, 365(23), 2205–2219. <https://doi.org/10.1056/NEJMra1004965>
- Mellado, M., Martínez-Muñoz, L., Cascio, G., Lucas, P., Pablos, J. L., & Rodríguez-Frade, J. M. (2015). T cell migration in rheumatoid arthritis. *Frontiers in Immunology*, 6, 384. <https://doi.org/10.3389/fimmu.2015.00384>
- Mendoza, L., & Xenarios, I. (2006). A method for the generation of standardized qualitative dynamical systems of regulatory networks. *Theoretical Biology & Medical Modelling*, 3, 13. <https://doi.org/10.1186/1742-4682-3-13>
- Mi, H., Schreiber, F., Moodie, S., Czauderna, T., Demir, E., Haw, R., Luna, A., Le Novère, N., Sorokin, A., & Villéger, A. (2015). Systems Biology Graphical Notation: Activity Flow language Level 1 Version 1.2. *Journal of Integrative Bioinformatics*, 12(2), 265. <https://doi.org/10.2390/biecoll-jib-2015-265>
- Mizoguchi, F., Slowikowski, K., Wei, K., Marshall, J. L., Rao, D. A., Chang, S. K., Nguyen, H. N., Noss, E. H., Turner, J. D., Earp, B. E., Blazar, P. E., Wright, J., Simmons, B. P., Donlin, L. T., Kalliolias, G. D., Goodman, S. M., Bykerk, V. P., Ivashkiv, L. B., Lederer, J. A., ... Brenner, M. B. (2018). Functionally distinct disease-associated fibroblast subsets in rheumatoid arthritis. *Nature Communications*, 9(1), 789. <https://doi.org/10.1038/s41467-018-02892-y>

- Mizuno, S., Iijima, R., Ogishima, S., Kikuchi, M., Matsuoka, Y., Ghosh, S., Miyamoto, T., Miyashita, A., Kuwano, R., & Tanaka, H. (2012). AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. *BMC Systems Biology*, 6, 52. <https://doi.org/10.1186/1752-0509-6-52>
- Mongan, E. S., & Jacox, R. F. (1964). Erythrocyte survival in rheumatoid arthritis. *Arthritis and Rheumatism*, 7, 481–489. <https://doi.org/10.1002/art.1780070504>
- Moodie, S., Le Novère, N., Demir, E., Mi, H., & Villéger, A. (2015). Systems Biology Graphical Notation: Process Description language Level 1 Version 1.3. *Journal of Integrative Bioinformatics*, 12(2), 263. <https://doi.org/10.2390/biecoll-jib-2015-263>
- Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., Bader, G. D., & Ferrin, T. E. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, 12, 436. <https://doi.org/10.1186/1471-2105-12-436>
- Mountz, J. D., Hsu, H. C., Matsuki, Y., & Zhang, H. G. (2001). Apoptosis and rheumatoid arthritis: past, present, and future directions. *Current Rheumatology Reports*, 3(1), 70–78. <https://doi.org/10.1007/s11926-001-0053-y>
- Müller-Ladner, U., Ospelt, C., Gay, S., Distler, O., & Pap, T. (2007). Cells of the synovium in rheumatoid arthritis. Synovial fibroblasts. *Arthritis Research & Therapy*, 9(6), 223. <https://doi.org/10.1186/ar2337>
- Namba, S., Nakano, R., Kitanaka, T., Kitanaka, N., Nakayama, T., & Sugiya, H. (2017). ERK2 and JNK1 contribute to TNF- α -induced IL-8 expression in synovial fibroblasts. *Plos One*, 12(8), e0182923. <https://doi.org/10.1371/journal.pone.0182923>
- Nanki, T., Nagasaka, K., Hayashida, K., Saita, Y., & Miyasaka, N. (2001). Chemokines regulate IL-6 and IL-8 production by fibroblast-like synoviocytes from patients with rheumatoid arthritis. *Journal of Immunology*, 167(9), 5381–5385. <https://doi.org/10.4049/jimmunol.167.9.5381>
- Niarakis, A., Bounab, Y., Grieco, L., Roncagalli, R., Hesse, A.-M., Garin, J., Malissen, B., Daëron, M., & Thieffry, D. (2014). Computational modeling of the main signaling pathways involved in mast cell activation. *Current Topics in Microbiology and Immunology*, 382, 69–93. https://doi.org/10.1007/978-3-319-07911-0_4
- Noack, M., & Miossec, P. (2017). Selected cytokine pathways in rheumatoid arthritis. *Seminars in Immunopathology*, 39(4), 365–383. <https://doi.org/10.1007/s00281-017-0619-z>

- Noort, A. R., Tak, P. P., & Tas, S. W. (2015). Non-canonical NF- κ B signaling in rheumatoid arthritis: Dr Jekyll and Mr Hyde? *Arthritis Research & Therapy*, 17, 15. <https://doi.org/10.1186/s13075-015-0527-3>
- Ogishima, S., Mizuno, S., Kikuchi, M., Miyashita, A., Kuwano, R., Tanaka, H., & Nakaya, J. (2016). Alzpathway, an updated map of curated signaling pathways: towards deciphering alzheimer's disease pathogenesis. *Methods in Molecular Biology*, 1303, 423–432. https://doi.org/10.1007/978-1-4939-2627-5_25
- Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R. R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J. C., Carroll, R. J., Eyler, A. E., Greenberg, J. D., Kremer, J. M., ... Plenge, R. M. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488), 376–381. <https://doi.org/10.1038/nature12873>
- Ostaszewski, M., Gebel, S., Kuperstein, I., Mazein, A., Zinovyev, A., Dogrusoz, U., Hasenauer, J., Fleming, R. M. T., Le Novère, N., Gawron, P., Ligon, T., Niarakis, A., Nickerson, D., Weindl, D., Balling, R., Barillot, E., Auffray, C., & Schneider, R. (2019). Community-driven roadmap for integrated disease maps. *Briefings in Bioinformatics*, 20(2), 659–670. <https://doi.org/10.1093/bib/bby024>
- Ostaszewski, M., Mazein, A., Gillespie, M. E., Kuperstein, I., Niarakis, A., Hermjakob, H., Pico, A. R., Willighagen, E. L., Evelo, C. T., Hasenauer, J., Schreiber, F., Dräger, A., Demir, E., Wolkenhauer, O., Furlong, L. I., Barillot, E., Dopazo, J., Orta-Resendiz, A., Messina, F., ... Schneider, R. (2020). Author Correction: COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Scientific Data*, 7(1), 247. <https://doi.org/10.1038/s41597-020-00589-w>
- Ostrowska, M., Maśliński, W., Prochorec-Sobieszek, M., Nieciecki, M., & Sudoł-Szopińska, I. (2018). Cartilage and bone damage in rheumatoid arthritis. *Reumatologia*, 56(2), 111–120. <https://doi.org/10.5114/reum.2018.75523>
- Panagopoulos, P. K., & Lambrou, G. I. (2018). Bone erosions in rheumatoid arthritis: recent developments in pathogenesis and therapeutic implications. *Journal of Musculoskeletal & Neuronal Interactions*, 18(3), 304–319.
- Parton, A., McGilligan, V., Chemaly, M., O’Kane, M., & Watterson, S. (2019). New models of atherosclerosis and multi-drug therapeutic interventions. *Bioinformatics*, 35(14), 2449–2457. <https://doi.org/10.1093/bioinformatics/bty980>

- Pettit, A. R., Ji, H., von Stechow, D., Müller, R., Goldring, S. R., Choi, Y., Benoist, C., & Gravallese, E. M. (2001). TRANCE/RANKL knockout mice are protected from bone erosion in a serum transfer model of arthritis. *The American Journal of Pathology*, 159(5), 1689–1699. [https://doi.org/10.1016/S0002-9440\(10\)63016-7](https://doi.org/10.1016/S0002-9440(10)63016-7)
- Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., & Furlong, L. I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1), D833–D839. <https://doi.org/10.1093/nar/gkw943>
- Punta, M., Coghill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., & Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, 40(Database issue), D290-301. <https://doi.org/10.1093/nar/gkr1065>
- Rein, P., & Mueller, R. B. (2017). Treatment with Biologicals in Rheumatoid Arthritis: An Overview. *Rheumatology and Therapy*, 4(2), 247–261. <https://doi.org/10.1007/s40744-017-0073-3>
- Rheumatoid arthritis. (2018). *Nature Reviews. Disease Primers*, 4, 18002. <https://doi.org/10.1038/nrdp.2018.2>
- Rohn, H., Junker, A., Hartmann, A., Grafahrend-Belau, E., Treutler, H., Klapperstück, M., Czauderna, T., Klukas, C., & Schreiber, F. (2012). VANTED v2: a framework for systems biology applications. *BMC Systems Biology*, 6, 139. <https://doi.org/10.1186/1752-0509-6-139>
- Rosengren, S., Corr, M., & Boyle, D. L. (2010). Platelet-derived growth factor and transforming growth factor beta synergistically potentiate inflammatory mediator synthesis by fibroblast-like synoviocytes. *Arthritis Research & Therapy*, 12(2), R65. <https://doi.org/10.1186/ar2981>
- Rougny, A., Touré, V., Moodie, S., Balaur, I., Czauderna, T., Borlinghaus, H., Dogrusoz, U., Mazein, A., Dräger, A., Blinov, M. L., Villéger, A., Haw, R., Demir, E., Mi, H., Sorokin, A., Schreiber, F., & Luna, A. (2019). Systems Biology Graphical Notation: Process Description language Level 1 Version 2.0. *Journal of Integrative Bioinformatics*, 16(2). <https://doi.org/10.1515/jib-2019-0022>
- Sandler, C., Joutsiniemi, S., Lindstedt, K. A., Juutilainen, T., Kovanen, P. T., & Eklund, K. K.

- (2006). Imatinib mesylate inhibits platelet derived growth factor stimulated proliferation of rheumatoid synovial fibroblasts. *Biochemical and Biophysical Research Communications*, 347(1), 31–35. <https://doi.org/10.1016/j.bbrc.2006.06.052>
- Sato, K., & Takayanagi, H. (2006). Osteoclasts, rheumatoid arthritis, and osteoimmunology. *Current Opinion in Rheumatology*, 18(4), 419–426. <https://doi.org/10.1097/01.bor.0000231912.24740.a5>
- Savageau, M. A. (1970). Biochemical systems analysis. 3. Dynamic solutions using a power-law approximation. *Journal of Theoretical Biology*, 26(2), 215–226. [https://doi.org/10.1016/s0022-5193\(70\)80013-3](https://doi.org/10.1016/s0022-5193(70)80013-3)
- Schett, G., Zwerina, J., & Firestein, G. (2008). The p38 mitogen-activated protein kinase (MAPK) pathway in rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 67(7), 909–916. <https://doi.org/10.1136/ard.2007.074278>
- Schinnerling, K., Aguillón, J. C., Catalán, D., & Soto, L. (2017). The role of interleukin-6 signalling and its therapeutic blockage in skewing the T cell balance in rheumatoid arthritis. *Clinical and Experimental Immunology*, 189(1), 12–20. <https://doi.org/10.1111/cei.12966>
- Schramek, H. (2002). MAP kinases: from intracellular signals to physiology and disease. *News in Physiological Sciences : An International Journal of Physiology Produced Jointly by the International Union of Physiological Sciences and the American Physiological Society*, 17, 62–67. <https://doi.org/10.1152/nips.01365.2001>
- Schrodi, S. J., Mukherjee, S., Shan, Y., Tromp, G., Sninsky, J. J., Callear, A. P., Carter, T. C., Ye, Z., Haines, J. L., Brilliant, M. H., Crane, P. K., Smelser, D. T., Elston, R. C., & Weeks, D. E. (2014). Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. *Frontiers in Genetics*, 5, 162. <https://doi.org/10.3389/fgene.2014.00162>
- Scott, D. L., Wolfe, F., & Huizinga, T. W. J. (2010). Rheumatoid arthritis. *The Lancet*, 376(9746), 1094–1108. [https://doi.org/10.1016/S0140-6736\(10\)60826-4](https://doi.org/10.1016/S0140-6736(10)60826-4)
- Seemayer, C. A., Kuchen, S., Neidhart, M., Kuenzler, P., Rihosková, V., Neumann, E., Pruschy, M., Aicher, W. K., Müller-Ladner, U., Gay, R. E., Michel, B. A., Firestein, G. S., & Gay, S. (2003). p53 in rheumatoid arthritis synovial fibroblasts at sites of invasion. *Annals of the Rheumatic Diseases*, 62(12), 1139–1144.

<https://doi.org/10.1136/ard.2003.007401>

- Setoguchi, R., Kinashi, T., Sagara, H., Hirosawa, K., & Takatsu, K. (1998). Defective degranulation and calcium mobilization of bone-marrow derived mast cells from Xid and Btk-deficient mice. *Immunology Letters*, 64(2–3), 109–118. [https://doi.org/10.1016/s0165-2478\(98\)00086-8](https://doi.org/10.1016/s0165-2478(98)00086-8)
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504. <https://doi.org/10.1101/gr.1239303>
- Sharp, S., & Workman, P. (2006). Inhibitors of the HSP90 molecular chaperone: current status. *Advances in Cancer Research*, 95, 323–348. [https://doi.org/10.1016/S0065-230X\(06\)95009-X](https://doi.org/10.1016/S0065-230X(06)95009-X)
- Shiozawa, S., Tsumiyama, K., Yoshida, K., & Hashiramoto, A. (2011). Pathogenesis of joint destruction in rheumatoid arthritis. *Archivum Immunologiae et Therapiae Experimentalis*, 59(2), 89–95. <https://doi.org/10.1007/s00005-011-0116-3>
- Shmulevich, I., Dougherty, E. R., & Wei Zhang. (2002). From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proceedings of the IEEE*, 90(11), 1778–1792. <https://doi.org/10.1109/JPROC.2002.804686>
- Simmonds, R. E., & Foxwell, B. M. (2008). Signalling, inflammation and arthritis: NF-kappaB and its relevance to arthritis and inflammation. *Rheumatology (Oxford, England)*, 47(5), 584–590. <https://doi.org/10.1093/rheumatology/kem298>
- Singh, V., Kallioliass, G. D., Ostaszewski, M., Veyssiere, M., Pilalis, E., Gawron, P., Mazein, A., Bonnet, E., Petit-Teixeira, E., & Niarakis, A. (2020). RA-map: building a state-of-the-art interactive knowledge base for rheumatoid arthritis. *Database: The Journal of Biological Databases and Curation*, 2020. <https://doi.org/10.1093/database/baaa017>
- Singh, V., Ostaszewski, M., Kallioliass, G. D., Chiocchia, G., Olaso, R., Petit-Teixeira, E., Helikar, T., & Niarakis, A. (2018). Computational systems biology approach for the study of rheumatoid arthritis: from a molecular map to a dynamical model. *Genomics and Computational Biology*, 4(1). <https://doi.org/10.18547/gcb.2018.vol4.iss1.e100050>
- Smith, M. D., Weedon, H., Papangelis, V., Walker, J., Roberts-Thomson, P. J., & Ahern, M. J. (2010). Apoptosis in the rheumatoid arthritis synovial membrane: modulation by disease-

- modifying anti-rheumatic drug treatment. *Rheumatology (Oxford, England)*, 49(5), 862–875. <https://doi.org/10.1093/rheumatology/kep467>
- Smolen, J. S., Aletaha, D., Barton, A., Burmester, G. R., Emery, P., Firestein, G. S., Kavanaugh, A., McInnes, I. B., Solomon, D. H., Strand, V., & Yamamoto, K. (2018). Rheumatoid arthritis. *Nature Reviews. Disease Primers*, 4, 18001. <https://doi.org/10.1038/nrdp.2018.1>
- Song, G., Ouyang, G., & Bao, S. (2005). The activation of Akt/PKB signaling pathway and cell survival. *Journal of Cellular and Molecular Medicine*, 9(1), 59–71. <https://doi.org/10.1111/j.1582-4934.2005.tb00337.x>
- Sparks, J. A. (2019). Rheumatoid Arthritis. *Annals of Internal Medicine*, 170(1), ITC1–ITC16. <https://doi.org/10.7326/AITC201901010>
- Stoll, G., Caron, B., Viara, E., Dugourd, A., Zinovyev, A., Naldi, A., Kroemer, G., Barillot, E., & Calzone, L. (2017). MaBoSS 2.0: an environment for stochastic Boolean modeling. *Bioinformatics*, 33(14), 2226–2228. <https://doi.org/10.1093/bioinformatics/btx123>
- Stoll, G., Viara, E., Barillot, E., & Calzone, L. (2012). Continuous time Boolean modeling for biological signaling: application of Gillespie algorithm. *BMC Systems Biology*, 6, 116. <https://doi.org/10.1186/1752-0509-6-116>
- Su, G., Kuchinsky, A., Morris, J. H., States, D. J., & Meng, F. (2010). GLay: community structure analysis of biological networks. *Bioinformatics*, 26(24), 3135–3137. <https://doi.org/10.1093/bioinformatics/btq596>
- Szekanecz, Z., Kim, J., & Koch, A. E. (2003). Chemokines and chemokine receptors in rheumatoid arthritis. *Seminars in Immunology*, 15(1), 15–21.
- Takekawa, M., Kubota, Y., Nakamura, T., & Ichikawa, K. (2011). Regulation of stress-activated MAP kinase pathways during cell fate decisions. *Nagoya Journal of Medical Science*, 73(1–2), 1–14.
- Tak, P. P., Zvaifler, N. J., Green, D. R., & Firestein, G. S. (2000). Rheumatoid arthritis and p53: how oxidative stress might alter the course of inflammatory diseases. *Immunology Today*, 21(2), 78–82.
- Taranto, E., Xue, J. R., Lacey, D., Hutchinson, P., Smith, M., Morand, E. F., & Leech, M. (2005). Detection of the p53 regulator murine double-minute protein 2 in rheumatoid arthritis. *The Journal of Rheumatology*, 32(3), 424–429.

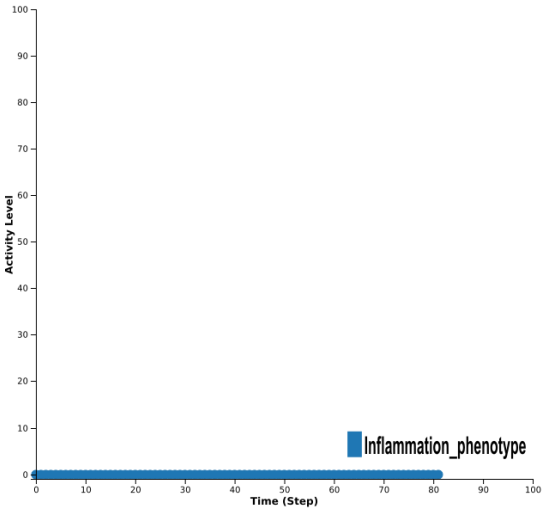
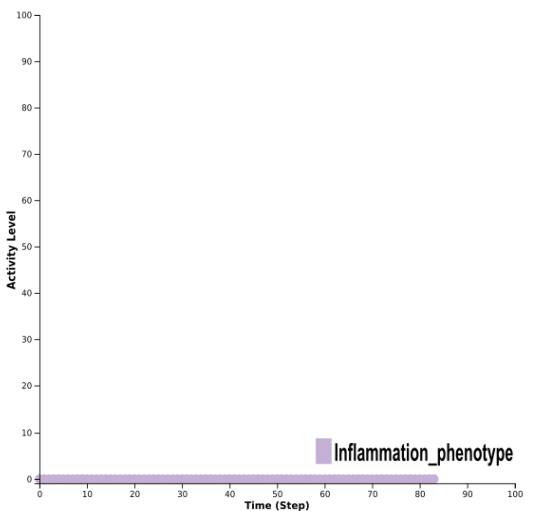
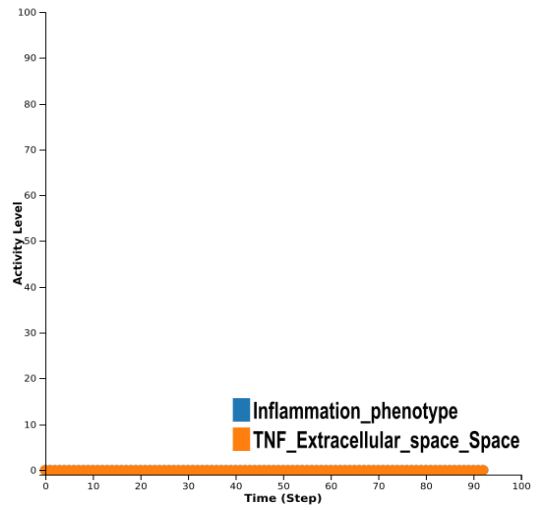
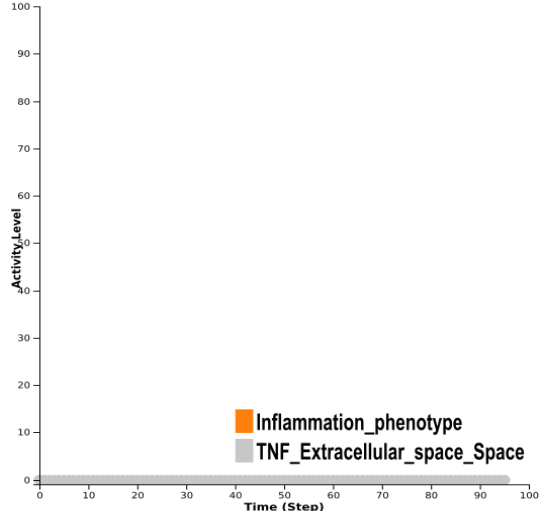
- Terabe, F., Kitano, M., Kawai, M., Kuwahara, Y., Hirano, T., Arimitsu, J., Hagihara, K., Shima, Y., Narazaki, M., Tanaka, T., Kawase, I., Sano, H., & Ogata, A. (2009). Imatinib mesylate inhibited rat adjuvant arthritis and PDGF-dependent growth of synovial fibroblast via interference with the Akt signaling pathway. *Modern Rheumatology*, 19(5), 522–529. <https://doi.org/10.1007/s10165-009-0193-x>
- Thalhamer, T., McGrath, M. A., & Harnett, M. M. (2008). MAPKs and their relevance to arthritis and inflammation. *Rheumatology (Oxford, England)*, 47(4), 409–414. <https://doi.org/10.1093/rheumatology/kem297>
- Thomas, R. (1973). Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42(3), 563–585. [https://doi.org/10.1016/0022-5193\(73\)90247-6](https://doi.org/10.1016/0022-5193(73)90247-6)
- Thomas, R. (1978). Logical analysis of systems comprising feedback loops. *Journal of Theoretical Biology*, 73(4), 631–656. [https://doi.org/10.1016/0022-5193\(78\)90127-3](https://doi.org/10.1016/0022-5193(78)90127-3)
- Thomas, R., Gathoye, A. M., & Lambert, L. (1976). A complex control circuit. Regulation of immunity in temperate bacteriophages. *European Journal of Biochemistry / FEBS*, 71(1), 211–227. <https://doi.org/10.1111/j.1432-1033.1976.tb11108.x>
- Tozluoğlu, M., Karaca, E., Haliloglu, T., & Nussinov, R. (2008). Cataloging and organizing p73 interactions in cell cycle arrest and apoptosis. *Nucleic Acids Research*, 36(15), 5033–5049. <https://doi.org/10.1093/nar/gkn481>
- Tripathi, S., Flobak, Å., Chawla, K., Baudot, A., Bruland, T., Thommesen, L., Kuiper, M., & Lægreid, A. (2015). The gastrin and cholecystokinin receptors mediated signaling network: a scaffold for data analysis and new hypotheses on regulatory mechanisms. *BMC Systems Biology*, 9, 40. <https://doi.org/10.1186/s12918-015-0181-z>
- Turner, J. D., & Filer, A. (2015). The role of the synovial fibroblast in rheumatoid arthritis pathogenesis. *Current Opinion in Rheumatology*, 27(2), 175–182. <https://doi.org/10.1097/BOR.0000000000000148>
- van der Laan, W. H., Quax, P. H. A., Seemayer, C. A., Huisman, L. G. M., Pieterman, E. J., Grimbergen, J. M., Verheijen, J. H., Breedveld, F. C., Gay, R. E., Gay, S., Huizinga, T. W. J., & Pap, T. (2003). Cartilage degradation and invasion by rheumatoid synovial fibroblasts is inhibited by gene transfer of TIMP-1 and TIMP-3. *Gene Therapy*, 10(3), 234–242. <https://doi.org/10.1038/sj.gt.3301871>
- Vidal, M., Cusick, M. E., & Barabási, A.-L. (2011). Interactome networks and human disease.

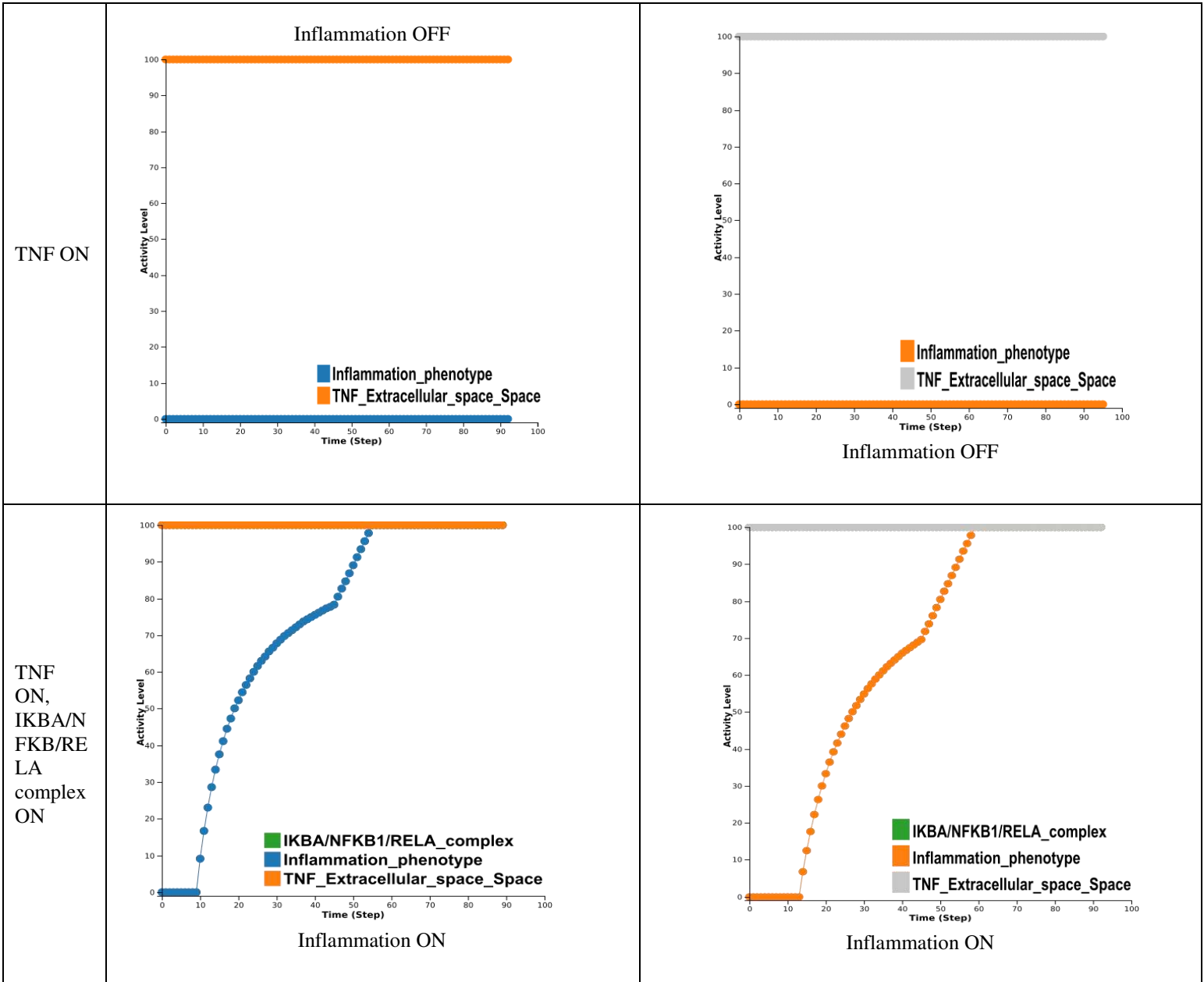
- Cell*, 144(6), 986–998. <https://doi.org/10.1016/j.cell.2011.02.016>
- Vogt, T., Czauderna, T., & Schreiber, F. (2013). Translation of SBGN maps: Process Description to Activity Flow. *BMC Systems Biology*, 7, 115. <https://doi.org/10.1186/1752-0509-7-115>
- von Dassow, G., Meir, E., Munro, E. M., & Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature*, 406(6792), 188–192. <https://doi.org/10.1038/35018085>
- Waler, E. (2007). On the occurrence of a factor in human serum activating the specific agglutination of sheep blood corpuscles. 1939. *Acta Pathologica, Microbiologica, et Immunologica Scandinavica*, 115(5), 422–438; discussion 439. https://doi.org/10.1111/j.1600-0463.2007.apm_682a.x
- Wang, P., Lü, J., & Yu, X. (2014). Identification of important nodes in directed biological networks: a network motif approach. *Plos One*, 9(8), e106132. <https://doi.org/10.1371/journal.pone.0106132>
- Wang, R.-S., Saadatpour, A., & Albert, R. (2012). Boolean modeling in systems biology: an overview of methodology and applications. *Physical Biology*, 9(5), 055001. <https://doi.org/10.1088/1478-3975/9/5/055001>
- Wijbrandts, C. A., & Tak, P. P. (2017). Prediction of response to targeted treatment in rheumatoid arthritis. *Mayo Clinic Proceedings*, 92(7), 1129–1143. <https://doi.org/10.1016/j.mayocp.2017.05.009>
- Woetzel, D., Huber, R., Kupfer, P., Pohlers, D., Pfaff, M., Driesch, D., Häupl, T., Koczan, D., Stiehl, P., Guthke, R., & Kinne, R. W. (2014). Identification of rheumatoid arthritis and osteoarthritis patients by transcriptome-based rule set generation. *Arthritis Research & Therapy*, 16(2), R84. <https://doi.org/10.1186/ar4526>
- Wu, G., Zhu, L., Dent, J. E., & Nardini, C. (2010). A comprehensive molecular interaction map for rheumatoid arthritis. *Plos One*, 5(4), e10137. <https://doi.org/10.1371/journal.pone.0010137>
- Xu, L., Feng, X., Tan, W., Gu, W., Guo, D., Zhang, M., & Wang, F. (2013). IL-29 enhances Toll-like receptor-mediated IL-6 and IL-8 production by the synovial fibroblasts from rheumatoid arthritis patients. *Arthritis Research & Therapy*, 15(5), R170. <https://doi.org/10.1186/ar4357>

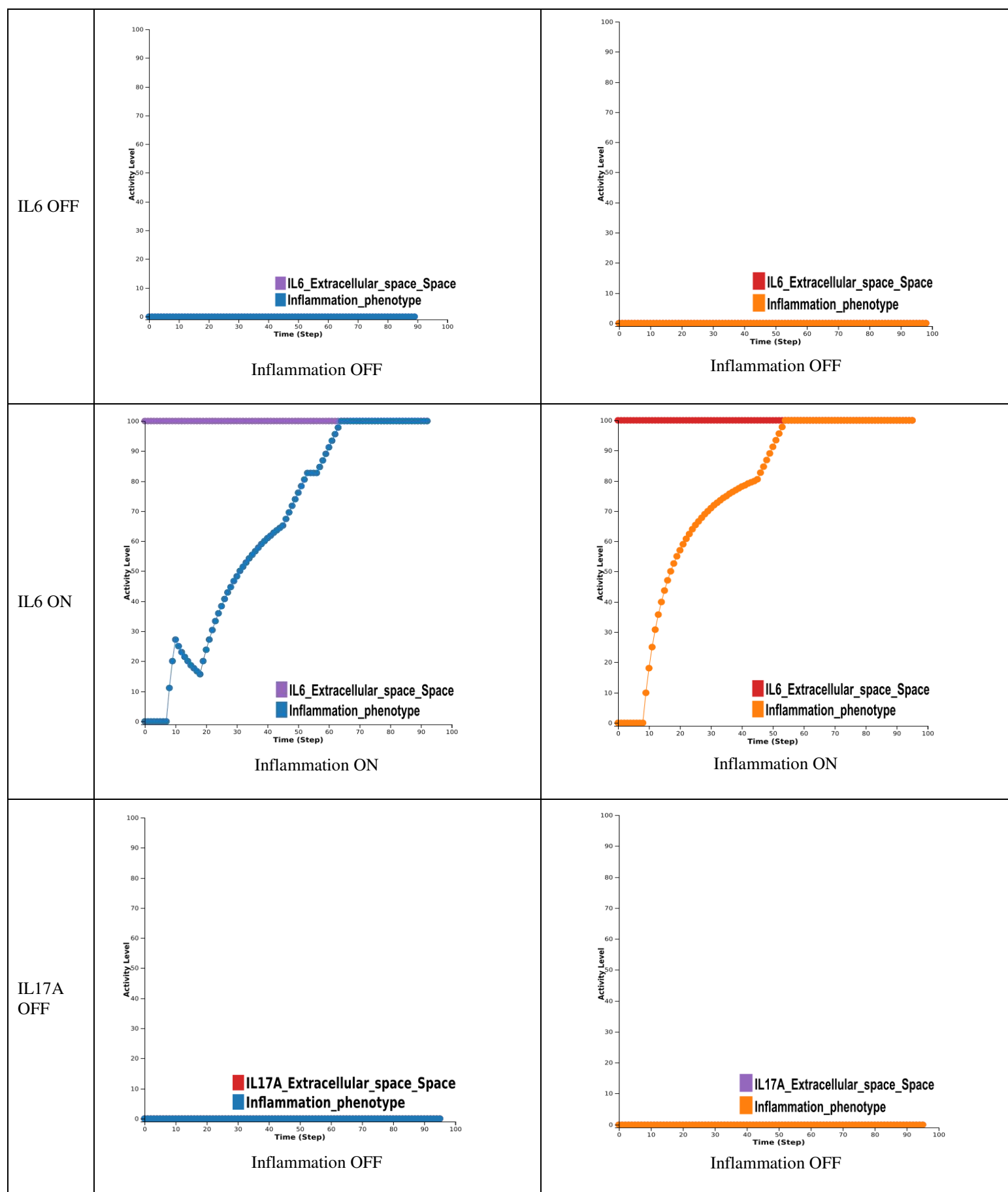
- Yasuda, T. (2006). Cartilage destruction by matrix degradation products. *Modern Rheumatology*, 16(4), 197–205. <https://doi.org/10.1007/s10165-006-0490-6>
- Yi, A.-K., Yoon, T. W., Cho, H., Brand, D., Kim-Hoehamer, Y.-I., Park, J.-E., Hasty, K., & Stuart, J. (2015). Contribution of MyD88-dependent PKD1 to the development of arthritis (THER2P.954). *The Journal of Immunology*, 194(1 Supplement), 67.5-67.5.
- Yoshihara, Y., & Yamada, H. (2007). [Matrix metalloproteinases and cartilage matrix degradation in rheumatoid arthritis]. *Clinical Calcium*, 17(4), 500–508. <https://doi.org/CliCa0704500508>
- Zerrouk, N., Miagoux, Q., Dispot, A., Elati, M., & Niarakis, A. (2020). Identification of putative master regulators in rheumatoid arthritis synovial fibroblasts using gene expression data and network inference. *Scientific Reports*, 10(1), 16236. <https://doi.org/10.1038/s41598-020-73147-4>
- Zhang, B., Tian, Y., & Zhang, Z. (2014). Network biology in medicine and beyond. *Circulation. Cardiovascular Genetics*, 7(4), 536–547. <https://doi.org/10.1161/CIRCGENETICS.113.000123>
- Zhang, T., Li, H., Shi, J., Li, S., Li, M., Zhang, L., Zheng, L., Zheng, D., Tang, F., Zhang, X., Zhang, F., & You, X. (2016). p53 predominantly regulates IL-6 production and suppresses synovial inflammation in fibroblast-like synoviocytes and adjuvant-induced arthritis. *Arthritis Research & Therapy*, 18(1), 271. <https://doi.org/10.1186/s13075-016-1161-4>
- Zhu, L.-J., Yang, T.-C., Wu, Q., Yuan, L.-P., Chen, Z.-W., Luo, M.-H., Zeng, H.-O., He, D.-L., & Mo, C.-J. (2017). Tumor necrosis factor receptor-associated factor (TRAF) 6 inhibition mitigates the pro-inflammatory roles and proliferation of rheumatoid arthritis fibroblast-like synoviocytes. *Cytokine*, 93, 26–33. <https://doi.org/10.1016/j.cyto.2017.05.001>
- Zhu, W., Meng, L., Jiang, C., He, X., Hou, W., Xu, P., Du, H., Holmdahl, R., & Lu, S. (2011). Arthritis is associated with T-cell-induced upregulation of Toll-like receptor 3 on synovial fibroblasts. *Arthritis Research & Therapy*, 13(3), R103. <https://doi.org/10.1186/ar3384>

ANNEX A

Inflammation

Initial conditions of inputs	Module behaviour	Model behaviour
No condition	 <p>Inflammation OFF</p>	 <p>Inflammation OFF</p>
TNF OFF	 <p>Inflammation OFF</p>	 <p>Inflammation OFF</p>





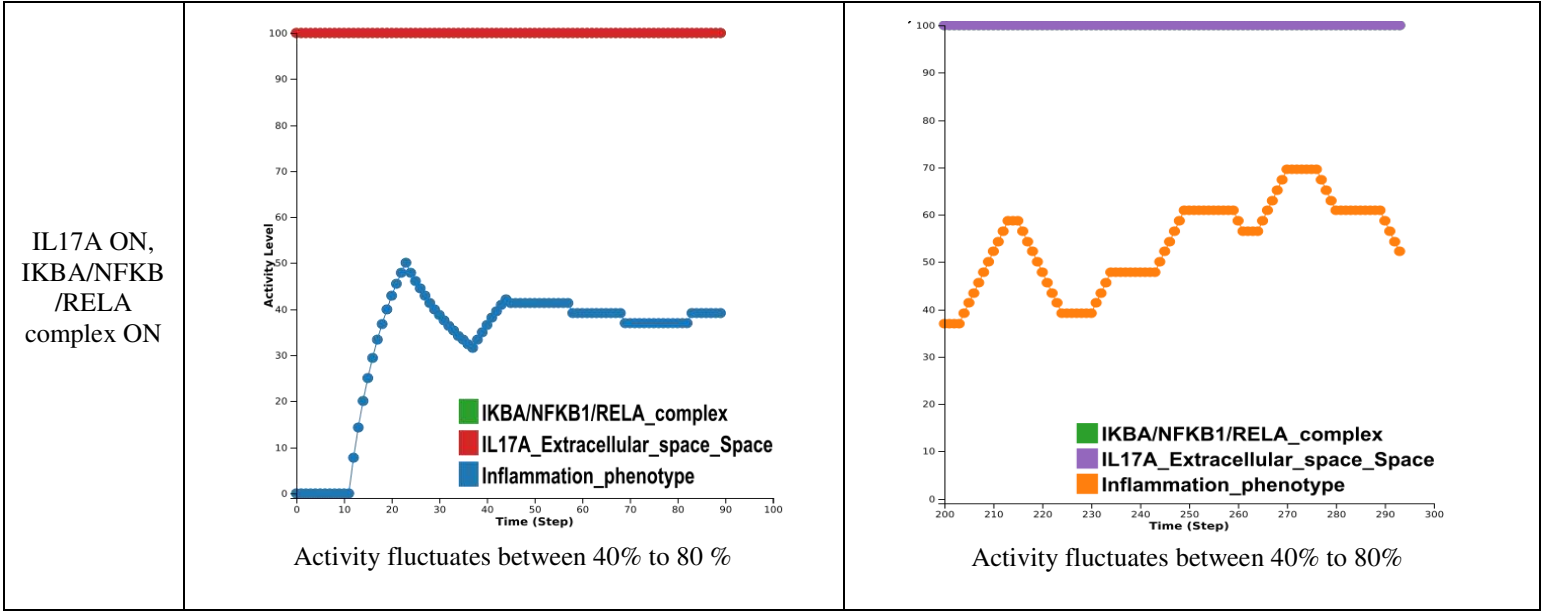
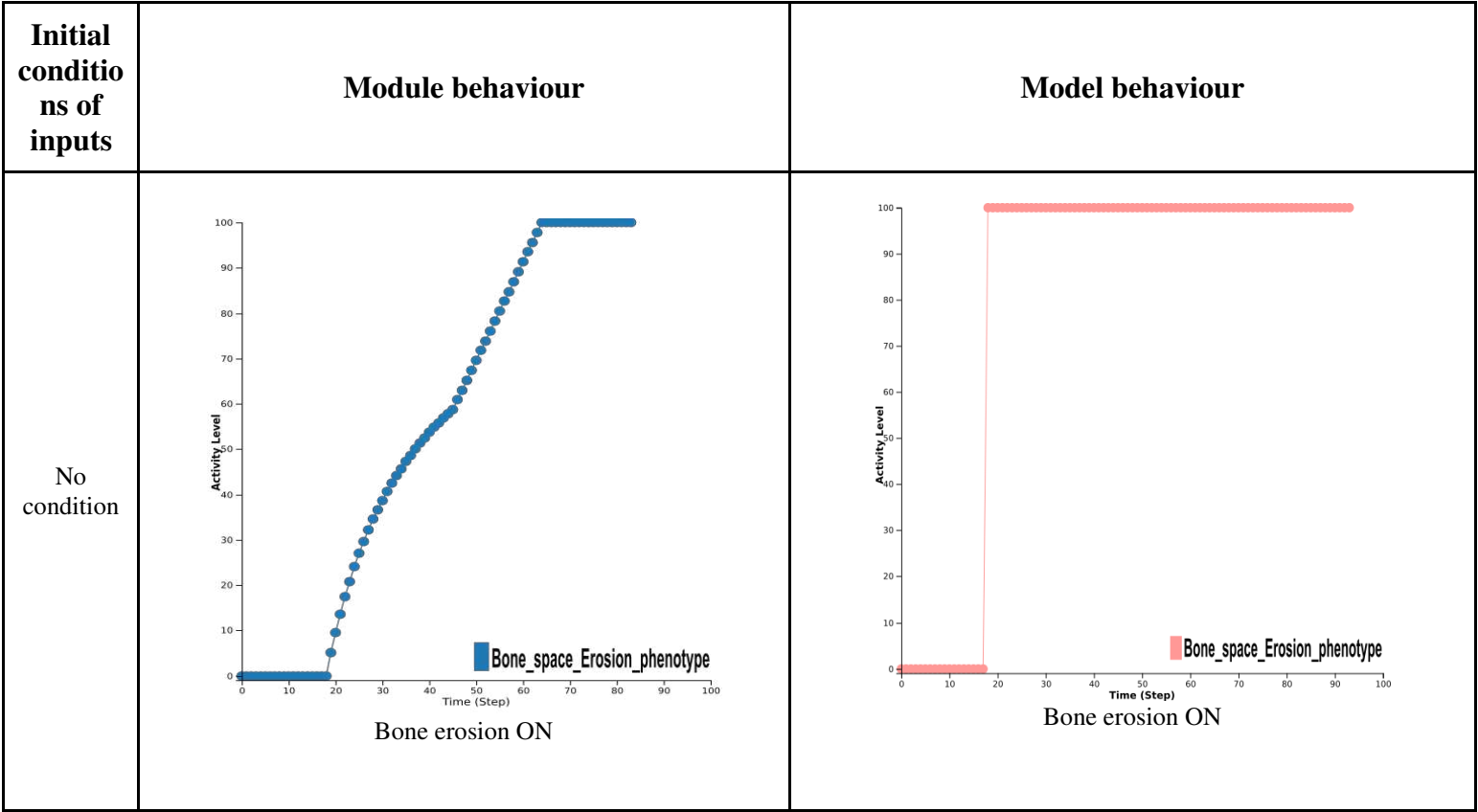
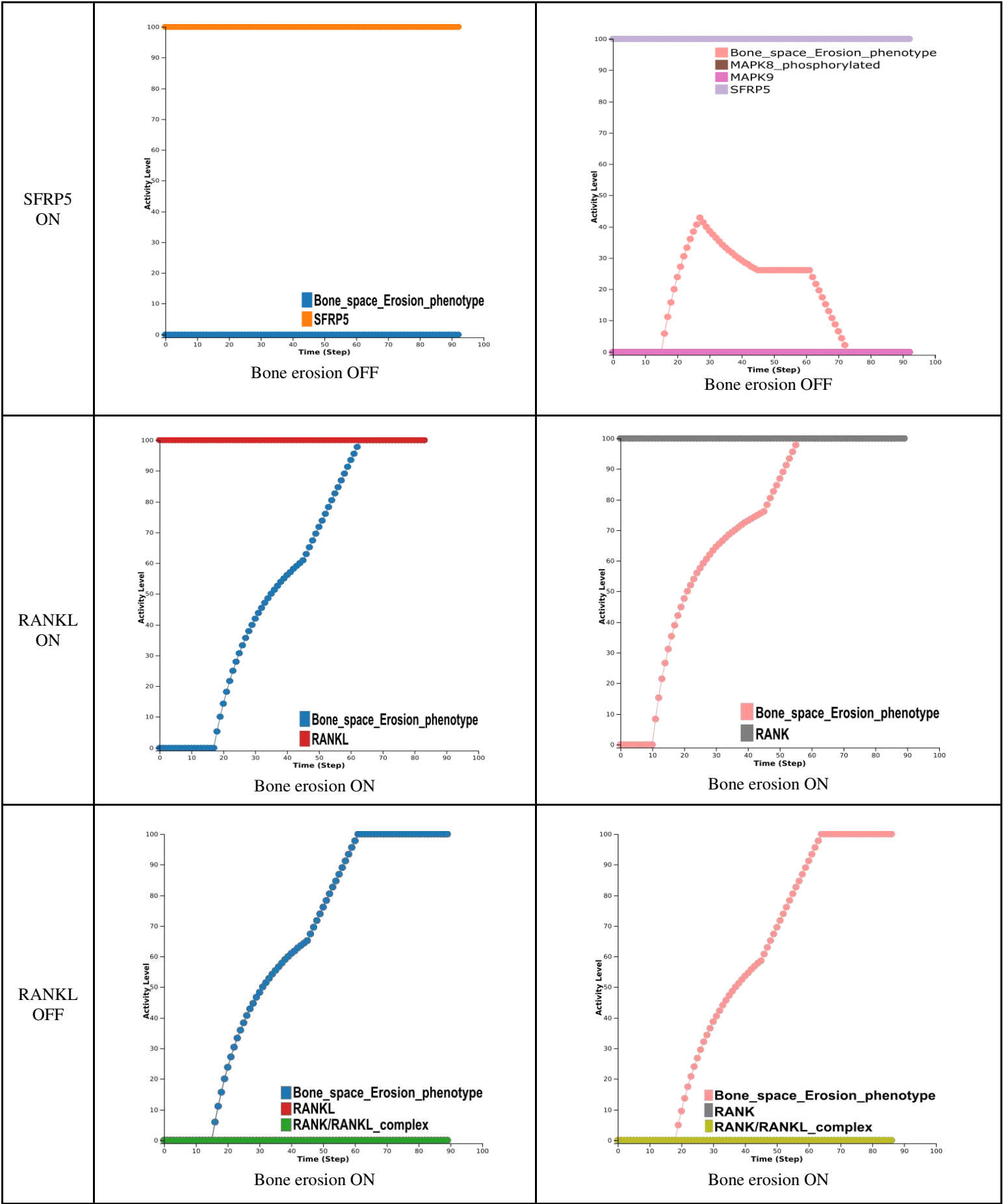


Figure A1: *In silico* simulation of the Inflammation module and the merged model.

Bone Erosion





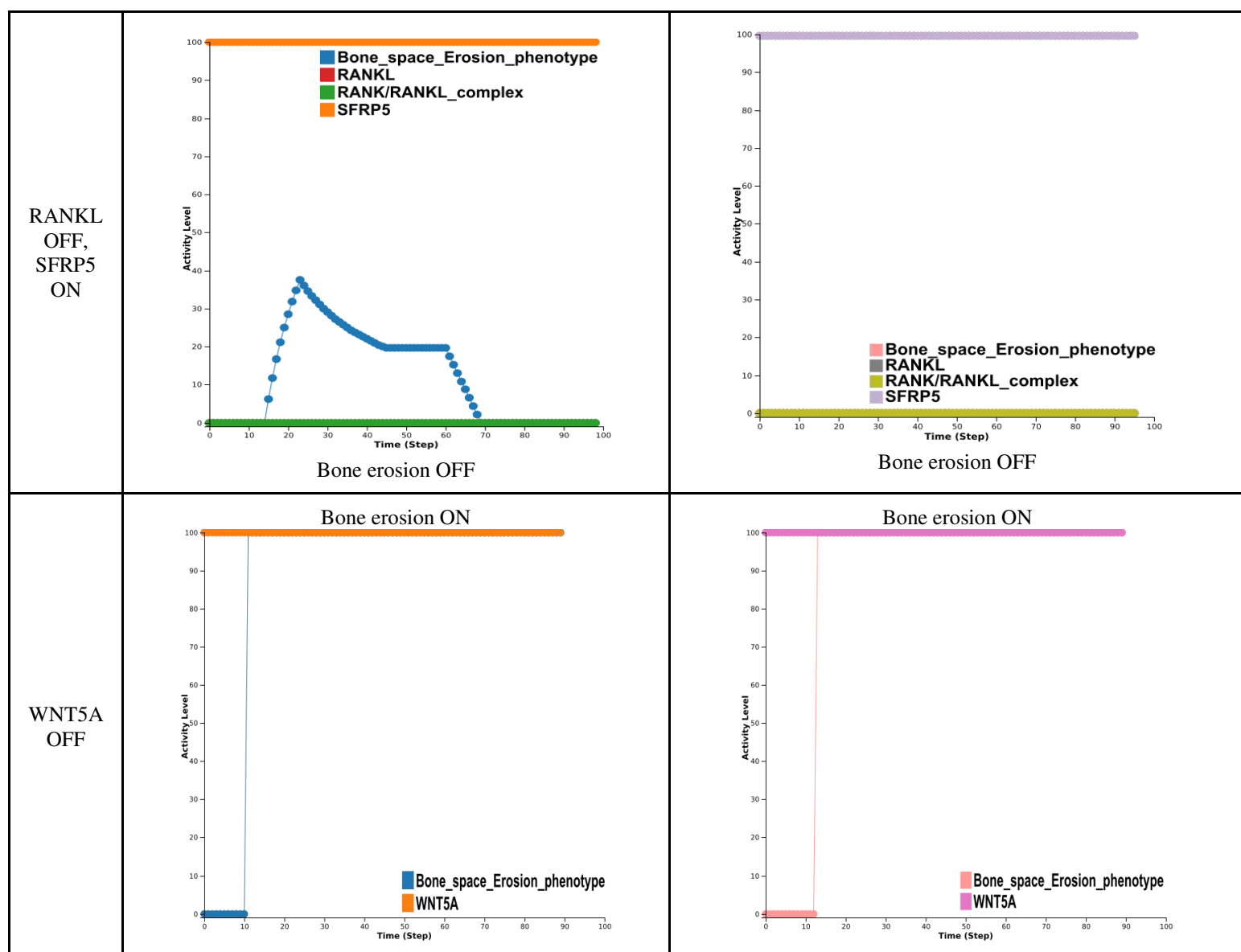
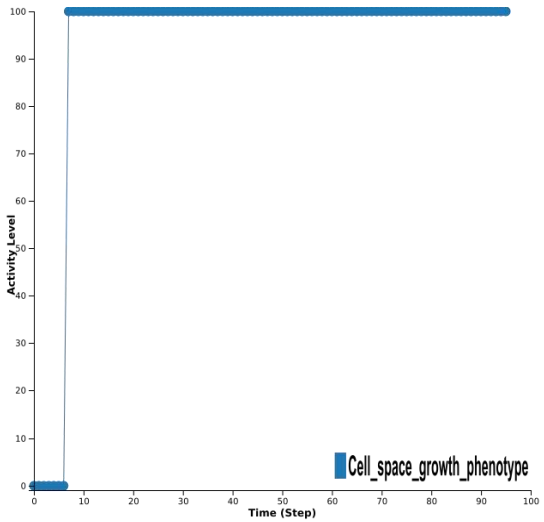
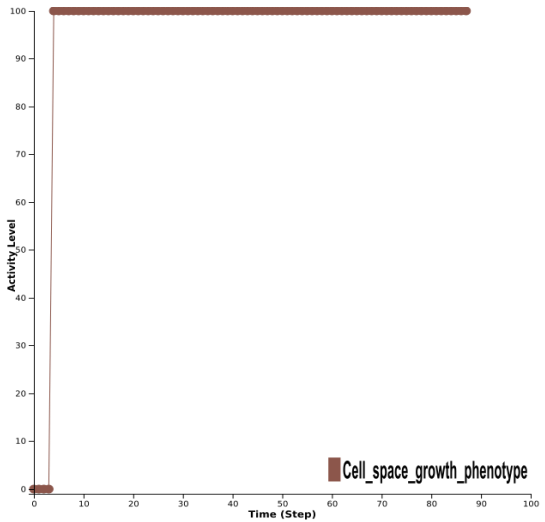
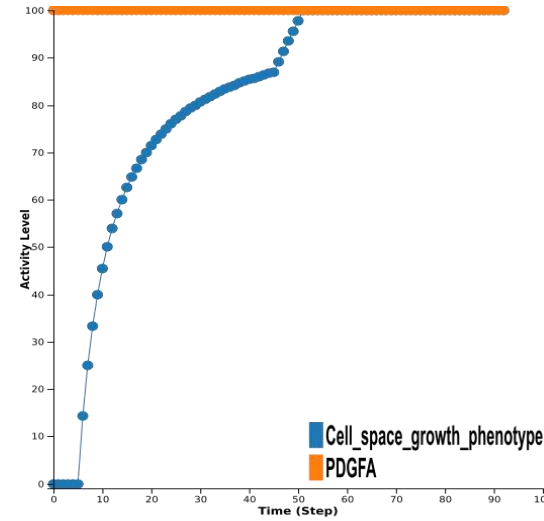
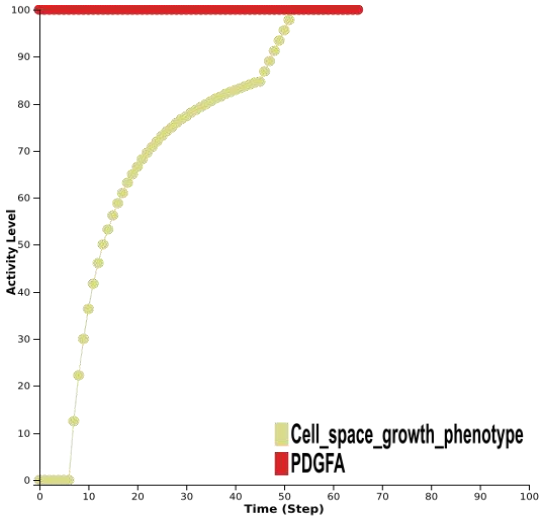
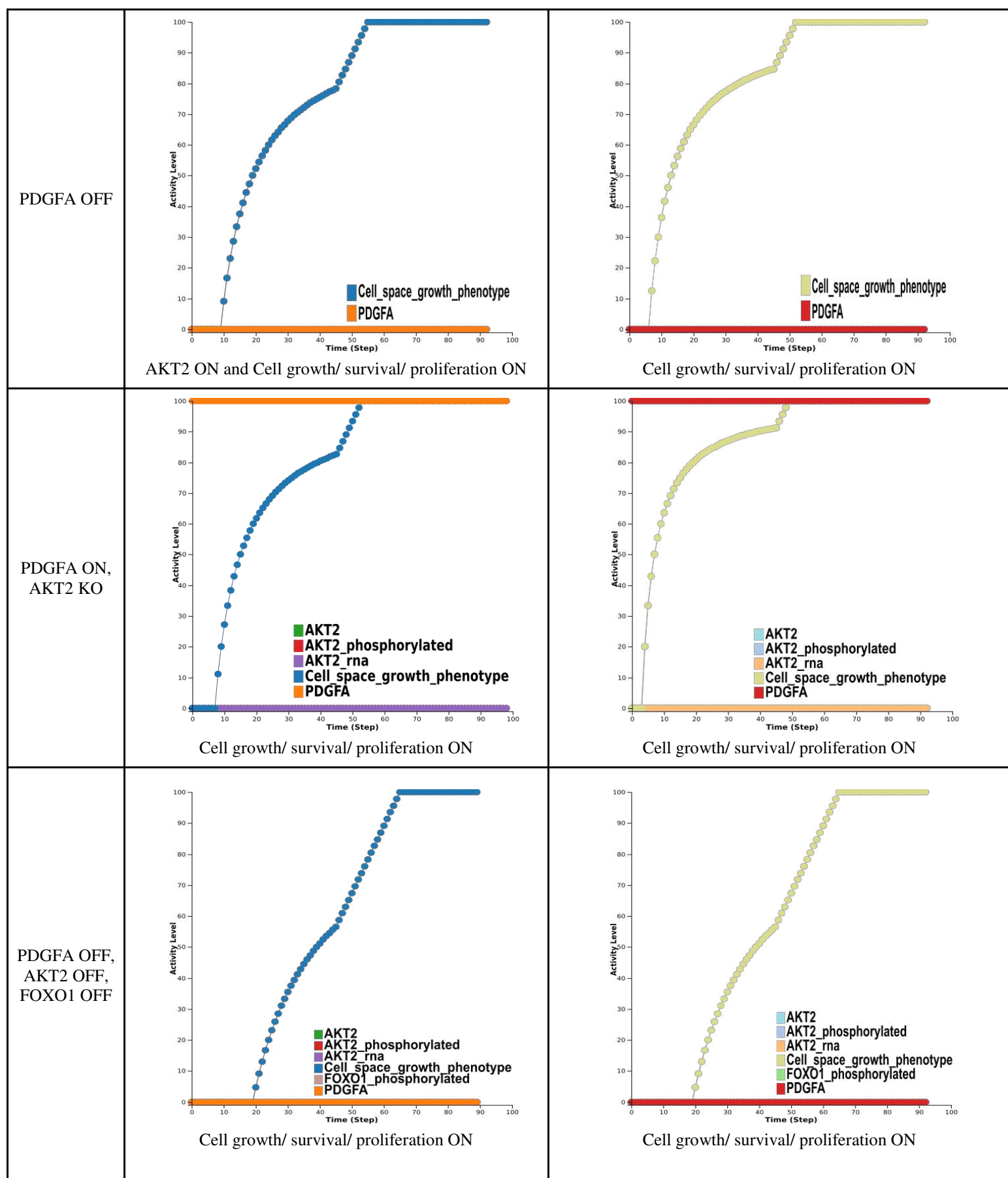


Figure A2: *In silico* simulations of the Bone Erosion module and the merged model.

Cell growth/survival/proliferation

Initial conditions of inputs	Module behaviour	Model behaviour
No condition	 <p>Cell growth/ survival/ proliferation ON</p>	 <p>Cell growth/ survival/ proliferation ON</p>
PDGFA ON	 <p>Cell growth/ survival/ proliferation ON</p>	 <p>Cell growth/ survival/ proliferation ON</p>



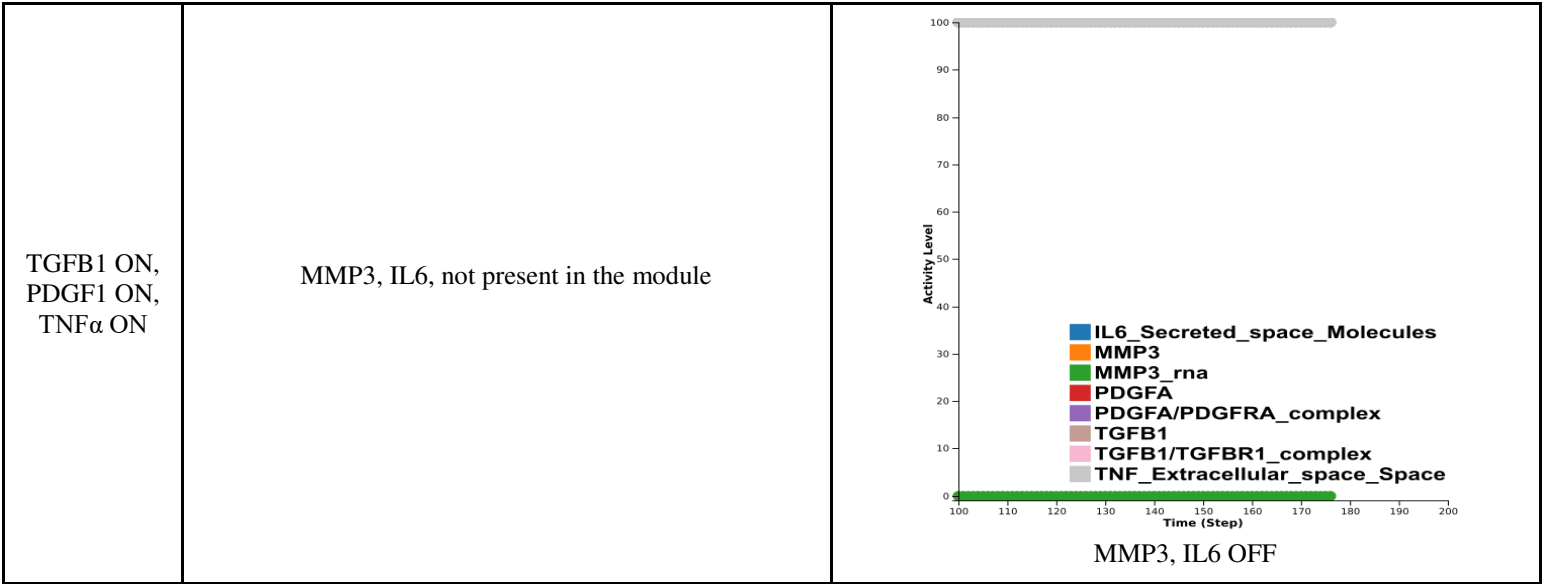
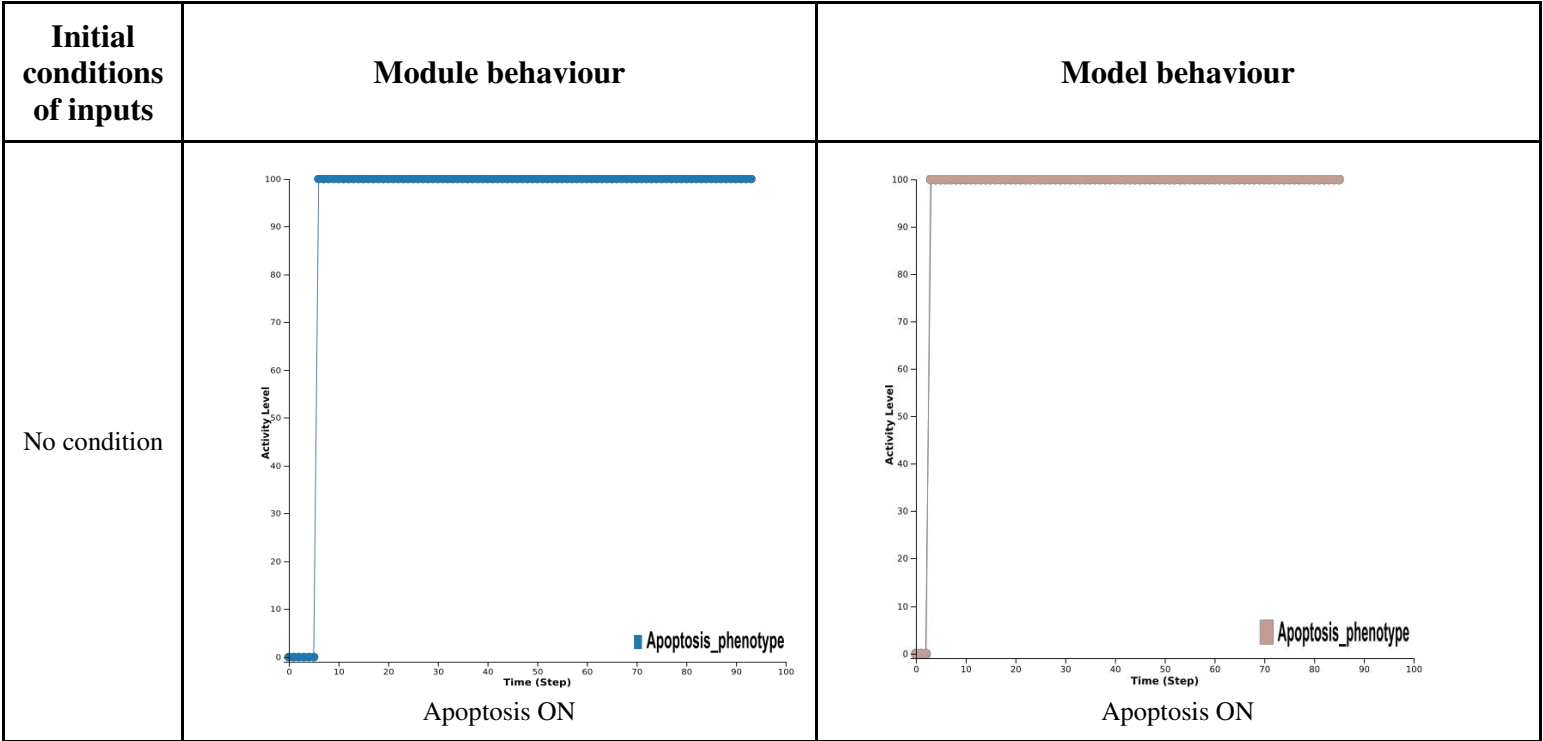


Figure A3: *In silico* simulation of Cell Growth module and merged model.

Apoptosis



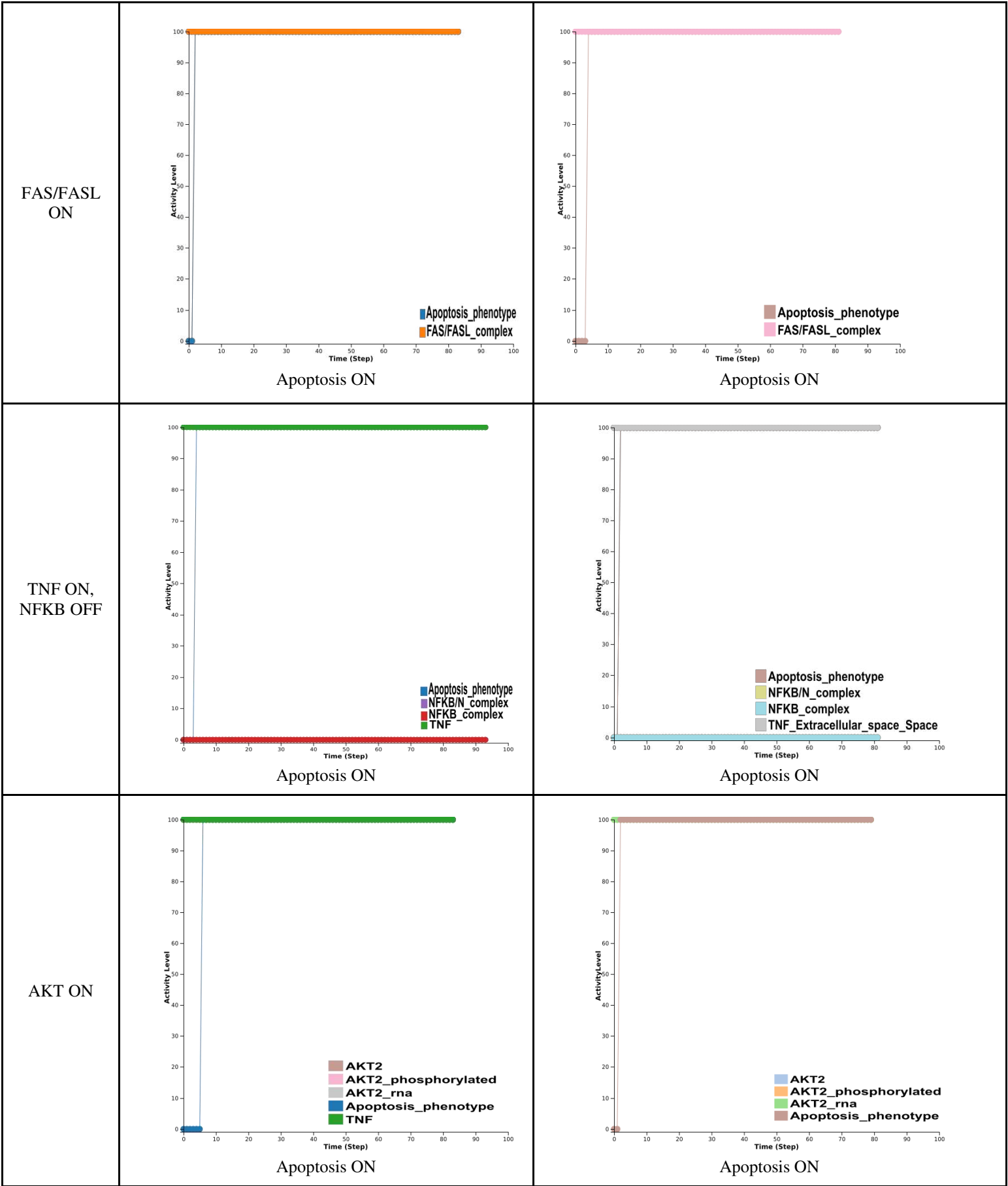
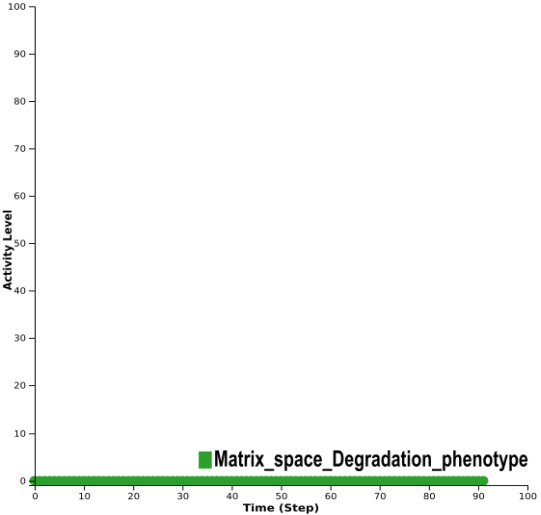
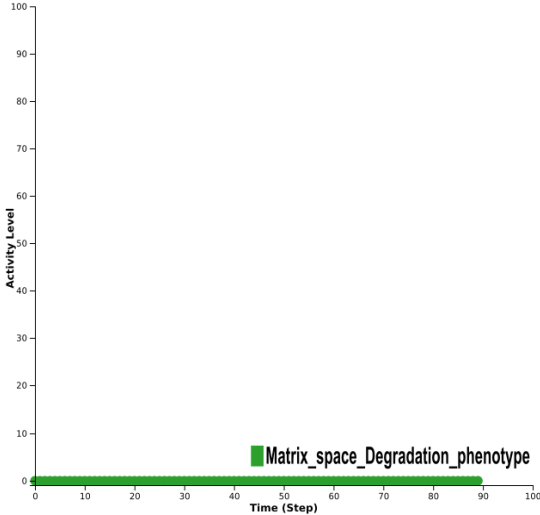
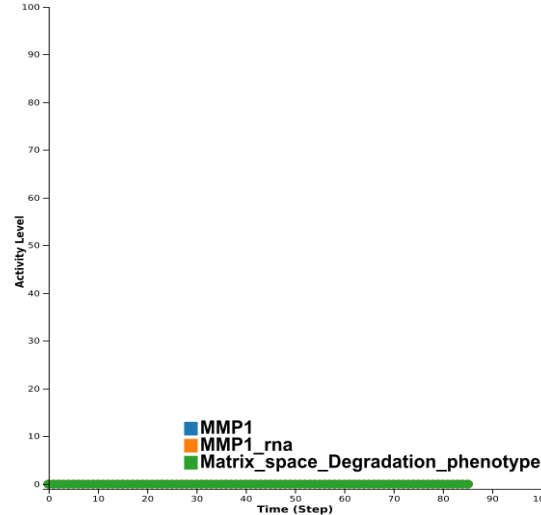
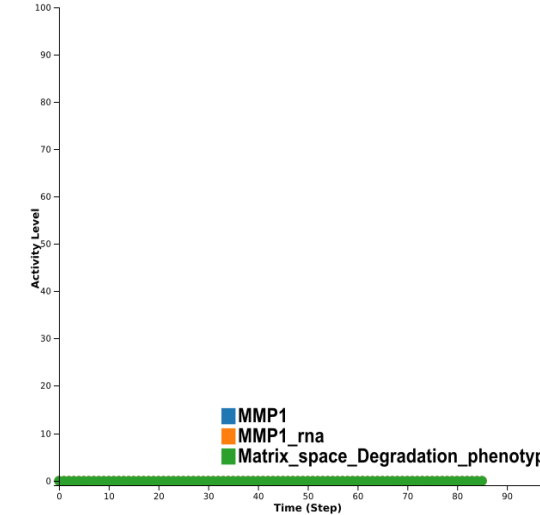


Figure A4: *In silico* simulation of the Apoptosis module and the merged model.

Matrix degradation

Initial conditions of inputs	Module behaviour	Model behaviour
No condition	 <p>Matrix degradation OFF</p>	 <p>Matrix degradation OFF</p>
MMP1 OFF	 <p>Matrix degradation OFF</p>	 <p>Matrix degradation OFF</p>

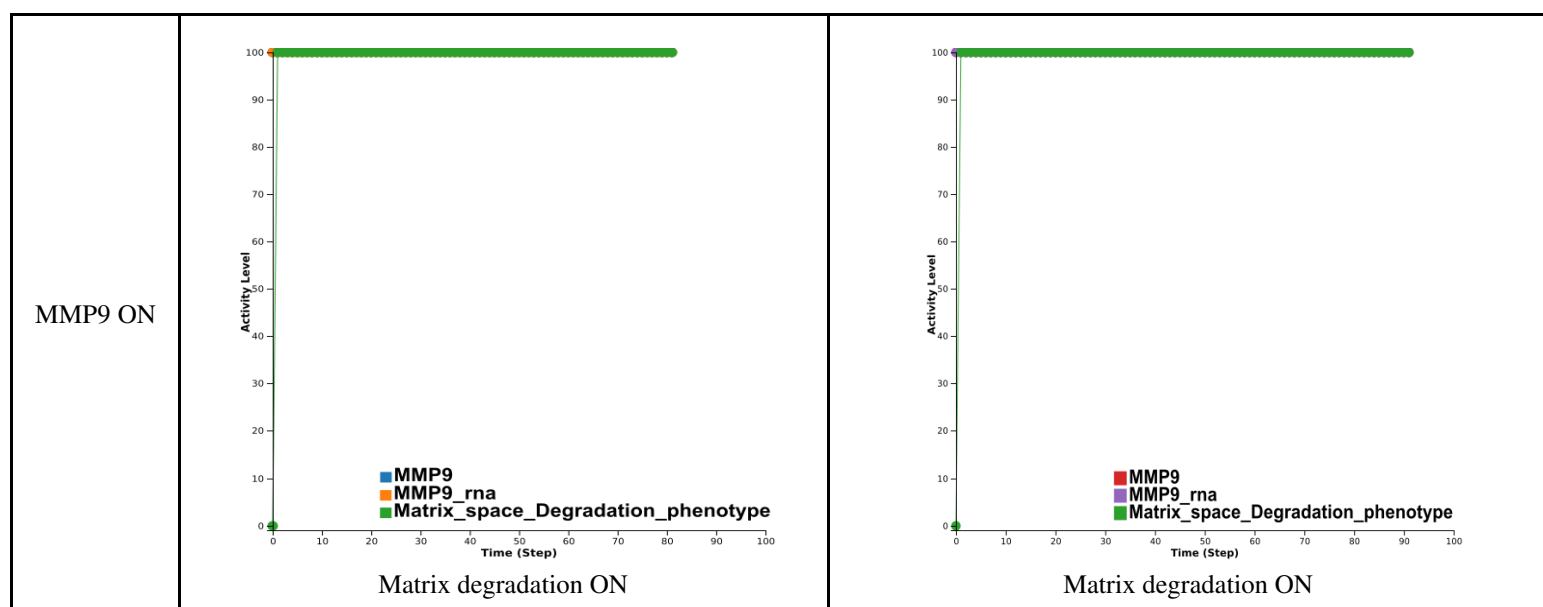


Figure A5. *In silico* simulation of the Matrix degradation module and the merged model.

ANNEX B

All used data, executable models, and attractors are available from the following link.

<https://github.com/vidisha007/RA-map-and-boolean-modeling>

ANNEX C

Publications



Published in final edited form as:

Genom Comput Biol. 2018 ; 4(1): . doi:10.18547/gcb.2018.vol4.iss1.e100050.

Computational Systems Biology Approach for the Study of Rheumatoid Arthritis: From a Molecular Map to a Dynamical Model

Vidisha Singh¹, Marek Ostaszewski², George D. Kalliolias³, Gilles Chiocchia⁴, Robert Olaso⁵, Elisabeth Petit-Teixeira¹, Tomáš Helikar⁶, and Anna Niarakis^{1,*}

¹GenHotel EA3886, Univ Evry, Université Paris-Saclay, 91025, Evry, France

²Luxembourg Centre for Systems Biomedicine, Université du Luxembourg, Esch-sur-Alzette, Luxembourg

³Arthritis and Tissue Degeneration Program, Hospital for Special Surgery, New York, USA; Department of Medicine, Weill Cornell Medical College, New York City, USA

⁴Faculty of Health Sciences Simone Veil, INSERM U1173, University of Versailles Saint-Quentin-en-Yvelines, Montigny-le-Bretonneux, France

⁵Centre National de Recherche en Génomique Humaine (CNRGH), CEA, Evry, France

⁶Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, NE, USA

SUMMARY

In this work we present a systematic effort to summarize current biological pathway knowledge concerning Rheumatoid Arthritis (RA). We are constructing a detailed molecular map based on exhaustive literature scanning, strict curation criteria, re-evaluation of previously published attempts and most importantly experts' advice. The RA map will be web-published in the coming months in the form of an interactive map, using the MINERVA platform, allowing for easy access, navigation and search of all molecular pathways implicated in RA, serving thus, as an on line knowledgebase for the disease. Moreover the map could be used as a template for Omics data visualization offering a first insight about the pathways affected in different experimental datasets. The second goal of the project is a dynamical study focused on synovial fibroblasts' behavior under different initial conditions specific to RA, as recent studies have shown that synovial fibroblasts play a crucial role in driving the persistent, destructive characteristics of the disease. Leaning on the RA knowledgebase and using the web platform Cell Collective, we are currently building a Boolean large scale dynamical model for the study of RA fibroblasts' activation.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

*Correspondence: anna.niaraki@univ-evry.fr.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Keywords

Complex human disease; Rheumatoid arthritis; Computational systems biology; Interactive molecular map; Signaling network; Dynamical modelling

Protein-protein interactions are a major driving force behind most biological processes. They play a pivotal role in intra- and extra-cellular functions, and especially in the propagation of signals and cellular regulation. Signal transduction is a fundamental process for the communication of the cell with its environment, comprising several interacting receptors, proteins, enzymes, second messengers and transcription factors. Disruption and dysregulation of these complex molecular and signaling networks can lead to disease. Therefore, the mapping and accurate representation of pathways implicated is a primary but essential step for elucidating the mechanisms underlying disease pathogenesis.

The release of various molecular maps dealing with obesity [1], gastrin and cholecystokinin receptors signaling [2], FcεRI receptor signaling in allergy [3], MAPKs [4], mTOR signaling [5] to name a few, corroborates the fact that pathway assembly in the form of network is gaining ground in systems biology. As more scientists invest time and effort to construct large molecular networks, there is an increasing need for practical guidelines and a standardized framework. Toward this direction, initiatives have emerged, such as The Cancer Cell Map Initiative [6], the Atlas of Cancer Signaling Networks (<http://acsn.curie.fr>) concerning cancer, and the Disease Maps Project (<http://disease-maps.org>), an open, large-scale community effort that consists of a network of groups working together for developing best practices, standards and tools in order to better represent disease mechanisms.

However, as all living systems are dynamic in nature, static representations of molecular networks can provide useful but relatively limited understanding. A dynamical study can reveal information about the system's behavior under different conditions by *in silico* simulations, perturbations, hypotheses testing and predictions. Quantitative kinetic modelling approaches using differential or stochastic equations can provide a detailed analysis of a network's dynamics, but the large number of parameters required make them less appropriate for large scale signaling networks. In order to address the lack of kinetic data, discrete logical modelling can be used as an alternative way to study the system's qualitative dynamic behavior [7, 8].

In this work we present a systematic effort to summarize current biological pathway knowledge concerning Rheumatoid Arthritis (RA), a multifactorial autoimmune disease that causes chronic inflammation of the synovial joints with an etiology that still remains unclear. With the use of the software CellDesigner [9], we are constructing a detailed molecular map based on exhaustive literature scanning, strict curation criteria, re-evaluation of previously published attempts [10] and most importantly experts' advice (Figure 1).

In 2010 Wu *et al.* published a detailed molecular map concerning rheumatoid arthritis using the software CellDesigner. We decided to use this map as a basis, and expand. The map has been updated with information published after 2010 by exhaustive manual curation and the help of data mining tools. Only experimentally validated interactions in at least two peer

reviewed scientific publications are kept. Due to the fact that the initial map was based on high throughput gene expression data from 28 studies and interactions inferred from KEGG database, all nodes and interactions are re-evaluated carefully in an effort to limit false positives. When validation with small scale experiments is not possible, we keep nodes that appear in at least two different high throughput studies. Detailed annotation including PubMed IDs and HUGO names is also added in the MIRIAM section of the CellDesigner file. As far as context representation and overall structure of the map, expert's advice has been taken into account along with an effort to comply with SBGN standards.

The RA map will be web-published in the coming months (a full length manuscript is under preparation) in the form of an interactive map, using the platform MINERVA [11], allowing for easy access, navigation and search of all molecular pathways implicated in RA, serving thus, as an on line knowledge base for the disease. The user will have access to all literature used, with detailed annotations for every component and reaction, including PubMed IDs, and a list of identifiers such as Uniprot, EntrezGene, Ensembl, HGNC and RefSeq. As the map is constructed using information from various experimental studies, the user will also be able to opt for visualization of data with specific cell origin, highlighting cell-specific sub-networks within the global one. Moreover, the user will have the possibility to spot all known drug targets, and the corresponding drugs up to date for RA. Detailed view of an element will allow the search for drugs, chemicals and miRNAs targeting this particular element. Additionally, user-provided omic datasets could be displayed as overlay, giving a first estimation of affected pathways and components. Lastly, the map will provide feedback about the unmapped molecules from the dataset, allowing for better understanding of the experimental results and for further development of the map's contents. We have used public datasets from proteomic and transcriptomic studies [12–14] to demonstrate how the map can be used as a template for separate or simultaneous visualization of different experimental results. The map will also be used for the mapping of in-house data concerning the transcriptome analysis of ten individuals that developed RA (measurements before the onset of the disease and early after) (Teixeira *et al.*, under preparation).

The RA map so far includes information derived from more than 100 scientific papers. It has six distinct compartments, namely extracellular space (with extracellular proteins), plasma membrane (with membrane receptors and ligand proteins), cytoplasm (with proteins, miRNAs, small molecules and the sub-compartments of mitochondrion, Golgi apparatus and endoplasmic reticulum), nucleus (with genes, RNAs and transcription factors), a compartment for the secreted molecules and a phenotype compartment including more than ten cellular fates. It comprises more than 400 components and a total of 324 reactions. Each component and reaction in the map is referenced with at least two PubMed IDs or database identifiers if inferred from a specific database.

Topological analysis of the RA map using the software Cytoscape [15] and relevant plugins reveals unconnected or loosely connected parts that reflect our fragmented knowledge about physical and/or genetic interactions, posing thus obstacles in the subsequent derivation of a reliable dynamical model. To improve connectivity we use dedicated PPI databases (through <http://www.imexconsortium.org>), pathway databases (e.g. KEGG, SIGNOR or REACTOME) and the commercial software Ingenuity Pathway Analysis (IPA, <http://>

www.ingenuity.com) in order to investigate potential co-players of the proteins of interest. For the time being, we do not make use of simulated/computationally inferred interactions or interactions inferred from other species (i.e. mice), restricting our search to experimentally validated data of human origin.

Characteristic features of RA include synovial inflammation that can lead to bone erosion and permanent deformity. It is broadly recognized that in RA, synovial inflammation results from complex interactions between haematopoietic and stromal cells. Recent studies have shown that RA synovial fibroblasts play a crucial role in driving the persistent, destructive characteristics of the disease [16].

The second scope of the project is to model synovial fibroblasts behavior under different initial conditions specific to RA, in order to see if we could influence the cellular fate (e.g. enhancing an apoptotic phenotype) or understand what could lead to patient's resistance to a certain drug and how to overcome it (e.g. presence of rescue pathways, complex feedback mechanisms).

In general, pathway representation and modelling can be seen as two separate tasks with different primary objectives. The first is to draw an accurate, comprehensive diagram depicting current biological knowledge while the second is to study the emergent behavior of the system under different conditions. However, a detailed, fully annotated molecular map works as an excellent scaffold for the building of a regulatory graph and the subsequent derivation of the logical model. This process, that involves many iterations, obliges one to look meticulously into the mapped pathways, spotting potentially problematic or ambiguous aspects of the map. Model simulations can also reveal inconsistencies concerning the global behavior, advocating the necessity for further revisions and refinements. Leaning on the RA knowledge base and using the web platform Cell Collective [17], we are currently building a Boolean dynamical model for the study of RA fibroblasts' activation. The model is based on a previously published, more generic model on fibroblasts [18] that is being modified accordingly in order to be RA specific.

In Boolean formalism, nodes represent regulatory components (proteins, complexes, transcription factors, etc.) and arcs represent their interactions. Each regulatory component is associated with a Boolean variable (taking the values 0 or 1) denoting its qualitative concentration or level of activity (0 for absent or inactive, 1 for present or active). The future state of each node depends on the states of its upstream regulators and is defined by a Boolean function, expressed in the form of a rule using the logical operators AND, OR and NOT.

The tuning of the model includes testing against published data and appropriate modifications of logical rules and/or addition/deletion of interactions/components until it is able to reproduce well established input-output relations (global behavior). This process will inevitably lead back to the re-evaluation of the molecular map and further discussions with experts until all issues are resolved in a biologically sound way. The model will then be used to systematically test different initial conditions and stimuli (presence or absence of different cytokines and growth factors, and their combinations). The aim is to predict the system's

response to single or combined perturbations, and identify novel targets for pharmacological intervention.

The Boolean model will be made publicly available in Cell Collective for further contributions, simulations, and analyses by the research community, hopefully within 2018. The web based platform Cell Collective allows real time simulations without the need for software installation making the model more accessible to a wider audience. Moreover, the platform supports annotation, so the user can have simultaneous access to the model description, the rules and the literature used for the rules' inference.

Lastly, the resulting logical model for RA fibroblasts could be further analyzed with the software GINsim [19] and also serve as a template for the derivation of a continuous model using the software MaBoSS [20] allowing the computation of phenotype probabilities.

References

1. Jagannadham J, Jaiswal HK, Agrawal S, Rawal K. Comprehensive Map of Molecules Implicated in Obesity. *PLoS ONE*. 2016; 11(2):e0146759.doi: 10.1371/journal.pone.0146759 [PubMed: 26886906]
2. Tripathi S, Flobak Å, Chawla K, Baudot A, Bruland T, Thommesen L, et al. The gastrin and cholecystokinin receptors mediated signaling network: a scaffold for data analysis and new hypotheses on regulatory mechanisms. *BMC Systems Biology*. 2015; 9:40.doi: 10.1186/s12918-015-0181-z [PubMed: 26205660]
3. Niarakis A, Bounab Y, Grieco L, Roncagalli R, Hesse AM, G J, et al. Computational modeling of the main signaling pathways involved in mast cell activation. *Curr Top Microbiol Immunol*. 2014; 382:69–93. DOI: 10.1007/978-3-319-07911-04 [PubMed: 25116096]
4. Grieco L, Calzone L, Bernard-Pierrot I, Radvanyi F, Kahn-Perlès B, Thieffry D. Integrative Modelling of the Influence of MAPK Network on Cancer Cell Fate Decision. *PLoS Computational Biology*. 2013; 9(10):e1003286.doi: 10.1371/journal.pcbi.1003286 [PubMed: 24250280]
5. Caron E, Ghosh S, Matsuoka Y, Ashton-Beaucage D, Therrien M, Lemieux S, et al. A comprehensive map of the mTOR signaling network. *Molecular Systems Biology*. 2010; 6:453.doi: 10.1038/msb.2010.108 [PubMed: 21179025]
6. Krogan NJ, Lippman S, Agard DA, Ashworth A, Ideker T. The Cancer Cell Map Initiative: Defining the Hallmark Networks of Cancer. *Molecular Cell*. 2015; 58(4):690–698. DOI: 10.1016/j.molcel.2015.05.008 [PubMed: 26000852]
7. Wynn ML, Consul N, Merajver SD, Schnell S. Logic-based models in systems biology: a predictive and parameter-free network analysis method. *Integrative Biology*. 2012; 4(11)doi: 10.1039/c2ib20193c
8. Abou-Jaoudé W, Traynard P, Monteiro PT, Saez-Rodriguez J, Helikar T, Thieffry D, et al. Logical Modeling and Dynamical Analysis of Cellular Networks. *Frontiers in Genetics*. 2016; 7:94.doi: 10.3389/fgene.2016.00094 [PubMed: 27303434]
9. Funahashi A, Morohashi M, Kitano H. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico*. 2003; 1(5):159–62. DOI: 10.1016/S1478-5382(03)02370-9
10. Wu G, Zhu L, Dent JE, Nardini C. A Comprehensive Molecular Interaction Map for Rheumatoid Arthritis. *PLOS ONE*. 2010; 5(4):1–16. 04. DOI: 10.1371/journal.pone.0010137
11. Gawron P, Ostaszewski M, Satagopam V, Gebel S, Mazein A, Kuzma M, et al. MINERVA—a platform for visualization and curation of molecular interaction network. *npj Systems Biology and Applications*. 2016; 2:16020.doi: 10.1038/npjsba.2016.20 [PubMed: 28725475]
12. Dasuri K, Antonovici M, Chen K, Wong K, Standing K, Ens W, et al. The synovial proteome: analysis of fibroblast-like synoviocytes. *Arthritis Research and Therapy*. 2004; 6(2):R161.doi: 10.1186/ar1153 [PubMed: 15059280]

13. Teixeira VH, Olaso R, Martin-Magniette ML, Lasbleiz S, Jacq L, Oliveira CR, et al. Transcriptome Analysis Describing New Immunity and Defense Genes in Peripheral Blood Mononuclear Cells of Rheumatoid Arthritis Patients. *PLoS ONE*. 2009; 4(8):0006803.doi: 10.1371/journal.pone.0006803
14. Heruth DP, Gibson M, Grigoryev DN, Zhang LQ, Ye SQ. RNA-seq analysis of synovial fibroblasts brings new insights into rheumatoid arthritis. *Cell and bioscience*. 2012; 2(1):43.doi: 10.1186/2045-3701-2-43 [PubMed: 23259760]
15. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*. 2003; 13(11):2498–2504. DOI: 10.1101/gr.1239303 [PubMed: 14597658]
16. Juarez M, Filer A, Buckley C. Fibroblasts as therapeutic targets in rheumatoid arthritis and cancer. *Swiss Medical Weekly*. 2012; 142:w13529.doi: 10.4414/smw.2012.13529 [PubMed: 22367980]
17. Helikar T, Kowal B, McClenathan S, Bruckner M, Rowley T, Madrahimon A, et al. The Cell Collective: Toward an open and collaborative approach to systems biology. *BMC Systems Biology*. 2012; 6(96)doi: 10.1186/1752-0509-6-96
18. Helikar T, Konvalina J, Heidel J, Rogers JA. Emergent decision-making in biological signal transduction networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105(6):1913–1918. DOI: 10.1073/pnas.0705088105 [PubMed: 18250321]
19. Chaouiya, C., Naldi, A., Thieffry, D., Logical Modelling of Gene Regulatory Networks with GINsim. *Bacterial Molecular Networks: Methods and Protocols*. New York, NY: Springer New York; 2012. p. 463-479.
20. Stoll G, Caron B, Viara E, Dugourd A, Zinovyev A, Naldi A, et al. MaBoSS 2.0: an environment for stochastic Boolean modeling. *Bioinformatics*. 2017; 33(14):2226–2228. DOI: 10.1093/bioinformatics/btx123 [PubMed: 28881959]

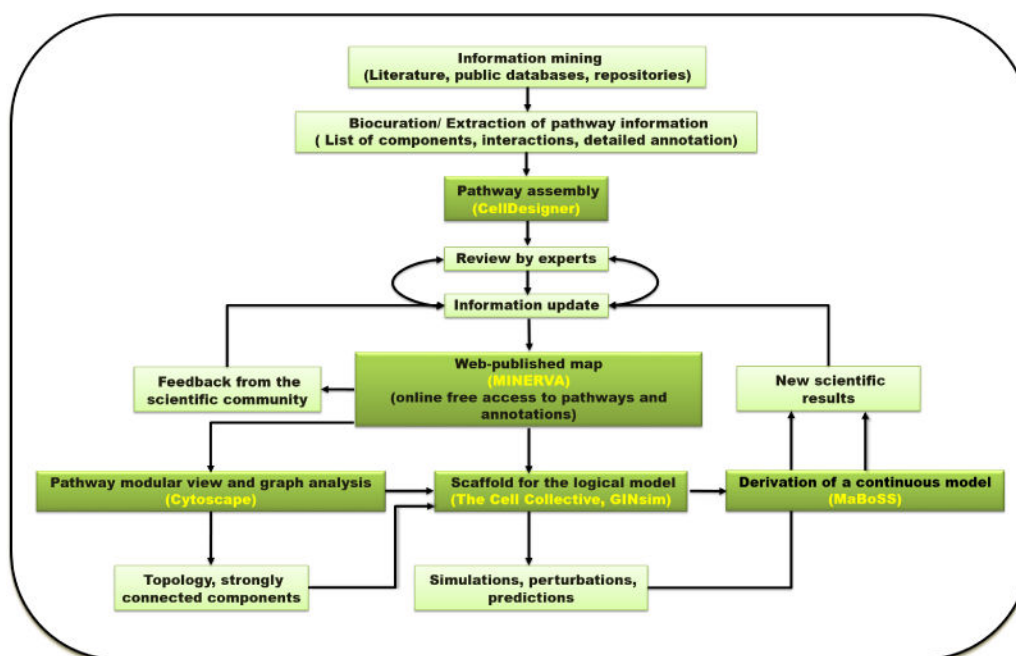


Figure 1. Data integration workflow

The building of a logical model is an iterative multistep process. The assembly of a molecular map comprising biological pathways of interest and integrating information from literature and public databases could serve as the first step. Experts' feedback assures the quality of the map and the accuracy of the knowledge represented, along with strict curation criteria and standards for the graphical representation. Web publication facilitates community feedback and transforms the map in a powerful data analysis and visualization tool. The network can be further exploited using graph analysis tools to identify important nodes and pathways, or it can serve as a scaffold for dynamical models allowing simulations. Interesting predictions can then be experimentally tested, contributing to the validation and refinement of the map. Regular revisions are also necessary to ensure the incorporation of novel data (Figure adapted from Niarakis *et al.*, 2014 [3]).



Original article

RA-map: building a state-of-the-art interactive knowledge base for rheumatoid arthritis

Vidisha Singh¹, George D. Kalliolias^{2,3}, Marek Ostaszewski⁴,
Maëva Veyssiere¹, Eleftherios Pilalis⁵, Piotr Gawron⁴,
Alexander Mazein⁴, Eric Bonnet⁶, Elisabeth Petit-Teixeira¹ and
Anna Niarakis^{1,*}

¹Laboratoire Européen de Recherche pour la Polyarthrite Rhumatoïde - Genhotel, Univ Evry, Université Paris-Saclay, 2, rue Gaston Crémieux, 91057 EVRY-GENOPOLE cedex, Evry, France, ²Arthritis and Tissue Degeneration Program, Hospital for Special Surgery, 535 East 70th Street, New York, NY 10021, USA, ³Weill Cornell Medical Center, Weill Department of Medicine, 525 East 68th Street, New York, NY 10065, USA, ⁴Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 Avenue du Swing, L-4367 Belvaux, Luxembourg, ⁵eNIOS Applications P.C., R&D department, Alexandrou Pantou 25, 17671, Kallithea-Athens, Greece and ⁶Centre National de Recherche en Génomique Humaine (CNRGH), CEA, 2 rue Gaston Crémieux, CP5706 91057 EVRY-GENOPOLE cedex, Evry, France

*Corresponding author: Email: anna.niaraki@univ-evry.fr

Citation details: Singh, V., Kalliolias, G. D., Ostaszewski, M. *et al.* RA-map: building a state-of-the-art interactive knowledge base for rheumatoid arthritis. *Database* (2020) Vol. 2020: article ID baaa017; doi:10.1093/database/baaa017

Received 7 December 2019; Revised 21 January 2020; Accepted 13 February 2020

Abstract

Rheumatoid arthritis (RA) is a progressive, inflammatory autoimmune disease of unknown aetiology. The complex mechanism of aetiopathogenesis, progress and chronicity of the disease involves genetic, epigenetic and environmental factors. To understand the molecular mechanisms underlying disease phenotypes, one has to place implicated factors in their functional context. However, integration and organization of such data in a systematic manner remains a challenging task. Molecular maps are widely used in biology to provide a useful and intuitive way of depicting a variety of biological processes and disease mechanisms. Recent large-scale collaborative efforts such as the Disease Maps Project demonstrate the utility of such maps as versatile tools to organize and formalize disease-specific knowledge in a comprehensive way, both human and machine-readable. We present a systematic effort to construct a fully annotated, expert validated, state-of-the-art knowledge base for RA in the form of a molecular map. The RA map illustrates molecular and signalling pathways implicated in the disease. Signal transduction is depicted from receptors to the nucleus using the Systems Biology Graphical Notation (SBGN) standard representation. High-quality manual curation, use of only human-specific studies and focus on small-scale experiments aim to limit false positives in the map. The state-of-the-art molecular map for RA, using information from 353 peer-reviewed scientific publications, comprises 506 species, 446 reactions and 8 phenotypes. The species in the map are classified to 303 proteins, 61 complexes, 106

genes, 106 RNA entities, 2 ions and 7 simple molecules. The RA map is available online at ramap.elixir-luxembourg.org as an open-access knowledge base allowing for easy navigation and search of molecular pathways implicated in the disease. Furthermore, the RA map can serve as a template for *omics* data visualization.

Introduction

Rheumatoid arthritis (RA) is a progressive inflammatory and autoimmune disease with unknown aetiology. It affects 0.5–1% of the world population, and disease characteristics involve synovial inflammation and hyperplasia, cartilage and bone destruction, production of autoantibodies like rheumatoid factor (RF) and anti-citrullinated protein (ACPA), and various systemic features such as cardiovascular, pulmonary, psychological and skeletal disorders (1). The pathogenesis of RA is a multistep process involving an intricate interplay between genetic, environmental and epigenetic mechanisms, a variety of intertwined signalling cascades and the expression of pro-inflammatory mediators (1, 2).

Systems Biology allows deciphering complex disease mechanisms by treating biological processes in living organisms as coordinated and interdependent events. Especially in human diseases, genes and proteins rarely act alone when affecting implicated cells, tissues or organs. To understand the molecular mechanisms underlying these phenotypes, one has to place the implicated biomolecules in their functional context and interconnect them. This way, a graphical representation of disease mechanisms is established and can be refined, validated and interpreted using the wealth of high-throughput biological data. Nevertheless, integration and organization of both graph and data in a systematic and standardized manner remains a challenge.

Molecular maps are widely used in biology to provide a useful and intuitive way of depicting a variety of biological processes and disease mechanisms. Examples of such maps include the gastrin and cholecystokinin receptor signalling (3), yeast stress response pathways (4), FcεRI receptor signalling in allergy (5), mitogen-activated protein kinase (MAPK) pathways (6), Parkinson's disease (7), Alzheimer's disease (8), influenza A virus (9), asthma (10), cancer (11) and RA (12). Recent large-scale collaborative efforts such as the Disease Maps Project (13, 14), demonstrate the utility of such maps as versatile tools to organize and formalize disease-specific knowledge in a comprehensive way, both human and machine-readable.

In this work, we present a systematic effort to construct a fully annotated, expert validated, state-of-the-art knowledge base for RA in the form of a molecular map. The RA map illustrates molecular and signalling pathways implicated in the disease. Signal transduction is depicted

from receptors to the nucleus in a systematic fashion using the Systems Biology Graphical Notation (SBGN) standard representation (15). High-quality manual curation, use of only human-specific studies and focus on small-scale experiments aim to limit false positives in the map. The RA map serves as an interactive knowledge base but also as a template for *omic* data visualization. *Omic* datasets can be superimposed on the map, pinpointing affected areas in different samples.

Furthermore, the map is a good starting point for the development of a computational model, providing an intermediate step between a conceptual, mechanistic graph and an executable mathematical model (12). The article comprises three parts. In the first part, we present the process of constructing the RA map, highlighting the most critical pathways. In the second part, we transform the RA map into a state-of-the-art interactive knowledge base for the disease, which interfaces with various databases for content annotation and enrichment analysis of experimental results. In the third part, we use bioinformatics tools such as BioInfoMiner (16) (<https://bioinforminer.com>) and Cytoscape (17) for the analysis of the RA map as a complex biological network, revealing topological and functional aspects of the map (Figure 1).

Methods

Construction of the RA map

CellDesigner (18) is a structured diagram editor for the creation of gene-regulatory and biochemical networks. Networks are drawn using the Process Description visual language of SBGN, and are stored using the Systems Biology Markup Language (SBML) (20), a standard for representing models of biochemical and gene-regulatory networks. In a CellDesigner diagram, nodes represent species like proteins, genes, complexes and other molecules, and the edges denote the interaction between the nodes, which can be activation, inhibition, catalysis and state transition among other possible interactions (21, 22). A comprehensive molecular interaction map for RA was published in 2010 (23) with information derived from high-throughput data combined with interaction data from the KEGG pathway database (24–27) (<http://www.genome.jp/kegg/pathway.html>). The researchers of this study used 28 published studies for the construction of the first RA

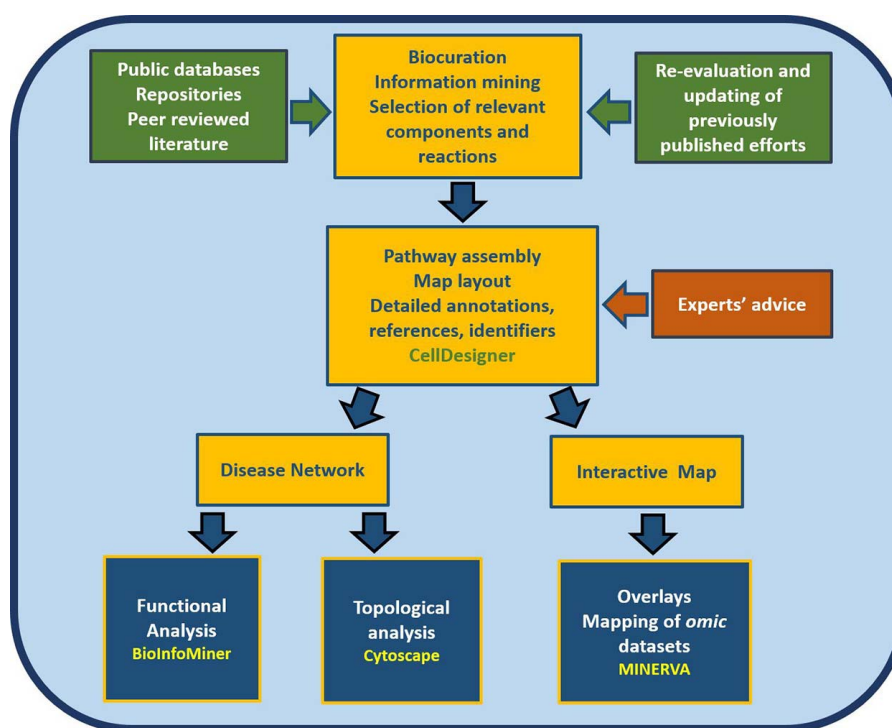


Figure 1. Workflow for the construction and use of the RA map. The assembly of the signalling and molecular pathways implicated in RA involves exhaustive manual curation and information mining from literature, public databases and repositories and the use of the software CellDesigner (18). The RA map contains mechanisms reported in the most recently published studies, after validation from RA experts. The map can be transformed into an online interactive knowledge base using the platform MINERVA (19). Functional enrichment and topological analysis is possible using the software BioInfoMiner (16) (<https://bioinforminer.com>) and Cytoscape (17), respectively.

map that included experiments performed in different cell types/tissues/fluids such as the peripheral blood mononuclear cells, synovial fibroblasts, macrophages, chondrocytes, synovial tissues, bone, blood, and synovial fluid (Figure S1, Table S1). We used this RA map as a basis and extended it to create a state of the art map for RA. However, apart from updates, the first map has been significantly modified. A systematic effort was made to create an SBGN-compliant map, the first to our knowledge. We also removed from the map many factors and reactions that were either not disease-specific or did not follow the curation criteria (discussed in section [Annotation and curation criteria](#)). The map was restructured to depict a cell layout. We grouped the receptors by category (growth factors, cytokines, chemokines, integrins and Toll-like receptors). For the updating, keywords like ‘rheumatoid arthritis’, ‘pathogenesis of rheumatoid arthritis’, ‘cytokines involved in rheumatoid arthritis’, ‘factors involved in rheumatoid arthritis’, ‘signalling pathways in rheumatoid arthritis’ were used to select relevant literature after 2010 (or older than 2010 that would correspond to small-scale experiments, in order to annotate nodes and reactions already present in the map) with emphasis given on recent review articles and their reference lists. We added proteins, genes and cellular phenotypes to

the map and used databases like KEGG pathway (24–27) (<http://www.genome.jp/kegg/pathway.html>), and Ingenuity Pathway Analysis (IPA) (28) to retrieve connections among them, where it was not possible to retrieve the links directly from the corresponding articles. All added factors were discussed thoroughly with RA experts before addition to the map and advice was taken for the best possible representation of their mechanism of action.

Annotation and curation criteria

We carried out an exhaustive literature search for new proteins, genes and other molecules involved in the pathogenesis of RA. Relevant keywords and key phrases like ‘Pathogenesis of RA’, ‘Cytokines in the pathogenesis of RA’, ‘Therapeutic targets in RA’ among many others were used to filter the literature abstracts and studies in PubMed and Google Scholar. Along with it, we used peer reviewed articles concerning RA and searched their bibliographies to mine relevant information. We focused only on studies based on cells, fluids and tissues of human origin using small-scale experiments, in an attempt to limit false positives from gene expression data used to construct the first RA map. New RA mediators were added and referenced

with at least two PubMed IDs. However, we made some exceptions during the building of the map. For molecules that were either published very recently (since January 2018) or were part of well-characterized pathways involved in RA, we used one PubMed or KEGG ID. For the purposes of this project, we aimed to be inclusive of the whole spectrum of RA. In this context, we used RA as a defining criterion and did not make the distinction between sero-negative and sero-positive RA when reviewing the literature.

We added annotations for all the components (proteins, RNAs and genes) and reactions present in the CellDesigner XML file using the sections text NOTE and Minimal Information Requested In the Annotation of Models (MIRIAM) (29), which are human and machine-readable formats respectively (Figure S2). In the MIRIAM segment, we added PubMed IDs for different cell types with the tag 'bqbiol: is described by'. In the NOTE section, we added text information about KEGG pathway identifiers used to cross-validate interactions.

Evaluation of components and reactions

We carefully evaluated all elements and reactions of the previous RA map and added annotations concerning experimental validation with small-scale experiments where possible. Molecules, for which we could not find small-scale experiments, were kept if appeared in at least two high-throughput studies. We removed from the map molecules that failed to fulfil the above criteria.

Compartments, structure and layout

To improve the layout of the molecular map, we used the CellDesigner plugin Relayout Model (<http://www.celldesigner.org/plugins.html>). The RA map includes six compartments, namely extracellular space, plasma membrane, cytoplasm (including Golgi apparatus, endoplasmic reticulum, and mitochondria), nucleus, secreted molecules and cellular phenotypes.

A cellular phenotype can be viewed as the endpoint of multiple cellular processes that define and shape the morphology and function of the cell, dictating its fate. Extracellular space includes the protein ligands outside the cell that can form a complex with the plasma membrane receptors and proteins resulting in the activation of several signalling cascades. Cytoplasm compartment includes the signalling proteins, enzymes, small molecules and transcription factors, which are subsequently transported to the nucleus and are involved in gene expression regulation. The nucleus compartment includes transcription factors transported from the cytoplasm, genes and RNAs (miRNA

and mRNA). A separate compartment contains proteins secreted out of the cell and, finally, a dedicated compartment contains cellular phenotypes relevant for RA. The RA map has the form of a cell with surrounding extracellular space, the cytoplasmic area containing organelles, proteins and small molecules, the nucleus with gene-regulatory mechanisms, secreted molecules and cellular phenotypes. We used a distinct colour code for the components in the RA map: plasma membrane receptors in peach, proteins in purple, genes in green, RNAs in red and cellular phenotypes in yellow. Inhibition edges are represented in red colour, while for all others like state transition, catalysis, transport, reduced physical stimulation and heterodimer association we used black colour.

Experts' advice and feedback

Experts' curation is critical to reconstructing molecular and cellular interactions from the available literature. Due to the complexity of RA regarding cell types (macrophages, lymphocytes, endothelial cells, synovial fibroblasts), mediators of inflammation (cytokines, chemokines, growth factors, tissue-degrading enzymes) and the variety of biological processes implicated in the disease, the review of the map by RA experts was necessary for an accurate representation of disease hallmarks. To provide a systematic and comprehensive molecular map, we used SBGN standards and a cell layout. We took advice from experienced scientists in both biological and computational domains to make the content comprehensive and functional for different types of users such as experimental biologists, clinicians, computational modellers and bioinformaticians. The RA map layout, the representation of various levels of information and the validity of molecules and pathways included in the RA map, were carefully examined in this context.

SBGN standards and process description map validation

The SBGN (15) is a standard for the visual representation of biological/biochemical processes as networks. Three types of SBGN languages cover different ways to represent biological networks, Process Description (PD), Entity-Relationship (ER) and Activity Flow (AF) (30). The RA map is a PD map showing the detailed biological processes implicated in RA. We systematically checked the compliance to the SBGN standard. For keeping the diagram compact and avoid repeating the same pattern multiple times (activation of protein production from an empty set), we used the translation connectors. VANTED (Visualisation and Analysis of Networks containing Experimental Data) (31), is a framework for systems biology applications

with functionalities ranging from network reconstruction, data visualization, integration of various data types to network simulation using systems biology standards for visualization and data exchange. We used SBGN-ED (an add-on for VANTED for editing, validating and translating of SBGN maps) (32) to validate our SBGN PD encoding of the RA map. As this tool works with SBGN-ML file format, we utilized the CellDesigner to SBGN converter (<https://royludo.github.io/cd2sbgnml>) for converting the CellDesigner XML file into SBGN-ML format and subsequently import the file to VANTED for further analysis.

Web-based MINERVA map The RA map is available as an online interactive map using MINERVA (Molecular Interaction NEtworkKs VisuAlization) platform (19). MINERVA is a web service that supports curation, annotation and visualization of molecular interaction networks in the SBGN-compliant format. MINERVA provides automated content annotation and verification, along with mapping of drug targets and overlaying experimental data on the visualized networks. Automated annotations (HGCN) and curator's annotations for every component and reaction are displayed in the left panel (see Figure 3A). The user can also visualize cell-specific data based on curated overlays or analyse patients' *omic* datasets (see Figure 8). Moreover, MINERVA provides an interface for interrogating several other databases such as DrugBank (33) (<https://www.drugbank.ca/>), ChEMBL (34) (<https://www.ebi.ac.uk/chembl/>), CTD (35) (<http://ctdbase.org>) and miRTarBase (36) (<http://mirtarbase.mbc.nctu.edu.tw>).

Overlays We provide three different types of overlays with the RA map. The first type corresponds to cell, tissue and fluid specific overlays. The RA map is a global map, integrating data and information from various sources. As a result, it has reactions and components that come from different cell or tissue types. We have grouped the sources into seven distinct groups that we provide as overlays. The groups are synovial fibroblasts, synovial tissue, peripheral blood mononuclear cells, blood, synovial fluid, chondrocytes and macrophages (Table S1). These overlays allow visualizing cell or tissue-specific interactions and molecules. The second type of overlay comes from publicly available datasets and facilitates visualization of mapping components onto the RA map. The third type of overlays concerns canonical pathways retrieved from REACTOME, EBI for TNF, IL6, MAPK and Interferon signalling (Table S2).

BioInfoMiner analysis The algorithm performs a topological analysis of semantic networks, derived from ontologies

(Gene Ontology (37, 38), Human Phenotype Ontology (39) and Mammalian Phenotype Ontology (40)) and pathway databases with hierarchical structure, like REACTOME (41–43). It employs a graph-theoretical method that corrects the annotation bias of community ontologies, performs enrichment analysis to assess the over-representation of terms and ranks the related genes according to their connectivity in the corrected semantic network (44, 45). Systemic processes are clusters of terms that share maximum semantic similarity among them, but minimal similarity among other clusters. The highly ranked genes are those associated with many systemic processes, and thus, they are considered hub genes in the semantic network, assuring cross-talking among distinct, orthogonal (inter-independent) processes. Finally, the application derives a signature, consisting of the mapping of the prioritized genes to a minimal set of clustered systemic processes. Furthermore, BioInfoMiner provides a pharmacogenomic analysis, as the derived hub genes constitute putative drug targets.

Topological and gene ontology enrichment analysis with Cytoscape

The RA map XML file was imported in Cytoscape, version 3.5.0, and the built-in NetworkAnalyzer function was used for topological analysis (17).

Results

A comprehensive molecular interaction map for RA

The RA map graphically illustrates signalling pathways, gene expression regulation, molecular mechanisms and cellular phenotypes involved in the pathogenesis of the disease. As shown in Figure 1, and explained in detail in the methodology section, the RA map requires exhaustive literature curation, information mining from relevant databases along with continuous updating and advice from domain experts. Importantly, the interactions shown in the diagram represent a graphical model encoded using a standardized format, making the map computationally tractable.

For the construction of the map, we used the graphical editor CellDesigner (18). In Figure 2, one can see an overview of the RA map. We constructed the RA map following the SBGN Process Description format (46). We made only one exception concerning the choice of the translation and transcription representation, for which we used the CellDesigner's system of symbols. The RA Map features 506 species, 449 reactions and 8 cellular phenotypes. The biomolecules in the map are 303 proteins, 61 molecular

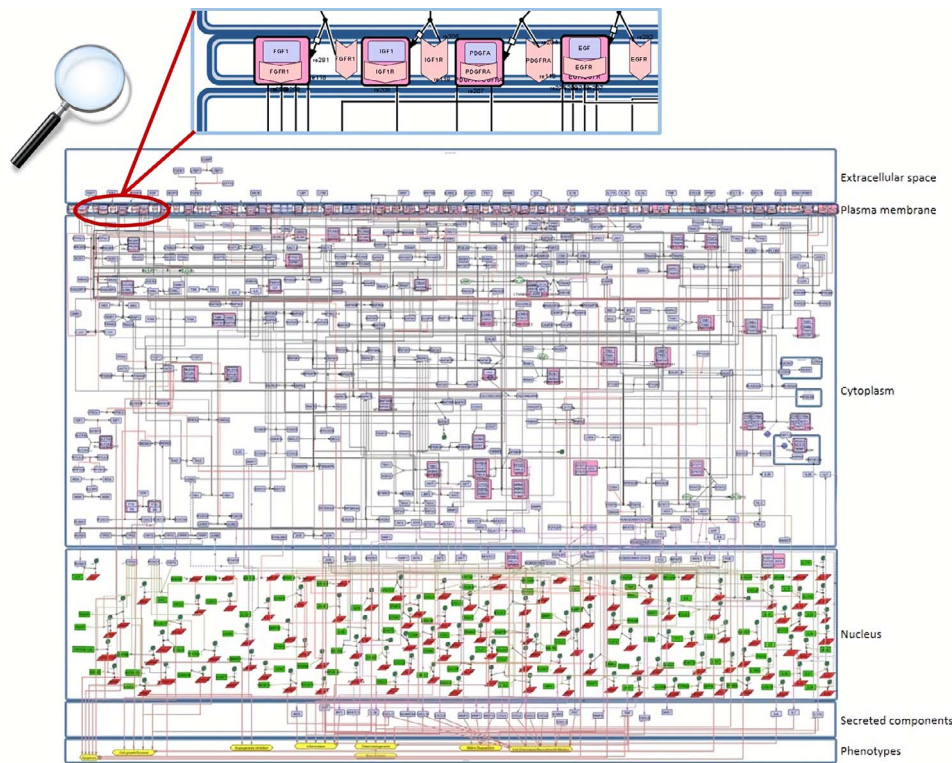


Figure 2. Snapshot of the SBGN-compliant RA map. The map is colour-coded with proteins in purple, genes in green, RNAs in red and phenotypes in yellow. State transitions and catalysis reactions are displayed in black, and the inhibitions are in red. Compartments are distinguished as bounding boxes. The map was built using CellDesigner, version 4.4 (18). Modifications to the SBGN format: translation arcs are used to keep the representation compact, as well as the gene and RNA shapes.

complexes, 106 genes, 106 RNA entities, 2 ions and 7 simple chemical species like for example cAMP, H_2O_2 or PIP_3 . Proteins include extracellular, membrane and cytoplasmic proteins comprising signalling proteins, enzymes and transcription factors. The reactions are classified as state transitions, catalyses, inhibitions, transports, heterodimer associations, dissociations, Boolean AND gates and reduced physical stimulations. All the components in the map have at least two manually curated PubMed references, giving overall 353 publications covering a period from 1973 to 2019 (Figure S3).

The RA map is organized in the form of a cell representing the flow of information from the extracellular space (ligands) to the plasma membrane (ligand–receptor complexes) and then to the cytoplasm (signalling pathways), the nucleus (gene regulation) and the secreted compartment or cellular phenotypes (Figure 2).

Molecular pathways covered in the RA map

The RA map contains hallmark cellular and molecular pathways that participate in disease pathogenesis. In signalling cascades, the activation occurs as a response to an upstream stimulus. After activation, the signal propagates

through a series of coupled reactions from the plasma membrane to the cytoplasm, to regulate key factors that are responsible for gene regulation and different cellular phenotypes. The RA map includes the following upstream stimuli:

- (i) Cytokines and chemokines: a diverse group of proteins like tumour necrosis factor (TNF) and interleukins to list a few, implicated in various phases of RA pathogenesis by promoting autoimmunity, initiating and maintaining chronic inflammatory synovitis and driving cartilage and bone destruction (47–49);
- (ii) Growth factors: such as epidermal growth factor (EGF), fibroblast growth factor (FGF), insulin-like growth factor (IGF), vascular endothelial growth factor (VEGF), platelet-derived growth factor (PDGF), activate intracellular signalling pathways (such as PI3K-AKT pathway) and regulate a broad range of cellular functions like cell growth, survival, cell motility and apoptosis (50, 51);
- (iii) Toll-like receptors (TLRs): TLR2 and TLR4 are primarily expressed in synovial fibroblasts and macrophages in human RA joints (52–54). Activation of TLR2 and TLR4 results in recruitment of adaptor

molecules such as MyD88, IRAK, TRAF6 and TANK-binding kinase (TBK)-1 and leads to the activation of MAPKs and NF- κ B and the increased expression of various pro-inflammatory and tissue-destructive mediators (such as TNF, IL-6, chemokines and MMPs) (55, 56).

The activation of these upstream components leads to the activation of downstream pathways that include:

- (i) The JAK-STAT pathway: this is an effective target in RA therapy. Many cytokines, including IL-6 and TNF, which are validated therapeutic targets in RA, activate directly (for example IL-6) or indirectly (for example TNF) this pathway by phosphorylating JAK proteins. JAKs, in turn, phosphorylate STATs, which then dimerize and translocate to the nucleus and bind to regulatory elements of DNA modulating the expression of target genes (57, 58). Activation of JAK-STAT pathway also results in the activation of suppression of cytokine signalling (SOCS), which operates as a feedback inhibitory loop aiming to terminate excessive activation of JAK-STAT (59).
- (ii) The NF- κ B pathway: it is involved in inflammation, cell survival and proliferation. Activated NF- κ B is detected in immune cells (such as macrophages and lymphocytes) as well as in stromal cells (such as FLS and endothelial cells) and stimulates the transcription of arthritogenic mediators like IL-1, TNF, RANKL, PTGS2 and IL-6 in RA synovium. TNF, IL-1 and RANKL are key upstream RA-relevant triggers of the activation of the NF- κ B pathway (60).
- (iii) The MAPK pathway: all the three classes of MAPKs, namely ERK, JNK and p38, are found to be expressed and activated in synovial tissue in RA. A series of cytokines including among others TNF, IL-1 and IL-6 activate ERK, JNK and p38 MAPK in synovial tissue with successive induction of proinflammatory mediators such as cytokines and tissue destructive enzymes (e.g. MMP-1 and MMP-13) (61, 62). Negative feedbacks are required to keep in check the constitutive activation of MAPK proteins in order to control the excessive prolonged expression of pro-inflammatory genes (61).
- (iv) The PI3K-AKT pathway: growth factors like VEGF and FGF induce the PI3K-AKT pathway (50, 63–65). Activated cellular AKT regulates immune cells, and survival of synoviocytes and chondrocytes by phosphorylating several downstream signalling proteins modulating mTOR, BAD, FOXO3 and tumour protein-73 (TP-73) (63).

All signalling cascades end at specific cellular outcomes grouped in eight distinct phenotypes such as inflammation (1, 51, 66, 67), cell chemotaxis/recruitment/infiltration (68, 69), matrix degradation (66, 70–73), osteoclastogenesis (66, 74, 75) and bone erosion (1, 66, 76, 77), angiogenesis (51, 66, 78, 79), apoptosis (66, 80–83) and finally cell survival/growth/proliferation (51, 84–86).

Transforming RA map into a state of the art knowledge base using MINERVA

The RA map is available at ramap.elixir-luxembourg.org in the form of an interactive diagram, using the platform MINERVA (Molecular Interaction NETworks VisuAlization) (Figure 3). Clicking on a biomolecule in the map, the user can choose to visualize interacting drugs, chemicals and miRNAs. The RA map interfaces with DrugBank (<https://www.drugbank.ca/>), ChEMBL (<https://www.ebi.ac.uk/chembl/>), CTD (<http://ctdbase.org>) and miRTarBase (<http://mi.rtarbase.mbc.nctu.edu.tw>).

RA map offers custom visualization and export capabilities via MINERVA plugins (87). For instance, users can explore the RA map starting from a molecule of interest and easily follow its interactions, even throughout a dense and complex network. This functionality facilitates navigating through the contents and tracking the flow of the signal from the ligand to the corresponding phenotype (Figure 4A). Another feature of the RA map is the stream plugin, allowing for highlight and export of entire subnetworks in the map in one click. This feature is especially important to visualize the ensemble of signalling pathways converging on the same disease-related phenotype (Figure 4B).

The RA map as a template for visualizing cell-specific overlays

The RA map contains information from various sources serving as a generic blueprint for disease mechanisms. However, due to extensive annotation and reference, the user can opt for visualizing cell-specific nodes and interactions. In the RA map, we have grouped our sources in seven distinct groups: synovial fibroblasts, synovial tissue, peripheral blood mononuclear cells (including PMNs), blood (including T and B cells), synovial fluid, chondrocytes and macrophages (Table S1). Synovial fibroblasts are the most frequent cell type in the RA map covering a total of 45%, followed by synovial tissue with 36% (Figure S1). In the RA map, the user can select to visualize one of the corresponding overlays, for example, synovial tissue overlay (Figure 5).

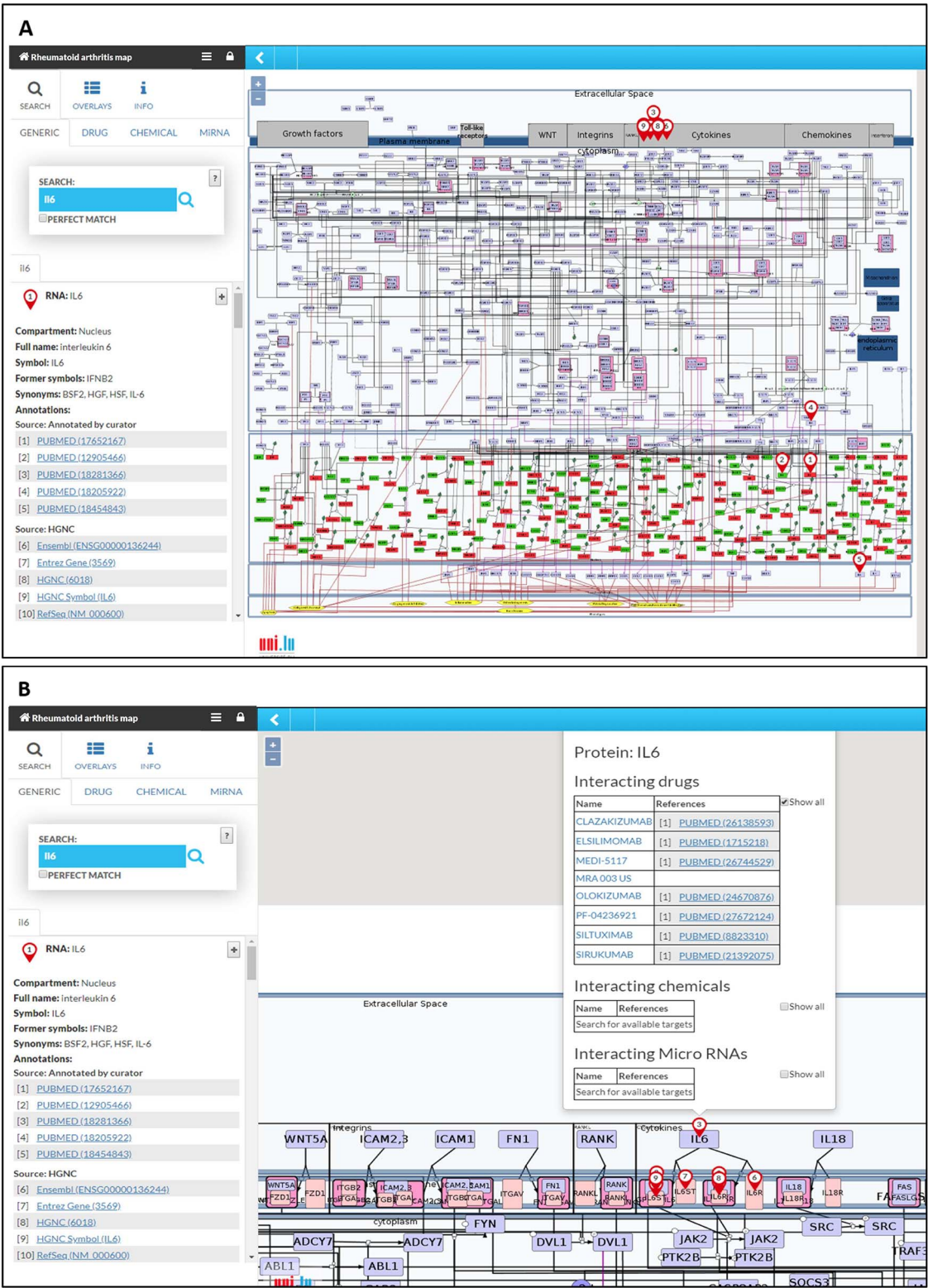


Figure 3. The RA map in MINERVA platform. (A) Users can use the search box to type in the element of interest. The resulting element shows up as pins on the map. Corresponding annotations of the searched element, like HGNC, Entrez Gene, RefSeq and Ensembl identifiers are displayed on the left panel along with the PubMed identifiers of the manually curated annotations. **(B)** Further clicking on the pin will display additional information about interacting drugs, chemicals and microRNAs for the element.

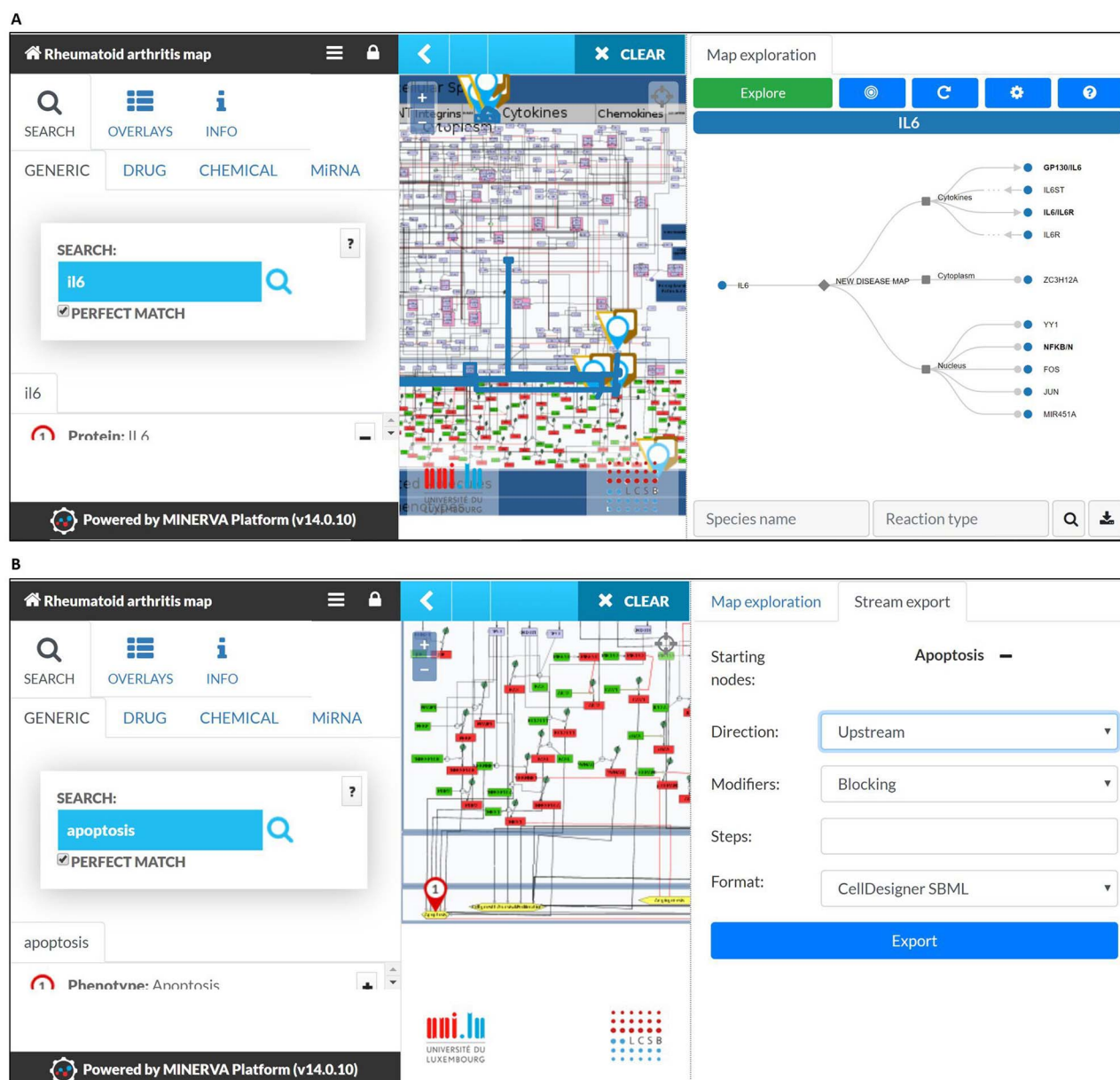


Figure 4. MINERVA plugins. (A) The tree plugin allows to navigate in dense networks by following interactions in a tree-like manner. (B) The stream plugin allows for downstream or upstream expansion when selecting a node of interest.

Visualizing various datasets

We used publicly available datasets for visualization with the RA map. Our goal was to compare the differentially expressed pathways or map regions in different datasets. For this purpose, we used the datasets from transcriptomic data of synovial tissue (88). We performed differential expression analysis between Berlin, Leipzig and Jena datasets using osteoarthritis as control and visualized the mapping of 122 molecules to the RA map. Most pathways were highlighted, as molecules that lead to most phenotypes were present. Interestingly, we found enrichment for almost all cellular phenotypes except for apoptosis and angiogen-

esis. Molecules leading to six out of eight phenotypes were expressed, while molecules linked to the two mentioned phenotypes were absent (Figure 6).

Systemic interpretation and pharmacogenomics analysis using BioInfoMiner

We also used the BioInfoMiner web application (16) (<https://bioinforminer.com>) to perform a functional analysis of the RA map. The application performs a biological interpretation of gene sets, which comprises detection and prioritization of systemic processes and pathways, as well as prioritization of genes based on their mapping to those

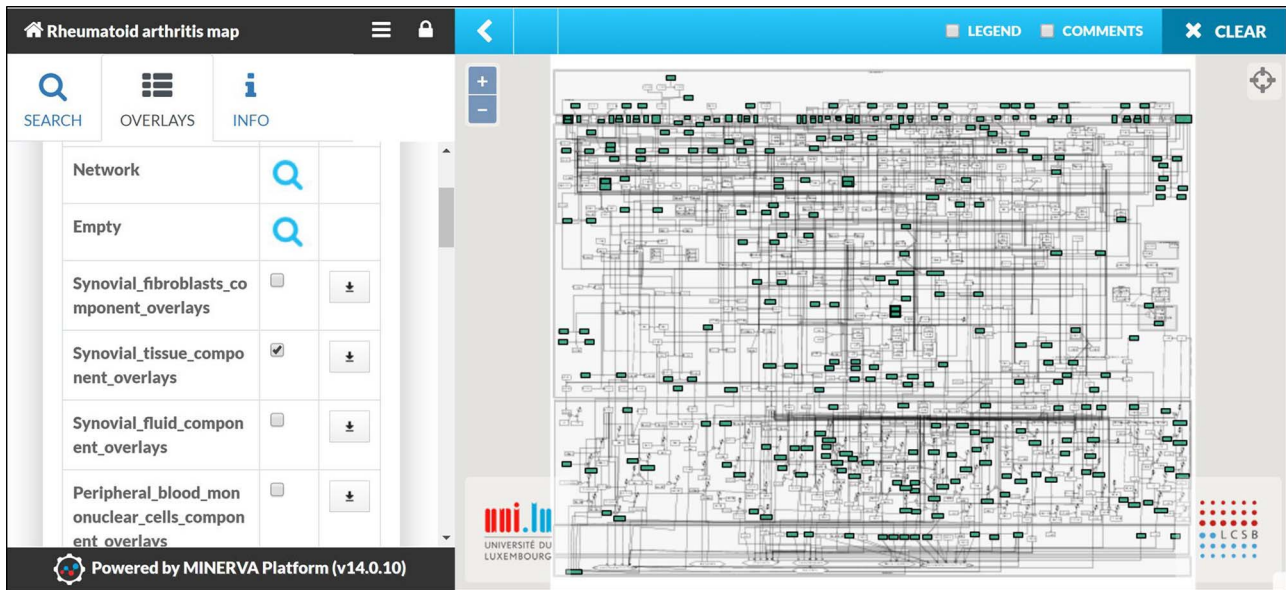


Figure 5. Visualizing cell/tissue/fluid-specific parts of the RA map using dedicated overlays. Snapshot of the visualization of the Synovial Tissue overlay.

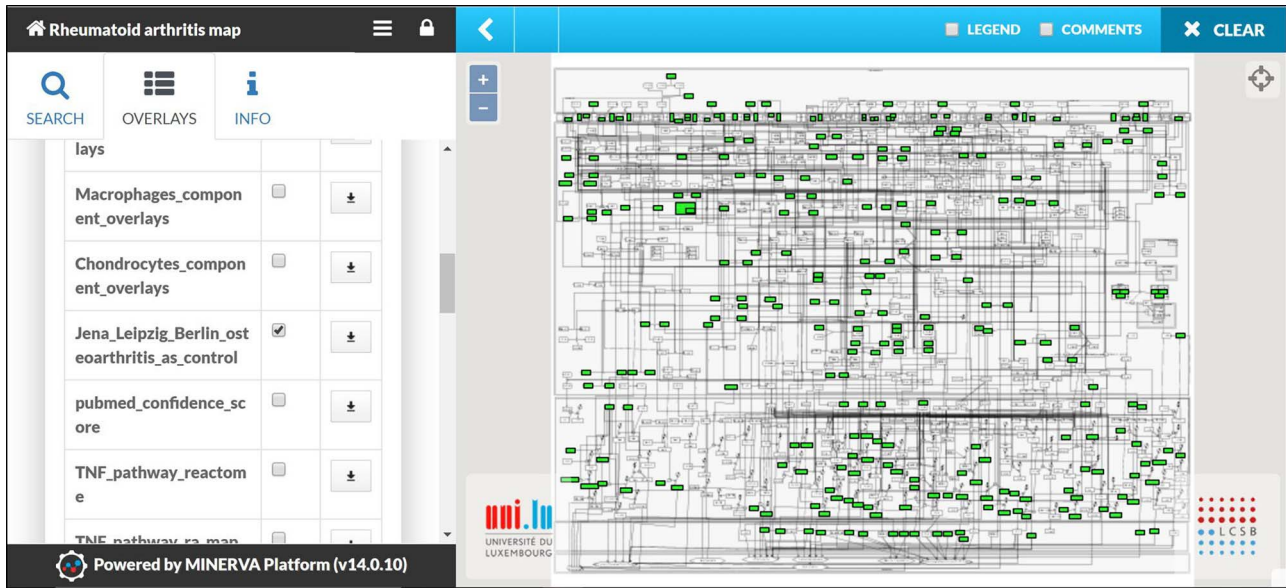


Figure 6. Mapping of Omic datasets from RA synovial tissue. The apoptosis and angiogenesis phenotypes appear to be inactive as no molecule leading to these cellular phenotypes is mapped.

processes. We used BioInfoMiner as a second layer of analysis to see if the functional enrichment would give results relevant to the autoimmune process and RA. We performed two sets of analyses using gene ontology (GO) and human phenotype ontology (PHO) terms. The first analysis using GO gave enrichment of terms like Inflammatory response, Regulation of cytokine production and Activation of MAPK activity, all relevant to pathways included in the RA map. The top five GO terms included apoptotic signalling pathway, positive regulation of cell death, negative regulation of apoptotic signalling pathway, positive regulation of

NF-kappaB transcription factor activity and regulation of I-kappaB kinase/NF-kappaB signalling. It also gave a list of 48 prioritized genes (Table S3). The top 10 priority genes obtained were TNF, toll-like receptor 4 (TLR4), receptor-interacting serine/threonine kinase 2 (RIPK2), interleukin 1 beta (IL1B), receptor-interacting serine/threonine kinase 1 (RIPK1), fas-associated via death domain (FADD), Janus kinase 2 (JAK2), wnt family member 5A (WNT5A), TNF receptor-associated factor 6 (TRAF6) and innate immune signal transduction adaptor (MYD88). The signature we obtain using GO consists of the ranked systemic processes

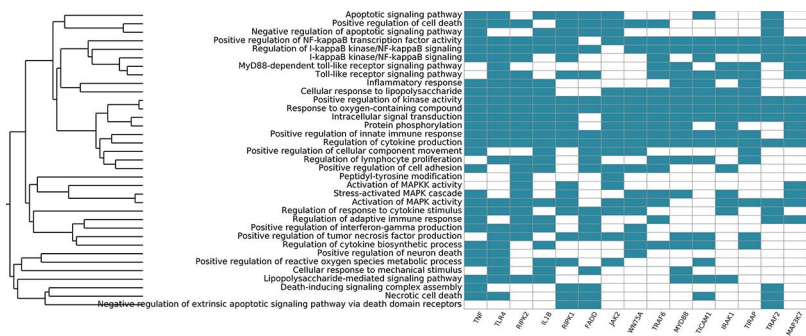


Figure 7. Systemic functional analysis of the RA map using GO terms. Heat map of the top 15 priority genes and their systemic interpretation using BioInfoMiner and GO terms.

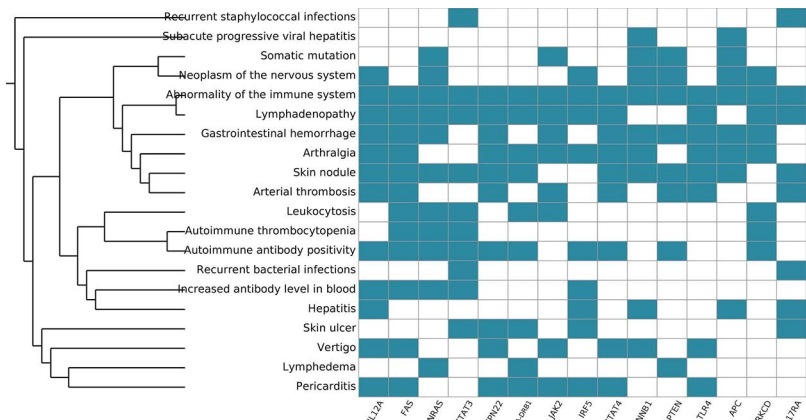


Figure 8. Systemic functional analysis of the RA map using HPO terms. Heat map of the top 15 priority genes and their systemic interpretation using BioInfoMiner and HPO terms.

(y-axis) and prioritized genes (x-axis) (Figure 7). The first most prioritized gene was TNF, a prevalent target for many approved drugs such as anti-TNF agents, while all other nine genes have been implicated in studies for drug targeting in RA (89, 90).

Functional analysis with BioInfoMiner using the Human Phenotype Ontology gave 32 priority genes (Table S4) and enrichment in terms containing arthralgia, skin nodule, abnormality of the immune system, among others (Table S5), as we can see in Figure 8. Overall, the systemic functional analysis with BioInfoMiner further confirmed the validity of the model at the semantic level, complementary to the mechanistic one. The top 10 priority genes using PHO terms are interleukin 12A (IL12A), Fas cell surface death receptor (FAS), NRAS proto-oncogene (NRAS) GTPase, signal transducer and activator of transcription 3 (STAT3), protein tyrosine phosphatase (PTPN22), non-receptor type 22, major histocompatibility complex, class II, DR beta 1 (HLA-DRB1), Janus kinase 2 (JAK2), interferon regulatory factor 5 (IRF5), signal transducer and activator of transcription 4 (STAT4), catenin beta 1 (CTNNB1). All of these genes have been considered as putative drug targets in RA.

Topological analysis of the RA map as a complex network

We imported the RA map to Cytoscape 3 to perform network analysis. The RA network comprises 1225 nodes and 1471 interactions (Figure 9). The analysis using Network Analyzer, a built-in tool of Cytoscape, revealed that the RA network consists of 30 connected components. These connected components correspond to the connected subgraphs, i.e. parts of the graph in which any node is accessible from any other node by a path, with a core subgraph of 1106 nodes and 1379 reactions and 29 smaller ones.

Node degree is a characteristic of the nodes of a network that describes the number of adjacent nodes (nodes directly connected to them). In directed networks such as signalling networks where the reactions are oriented (i.e. from the ECM to the nucleus) we can distinguish two types of node degree: the in-degree, meaning the number of directed edges that have the node as target, and the out-degree that is the number of directed edges that have the node as source. Node degree is an individual characteristic for each node, but a degree distribution can be computed to assess the diversity of the whole network.

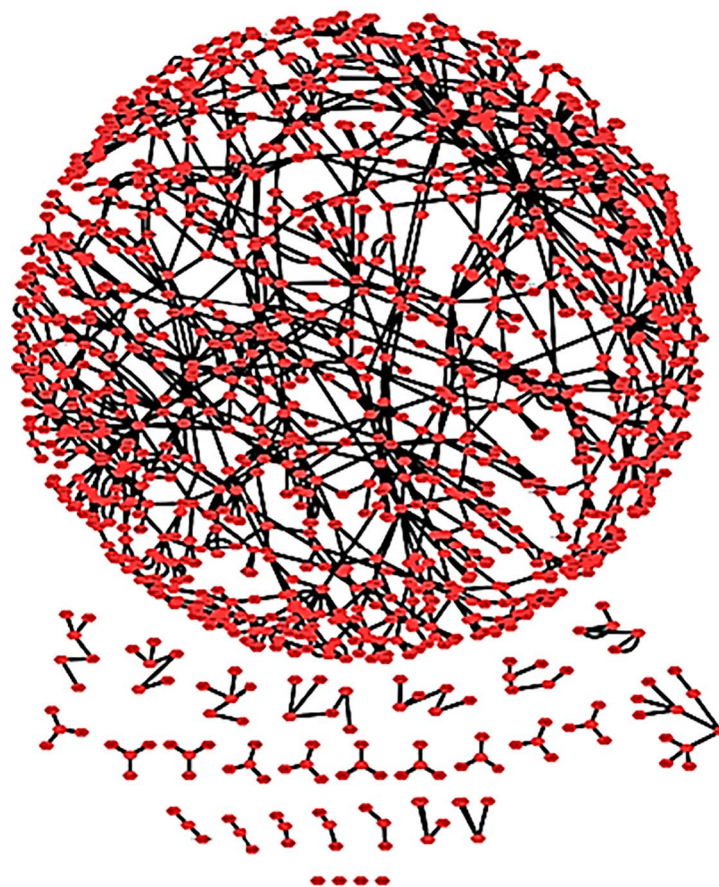


Figure 9. The RA map as a complex network. The RA network with spring embedded layout. One connected core and several smaller unconnected parts are shown.

The majority of biological networks display scale-free properties (91), which means that they contain a few central nodes that are highly connected (hubs) and several other loosely connected peripheral nodes. These networks follow a power law. This function indicates that there is a high diversity of node degrees which is why we describe these networks as ‘scale-free’.

First, we performed the analysis considering the network as undirected to obtain the overall degree distribution (in and out) and then as directed to get the in-degree and out-degree distributions. All node degree distributions follow a power law, showing that the RA network is indeed a scale-free network (91) (Figure 10).

In Table 1, we can see some of the topological characteristics of the RA network, analysed in Cytoscape. Each node has an average of 2.299 neighbours (nodes to which it is connected). We used the degree distribution to obtain the hubs of the RA network, and in Table 2, we display the top 10 hubs. The network diameter of the RA network that corresponds to the maximum length of shortest paths between two nodes is 24 suggesting that the signal starting from ligand–receptor complexes in the membrane reaches

most of the network within 24 steps. The characteristic path length of the network that corresponds to the expected distance between two connected nodes is approximately 10, meaning that the response to a signal and its propagation can occur relatively rapidly.

Discussion

Visual representation of complex pathways and biological processes involved in a disease allows clinical and life sciences researchers to explore relevant mechanisms, which are often intricate and intertwined. Standardized representation and formalization of knowledge in the form of disease maps create an interface to a broad range of bioinformatics and modelling workflows. We present here a state-of-the-art, large-scale molecular interaction map for RA, which is to our knowledge the first SBGN-compliant Process Description disease map. While other efforts, such as the Asthma map, follow the SBGN format, their approach is different as they use three levels of granularity and different SBGN representations for every layer of information. The Process Description level for Asthma map

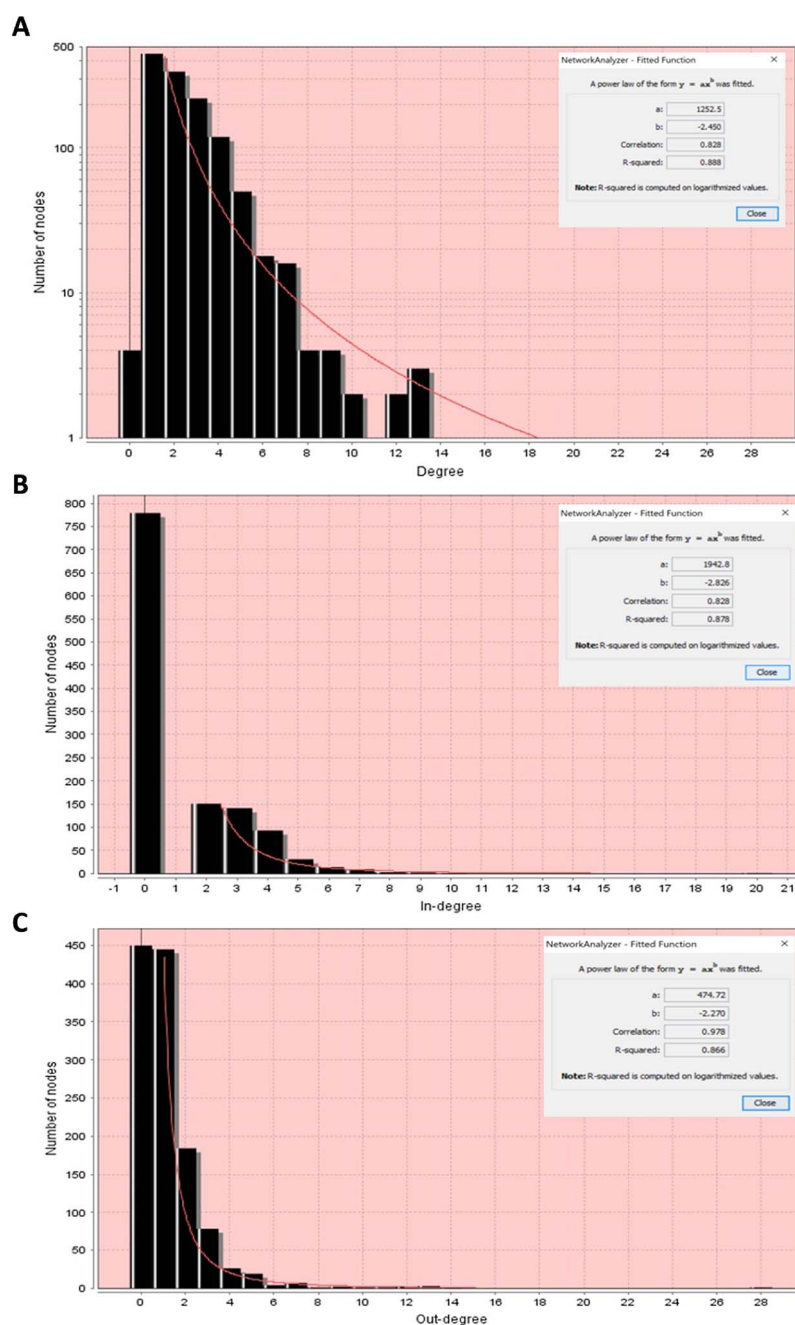


Figure 10. Node degree distributions of the RA map with a fitted power law. (A) Overall degree distribution. (B) In-degree distribution. (C) Out-degree distribution.

consists of a set of separate modules that correspond to an Activity Flow layer, while the RA map is a global Process Description disease map.

All the components and reactions are annotated using only RA and human-specific studies. The RA map is part of the Disease Maps Project, a large scale community effort to comprehensively represent mechanisms for various diseases (13, 14) (<http://disease-maps.org/>). The community fosters the exchange of good practices and promotes the use of standards for the development of disease maps. The

standards of curation and graphical representation, as well as the extensive annotation in both human and machine-readable formats of the RA map, ensure transparency, reproducibility and reusability of its content.

In 2010 the first RA map was published by Wu et al. They used exclusively high-throughput RA experiments (mRNA, miRNA) described in 28 studies combined with data available in the KEGG database. A total of 435 species (263 proteins, 58 genes, 48 RNAs, seven simple molecules, one ion, one antisense RNA, 47 complexes), 265 reactions

Table 1. Example of simple topological parameters obtained with Network Analyzer for the RA network

Topological parameters	Corresponding values
Connected component	30
Network diameter	24
Characteristic path length	10.099
Average number of neighbours	2.299
Number of nodes	1225
Isolated nodes	4

Table 2. Top 10 hubs of the RA map

RA map nodes	Node degree	Role	Reference
NFKB	28	Implicated in RA and inflammation	(92, 93)
Inflammation	14	A major characteristic of RA	(1, 66)
AKT	13	Regulates apoptosis in RA	(94, 95)
Cell chemotaxis/recruitment/infiltration	13	Implicated in RA	(96, 97)
JUN	13	Implicated in RA	(98, 99)
MAPK1	12	Implicated in RA	(100, 101)
RAC1,2	12	Implicated in RA	(102, 103)
Cell growth/survival	11	Major characteristic of RA	(66, 85)
Osteoclastogenesis	10	Results in bone damage in RA	(74, 104)
TP53	9	Involved in the apoptosis pathway implicated in RA	(105, 106)

and 10 phenotypes involved in RA were identified using this approach. We decided to follow a different approach as described in the methodology section, in an attempt to limit false positives, increase confidence by incorporating experts' advice and promote the use of SBGN standards for representation to assure reusability of the map. The new RA map we present here includes information from 353 peer-reviewed publications, and it has a significantly bigger size, as it features 506 species, 446 reactions and 8 phenotypes. The species in the map are classified to 303 proteins, 61 complexes, 106 genes, 106 RNA entities, 2 ions and 7 simple molecules.

The RA map can also be used as an interactive knowledge base, using the platform MINERVA and serve as a template for overlaying multiple datasets. Visualization of experimental data could help highlight aspects of the affected biological process and make differences between experimental conditions more evident. Visualizing the results of differential expression analysis of three datasets of gene expression of RA synovial tissues showed enrichment in all cellular phenotypes but not in apoptosis. This finding is in line with the fact that fibroblasts, which constitute a large percentage of the RA synoviocytes, have an apoptosis-resistant phenotype (107, 108).

We performed functional analysis and gene prioritization using BioInfoMiner (16). The genes that rank higher in this analysis are associated with many systemic pro-

cesses and are considered as hubs in the semantic network. Along with prioritization, a pharmacogenomic analysis is provided since the hubs proposed are considered as putative drug targets. The results of the analyses using GO and PHO terms revealed known RA players, most of which have been already used as drug targets demonstrating that the RA map comprises well-characterized factors and captures most of the relevant systemic processes implicated in the disease.

The RA map serves as a curated knowledge base, but it can also be analysed as a complex network. Topological analysis can reveal underlying structural features of the RA map like unconnected parts of the network, or important hubs (well-connected nodes) which are otherwise hard to perceive in large-scale networks. The topological analysis performed in this study revealed connected and unconnected parts of the network. This result reflects our fragmented knowledge on the one hand, but also the use of stringent criteria for the nodes included in the map: experimentally validated interactions in at least two published studies, use of data of strictly human origin and disease-specific.

Another reason that contributes to the limited wiring of some of the RA map components is the unavailability of known interactions for newly discovered factors for RA. However, we keep them present because the RA map also works as an encyclopaedia for the disease, even if some parts of the puzzle are still missing.

The topological analysis also assists in the understanding of significantly connected nodes (hubs), placing them in their functional context. The top ten hubs of the RA map as seen in Table 1 (NFKB, AKT, Inflammation, Cell chemotaxis/recruitment/infiltration, JUN, MAPK1, RAC1,2 Cell growth/Survival, Osteoclastogenesis, TP53) are well-characterized factors implicated in the disease. Not surprisingly, four of them (AKT, MAPK1, RAC1,2, TP53) were also characterized as hubs in the first RA map by Wu *et al.*, based on high-throughput data.

Conclusion

The RA map is the fruit of interdisciplinary collaborations between clinicians, biologists and bioinformaticians. The aim was to build not only a knowledge repository but a versatile tool that can be used for various purposes. The RA map can offer to experimental biologists and clinicians easy access to all molecular pathways implicated in the disease along with references and annotations, to bioinformaticians a template for disease-specific pathway enrichment of *omic* datasets and finally, to computational modellers a mechanistic scaffold for the inference of a computational model (5, 6, 109), providing an intermediate step between a conceptual and an executable model.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The RA map is freely accessible at ramap.elixir-luxembourg.org

The original CellDesigner XML file of the whole map can be downloaded from MINERVA from the INFO section by clicking on the source file (third tab in the left panel of MINERVA website). Right clicking on the main screen also gives an option to export the visible content in three formats – SBML, CellDesigner SBML and SBGN-ML.

Competing interests

The authors declare that they have no competing interests. G.D.K. is currently also an employee at Regeneron Pharmaceuticals Inc. and declares no conflict of interest regarding the content of this manuscript.

Authors' contributions

AN designed the study, A.N., V.S. and G.D.K. built the new RA map content, V.S. drew the map including the addition of annotations and performed literature mining and topological analysis. G.D.K. validated the content, V.S. and M.V. performed DEG in datasets, M.V. produced scripts for all graphs, A.N. and E.P. performed functional and systemic analysis with BioInfoMiner, V.S. and A.M. validated the SBGN Process Description format of the RA map, M.O. and P.G. set up the MINERVA online version of the RA map, M.O. and V.S. added annotations to reactions of the RA map, V.S. and E.B. produced IPA RA datasets, E.P.T. helped with the validation of functional analyses results and A.N. and V.S. wrote the manuscript, all authors read and suggested modifications, all authors read and approved the final manuscript. **Corresponding author:** Dr Anna Niarakis.

Acknowledgements

The authors would like to thank the ELIXIR node in Luxembourg for hosting and providing technical support for the RA map and all the members of the Disease Maps Project consortium <https://disease-maps.org/> for stimulating discussions and valuable feedback.

Funding

The project is supported by The Open Health Institute, <https://www.openhealth-institute.org/> for Open Access publishing and FONDAGEN for scholarships received by V.S.

References

- McInnes, I.B. and Schett, G. (2011) The pathogenesis of rheumatoid arthritis. *N. Engl. J. Med.*, **365**, 2205–2219.
- Frank-Bertoncelj, M., Klein, K. and Gay, S. (2017) Interplay between genetic and epigenetic mechanisms in rheumatoid arthritis. *Epigenomics*, **9**, 493–504.
- Tripathi, S., Flobak, Å., Chawla, K. *et al.* (2015) The gastrin and cholecystokinin receptors mediated signaling network: a scaffold for data analysis and new hypotheses on regulatory mechanisms. *BMC Syst. Biol.*, **9**, 40.
- Kawakami, E., Singh, V.K., Matsubara, K. *et al.* (2016) Network analyses based on comprehensive molecular interaction maps reveal robust control structures in yeast stress response pathways. *npj Syst. Biol. Appl.*, **2**, 15018.
- Niarakis, A., Bounab, Y., Grieco, L. *et al.* (2014) Computational modeling of the main signaling pathways involved in mast cell activation. *Curr. Top. Microbiol. Immunol.*, **382**, 69–93.
- Grieco, L., Calzone, L., Bernard-Pierrot, I. *et al.* (2013) Integrative modelling of the influence of MAPK network on cancer cell fate decision. *PLoS Comput. Biol.*, **9**, e1003286.
- Fujita, K.A., Ostaszewski, M., Matsuoka, Y. *et al.* (2014) Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol. Neurobiol.*, **49**, 88–102.

8. Mizuno,S., Iijima,R., Ogishima,S. *et al.* (2012) AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. *BMC Syst. Biol.*, **6**, 52.
9. Matsuoka,Y., Matsumae,H., Katoh,M. *et al.* (2013) A comprehensive map of the influenza a virus replication cycle. *BMC Syst. Biol.*, **7**, 97.
10. Mazein,A., Knowles,R.G., Adcock,I. *et al.* (2018) AsthmaMap: an expert-driven computational representation of disease mechanisms. *Clin. Exp. Allergy*, **48**, 916–918.
11. Kuperstein,I., Bonnet,E., Nguyen,H.A. *et al.* (2015) Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, **4**, e160.
12. Singh,V., Ostaszewski,M., Kalliolias,G.D. *et al.* (2018) Computational systems biology approach for the study of rheumatoid arthritis: from a molecular map to a dynamical model. *Genomics Comput. Biol.*, **4**.
13. Mazein,A., Ostaszewski,M., Kuperstein,I. *et al.* (2018) Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *npj Syst. Biol. Appl.*, **4**, 21.
14. Ostaszewski,M., Gebel,S., Kuperstein,I. *et al.* (2019) Community-driven roadmap for integrated disease maps. *Brief. Bioinformatics*, **20**, 659–670.
15. Le Novère,N., Hucka,M., Mi,H. *et al.* (2009) The systems biology graphical notation. *Nat. Biotechnol.*, **27**, 735–741.
16. Lhomond,S., Avril,T., Dejeans,N. *et al.* (2018) Dual IRE1 RNase functions dictate glioblastoma development. *EMBO Mol. Med.*, **10**.
17. Shannon,P., Markiel,A., Ozier,O. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
18. Matsuoka,Y., Funahashi,A., Ghosh,S. *et al.* (2014) Modeling and simulation using CellDesigner. *Methods Mol. Biol.*, **1164**, 121–145.
19. Gawron,P., Ostaszewski,M., Satagopam,V. *et al.* (2016) MINERVA-a platform for visualization and curation of molecular interaction networks. *npj Syst. Biol. Appl.*, **2**, 16020.
20. Hucka,M., Finney,A., Sauro,H.M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
21. Kitano,H., Funahashi,A., Matsuoka,Y. *et al.* (2005) Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.*, **23**, 961–966.
22. Wang,P., Lü,J. and Yu,X. (2014) Identification of important nodes in directed biological networks: a network motif approach. *PLoS ONE*, **9**, e106132.
23. Wu,G., Zhu,L., Dent,J.E. *et al.* (2010) A comprehensive molecular interaction map for rheumatoid arthritis. *PLoS ONE*, **5**, e10137.
24. Kanehisa,M. (2009) Representation and analysis of molecular networks involving diseases and drugs. *Genome Inform.*, **23**, 212–213.
25. Kanehisa,M., Sato,Y., Furumichi,M. *et al.* (2019) New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*, **47**, D590–D595.
26. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
27. Kanehisa (2019) Toward understanding the origin and evolution of cellular organisms. *Protein Sci.*, **28**, 1947–1951.
28. Krämer,A., Green,J., Pollard,J. *et al.* (2014) Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, **30**, 523–530.
29. Le Novère,N., Finney,A., Hucka,M. *et al.* (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.*, **23**, 1509–1515.
30. Le Novère (2015) Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.*, **16**, 146–158.
31. Rohn,H., Junker,A., Hartmann,A. *et al.* (2012) VANTED v2: a framework for systems biology applications. *BMC Syst. Biol.*, **6**, 139.
32. Czauderna,T., Klukas,C. and Schreiber,F. (2010) Editing, validating and translating of SBGN maps. *Bioinformatics*, **26**, 2340–2341.
33. Wishart, D. S., Feunang, Y. D., Guo, A. C., *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
34. Gaulton,A., Hersey,A., Nowotka,M. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.
35. Davis,A.P., Grondin,C.J., Johnson,R.J. *et al.* (2019) The comparative toxicogenomics database: update 2019. *Nucleic Acids Res.*, **47**, D948–D954.
36. Chou,C.-H., Shrestha,S., Yang,C.-D. *et al.* (2018) miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **46**, D296–D302.
37. Ashburner,M., Ball,C.A., Blake,J.A. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
38. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
39. Köhler,S., Carmody,L., Vasilevsky,N. *et al.* (2019) Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Res.*, **47**, D1018–D1027.
40. Smith,C.L. and Eppig,J.T. (2009) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **1**, 390–399.
41. Jassal,B., Matthews,L., Viteri,G. *et al.* (2019) The reactome pathway knowledgebase. *Nucleic Acids Res.*
42. Fabregat,A., Jupe,S., Matthews,L. *et al.* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
43. Fabregat,A., Sidiropoulos,K., Viteri,G. *et al.* (2017) Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics*, **18**, 142.
44. Moutselos,K., Maglogiannis,I. and Chatziioannou,A. (2011) GOREvenge: a novel generic reverse engineering method for the identification of critical molecular players, through the use of ontologies. *IEEE Trans Biomed Eng.*, **58**, 3522–3527.
45. Koutsandreas,T., Binenbaum,I., Pilalis,E. *et al.* (2016) Analyzing and visualizing genomic complexity for the derivation of the emergent molecular networks. *Int J Monit Surveill Technol Res IJMSTR*, **4**, 30–49.

46. Rougny,A., Touré,V., Moodie,S. *et al.* (2019) Systems biology graphical notation: process description language level 1 version 2.0. *J. Integr. Bioinform.*, **16**.
47. Kalliolias,G.D. and Ivashkiv,L.B. (2016) TNF biology, pathogenic mechanisms and emerging therapeutic strategies. *Nat. Rev. Rheumatol.*, **12**, 49–62.
48. Noack,M. and Miossec,P. (2017) Selected cytokine pathways in rheumatoid arthritis. *Semin. Immunopathol.*, **39**, 365–383.
49. Hwang,D. and Kim,W.-U. (2017) Rheumatoid arthritis: Modelling cytokine signalling networks. *Nat. Rev. Rheumatol.*, **13**, 5–6.
50. Song,G., Ouyang,G. and Bao,S. (2005) The activation of Akt/PKB signaling pathway and cell survival. *J. Cell. Mol. Med.*, **9**, 59–71.
51. Malemud,C.J. (2007) Growth hormone, VEGF and FGF: involvement in rheumatoid arthritis. *Clin. Chim. Acta*, **375**, 10–19.
52. Kim,S.-J., Chen,Z., Chamberlain,N.D. *et al.* (2014) Ligation of TLR5 promotes myeloid cell infiltration and differentiation into mature osteoclasts in rheumatoid arthritis and experimental arthritis. *J. Immunol.*, **193**, 3902–3913.
53. Zhu,W., Meng,L., Jiang,C. *et al.* (2011) Arthritis is associated with T-cell-induced upregulation of Toll-like receptor 3 on synovial fibroblasts. *Arthritis Res. Ther.*, **13**, R103.
54. Elshabrawy,H.A., Essani,A.E., Szekanecz,Z. *et al.* (2017) TLRs, future potential therapeutic targets for RA. *Autoimmun. Rev.*, **16**, 103–113.
55. Cho,M.-L., Ju,J.-H., Kim,H.-R. *et al.* (2007) Toll-like receptor 2 ligand mediates the upregulation of angiogenic factor, vascular endothelial growth factor and interleukin-8/CXCL8 in human rheumatoid synovial fibroblasts. *Immunol. Lett.*, **108**, 121–128.
56. Xu,L., Feng,X., Tan,W. *et al.* (2013) IL-29 enhances toll-like receptor-mediated IL-6 and IL-8 production by the synovial fibroblasts from rheumatoid arthritis patients. *Arthritis Res. Ther.*, **15**, R170.
57. Ivashkiv,L.B. and Hu,X. (2003) The JAK/STAT pathway in rheumatoid arthritis: pathogenic or protective? *Arthritis Rheum.*, **48**, 2092–2096.
58. Schinnerling,K., Aguilón,J.C., Catalán,D. *et al.* (2017) The role of interleukin-6 signalling and its therapeutic blockage in skewing the T cell balance in rheumatoid arthritis. *Clin. Exp. Immunol.*, **189**, 12–20.
59. Malemud,C.J. (2017) Negative regulators of JAK/STAT signaling in rheumatoid arthritis and osteoarthritis. *Int. J. Mol. Sci.*, **18**.
60. Han,Z., Boyle,D.L., Manning,A.M. *et al.* (1998) AP-1 and NF-kappaB regulation in rheumatoid arthritis and murine collagen-induced arthritis. *Autoimmunity*, **28**, 197–208.
61. Clark,A.R. and Dean,J.L. (2012) The p38 MAPK pathway in rheumatoid arthritis: a sideways look. *Open Rheumatol. J.*, **6**, 209–219.
62. Schett,G., Zwerina,J. and Firestein,G. (2008) The p38 mitogen-activated protein kinase (MAPK) pathway in rheumatoid arthritis. *Ann. Rheum. Dis.*, **67**, 909–916.
63. Malemud,C.J. (2013) Intracellular signaling pathways in rheumatoid arthritis. *J. Clin. Cell. Immunol.*, **4**, 160.
64. Higgs,R. (2010) Rheumatoid arthritis: synergistic effects of growth factors drive an RA phenotype in fibroblast-like synoviocytes. *Nat. Rev. Rheumatol.*, **6**, 383.
65. Rosengren,S., Corr,M. and Boyle,D.L. (2010) R65, platelet-derived growth factor and transforming growth factor beta synergistically potentiate inflammatory mediator synthesis by fibroblast-like synoviocytes. *Arthritis Res. Ther.*, **12**.
66. Müller-Ladner,U., Ospelt,C., Gay,S. *et al.* (2007) Cells of the synovium in rheumatoid arthritis. Synovial fibroblasts. *Arthritis Res. Ther.*, **9**, 223.
67. Demoruelle,M.K., Deane,K.D. and Holers,V.M. (2014) When and where does inflammation begin in rheumatoid arthritis? *Curr. Opin. Rheumatol.*, **26**, 64–71.
68. Mellado,M., Martínez-Muñoz,L., Cascio,G. *et al.* (2015) T cell migration in rheumatoid arthritis. *Front. Immunol.*, **6**, 384.
69. Goddard,D.H., Kirk,A.P., Kirwan,J.R. *et al.* (1984) Impaired polymorphonuclear leucocyte chemotaxis in rheumatoid arthritis. *Ann. Rheum. Dis.*, **43**, 151–156.
70. Yoshihara,Y. and Yamada,H. (2007) Matrix metalloproteinases and cartilage matrix degradation in rheumatoid arthritis. *Clin. Calcium*, **17**, 500–508.
71. Ainola,M.M., Mandelin,J.A., Liljeström,M.P. *et al.* (2005) Pan-nus invasion and cartilage degradation in rheumatoid arthritis: involvement of MMP-3 and interleukin-1beta. *Clin. Exp. Rheumatol.*, **23**, 644–650.
72. Yasuda (2006) Cartilage destruction by matrix degradation products. *Mod. Rheumatol.*, **16**, 197–205.
73. Shiozawa,S., Tsumiyama,K., Yoshida,K. *et al.* (2011) Pathogenesis of joint destruction in rheumatoid arthritis. *Arch Immunol Ther Exp (Warsz)*, **59**, 89–95.
74. Sato,K. and Takayanagi,H. (2006) Osteoclasts, rheumatoid arthritis, and osteoimmunology. *Curr. Opin. Rheumatol.*, **18**, 419–426.
75. Schett,G. (2007) Cells of the synovium in rheumatoid arthritis. *Osteoclasts. Arthritis Res. Ther.*, **9**, 203.
76. Goldring,S.R. (2002) Pathogenesis of bone erosions in rheumatoid arthritis. *Curr. Opin. Rheumatol.*, **14**, 406–410.
77. Panagopoulos,P.K. and Lambrou,G.I. (2018) Bone erosions in rheumatoid arthritis: recent developments in pathogenesis and therapeutic implications. *J. Musculoskelet. Neuronal Interact.*, **18**, 304–319.
78. Elshabrawy,H.A., Chen,Z., Volin,M.V. *et al.* (2015) The pathogenic role of angiogenesis in rheumatoid arthritis. *Angiogenesis*, **18**, 433–448.
79. Szekanecz,Z., Besenyei,T., Paragh,G. *et al.* (2009) Angiogenesis in rheumatoid arthritis. *Autoimmunity*, **42**, 563–573.
80. Li,H. and Wan,A. (2013) Apoptosis of rheumatoid arthritis fibroblast-like synoviocytes: possible roles of nitric oxide and the thioredoxin 1. *Mediators Inflamm.*, **2013**, 953462.
81. Ichikawa,K., Liu,W., Fleck,M. *et al.* (2003) TRAIL-R2 (DR5) mediates apoptosis of synovial fibroblasts in rheumatoid arthritis. *J. Immunol.*, **171**, 1061–1069.
82. Firestein,G.S., Yeo,M. and Zvaifler,N.J. (1995) Apoptosis in rheumatoid arthritis synovium. *J. Clin. Invest.*, **96**, 1631–1638.
83. Korb,A., Pavenstädt,H. and Pap,T. (2009) Cell death in rheumatoid arthritis. *Apoptosis*, **14**, 447–454.

84. Kramer,I., Wibulsas,A., Croft,D. *et al.* (2003) Rheumatoid arthritis: targeting the proliferative fibroblasts. *Prog. Cell Cycle Res.*, **5**, 59–70.
85. Jacobs,R.A., Perrett,D., Axon,J.M. *et al.* (1995) Rheumatoid synovial cell proliferation, transformation and fibronectin secretion in culture. *Clin. Exp. Rheumatol.*, **13**, 717–723.
86. Mongan,E.S. and Jacox,R.F. (1964) Erythrocyte survival in rheumatoid arthritis. *Arthritis Rheum.*, **7**, 481–489.
87. Hoksza,D., Gawron,P., Ostaszewski,M. *et al.* (2019) MINERVA API and plugins: opening molecular network analysis and visualization to the community. *Bioinformatics*.
88. Woetzel,D., Huber,R., Kupfer,P. *et al.* (2014) Identification of rheumatoid arthritis and osteoarthritis patients by transcriptome-based rule set generation. *Arthritis Res. Ther.*, **16**, R84.
89. Ma,X. and Xu,S. (2013) TNF inhibitor therapy for rheumatoid arthritis. *Biomed. Rep.*, **1**, 177–184.
90. Monaco,C., Nanchahal,J., Taylor,P. *et al.* (2015) Anti-TNF therapy: past, present and future. *Int. Immunol.*, **27**, 55–62.
91. Barabási,A.-L. (2009) Scale-free networks: a decade and beyond. *Science*, **325**, 412–413.
92. Liu,T., Zhang,L., Joo,D. *et al.* (2017) NF- κ B signaling in inflammation. *Signal Transduct. Target. Ther.*, **2**.
93. Noort,A.R., Tak,P.P. and Tas,S.W. (2015) Non-canonical NF- κ B signaling in rheumatoid arthritis: Dr Jekyll and Mr Hyde? *Arthritis Res. Ther.*, **17**, 15.
94. García,S., Liz,M., Gómez-Reino,J.J. *et al.* (2010) Akt activity protects rheumatoid synovial fibroblasts from Fas-induced apoptosis by inhibition of bid cleavage. *Arthritis Res. Ther.*, **12**, R33.
95. Mountz,J.D., Hsu,H.C., Matsuki,Y. *et al.* (2001) Apoptosis and rheumatoid arthritis: past, present, and future directions. *Curr. Rheumatol. Rep.*, **3**, 70–78.
96. Nevius,E., Gomes,A.C. and Pereira,J.P. (2016) Inflammatory cell migration in rheumatoid arthritis: a comprehensive review. *Clin. Rev. Allergy Immunol.*, **51**, 59–78.
97. Hanlon,S.M., Panayi,G.S. and Laurent,R. (1980) Defective polymorphonuclear leucocyte chemotaxis in rheumatoid arthritis associated with a serum inhibitor. *Ann. Rheum. Dis.*, **39**, 68–74.
98. Hannemann,N., Jordan,J., Paul,S. *et al.* (2017) The AP-1 transcription factor c-Jun promotes arthritis by regulating cyclooxygenase-2 and arginase-1 expression in macrophages. *J. Immunol.*, **198**, 3605–3614.
99. Han,Z., Boyle,D.L., Chang,L. *et al.* (2001) C-Jun N-terminal kinase is required for metalloproteinase expression and joint destruction in inflammatory arthritis. *J. Clin. Invest.*, **108**, 73–81.
100. Thalhamer,T., McGrath,M.A. and Harnett,M.M. (2008) MAPKs and their relevance to arthritis and inflammation. *Rheumatology (Oxford)*, **47**, 409–414.
101. Namba,S., Nakano,R., Kitanaka,T. *et al.* (2017) ERK2 and JNK1 contribute to TNF- α -induced IL-8 expression in synovial fibroblasts. *PLoS ONE*, **12**, e0182923.
102. Bartok,B., Hammaker,D. and Firestein,G.S. (2014) Phosphoinositide 3-kinase δ regulates migration and invasion of synoviocytes in rheumatoid arthritis. *J. Immunol.*, **192**, 2063–2070.
103. Firestein,G.S. (2010) ‘Rac’-ing upstream to treat rheumatoid arthritis. *Arthritis Res. Ther.*, **12**, 109.
104. Shigeyama,Y., Pap,T., Kunzler,P. *et al.* (2000) Expression of osteoclast differentiation factor in rheumatoid arthritis. *Arthritis Rheum.*, **43**, 2523–2530.
105. Seemayer,C.A., Kuchen,S., Neidhart,M. *et al.* (2003) p53 in rheumatoid arthritis synovial fibroblasts at sites of invasion. *Ann. Rheum. Dis.*, **62**, 1139–1144.
106. Tak,P.P., Zvaifler,N.J., Green,D.R. *et al.* (2000) Rheumatoid arthritis and p53: how oxidative stress might alter the course of inflammatory diseases. *Immunol. Today*, **21**, 78–82.
107. Malemud,C.J. (2018) Defective T-cell apoptosis and T-regulatory cell dysfunction in rheumatoid arthritis. *Cells*, **7**.
108. Baier,A., Meineckel,I., Gay,S. *et al.* (2003) Apoptosis in rheumatoid arthritis. *Curr. Opin. Rheumatol.*, **15**, 274–279.
109. Rodríguez-Jorge,O., Kempis-Calanis,L.A., Abou-Jaoudé,W. *et al.* (2019) Cooperation between T cell receptor and toll-like receptor 5 signaling for CD4+ T cell activation. *Sci. Signal.*, **12**.

Systems biology

Automated inference of Boolean models from molecular interaction maps using CaSQ

Sara Sadat Aghamiri ^{1,†}, Vidisha Singh^{1,†}, Aurélien Naldi ², Tomáš Helikar ³, Sylvain Soliman^{4,*} and Anna Niarakis ^{1,*}

¹GenHotel, Département de Biologie, Univ. èvry, Université Paris-Saclay, Genopole, èvry 91025, France, ²Département de Biologie, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), école Normale Supérieure, CNRS, INSERM, Université PSL, Paris 75005, France, ³Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, NE 68588, USA and ⁴Lifeware Group, Inria Saclay-île de France, Palaiseau 91120, France

Associate Editor: Jinbo Xu

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint Authors.

Received on October 30, 2019; revised on April 17, 2020; editorial decision on May 4, 2020; accepted on May 6, 2020

Abstract

Motivation: Molecular interaction maps have emerged as a meaningful way of representing biological mechanisms in a comprehensive and systematic manner. However, their static nature provides limited insights to the emerging behaviour of the described biological system under different conditions. Computational modelling provides the means to study dynamic properties through *in silico* simulations and perturbations. We aim to bridge the gap between static and dynamic representations of biological systems with CaSQ, a software tool that infers Boolean rules based on the topology and semantics of molecular interaction maps built with CellDesigner.

Results: We developed CaSQ by defining conversion rules and logical formulas for inferred Boolean models according to the topology and the annotations of the starting molecular interaction maps. We used CaSQ to produce executable files of existing molecular maps that differ in size, complexity and the use of Systems Biology Graphical Notation (SBGN) standards. We also compared, where possible, the manually built logical models corresponding to a molecular map to the ones inferred by CaSQ. The tool is able to process large and complex maps built with CellDesigner (either following SBGN standards or not) and produce Boolean models in a standard output format, Systems Biology Marked Up Language-qualitative (SBML-qual), that can be further analyzed using popular modelling tools. References, annotations and layout of the CellDesigner molecular map are retained in the obtained model, facilitating interoperability and model reusability.

Availability and implementation: The present tool is available online: <https://lifeware.inria.fr/~soliman/post/casq/> and distributed as a Python package under the GNU GPLv3 license. The code can be accessed here: <https://gitlab.inria.fr/soliman/casq>.

Contact: sylvain.soliman@inria.fr or anna.niaraki@univ-evry.fr

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

1.1 Biological network representations and molecular interaction maps

Biological phenomena can be viewed in the form of interaction networks where components (genes, proteins) are represented as ‘nodes’ and the interactions between components are represented as ‘edges’. Network interactions can be directed or undirected, depending on the biological information available that allows the characterization of the interaction (inhibition or activation) and also the

source and the target node. Representing the complexity of biological regulatory systems using networks enables the analysis of their topology, identifying distinct clusters that may correspond to specific biological processes (‘modules’) and nodes with a high degree of connectivity (‘hubs’), exercising a significant influence on the propagation of biological information (i.e. signal, regulation) (Barabási and Oltvai, 2004; Ideker and Nussinov, 2017; Zhang *et al.*, 2014).

The Systems Biology Graphical Notation (SBGN) scheme uses three different languages for network representation (Le Novère, 2015). First, the activity flow (AF) diagram that is an interaction

network, which includes influence direction and mode of regulation, such as activation and inhibition. Second, the entity-relationship (ER) representation that includes mechanistic details, the direction of influences but no sequential information and third, the process description diagram (PD) which is the most detailed of all, including details of the direction of influences, mechanism of action and the order of events. The SBGN-PD notation scheme is based on ideas first introduced to the field by Hiroaki Kitano and co-workers (2003).

Molecular interaction maps can be used to describe biological mechanisms concisely and effectively. Various molecular maps describing different biological processes (Caron et al., 2010; Fujita et al., 2014; Grieco et al., 2013; Jagannadham et al., 2016; Kuperstein et al., 2015; Mazein et al., 2018; Niarakis et al., 2014; Ogishima et al., 2016; Singh et al., 2018; Tripathi et al., 2015; Singh et al., 2020) have been published, and initiatives have emerged, such as the Disease Maps Project (<http://disease-maps.org>), demonstrating the utility and need of this type of representation of biological knowledge (Mazein et al., 2018; Ostaszewski et al., 2019). Molecular interaction maps can serve as a stand-alone knowledge base, or they can be used as a scaffold for building computational models. Based on information mining, human curation and expert advice, these maps summarize current knowledge about biological pathways in a process description representation, while accounting for as many mechanistic details as possible. They provide a comprehensive template for visualization and analysis of omics datasets, and can also be analyzed in terms of the underlying network structure. However, their static nature cannot account for the coordination of multiple biological processes, or how the regulation of several nodes due to the presence or absence of certain factors can alter the functional outcome (i.e. activation of a particular pathway following the repression of a given factor). These regulations that fine-tune the molecular interactions are of great importance as dysregulation or disruption can lead to disease (Cho et al., 2012; Furlong, 2013).

1.2 Boolean models for dynamical studies

Systems Biology approaches and especially computational modelling can be used to provide an executable, dynamic network that can reveal hidden properties and account for emerging system-level behaviours through *in silico* simulations and perturbations (Azeloglu and Iyengar, 2015; Helikar et al., 2008). Each interaction is described using mathematical formalism and the obtained machine-readable model can be used to test novel hypotheses and predict new features of the system of interest. Boolean models are well suited for addressing the lack of kinetic data and handling the large size of the biological pathways described in molecular interaction maps. These models are parameter-free; nevertheless, their simplistic nature can provide a powerful tool for dynamic analysis (Abou-Jaoudé et al., 2016; Furlong, 2013). In Boolean formalism, the simplest form of logical models, nodes represent regulatory components (proteins, enzymes, complexes, transcription factors, genes, to name a few) and arcs represent their interactions. Each regulatory component is associated with a Boolean variable (taking the values 0 or 1) denoting either its qualitative concentration (0 for absent or 1 for present) or its level of activity (0 for inactive or 1 for active). The future state of each node depends on the state of its upstream regulators and is defined by a Boolean function. The function is expressed in the form of a rule using the logical operators AND, OR and NOT. The updating of the rules can be in a synchronous, deterministic mode where all nodes are updated at the same time (Glass and Kauffman, 1973; Kauffman, 1969) or in an asynchronous mode, where only one node can be updated every time (Thomas, 1973, 1978; Thomas et al., 1976).

1.3 Bridging the gap between static and dynamic representations

The construction of a molecular interaction map and a dynamic model are two tasks that can serve different purposes and are usually performed independently. On the one hand, it is a question of

creating a knowledge base in the form of a comprehensive molecular map, and on the other of defining the underlying mechanism that links the system components and captures its dynamic behaviour. Nevertheless, these two constructs share much information, including the mode of influence (e.g. activation or inhibition) and the topology of the network. Molecular maps can be built using a structured diagram editor for drawing gene-regulatory and biochemical networks, such as CellDesigner (Funahashi et al., 2003). Networks in CellDesigner are drawn as process description diagrams (PD) and are stored using the Systems Biology Markup Language (SBML), a standard for representing models of biochemical and gene-regulatory networks (Hucka et al., 2003).

The idea of obtaining executable models from a network topology is not new. In the study by Büchel et al. (2013), researchers proposed a pipeline for the automatic generation of models using KEGG pathways as a resource. They succeed in producing SBML and Systems Biology Marked Up Language-qualitative (SBML-qual) files but these constructs can be seen as model scaffolds as they require further parameterization to become executable. In Mendoza and Xenarios (2006), a Standardized QUALitative Dynamical system (SQUAD) is obtained directly from an input network that is already a regulatory network and not a molecular interaction map. Furthermore, the aim is to obtain a continuous system corresponding to it, implying a small-scale network (about 20–30 nodes). Regarding Biolayout, now Graphia (Livigni et al., 2018), researchers use the modified Edinburgh Pathway Notation scheme (mEPN) to create SBML-like maps that they interpret directly as Petri nets. This approach imposes that all ‘logics’ are conjunctive. There is no direct negation, no disjunction, whereas the only firing rule in a Petri net is that all input places should be filled in order for the reaction to fire. However, molecular maps contain much more precise information (e.g. inhibitions) that cannot be expressed directly within this framework. Moreover, Petri nets are by nature quantitative, requiring several tokens to be assigned to each place, and having the consumption of some tokens by each rule. The rxncon language (Romers and Krantz, 2017) also tackles the idea that there are standard features between maps as knowledge-bases and executable Boolean models. However, their approach is quite different from ours in that they bridge this gap through an intermediate language based on Boolean bipartite graphs. One of the most important consequences is that the logical rules (contingencies in rxncon) are already part of the input (the map being, in a way, already a model). Finally, the <http://pd2af.org/> initiative (Vogt et al., 2013) proposes to translate an SBGN-PD graph, similar to a CellDesigner map, into an SBGN-AF graph, similar to the structure of a Boolean model, but does not go further as to propose an executable model. We will detail in the discussion some specific rules for which we have made similar or opposite choices concerning the graph transformation. However, one should note that our method adds the layer of inferring logical rules for the obtained model based on the original topology and annotations, making possible immediate simulations and analyses using the corresponding tools [e.g. GINSim (Chaouiya et al., 2012) and Cell Collective (Helikar et al., 2012)].

In this work, we present CaSQ (CellDesigner as SBML-qual), a tool for automated inference of large-scale, parameter-free Boolean models, from molecular interaction maps with preliminary logic rules based on network topology and semantics. CaSQ is, to the best of our knowledge, the first tool that produces executable molecular networks of hundreds of nodes (at least up to eight hundred), in the SBML-qual format that can be further simulated and analyzed using popular modelling tools.

2 Materials and methods

2.1 CaSQ

CaSQ is a tool that can convert a molecular interaction map built with CellDesigner (Funahashi et al., 2003) to an executable Boolean model. The tool is developed in Python and uses as source the xml file produced by CellDesigner (SBML plus CellDesigner-specific annotations) to infer preliminary Boolean rules based solely on

network topology and semantic annotations (e.g. certain arcs are noted as catalysis, inhibition, etc.). The aim is to convert a Process Description (PD) representation, i.e. a reaction model, into a complete logical model. The resulting structure is closer to an AF diagram, though not in a strict SBGN-PD to SBGN-AF notion. Moreover, logical rules that make the model executable are also obtained. For illustrating the rules of the conversion, we use the repertoire of notation schemes in CellDesigner (Fig. 1).

The conversion of the graph to an executable model is a four-step process:

Step 1: First, the map is reduced through a pass of graph-rewriting rules. These rules are executed in order and in a single pass, so the rewriting is terminating and confluent. The reasoning behind this reduction is that a single qualitative species of the logical model often represents by its state (active/inactive) several species of the original map. Therefore, those species might need to be merged into a single component or some inactive forms to be completely discarded to avoid redundancy in the logical model. The rules are the following:

Rule 1: If two species of the map are only reactants in a single reaction, i.e. do not take part in any other reaction, if that reaction is annotated as heterodimer association, and if one of the reactants is annotated as a receptor, then the receptor is deleted from the map (its annotations are added to the product of the reaction) (Fig. 2);

Rule 2: If two species of the map take part in a reaction annotated as heterodimer association, **if none of them** are annotated as

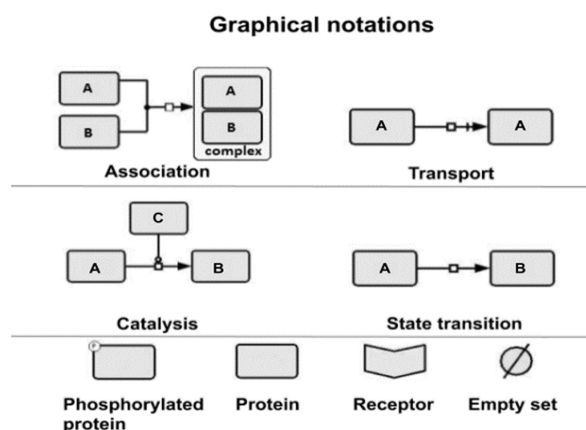


Fig. 1. The repertoire of CellDesigner graphical notation schemes used to illustrate CaSQ's rules. For CaSQ's conversion rules, we use the notation schemes for association, transport, catalysis, state transition and also the glyphs for receptor, protein, modified protein (here, we show phosphorylation as an example) and the empty set. The empty set can account for degradation or in SBGN-PD terms, can represent the creation (respectively, the disappearance) of an entity from an unspecified source (resp. sink) that we do not need or wish to explicit

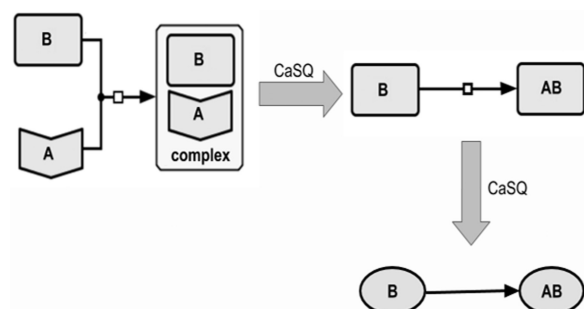


Fig. 2. Illustration of the 1st rule. If two species of the map are only reactants in a heterodimer association, and if one of the reactants is annotated as a receptor, then the receptor is deleted from the map (its annotations are added to the product of the reaction)

receptor, and **if both** do not take active part (i.e. reactant or modifier) in any other reaction, then both are merged into the complex, product of the reaction (their annotations are added to the product, and the reactions that had them as product are rewired to have the complex as product) (Fig. 3);

Rule 3: If one species only appears in a single reaction, if it appears there as a reactant if that reaction has a single product, and if both the reactant and the product have the same name, then the reactant is deleted (its annotations are merged into those of the product) (Fig. 4);

Rules 2 and 3 can be combined resulting in greater graph compression, as illustrated in Figure 5.

Rule 4: If one species only appears as a reactant in a single reaction (but maybe appearing as product in another reaction) that has a

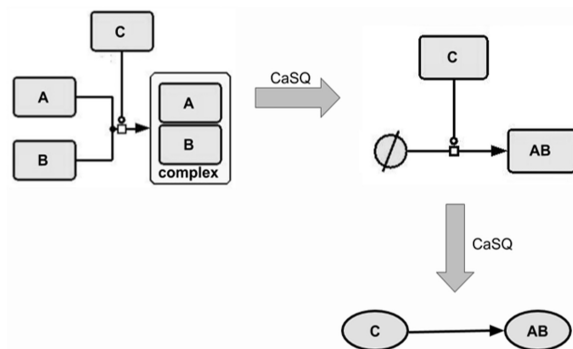


Fig. 3. Illustration of the 2nd rule. Compression of the complex formation, where none of the reactants is denoted as a receptor, and both reactants do not participate in any other reaction. As a result, both reactants are removed and modifiers are rewired to have the complex as a product

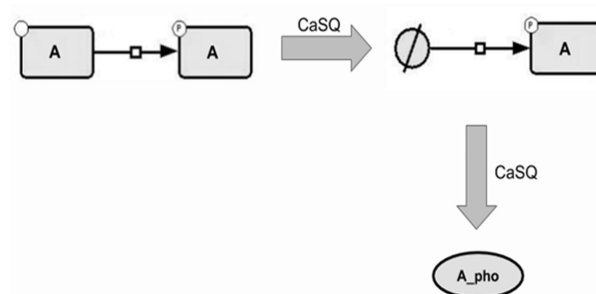


Fig. 4. Illustration of the 3rd rule. Removing inactive forms that do not participate in other reactions

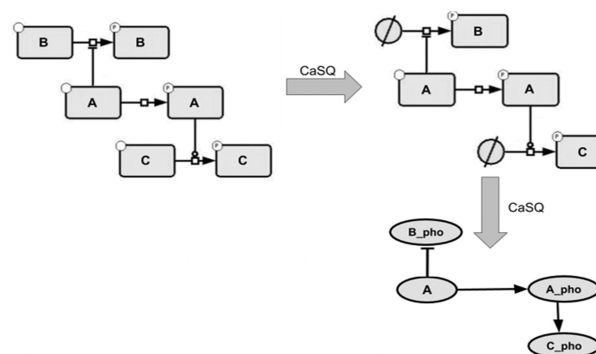


Fig. 5. Combination of rules 2 and 3. CaSQ retains components that contribute further to the propagation of the signal

single product and is annotated as transport, and if both the reactant and the product have the same name, then the reactant is merged into the product (its annotations are merged into those of the product, and the reactions producing it are rewired to the product) (Fig. 6).

The rationale of using the name to identify the same components in different states (gene, RNA, protein, transported/phosphorylated/methylated protein, etc.) is that we need to identify when species can be merged/discarded, to keep only what contributes further to signal propagation. However, relying on the *active* annotation (dotted circle) in CellDesigner maps proved to be insufficient: not all map curators use this notation, and it is not SBGN compliant.

Step 2: The topology of the model is then computed as a simple form of PD to AF conversion, with one qualitative species corresponding to each species in the reduced map obtained from Step 1. This species inherits the original map layout, using SBML3 Layout package, and MIRIAM annotations (e.g. PubMed IDs as bqbiol:isDescribedBy). The annotations have been associated with each regulated component rather than each regulation, mostly because tools supporting the latter are quite rare. All reactants and modifiers of a reaction exert a positive influence on all the products of that reaction, whereas all inhibitors exert a negative influence. Compared to the formal abstraction of influence graphs from reaction graphs (Rizk et al., 2011), note that, the mutual inhibition between reactants is purposely ignored as in Step 1 we already condense active and inactive forms of the same species.

Step 3: The logical rules of the model are computed. For each species, its logical rule is defined as the (i) disjunction (OR) for all reactions producing it, of (ii) the disjunction (OR) for all positive modifiers of a reaction being on and (iii) the conjunction (AND) of all products of that reaction being activated and all inhibitors being inactive. Therefore, a target is on if one of the reactions producing it is on, a reaction is on if all reactants are on, all inhibitors are off and one of the catalysts is on (Supplementary Fig. S1).

Step 4: Model refinement is performed through the optional removal of unconnected components. From our experience, keeping only the biggest connected component is what makes the most sense from a modelling perspective. However, it is possible to specify a

'minimum size' and keep all connected components above that size. Names of the qualitative species are also made more precise by adding the original type/modifications of the species (e.g. RNA, phosphorylated) and if there are still homonyms the original compartment is added too. More precisely, the name of the node in the model is, therefore, the name of the species in the map to which is added (separated by an underscore character '_'), its type as given in the map (RNA, Gene, etc.) unless that type is 'PROTEIN' and to which is added modifications given by the map (phosphorylation, methylation, etc.). If after that step, several species from the model are found to have the same name, the compartment is added too (once again, separated by an underscore) (Supplementary Tables S1 and S2).

CaSQ generates two output files; the proper logical model encoded in SBML-*qual*, a format that is compatible for further analysis with modelling tools such as GINsim (Chaouiya et al., 2012) or Cell Collective (Helikar et al., 2012), and a CSV file that contains information about the names, the logic formulae and the CellDesigner alias. The second file is mostly for automated treatment. The SBML-*qual* file can also be restricted to include only its biggest connected component (BCC), or only connected component above a given size threshold. This allows the modeller to obtain a more meaningful logical model even if the original map did contain several unconnected clusters corresponding to isolated pieces of information.

2.2 Molecular interaction maps and logic models

For testing the applicability of CaSQ, we used various molecular interaction maps that differ in size, complexity and use of SBGN notation, as shown in Table 1. Namely, we used one molecular interaction map comprising 125 nodes describing mast cell activation (Niarakis et al., 2014), one map comprising 232 nodes for MAPK activation (Grieco et al., 2013), one for cholecystokinin signaling with 530 nodes (Tripathi et al., 2015) and finally two large-scale molecular maps, one for rheumatoid arthritis (RA)—the only SBGN-compliant—(Singh et al., 2018, 2020) comprising 779 nodes, detailed annotations and references in the MIRIAM and text annotation section of the CellDesigner file (Funahashi et al., 2003) (Supplementary Fig. S2) and the Alzheimer's pathway map with 1361 nodes (Ogishima et al., 2016). The mast cell activation and the MAPK maps were published along with their corresponding manually built logical models.

2.3 Model comparison

For evaluating the performance of the tool, we compared size and shared nodes between manually built models that corresponded to the interaction maps (for mast cell and MAPK), with the CaSQ-inferred Boolean models. While size reduction is not the primary goal of the tool, it remains a measure of comparison between the process description static diagram of the original map and the regulatory graph that the tool produces after the conversion rules. Conversion from a process description to an AF diagram implies a more compact network. The comparison allows us to check if such compression was achieved. We also performed simulations to see if the CaSQ-inferred models were able to reproduce known biological

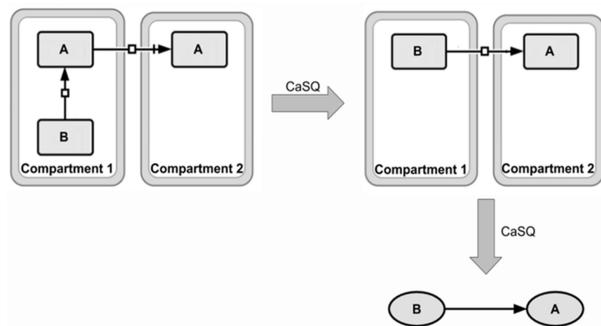


Fig. 6. Combination of the 2nd and the 4th rule. Components that are translocated across other compartments (e.g. transcription factors) are merged in one component that inherits all influences, provided that the original component does not participate in another reaction/regulation

Table 1. Size (number of components) of the CaSQ-inferred model using the default and BCC options

Map name	Map size	SBGN use	CaSQ-inferred model			
			Size	Graph reduction (%)	BCC size	Graph reduction(%)
Mast cell ^a	125	No	80	36	73	42
MAPK ^a	232	No	182	21	181	22
Cholecystokinin	530	No	404	24	383	28
RA	779	Yes	431	45	391	50
Alzheimer's	1361	No	1169	14	762	44

^aThe existence of a corresponding manually built logical model.

scenarios, and finally, we compared steady states, where feasible, between the inferred and the manually built models.

2.4 *In silico* simulations and calculation of stable states

For the simulations of the CaSQ-derived models, we used Cell Collective, a web-based, modelling platform for the collaborative construction, simulation and analyses of large-scale dynamic models (Helikar *et al.*, 2012). Models in Cell Collective can be created either *de novo* or they can be imported using the SBML-*qual* standard. Cell Collective SBML-*qual* import supports network layout, as well as model annotations. References stored in the MIRIAM section of the xml file of CellDesigner can be retrieved and visualized in the platform (Supplementary Fig. S3).

For the computation of stable states, we used GINsim (Chaouiya *et al.*, 2012), powerful software for constructing and analyzing logical models. GINsim can import SBML-*qual* files; however, it needs a pre-processing step to display the name and not the species IDs. Imported models retain their formulae, as well as the layout but are currently stripped from annotations during pre-processing.

3 Results

3.1 Graph reduction and model inference

We first tested the tool with different molecular maps of various sizes, complexities and use of standards to see if CaSQ was able to produce corresponding executable models. We performed the analysis with CaSQ first by default and then using the BCC option. While a model should be connected to be useful, a map can include unconnected parts as the objective of a map is to represent all current knowledge for the studied biological process and this knowledge is more likely to be fragmented. The purpose of using CaSQ with default and BCC options was also to evaluate the graph reduction capacities of the tool. The size was defined by the number of nodes included in the map (number of species in the CellDesigner files), and the number of components included in the published, manually built or CaSQ-inferred models.

CaSQ was able to handle small-, medium- and large-scale maps (ranging from 125 to 1361 nodes) with or without SBGN standards, and produce executable models smaller in size, offering a graph reduction of 21–45%. Using the BCC option that allows keeping the biggest connected component, the resulting models are slightly smaller. The size of the produced model—in terms of the number of components included—using BCC option is highly dependable on the connectivity of the initial map (Table 1).

3.2 CaSQ run time

The analysis was performed on a Dell working station with Windows 7, 64-bit Operating System, Installed memory (RAM): 64.0 GB and Processor: Intel (R) Xeon (R) CPU E5-1650 v4 @ 3.60 GHz. The run times of CaSQ for producing executable SBML-*qual* files with default options are 1.42 s for the mast cell activation map, 1.10 s for the MAPK map, 1.71 s for the Cholecystokinin, 2.29 s for the RA map and 5.24 s for the Alzheimer's map.

3.3 CaSQ-inferred Boolean models versus manually built models

3.3.1 Shared nodes

To evaluate the tool's ability to produce preliminary Boolean rules, we compared the CaSQ-inferred models with the manually built models (MM) published with the respective maps. First, we compared the size and graph reduction percentage (Table 2). For the size, we compared the shared nodes between the two models. The automated comparison gives the number of identical node names while the manual comparison accounts for differences in node names that derive from the fact that the manually built models do not correspond 100% to the maps. A modeller may choose to merge two nodes (i.e. receptor–ligand), change the name of one node (i.e. use capitals or add underscores for a complex), entirely skip it or add a node that does not exist in the initial map, making it difficult to evaluate in a fully automated way the correspondence between the manually built and the CaSQ-derived models. Manual comparison by visual inspection after the automated comparison revealed many cases where the node names were slightly different but corresponded to the exact protein or gene (Supplementary Tables S1 and S2). For example regarding the mast cell activation models, the manual model has RAS but the CaSQ model has H-RAS. Other cases concern grouping of instances, i.e. FYN in the manually built model corresponds to more instances in the CaSQ one, as the latter includes FYN with different modifications (phosphorylated, palmitoylated). For the MAPK model, an example is p53 in the manual model that corresponds to TP53 and TP53 phosphorylated in the CaSQ counterpart, or SMAD in the manually built that corresponds to a grouping of different SMAD proteins. An additional problem that made the comparison difficult was the fact that the researchers made different decisions concerning their map and model building. For instance, the receptor tyrosine kinase (RTK) component in the MAPK map represents several different receptors (e.g. EGFR, FGFR, VEGFR, etc.) while in the model they use explicitly the different receptors.

The two models used for CaSQ's benchmarking are medium-sized models (47–53 nodes). CaSQ models are twofold to fourfold bigger because they are inferred automatically from the corresponding maps (Table 2).

The CaSQ-inferred model for mast cell activation comprises 73 nodes while the manually built, 47 nodes. The authors of the manually built extracted information from the molecular map, but they also used proteomic data from bone marrow mononuclear cells (BMMCs) reported in Bounab *et al.* (2013) that focused on the SLP-76 protein and its partners. Node comparison revealed that 30 of these nodes are shared between the CaSQ inferred and the manually built models (Supplementary Table S1).

3.3.2 *In silico* simulations and dynamic analysis

Next, we simulated CaSQ-inferred models to see if they were capable of capturing the system's dynamics even though they were not identical with their manually built counterparts.

3.3.2.1 Comparison of the CaSQ-inferred model and the manually built model for mast cell activation. One important difference, besides size and logical formulae, is also the fact that the mast cell activation model contained one multivariate variable while CaSQ-inferred models are strictly Boolean. Despite the differences, CaSQ mast cell model was able to reproduce the Btk (Fig. 7a) and Syk (Fig. 7b) knockout experiments described in the publication (Niarakis *et al.*, 2014).

Table 2. Comparison of CaSQ-inferred Boolean models with manually built models (MM)

Map name	Map size	SBGN use	MM		CaSQ-inferred model BCC		Common nodes (%)
			Size	Graph reduction (%)	Size	Graph reduction (%)	
Mast cell	125	No	47	62	73	42	64
MAPK	232	No	53	77	181	22	79

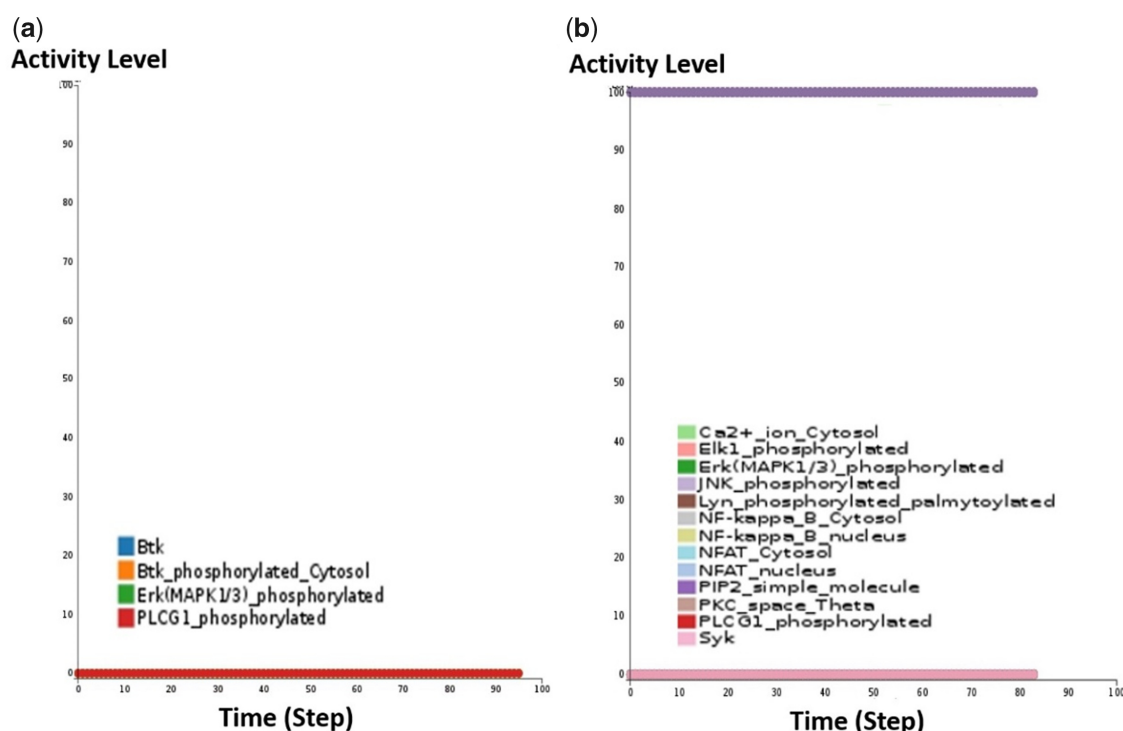


Fig. 7. (a) Screenshot of simulations for Btk knockout of the CaSQ-derived mast cell activation model using Cell Collective. When Btk is set to zero, Erk and PLCG1 are not expressed. (b) Screenshot of simulations for Syk knockout of the CaSQ-derived mast cell activation model using Cell Collective. When Syk is set to zero, Erk, JNK, NFAT, NFkB, Ca2+, PKC, Elk1, PLCG1 are not expressed

In Figure 7, we see simulation examples of the CaSQ-inferred model for mast cell activation in Cell Collective.

In the case of Btk knockout, a decrease in cytokine release and degranulation, as well as a decrease of PLCG1 and ERK levels have been observed (Kajita *et al.*, 2010; Setoguchi *et al.*, 1998). The simulation of Btk knockout using Cell Collective platform resulted in PLCG1 and ERK set to zero, a result that is directly comparable with the simulation described in Niarakis *et al.* (2014) (Fig. 7a).

In Syk knockout experiments, cytokine release and degranulation are both abolished (Gilfillan and Tkaczyk, 2006). We performed an *in silico* simulation of Syk knockout, with Lyn and PIP2 present at the initial state in Cell Collective as described in Niarakis *et al.* (2014) (Fig. 7b). In this condition, the CaSQ-inferred model reaches a state where ERK, JNK, Elk-1, NF-kB, NFAT, PKC, PLCG1, Ca2+ are all set to zero, in agreement with the simulation described in Niarakis *et al.* (2014).

3.3.2.2 Logical steady-state analysis for the mast cell activation models. We computed all the stable states of both the CaSQ-inferred model and the manually built one for mast cell activation using bioLQM java toolkit included in GINsim (<http://colomoto.org/biolqm/>). We obtained 18 stable states for the manually built model (Supplementary Fig. S4) and 524.288 for the CaSQ-inferred one. The difference in the number of stable states lies in the fact that the automatically inferred model is a close representation of the system as described in a molecular map and thus significantly bigger in size, including especially a much higher number of inputs. The manual counterpart is smaller in size and also of reduced complexity as several inputs are grouped and thus, the computation of stable states leads to considerably fewer solutions.

As shown in Supplementary Table S1, 30 components can be matched together between these two models. We then projected the identified stable states on these 30 components, which reduced the lists to nine stable states for the manually built model and 43.392 for the CaSQ-inferred one. Indeed, some of the original stable states only differ in the unmatched components and are thus projected on the same state. We found that three of the nine stable states of the manually built model are precisely reproduced in the CaSQ-inferred

model. If we accept a single difference between the states, we can recover four additional stable states, whereas the last two stable states can be recovered with two differences (Supplementary Table S3).

3.3.2.3 Comparison of the CaSQ-inferred model and the manually built model for MAPK. Concerning the MAPK manually built model, the authors produced a model that did not follow strictly the corresponding map (the model contained several merged inputs and merged outputs).

As stated above, the RTK component in the map represents several different receptors like EGFR, FGFR and VEGFR that the researchers decided to include in the model explicitly. Besides, to cope with simulations of their model, they used the model reduction option in GINsim (Grieco *et al.*, 2013) to produce different smaller sub-versions of the original model, each dedicated to a subset of simulations. In Table 3, we have regrouped biological scenarios modelled successfully with the MAPK manual model and the corresponding behaviour of the CaSQ counterpart. For the simulations of the CaSQ model, we used the platform Cell Collective as before (Fig. 8).

These reduced versions of the original MAPK model (52 components) ranged from 16 to 18 components. The CaSQ-inferred model for MAPK is inferred directly from the MAPK map and is thus significantly bigger in size and different in structure. However, comparison of the model's behaviour regarding its efficacy in capturing the systems dynamics, showed that the CaSQ model, was able to reproduce partially or completely known biological scenarios.

The size of the CaSQ-inferred MAPK model (181 nodes) made the calculation of stable states a non-realistic endeavour. Moreover, the fact that the manually built counterpart had to undergo multiple reductions for the dynamic analysis, would not have made the comparison straightforward.

4 Discussion

Building large-scale dynamic models can be tedious and time-consuming work that requires not only the construction of the regulatory graph but also the writing and tuning of the logical formulae.

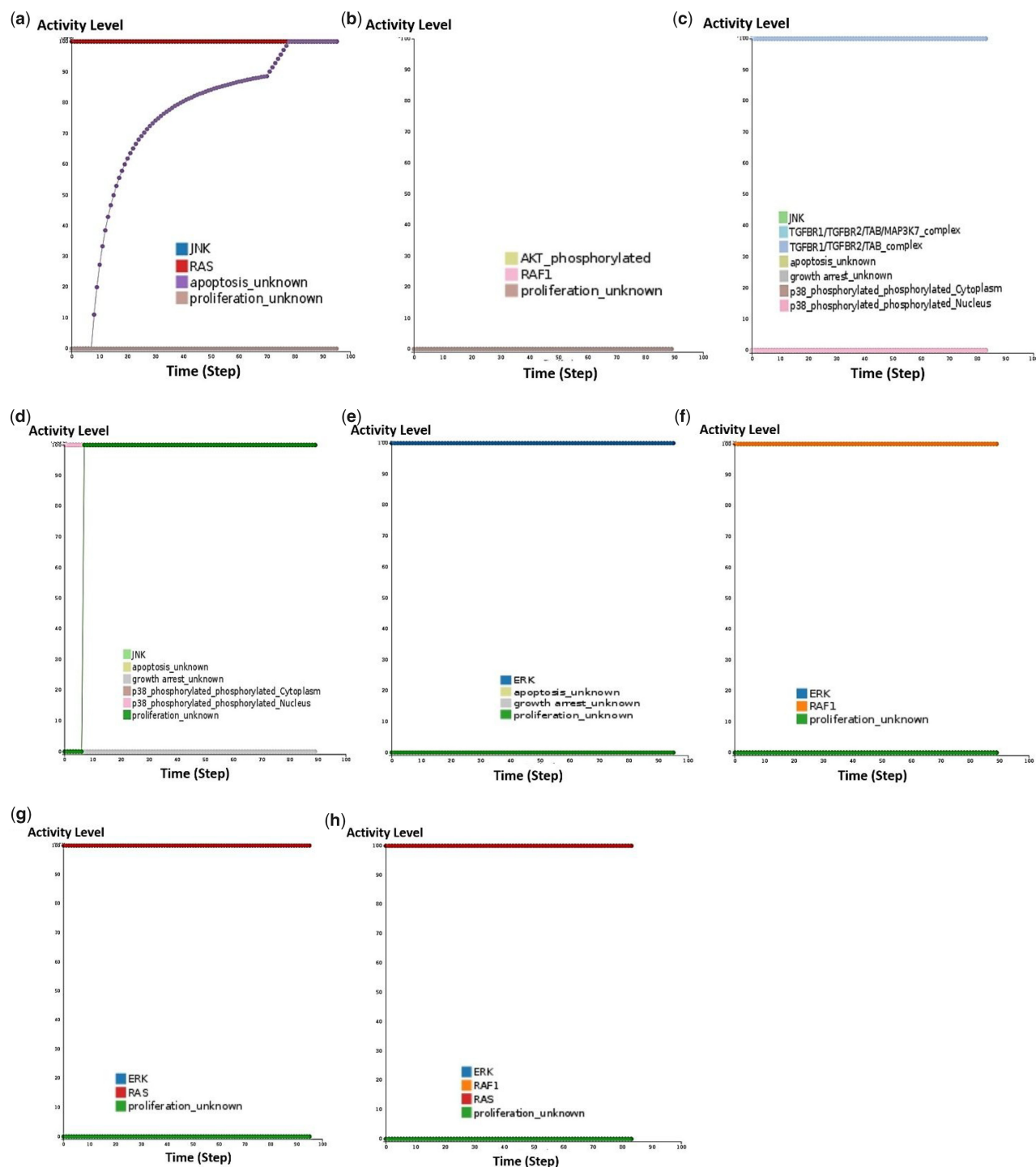


Fig. 8. Simulations of the CaSQ-inferred model using the modelling platform Cell Collective. The CaSQ-inferred model for MAPK was able to reproduce known biological scenarios, either completely or partially. The results of the *in silico* simulations for the three first biological conditions described in Table 3 showed perfect agreement with the results of manually built model, as depicted in a, b and c. For conditions described in scenarios 4 and 5 of Table 3, the CaSQ-inferred model could partially reproduce the attended behaviour (d and e) while simulation results for scenario 6, were inconsistent with the literature and the results of the manually built model (f, g and h)

CaSQ is a tool aiming to ease the construction of large-scale Boolean models, taking advantage of the similarities shared between molecular interaction maps and dynamic models. First of all, the molecular maps are process description representations that can be well annotated, providing a critical source of knowledge. The maps also contain information about the interactions, catalyzes, activations and inhibitions of the network, essential for the building of a computational model. In the framework proposed, we utilize

systems biology standards for model construction (SBML-qual), so that CaSQ tool can be interoperable with other tools and modelling software.

An attempt to produce automatically large-scale models (kinetic and logical) has been made with the Path2Models (Büchel *et al.*, 2013) where researchers proposed a pipeline for the automatic generation of models using KEGG pathways as a resource. For metabolic pathways, they produced SBML files which they complemented

Table 3. Biological data and corresponding behaviours of the manually built and the CaSQ-inferred models for MAPK

Biological data	Manually built MAPK model	CaSQ-inferred MAPK model	Agreement
1. JNK might reduce RAS-dependent tumour formation by inhibiting proliferation and promoting apoptosis (Kennedy and Davis, 2003)	<i>When JNK is always ON and RAS is always ON then proliferation is OFF and apoptosis is ON</i>	<i>When JNK is always ON and RAS is always ON then proliferation is OFF and apoptosis is ON (Fig. 8a)</i>	Yes
2. HSP90 inhibitor disrupts EGFR, RAF and AKT leading to successful cancer treatment (Sharp and Workman, 2006)	<i>Concomitant RAF, EGFR, AKT deletions block proliferation</i>	<i>There is no EGFR present in the model, RAF and AKT deletions lead to proliferation being OFF (Fig. 8b)</i>	Yes
3. P38 and JNK play important roles in stress responses such as cell cycle arrest and apoptosis (Kyriakis and Avruch, 2001; Takekawa et al., 2011)	<i>When p38/JNK are OFF (KOs) and TGB and DNA damage are ON then there is no growth arrest or apoptosis</i>	<i>There is no DNA damage present in the model, p38/JNK constitutively OFF and TGF stimuli ON, then Growth arrest is OFF and Apoptosis is OFF (Fig. 8c)</i>	Yes
4. P38 and JNK, especially in the absence of mitogenic stimuli, have been shown to induce apoptotic cell death (Kyriakis and Avruch, 2001; Takekawa et al., 2011)	<i>When P38/JNK are constitutively ON then Growth arrest is ON, Apoptosis is ON and proliferation is OFF</i>	<i>When p38/JNK are constitutively ON then Growth arrest is OFF, Apoptosis is ON and proliferation ON (Fig. 8d)</i>	Partial
5. ERK increases transcription of the cyclin genes and facilitates the formation of active Cdk/CDK complexes, leading to cell proliferation (Schramek, 2002)	<i>When ERK is always ON then Apoptosis and Growth arrest are OFF, and proliferation is ON</i>	<i>When ERK is constitutively ON then Apoptosis and Growth arrest are OFF, and proliferation is OFF (Fig. 8e)</i>	Partial
6. RAF or RAS overexpression can lead to constitutive activation of ERK (Dhillon et al., 2007)	<i>When either RAS or RAF are constitutively active then ERK is ON and proliferation is ON</i>	<i>When either RAF or RAS or both of them are constitutively active, then ERK is OFF and proliferation is OFF (Fig. 8f-h)</i>	No

where possible with kinetic data from respective databases, while for non-metabolic pathways, they produced SBML-qual files that could serve as scaffolds for logical models. These scaffolds do not contain logical rules, only topological relationships and interaction signs. In our pipeline, that requires only one tool, CaSQ, we start from detailed, mechanistic, process description diagrams and we produce fully executable large-scale logical models, with logical formulae for all components.

The methodology described in SQUAD (Mendoza and Xenarios, 2006) is complementary to what we propose and can be used in some parts of the obtained logical model if more quantitative evaluation is deemed necessary. For the inference of the logical formulae, we based our assumptions on topology and semantics of the molecular maps. More precisely, we decided to approach the conversion process using mostly OR gates over AND, so a target is on if one of the reactions producing it is on, a reaction is on if all reactants are on, all inhibitors are off and one of the catalysts is on. The idea behind this assumption is that very rarely we have exact information about the need for the presence of two or more activators for one target. Even if synergy is defined, very often a relative activation can happen even by the presence of one activator. Moreover, the number of events for which we do have such information is significantly lower than the uncertain ones and tuning the rules by hand should be a quick process.

The graph transformation rules that we use share some similarities with the rules used in <http://pd2af.org>, yet there exist significant differences: first, we do not address oligomerization as a specific case; instead, we chose to have a generic simplification for all complexes. On the contrary, we propose specific rules for receptors, as many of our use-cases have a signalling part which requires domain-specific rules. Concerning translocation, PD2AF does not make any simplification, whereas in our method, we have added a specific transport rule, as in the maps we treated we often encountered the case where an inactive form of a species is moving to another compartment and then becoming active (e.g. transcription factors). Ignoring the inactive version in the model did appear to correspond to what was done manually by the modellers in most of the cases studied.

Regarding activation and inhibition rules of PD2AF, our rules often agree except that we never extract the ‘hidden inhibition’ (or its converse): if there is an inhibition in the map, there will be an inhibition in the model, if there is an activation in the map, there is an

activation in the model. While we understand the idea behind the PD2AF reasoning for this rule, the fact that it results in deleting the products of some reactions is in contrast with the reasoning behind CaSQ, which only deletes inputs. This is linked to the fact that an ‘inactive’ product can be a meaningful output of the map/model.

Finally, the most common catalytic reaction rule of PD2AF is different from our choice on several accounts. First, it uses a single state transition for all products of the reaction, which is not in the SBGN-AF standard. Furthermore, this single transition with multiple outputs makes it impossible to obtain specific logical rules for each of the outputs. In contrast, our methodology will duplicate the effect of reactants, activators and inhibitors for all products, i.e. create as many copies of the transition as there are products, and then combine this transition with all other transitions on each of those products. Moreover, the case of several activators/inhibitors is not covered by PD2AF, whereas we made a specific choice on how to combine them in a logical rule (AND’ing the reactants, OR’ing the activators and AND’ing the NEGation of all inhibitors). Finally, the most significant contrast to PD2AF, as already stated above, is that our resulting model is executable since it has inferred logical rules for each node.

Manually built models that are based on corresponding molecular maps are usually small to medium size because simulating a large-scale Boolean model remains challenging, even if the model is parameter-free. This means that the modeller is obliged to prioritize and choose nodes over others in order to create abstractions that can be subsequently analyzed. With the use of CaSQ, as demonstrated in this study, we can now obtain large-scale Boolean models that can be executed using popular modelling software that can import SBML-qual files. However, challenges associated with the analysis of large-scale Boolean models exist, and are active topics of efforts in the field. For coping with size and complexity one can perform reductions and create different versions of the original model [as demonstrated in Grieco et al. (2013)].

In this work, for comparing the tool’s performance and accuracy, we compared the common nodes between the CaSQ inferred and the manually built models, their ability to reproduce biological scenarios performing simulations, and finally, we performed a comparison of stable states, where possible. One problem we encountered when searching for common nodes was that the automatic comparison was not sufficient as a human modeller may choose different naming (e.g. merge two or more components). The automated

comparison gave us an idea about the identical names and formulae, but a manual inspection was also compulsory as it revealed many cases where the corresponding nodes were present in both models, under slightly different naming. We also performed simulations to see if the CaSQ-inferred models could reproduce some of the dynamics of the original system. The next step was to perform logical steady-state analysis. For this purpose, we used GINsim, powerful software for logical modelling. The goal was to see if within the stable states of the CaSQ-inferred model, we could retrieve the stable states of the published manually built model.

We should note that CaSQ infers preliminary Boolean rules, so the modeller still needs to fine-tune the model and find the best logical rules to reproduce data accurately. Bekkar *et al.* (2018) show that logical models with added human curation perform better than models where rules are extracted automatically from a given topology. As demonstrated in the results, the CaSQ tool produces models that are largely in agreement with the model a human modeller would build, accelerating the time of model construction impressively.

This work was also a motivation for community work, as it addressed issues of model reusability, use of Systems Biology standard formats and interoperability between different tools that have complementary functionalities. As demonstrated, our method is scalable, and the large-scale SBML-qual models produced by CaSQ can be imported in Cell Collective and retain layout and annotations. However, the current import to GINsim requires a process that removes annotations and references before the analysis. Moreover, this process provides a solution for name display as GINsim displays species IDs that in our case make the model unreadable. The proper handling and reuse of annotations between different software tools could benefit from further interoperability work. The goal is to propose a seamless pipeline for producing executable Boolean models starting from molecular interaction maps which can be analyzed in depth using various tools for computational modelling. CaSQ tool can play the role of a bridge bringing together two distinct communities, curators and modellers to produce interoperable, annotated models of better quality, accuracy and reusability.

5 Conclusion—future prospects

CaSQ is a new tool for automated inference of Boolean models from CellDesigner molecular interaction maps. The rules defined for the translation have proven to be efficient to account for various biological scenarios, such as complex formation, protein activation, gene expression and transcription factor translocation. The obtained ‘raw’ models, with preliminary Boolean rules are able to reproduce complex behaviours and capture some of the systems dynamics. CaSQ can handle molecular maps varying significantly in terms of size, complexity, level of annotations and use of SBGN standards, with short run times. Finally, the obtained Boolean models retain the hierarchical layout of the map and its references in a standard format, SBML-qual, assuring model reusability and interoperability. The next step would be to use for downstream analysis of the CaSQ-inferred models, methods of probabilistic model checking to verify the correctness of our translation rules and the models’ sensitivity to their change (Abou-Jaoudé *et al.*, 2014; Bartocci and Lió, 2016; Traynard *et al.*, 2016). CaSQ-inferred models are compatible with tools like PRISM, a stochastic model checker (Kwiatkowska *et al.*, 2011) or MaBoSS, a software for simulating continuous/discrete time Markov processes, applied on a Boolean network (Stoll *et al.*, 2017). Performing in depth dynamical analysis of large-scale Boolean models and developing appropriate methodologies remain key challenges in the field of computational Systems Biology.

Acknowledgements

The authors would like to thank Denis Thieffry for advice and help with GINsim analysis; Laurence Calzone for advice and providing test files and Saran Pankaew for his preliminary work on CaSQ development.

Funding

A.N. was supported by UEVE funds, T.H. was supported by NIH grant #5R35GM119770-04 and S.S. was supported by ANR BIOPSY N°: ANR-16-CE18-0029.

Conflict of Interest: none declared.

References

- Abou-Jaoudé, W. *et al.* (2014) Model checking to assess T-helper cell plasticity. *Front. Bioeng. Biotechnol.*, 2, 86.
- Abou-Jaoudé, W. *et al.* (2016) Logical modeling and dynamical analysis of cellular networks. *Front. Genet.*, 7, 94.
- Azeloglu, E.U. and Iyengar, R. (2015) Good practices for building dynamical models in systems biology. *Sci. Signal.*, 8, fs8.
- Barabási, A.-L. and Oltvai, Z.N. (2004) Network biology: understanding the cell’s functional organization. *Nat. Rev. Genet.*, 5, 101–113.
- Bartocci, E. and Lió, P. (2016) Computational modeling, formal analysis, and tools for systems biology. *PLoS Comput. Biol.*, 12, e1004591.
- Bekkar, A. *et al.* (2018) Expert curation for building network-based dynamical models: a case study on atherosclerotic plaque formation. *Database (Oxford)*, 2018. 10.1093/database/bay031
- Bounab, Y. *et al.* (2013) Proteomic analysis of the SH2 domain-containing leukocyte protein of 76 kDa (SLP76) interactome in resting and activated primary mast cells [corrected]. *Mol. Cell. Proteomics*, 12, 2874–2889.
- Büchel, F. *et al.* (2013) Path2Models: large-scale generation of computational models from biochemical pathway maps. *BMC Syst. Biol.*, 7, 116.
- Caron, E. *et al.* (2010) A comprehensive map of the mTOR signaling network. *Mol. Syst. Biol.*, 6, 453.
- Chaouiya, C. *et al.* (2012) Logical modelling of gene regulatory networks with GINsim. *Methods Mol. Biol.*, 804, 463–479.
- Cho, D.-Y. *et al.* (2012) Chapter 5: network biology approach to complex diseases. *PLoS Comput. Biol.*, 8, e1002820.
- Dhillon, A.S. *et al.* (2007) MAP kinase signalling pathways in cancer. *Oncogene*, 26, 3279–3290.
- Fujita, K.A. *et al.* (2014) Integrating pathways of Parkinson’s disease in a molecular interaction map. *Mol. Neurobiol.*, 49, 88–102.
- Funahashi, A. *et al.* (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1, 159–162.
- Furlong, L.I. (2013) Human diseases through the lens of network biology. *Trends Genet.*, 29, 150–159.
- Gillfillan, A.M. and Tkaczuk, C. (2006) Integrated signalling pathways for mast-cell activation. *Nat. Rev. Immunol.*, 6, 218–230.
- Glass, L. and Kauffman, S.A. (1973) The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.*, 39, 103–129.
- Grieco, L. *et al.* (2013) Integrative modelling of the influence of MAPK network on cancer cell fate decision. *PLoS Comput. Biol.*, 9, e1003286.
- Helikar, T. *et al.* (2008) Emergent decision-making in biological signal transduction networks. *Proc. Natl. Acad. Sci. USA*, 105, 1913–1918.
- Helikar, T. *et al.* (2012) The cell collective: toward an open and collaborative approach to systems biology. *BMC Syst. Biol.*, 6, 96.
- Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19, 524–531.
- Ideker, T. and Nussinov, R. (2017) Network approaches and applications in biology. *PLoS Comput. Biol.*, 13, e1005771.
- Jagannadham, J. *et al.* (2016) Comprehensive map of molecules implicated in obesity. *PLoS ONE*, 11, e0146759.
- Kajita, M. *et al.* (2010) Interaction with surrounding normal epithelial cells influences signalling pathways and behaviour of Src-transformed cells. *J. Cell Sci.*, 123, 171–180.
- Kauffman, S.A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, 22, 437–467.
- Kennedy, N.J. and Davis, R.J. (2003) Role of JNK in tumor development. *Cell Cycle*, 2, 199–201.
- Kuperstein, I. *et al.* (2015) Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, 4, e160.
- Kwiatkowska, M. *et al.* (2011) PRISM 4.0: verification of probabilistic real-time systems. In: Gopalakrishnan, G. and Qadeer, S. (eds.) *Computer Aided Verification, Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, pp. 585–591.

- Kyriakis, J.M. and Avruch, J. (2001) Mammalian mitogen-activated protein kinase signal transduction pathways activated by stress and inflammation. *Physiol. Rev.*, **81**, 807–869.
- Le Novère, N. (2015) Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.*, **16**, 146–158.
- Livigni, A. et al. (2018) A graphical and computational modeling platform for biological pathways. *Nat. Protoc.*, **13**, 705–722.
- Mazein, A. et al. (2018) AsthmaMap: an expert-driven computational representation of disease mechanisms. *Clin. Exp. Allergy*, **48**, 916–918.
- Mendoza, L. and Xenarios, I. (2006) A method for the generation of standardized qualitative dynamical systems of regulatory networks. *Theor. Biol. Med. Modell.*, **3**, 13.
- Niarakis, A. et al. (2014) Computational modeling of the main signaling pathways involved in mast cell activation. *Curr. Top. Microbiol. Immunol.*, **382**, 69–93.
- Ogishima, S. et al. (2016) AlzPathway, an updated map of curated signaling pathways: towards deciphering Alzheimer's disease pathogenesis. *Methods Mol. Biol.*, **1303**, 423–432.
- Ostaszewski, M. et al. (2019) Community-driven roadmap for integrated disease maps. *Brief. Bioinf.*, **20**, 659–670.
- Rizk, A. et al. (2011) Continuous valuations of temporal logic specifications with applications to parameter optimization and robustness measures. *Theor. Comput. Sci.*, **412**, 2827–2839.
- Romers, J.C. and Krantz, M. (2017) rxncon 2.0: a language for executable molecular systems biology. 10.1101/107136.
- Schramek, H. (2002) MAP kinases: from intracellular signals to physiology and disease. *News Physiol. Sci.*, **17**, 62–67.
- Setoguchi, R. et al. (1998) Defective degranulation and calcium mobilization of bone-marrow derived mast cells from Xid and Btk-deficient mice. *Immunol. Lett.*, **64**, 109–118.
- Sharp, S. and Workman, P. (2006) Inhibitors of the HSP90 molecular chaperone: current status. *Adv. Cancer Res.*, **95**, 323–348.
- Singh, V. et al. (2081) Computational Systems Biology Approach for the Study of Rheumatoid Arthritis: From a Molecular Map to a Dynamical Model. *Genom. Comput. Biol.*, **4**, e100050.
- Singh, V. et al. (2020) RA-map: building a state-of-the-art interactive knowledge base for rheumatoid arthritis. *Database*, in press. 10.1093/database/baaa017.
- Stoll, G. et al. (2017) MaBoSS 2.0: an environment for stochastic Boolean modeling. *Bioinformatics*, **33**, 2226–2228.
- Takekawa, M. et al. (2011) Regulation of stress-activated MAP kinase pathways during cell fate decisions. *Nagoya J. Med. Sci.*, **73**, 1–14.
- Thomas, R. (1973) Boolean formalization of genetic control circuits. *J. Theor. Biol.*, **42**, 563–585.
- Thomas, R. (1978) Logical analysis of systems comprising feedback loops. *J. Theor. Biol.*, **73**, 631–656.
- Thomas, R. et al. (1976) A complex control circuit. Regulation of immunity in temperate bacteriophages. *Eur. J. Biochem.*, **71**, 211–227.
- Traynard, P. et al. (2016) Logical model specification aided by model-checking techniques: application to the mammalian cell cycle regulation. *Bioinformatics*, **32**, i772–i780.
- Tripathi, S. et al. (2015) The gastrin and cholecystokinin receptors mediated signaling network: a scaffold for data analysis and new hypotheses on regulatory mechanisms. *BMC Syst. Biol.*, **9**, 40.
- Vogt, T. et al. (2013) Translation of SBGN maps: process description to activity flow. *BMC Syst. Biol.*, **7**, 115.
- Zhang, B. et al. (2014) Network biology in medicine and beyond. *Circ. Cardiovasc. Genet.*, **7**, 536–547.

Titre (en français): Analyse intégrative et modélisation des voies moléculaires dérégulées dans la polyarthrite rhumatoïde

Mots clés : Biologie computationnelle des systèmes, polyarthrite rhumatoïde, réseaux booléens, modélisation informatique, auto-immunité, maladie humaine complexe, standards en biologie des systèmes

Résumé : La polyarthrite rhumatoïde (PR) est une maladie auto-immune complexe qui entraîne une inflammation synoviale et une hyperplasie pouvant provoquer une érosion osseuse et une destruction du cartilage dans les articulations. L'étiologie de la PR reste partiellement inconnue, mais elle implique de multiples cascades de signalisation croisées et l'expression de médiateurs pro-inflammatoires. Dans la première partie de mon projet de doctorat, nous présentons un effort systématique pour construire une base de connaissances sur la PR, entièrement annotée et validée par des experts. Cette carte de la PR illustre les voies moléculaires et de signalisation importantes impliquées dans la maladie. La transduction du signal est systématiquement représentée des récepteurs au noyau en utilisant la représentation standard de notation graphique en biologie des systèmes (SBGN). La curation manuelle est basée sur des critères stricts et spécifique aux études sur l'homme, limitant l'apparition de faux positifs sur la carte. Cette carte peut servir de base de connaissances interactive pour la maladie mais aussi de tableau pour la visualisation des données omiques. De plus, c'est une excellente base pour le développement d'un modèle informatique. La nature statique de la carte PR pourrait fournir une compréhension relativement limitée du comportement émergent du système dans différentes conditions. La modélisation informatique pourra révéler les propriétés dynamiques du réseau par le biais de perturbations in silico et peut être utilisée pour tester et prédire des hypothèses.

Dans la deuxième partie du projet, nous présentons un pipeline permettant la construction automatisée d'un grand modèle booléen, à partir d'une carte d'interactions

moléculaires. Pour cela, nous avons développé l'outil CaSQ (CellDesigner as SBML-qual), qui automatise la conversion des cartes moléculaires en modèles booléens exécutables basés sur la topologie et la sémantique des cartes. Le modèle booléen résultant pourrait être utilisé pour des simulations in silico afin de reproduire le comportement biologique connu du système et de prédire de nouvelles cibles thérapeutiques. Pour l'analyse de performance de l'outil, nous avons utilisé différentes cartes et modèles de maladies en mettant l'accent sur la grande carte moléculaire de la PR.

Dans la troisième partie du projet, nous présentons nos efforts pour créer un modèle dynamique (booléen) à grande échelle pour les synoviocytes de type fibroblaste de polyarthrite rhumatoïde (RA-FLS). Parmi de nombreuses cellules de l'articulation et du système immunitaire impliquées dans la pathogenèse de la PR, les RA-FLS jouent un rôle important dans l'initiation et la perpétuation de l'inflammation articulaire destructrice. Les RA-FLS expriment des cytokines immunomodulatrices, des molécules d'adhésion et des enzymes de modélisation matricielle. De plus, les RA-FLS présentent des taux de prolifération élevés et un phénotype résistant à l'apoptose. Les RA-FLS peuvent également se comporter comme les principaux moteurs de l'inflammation, et les thérapies dirigées contre les RA FLS pourraient devenir une approche complémentaire aux immunothérapies. Le défi est de prédire les conditions optimales qui favoriseraient l'apoptose des RA FLS, limiteraient l'inflammation, ralentiraient le taux de prolifération et minimiseraient l'érosion osseuse et la destruction du cartilage.

Title (en anglais): Integrative analysis and modeling of molecular pathways dysregulated in rheumatoid arthritis

Keywords : Computational Systems Biology, Rheumatoid Arthritis, Boolean Networks, Computational Modelling, Autoimmunity, Complex human disease, Systems Biology standards

Abstract : Rheumatoid arthritis (RA) is a complex autoimmune disease that results in synovial inflammation and hyperplasia leading to bone erosion and cartilage destruction in the joints. The aetiology of RA remains partially unknown, yet, it involves a variety of intertwined signalling cascades and the expression of pro-inflammatory mediators. In the first part of my PhD project, we present a systematic effort to construct a fully annotated, expert validated, state of the art knowledge-base for RA. The RA map illustrates significant molecular and signalling pathways implicated in the disease. Signal transduction is depicted from receptors to the nucleus systematically using the systems biology graphical notation (SBGN) standard representation. Manual curation based on strict criteria and restricted to only human-specific studies limits the occurrence of false positives in the map. The RA map can serve as an interactive knowledge base for the disease but also as a template for omic data visualization and as an excellent base for the development of a computational model. The static nature of the RA map could provide a relatively limited understanding of the emerging behavior of the system under different conditions. Computational modeling can reveal dynamic network properties through in silico perturbations and can be used to test and predict assumptions.

In the second part of the project, we present a pipeline allowing the automated construction of a large Boolean

model, starting from a molecular interaction map. For this purpose, we developed the tool CaSQ (CellDesigner as SBML-qual), which automates the conversion of molecular maps to executable Boolean models based on topology and map semantics. The resulting Boolean model could be used for in silico simulations to reproduce known biological behavior of the system and to further predict novel therapeutic targets. For benchmarking, we used different disease maps and models with a focus on the large molecular map for RA.

In the third part of the project we present our efforts to create a large scale dynamical (Boolean) model for rheumatoid arthritis fibroblast-like synoviocytes (RA-FLS). Among many cells of the joint and of the immune system involved in the pathogenesis of RA, RA FLS play a significant role in the initiation and perpetuation of destructive joint inflammation. RA-FLS are shown to express immuno-modulating cytokines, adhesion molecules, and matrix-modelling enzymes. Moreover, RA-FLS display high proliferative rates and an apoptosis-resistant phenotype. RA-FLS can also behave as primary drivers of inflammation, and RA FLS-directed therapies could become a complementary approach to immune-directed therapies. The challenge is to predict the optimal conditions that would favour RA FLS apoptosis, limit inflammation, slow down the proliferation rate and minimize bone erosion and cartilage destruction.