DOI:10.14753/SE.2023.2783

ANALYSIS AND ROLE OF TRANSLOCATING PROTEINS IN PROTEIN-PROTEIN INTERACTION NETWORKS

PhD thesis

Péter Mendik MD

Molecular Medicine Doctoral School

Semmelweis University

Supervisors:	Péter Csermely DSc Dániel Veres MD PhD
Official reviewers:	Orsolya Kapuy PhD Bence Szalai MD PhD
Head of the Complex Examination Committee:	Edit Buzás MD DSc
Members of the Complex Examination Committee:	Krisztina Ella PhD Illés Farkas DSc

Budapest

2022

Table of Contents

Disclosure

During my doctoral studies I worked as a member of the LINK-Group research group and I participated in diverse, cross functional teams where we worked on interesting scientific problems in a collaborative way. To honour the work of my colleagues and students and fellow researchers in this work I consciously use the phrase “we” when presenting results. A list of my individual contributions is summarized in Chapter 9: Individual contributions.

DOI:10.14753/SE.2023.2783

List of Abbreviations

ATO	Arsenic trioxide
DCS	Data complexity score
E-cadherin	Epithelial cadherin
EMT	Epithelial-mesenchymal transition
GANT	GLI antagonist
GO	Gene Ontology
HCC	Hepatocellular carcinoma
HPA	The Human Protein Atlas
IPO7	Importin-7
KI	Knock-in
KO	Knock-out
MAPK1 or ERK2	Mitogen-activated protein kinase 1
MEK	Dual specificity mitogen-activated protein kinase kinase 1
MET	Mesenchymal-epithelial transition
NF2	Merlin
NFκB	Nuclear factor NF-kappa-B
NLP	Natural language processing
NOTCH	Neurogenic locus notch homolog protein 1
ODE	Ordinary differential equation
p53	Cellular tumor antigen p53
p63	Tumor protein 63
PPI	Protein-protein interaction
PRKRA	Protein activator of the interferon-induced protein kinase
RNA	Ribonucleic acid
ROC AUC or AUC	Receiver operating characteristic curve
RTK	Receptor tyrosine kinase
SHH	Sonic Hedgehog
SNAI1	Zinc finger protein SNAI1
TES	Translocation evidence score
TGF-β	Transforming growth factor beta

DOI:10.14753/SE.2023.2783

1. Introduction

1.1 Subcellular organisation

The emergence of subcellular organisation was a major milestone during the development of life. The eukaryotic cell is divided into several intracellular compartments which enables the spatial separation of synchronous biochemical processes [1]. Intracellular organisation is a very subtly regulated system maintained in order to conserve physiological functions of cells [2]. In fact, the disruption of this homeostatic system is often the cause of certain pathologies, where the imperfect localisation of certain subcellular actors (mostly proteins) is joined by functional disturbances [3]. These "localisation-dependent" pathologies include several kinds of neoplastic diseases [3] which cause a significant burden on modern societies [4]. Improved understanding of governing forces of subcellular dynamics will open new horizons in the treatment options of these diseases [5-9].

Subcellular organelles are traditionally defined as compartments of the cells divided by membranes, such as the mitochondria, nucleus or the endoplasmic reticulum [10]. These subcellular organelles provide their own microenvironment and they enable the separation of different subcellular processes [11]. This separation enables that intracellular proteins can function at given times with different functions inside different organelles [12]. This altered functionality is partly due to the fact, that in different organelles the same protein may have very different interacting partners. This change of interactors naturally explains the observed functional changes [10].

The traditional organelles are relatively easy to study and they are known for several decades [13, 14]. Lately, subcellular dynamics became a broader topic and a topic of utmost importance [15]. This is due to the fact that with the advances of experimental procedures we started to understand more and more about these organelles, and about how complex their regulation is. An emerging topic is liquid-liquid phase separation [16] which is the formation of condensates inside compartments in which the concentration gradient of biophysical properties of certain proteins is changed. These condensates function as membraneless organelles, and provide a very precise regulation level for the cells to organise their processes [17].

1.2 Resources to study protein localisation

Protein localisation could be detected with several methods, here I only list some examples, since the detailed discussion of these methods and their characteristics is not a focus of this dissertation. Traditional experimental procedures like immunofluorescence, immunohistochemistry or immunocytochemistry provide great resolution, precise identification of localisation and the amount of proteins is also predictable, but these procedures usually require expensive and meticulous protocols thus they are not effective in studying large numbers of proteins [15, 18].

The other possibility is to utilise predictive algorithms that are able to predict protein localisation based on omics traits [19]. These computational methods drastically reduce the time of localisation prediction but their predictive power is not yet comparable with experimental studies [20]. A logical solution is to combine these different approaches and utilise the power of computational tools in the experimental setting as well. The Human Protein Atlas (HPA) successfully combines its immunohistochemistry and immunofluorescent based data with the help of image analysing software and the result is a comprehensive, system-level but experimentally also validated database of human protein localisations [21].

Beside the HPA database there are other available sources of large-scale protein localisation. The UniProt database [22] is probably the database with the highest coverage of proteins (565 928 Swiss-Prot entries (manually annotated and reviewed proteins) and 225 013 025 TrEMBL entries (automatically annotated and not reviewed); accessed on 25.02.2022), and the Gene Ontology (GO) database is also a frequently used source of information. The ComPPI [10] database is a unique option because it not only contains localisation data, but these data are also weighted based on the available evidence and this provides the opportunity to identify biologically relevant compartment-specific interactions.

1.3 Protein translocation

Until now I’ve covered the localisation of proteins and the intrinsic structure of cells, but this approach was rather static. In living cells there’s not only a high level of structure but this structure is constantly changing in response to external stimuli and environment [23].

This constant adaptation of the cells results in subcellular dynamics. One of the prominent players in this continuously changing cellular world are the translocating proteins. Proteins are sorted inside the cells and they are in a constant movement in order to reach their final destination, where they will function and execute their specific tasks [24]. Translocating proteins are "restless messengers" inside the cells, they constantly move between different organelles and their movement is a major source of information propagation. Thus, translocating proteins transfer information from one organelle to another and their appearance in certain organelles is able to change the behaviour of the whole cell [5, 24, 25].

In general, protein translocation is a process which refers to the alteration of a given protein's subcellular localisation. However, this phenomenon has no unified definition, and the word 'translocation' may also refer to gene translocation or RNA translocation at the ribosome. We defined protein translocation as a systems biology phenomenon, which refers to the regulated movement of a protein of a given post-translational state between subcellular compartments [24]. These subcellular compartments (cytoplasm, extracellular space, mitochondria, nucleus, membrane, secretory pathway) were defined following the logic of the ComPPI database [10]. This is due to the fact that in order to obtain a localisation-specific interactome, usually high-throughput methods are used and the widely available solutions offer this resolution as I will detail in the Discussion.

Protein translocation changes the interaction partners and leads to altered function(s) of translocating proteins. There are certain processes (such as co-translational, post-translational delivery-type, cell division induced, downregulation- or passive diffusion-related phenomena; discussed in details in Mendik et. al 2019: Supplementary Text S1 and S2 [24]) that may change the localisation of a protein, but to increase the focus and clarity of our work we did not consider them as translocation. Typical examples of translocating proteins include transcription factors shuttling between the cytoplasm and nucleus, as e.g. p53, NFκB or ERK2/MAPK1 (Figure 1).

The increasing availability of new spatial proteomics methods [26] will probably soon bring a new era in which we could define organelles on a way more precise level. This more accurate definition will naturally enhance the understanding of subcellular dynamics and will offer even more opportunities. We are aware of this upcoming

DOI:10.14753/SE.2023.2783

paradigm shift but during the years of our work this level of information was not yet available. When these methods will become general, we assume that the definition of translocation may be adjusted even from the systems biological point of view.

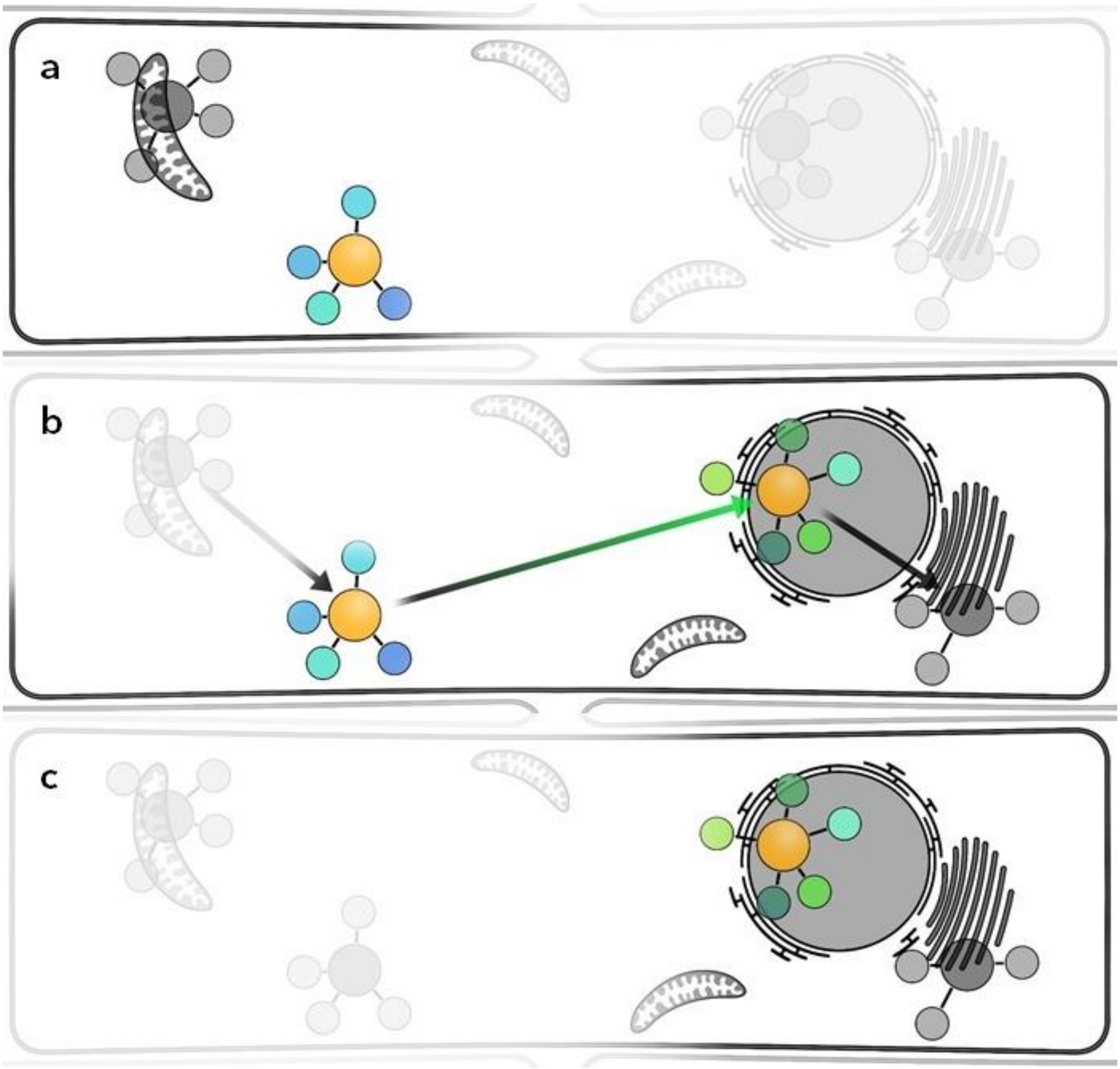


Figure 1. Translocation of the ERK2 (MAPK1) protein. ERK2 is a typical example of a translocating protein. **(a)** In the cytoplasm of resting cells ERK2 has kinase activity and exerts its functions via phosphorylating receptors, ion channels and other regulatory proteins [27]. **(b)** ERK2 has a nuclear translocation sequence. The translocation is initiated by the phosphorylation of Ser 244 and 246 in the kinase insert domain and after the subsequent binding to Importin-7 (IPO7) ERK2 is translocated to the nucleus [27]. **(c)** In the nucleus ERK2 regulates gene expression (by phosphorylating a number of transcription factors) and as a consequence the cells exhibit reduced anti-tumour activity and signs of epithelial-mesenchymal transition [27].

1.4 Epithelial-mesenchymal transition

Epithelial-mesenchymal transition (EMT) is a biological process which is important during early embryonic development, but it may also occur during cancer progression or tissue fibrosis. During EMT epithelial cells lose their apical-basal polarity and they acquire more stem cell like properties. The resulting phenotype is a mesenchymal like cell which is more motile [28]. EMT is not a one-way process: the mesenchymal cells can also undergo a mesenchymal-epithelial transition (MET) which is the inverse of EMT. These transitions (both EMT and MET) are triggered by cellular signals, e.g. transforming growth factor beta (TGF-β) is a potent inducer of EMT [29]. Historically, EMT was considered as a bimodal process, where cells either reside in the epithelial or in the mesenchymal state, but recently this view has changed. Now EMT is defined as a diverse palette of cellular states and cells can be found in any intermediary "hybrid" states between the well-defined epithelial and mesenchymal end states [30]. When we are defining EMT we must also take it into consideration that there is no dedicated marker of EMT (e.g. the loss of E-cadherin in itself), but one must always assess EMT from a complex approach and simultaneously interpret functional, morphological and transcriptional changes [30]. Prior computational studies of EMT often failed to capture this diverse aspect of EMT.

During the course of EMT a number of proteins gets activated and there are complex underlying regulatory processes [31]. Translocating proteins play an important role in this process, e.g. the translocation of β-catenin from the plasma membrane to the nucleus is an important regulatory step in the process [32]. Although we knew some translocating proteins that are important in the regulation of EMT, there were no studies that approached this process from the aspect of protein translocations as regulatory elements of EMT.

1.5 Computational network models of the EMT

The network representation of intracellular pathways is a standard and effective evaluation method for intracellular signalling processes [33-35]. These networks consist of nodes and edges. The nodes represent proteins and the edges between them stand for the interactions of proteins. In a signed network positive and negative edges represent the functional effect (i.e. activation or inhibition) of a certain interaction. These networks give us insight into the entire analysed system and not only about certain actors of a process.

Boolean models are similar to signed networks, but in this case each node also has a state variable [36]. This variable represents the active (referred to as TRUE, ON or 1) or inactive (referred to as FALSE, OFF or 0) state of a certain node. The state of a node is determined through Boolean equations where we use logical operators AND, OR and NOT. Computational programs are able to solve these equations very effectively [37], thus if we create Boolean rules that represent biological relationships we will be able to compute the state variable of a given system (Figure 2).

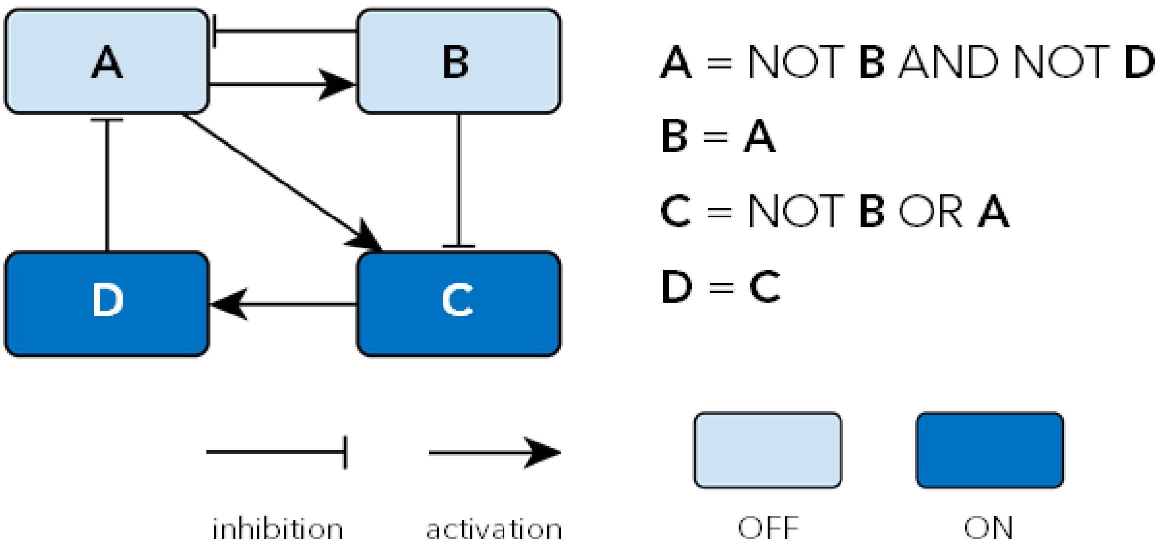


Figure 2. Boolean models in general. This is an example network demonstrating the basic Boolean operations. This model consists of 4 nodes and two of them are active (dark blue background) while two other nodes are inactive (light blue background). The state of each node is defined by the Boolean rules using the Boolean operators AND, OR and NOT. In the marked state each node's state is in accordance with its Boolean rule, so the system won't change (this is a stable state).

The use of Boolean networks became eminent because these are suited for mapping biological observations and hypotheses into a mathematical formalism which can then be computationally analysed [38]. The base concepts of Boolean modelling are understandable even without a background in quantitative sciences, but the combination of these concepts are suited to describe vastly complex biological scenarios. Thus

Boolean models are applicable to model biological processes. Furthermore, their computational needs are fairly low.

In order to enable the use of this Boolean approach for life scientists there are available software which can be used to build biological models. An available software is the BooleanNet [36] software which we used and updated during our research. This software is a Python (programming language) package, making the modelling of Boolean systems an available reality.

Previously Steinway et al. [39] created a Boolean model of the EMT. That model contained 70 nodes and 135 edges and properly uncovered the simultaneous activation of the sonic hedgehog and WNT pathways during TGF-β mediated EMT. Their model is based on experimental data and the *in silico* model rightly recapitulates experimental outcomes. More importantly, through some network reduction steps they have managed to significantly reduce the size of their network (from the 70 nodes and 135 edges to 19 nodes 70 edges) and this enables the analysis of EMT in a computationally affordable but still functionally rich manner. In their model Steinway et al. [39] already involved β-catenin as a translocating protein, but they did not address the role translocating proteins play in signalling processes and their underlying compartment-specific functions systematically. Although Boolean models are suitable to also assess these questions no previous work addressed this topic from a systematic point of view.

The work of Steinway et al. [39] served as baseline for our EMT related research, thus in this work I will usually refer to their simplified network model of 19 nodes and 70 edges as the “original EMT model” whereas our own EMT model will be named as a “compartmentalised EMT model”.

Though other computational models of EMT also exist, but these models often focus on a very limited number of nodes [40]. The advantage of these small models is that these can utilise kinetic and ordinary differential equations (ODE) to describe kinetic properties as well, but the computational burden of these models is huge, so larger networks cannot be evaluated with such methodology, preventing the execution of systematic studies. In conclusion, these ODE models are useful when one focuses on quantitatively describing parts of the EMT process or analyse the phase diagram or analyse bistable switches but they cannot be scaled up to conduct system level analyses and to predict complex interventions.

2. Objectives

This work can be divided into two complementary parts. During the first part we created a database of human translocating proteins, termed the Translocatome and then in the second part utilising data in the Translocatome database we implemented a compartmentalised *in silico* Boolean model.

As it is discussed in the introduction of this thesis, protein translocations are important regulatory events that govern the behaviour of cells. Several protein databases exist that classify proteins in a given way (e.g. UniProt database [22], MoonProt database [41], ComPPI database [10]), but none of the previously established databases focused on translocating proteins. Given the seminal role of translocating proteins in cellular signalling we aimed to create a database that fills this gap.

During the compilation of the Translocatome database [24] we wanted to create a database of human translocating proteins that extensively collects data on the translocation probability of human proteins. Thus we aimed to create a framework which enables information collection about translocating proteins, and to incorporate those data into a database that can be accessible via the internet and also supports the addition of future information. Experimental procedures can characterise protein translocations in a very complex way but our aim was to also have an extensive coverage of human proteins. So we wanted to utilise a machine learning based prediction tool (XGBoost [42]) which is able to classify proteins based on training sets. Overall we aimed to create a database of human translocating proteins which is based on a manually curated set of known translocating (and non-translocating) proteins, and relying on the predictive power of those datasets we wanted to predict the translocation probability of an extensive part of the human proteome.

During the second part we focused on utilising the data available in the Translocatome database, to create an *in silico* compartmentalised Boolean model where we can showcase the effect of subcellular dynamics and specifically the role of translocating proteins in cellular signalling. The observation of EMT seemed like a straightforward option as the EMT was previously analysed via systems biology approaches, so there were some previous studies as references to our compartmentalised model. Moreover, EMT is generally a well-studied process so *in vitro* comparisons were also available [31] and the translocation of certain factors of EMT were already proven [32]. We wanted to understand the role translocating proteins play in a subcellular process and how the compartmentalised functions govern certain cell processes from a dynamic perspective. In summary, our aim was to utilise the data of the Translocatome database and predict translocating proteins during EMT, then, after validating those translocations create a compartmentalised Boolean model where protein functions can be represented in a compartment-specific manner. Based on dynamic simulations we wanted to analyse the compartment-specific functions of translocating proteins.

3. Results

3.1 Description of the Translocatome database

We created the Translocatome [24] which is the first database that collects manually curated human translocating proteins and stores the predicted translocation probabilities of an extensive set of the human proteome. The database contains the manually curated translocating proteins' interacting partners in the localisations involved, translocation mechanism (including protein structure details if available), type of experimental evidence, affected signalling pathway(s) and pathological properties. The Translocatome database is based on a manually curated core dataset containing 213 manually curated human translocating proteins (http://translocatome.linkgroup.hu/coredata), which were collected via literature research of papers containing experimental evidence of protein translocations. Altogether the Translocatome contains 13 066 human proteins. These are all the human proteins with at least one available experimentally validated subcellular localisation in the ComPPI database (as accessed on 20th July 2018). The application of the gradient boosting machine learning tool XGBoost [42] enabled the prediction of translocation probabilities. This resulted in 1133 high-confidence translocating proteins, but all the 13 066 proteins of the Translocatome were characterized with a translocation likelihood, named as Translocation Evidence Score (TES; http://translocatome.linkgroup.hu/help/scores). Users can access the whole database online (http://translocatome.linkgroup.hu) and through various search and download options they can utilise the data according to their goals (Figure 3).

3.1.1 Content of the Translocatome database

The core of the Translocatome is an extensively curated set of 213 human translocating proteins (http://translocatome.linkgroup.hu/coredata). We aimed to collect detailed and experimentally validated information about every entry extracted from peer reviewed publications (details discussed in Mendik et al, 2019: Supplementary Text S3 [24]). For each protein between the 213 manually curated ones we collected the following data (if available):

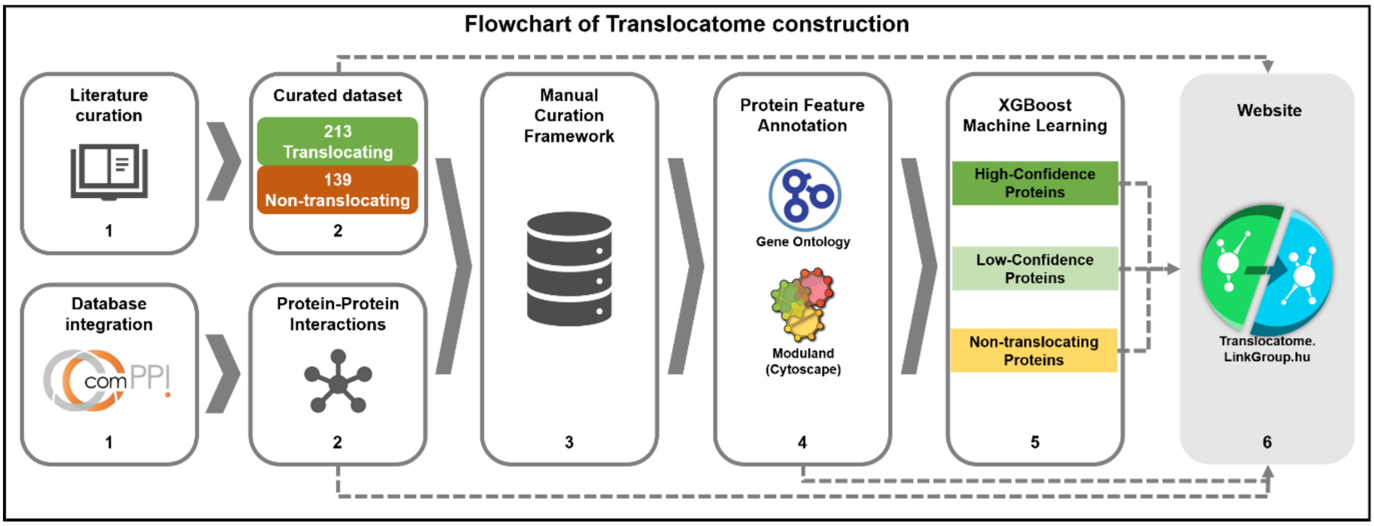


Figure 3. Schematic flowchart of the Translocatome database construction process highlighting 6 major steps. The main input sources of the Translocatome are manual curation of peer reviewed articles (Literature curation) and the ComPPI database (http://comppi.linkgroup.hu). During the manual curation process we recorded the source of experimental validation, several details of the translocation mechanism, the local compartmentalised interactome, as well as the involvement in signalling pathways and disease development **(1)**. This extensive manual curation resulted in a set of 213 translocating and another set of 139 non-translocating human proteins. To incorporate our data into a protein-protein interaction (PPI) network we imported the interaction data of 13 066 human proteins (151 889 interactions) from the ComPPI [10] database **(2)**. The Manual Curation Framework (MCF) is a user-friendly interface where the data of the Translocatome database is stored. Users can log in (after registration) to modify and update its data, which is published after expert cross-check **(3)**. We annotated each protein in our database with Gene Ontology [43, 44] functional and topological properties to enable a prediction process **(4)**. Based on the predictions of the XGBoost machine learning algorithm [42, 45, 46] we classified 13 066 human proteins into three sets: high- and low-confidence translocating proteins and non-translocating proteins **(5)**. On the http://translocatome.linkgroup.hu website the whole dataset is available for searching and downloading purposes freely and without registration. Translocatome can be updated by the community-based Manual Curation Framework. Moreover, Translocatome is linked to the ComPPI database [10] so in the case of its update Translocatome can be also updated **(6)**.

A. name set, gene name and UniProt accession number and link;
B. PubMed ID(s) and link(s) to peer-reviewed article(s) describing the experimental evidence of translocation;
C. initial and target localisations of the translocating protein;

D. interacting partners and biological functions (both in the initial and target compartments);
E. translocation mechanism;
F. detection method used;
G. protein structural information on translocation mechanism;
H. disease group, exact disease involved and pathological role;
I. signalling pathways affected.

For protein identification we used the terminology of the UniProt database [22], to describe localisations and biological processes the Gene Ontology [43, 44] terms and for the standardization of signalling pathways the KEGG naming convention [47]. Every protein was annotated to one of six major cellular localisations (cytoplasm, extracellular space, mitochondria, nucleus, membrane or secretory pathway), following the methodology of the ComPPI database [10]. When more precise localisation information was available the respective entry was also annotated with a minor localisation. All the manually curated translocating proteins are characterized by a Data Complexity Score (DCS), which described the amount of data annotated to each protein (details discussed in Mendik et al, 2019: Data complexity and translocation evidence scores [24], and also here: http://translocatome.linkgroup.hu/help/scores).

As detailed later, we utilised the XGBoost [42] algorithm to predict translocation probabilities of human proteins, and to enable this prediction we needed to rely on a training set. To create this training set we excluded 53 of the manually curated translocating proteins, as they were shown to translocate only under pathological conditions (e.g. during carcinogenesis). So the positive training set used to teach the XGBoost algorithm consisted of 160 physiologically translocating proteins.

Similarly to the positive training set, we needed to compile a manually curated negative dataset, so we collected 139 human non-translocating proteins (Mendik et al. 2019: Supplementary Table S3 [24]). Finding of these non-translocating entries is difficult as scientific publications, highlighting the absence of a trait (i.e.: the lack of translocation) are not common. So we had to define some baseline scenarios that we considered to be widespread reasons for the absence of protein translocation and we collected non-translocating proteins based on these considerations. So each protein in the negative training set is classified as a protein (a) with experimentally proved diffuse, multi-compartmental distribution, (b) with exclusive single-compartment localisation, (c) docked to DNA/RNA, (d) embedded in membranes or (e) attached to the cytoskeleton.

To understand the different datasets in the database please consider the Venn diagram shown in Figure 4.

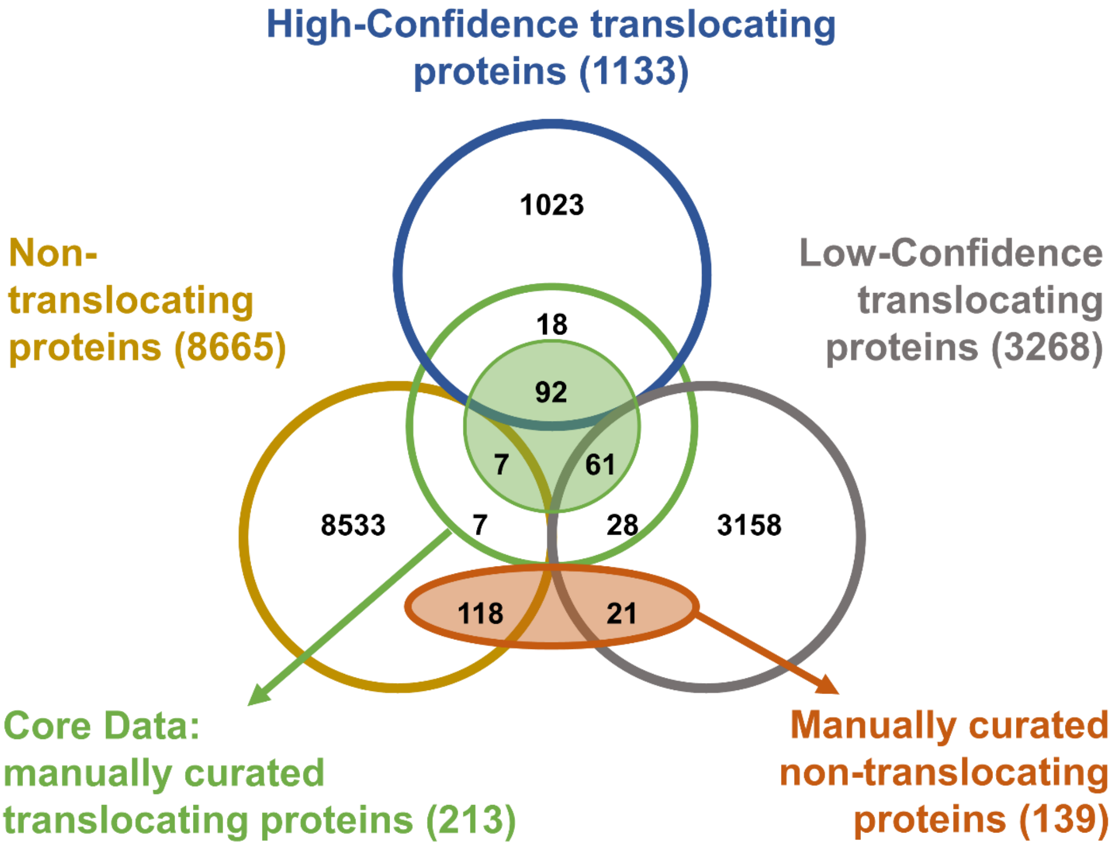


Figure 4. Structure of the Translocatome database. As shown in the above Venn diagram, the database consists of the Core Data of 213 manually curated translocating proteins, which are extended by 1133 and 3268 high- and low-confidence translocating proteins, respectively. Circles with green and red background represent the positive and negative training sets, respectively. The Core Data and positive learning set differ, since the latter does not contain the 53 proteins showing translocation exclusively under pathological conditions.

In summary, the Translocatome is a database that contains 13 066 human proteins, each protein is characterized with a translocation probability and the core of the database is a strongly manually curated dataset. The interactome data of these proteins are also imported from the ComPPI database adding 151 889 protein-protein interactions. These features make the Translocatome an extensive database of human translocating proteins and the data within can be utilised to create future studies that focus on the role of translocating proteins and their compartment-specific roles.

3.2 Predicting protein translocations with the Translocatome database

As discussed in the previous section, the Translocatome contains a great amount of manually curated data regarding translocating proteins. Such data could be utilised to conduct a supervised machine learning workflow: data collection, feature extraction, feature selection, classification, training, testing and interpretation. We used the well-established and widely used gradient boosting-type machine learning tool, XGBoost, which was already applied in previous studies to predict among others host-pathogen protein-protein interactions [48], microRNA disease association [49] or DNA methylation [50]. These studies showed that XGBoost gives the best performance if compared to other state-of-the-art machine learning methods. The data collection step is the manual curation process of translocating proteins (see Mendik et al, 2019: Supplementary Text S3 for details [24]) and the other steps are explained below.

During the training step we annotated each of the 13 066 proteins of the Translocatome with their relevant Gene Ontology (GO) terms (cellular component, biological process and molecular function terms specifically), including the ancestors of these terms. This process followed the methodology used by Kerepesi et al. [46] and is also detailed in Mendik et al. 2019: Supplementary Text S5 [24]. As a result a total of 21 020 GO terms were annotated to the proteins (see details in Mendik et al, 2019: Supplementary Text S6 [24] and here: https://github.com/kerepesi/translocatome_ml). Moreover, we annotated each protein with the network parameters degree and bridgeness (we used the interactome data of 151 889 interactions imported from the ComPPI database). Degree (the number of neighbours a protein has) and bridgeness [51] (nodes connecting different modules of a network have high bridgeness values) values were significantly higher among manually curated translocating proteins than between manually curated non-translocating proteins or the average (Figure 5). The underlying biological explanation behind these differences is that translocating proteins often play a central role in cellular regulation thus they have a lot of neighbours, or in network terms they act as hubs [52]. On the other hand, in light of the fact that translocating proteins can mean the “bridge” between distinct cellular organelles (e.g. nucleus and cytoplasm) it is not surprising that their bridgeness values are also significantly higher. GO terms, degree and bridgeness were selected by the XGBoost tool as feature sets.

Since the ComPPI database [10], from which we imported the protein-protein interaction data, does not contain interactions occurring under pathological conditions, we were prompted to exclude those 53 manually curated translocating proteins from the positive training set which translocated only under pathological conditions. The remaining 160 manually curated translocating proteins were used as a positive training set.

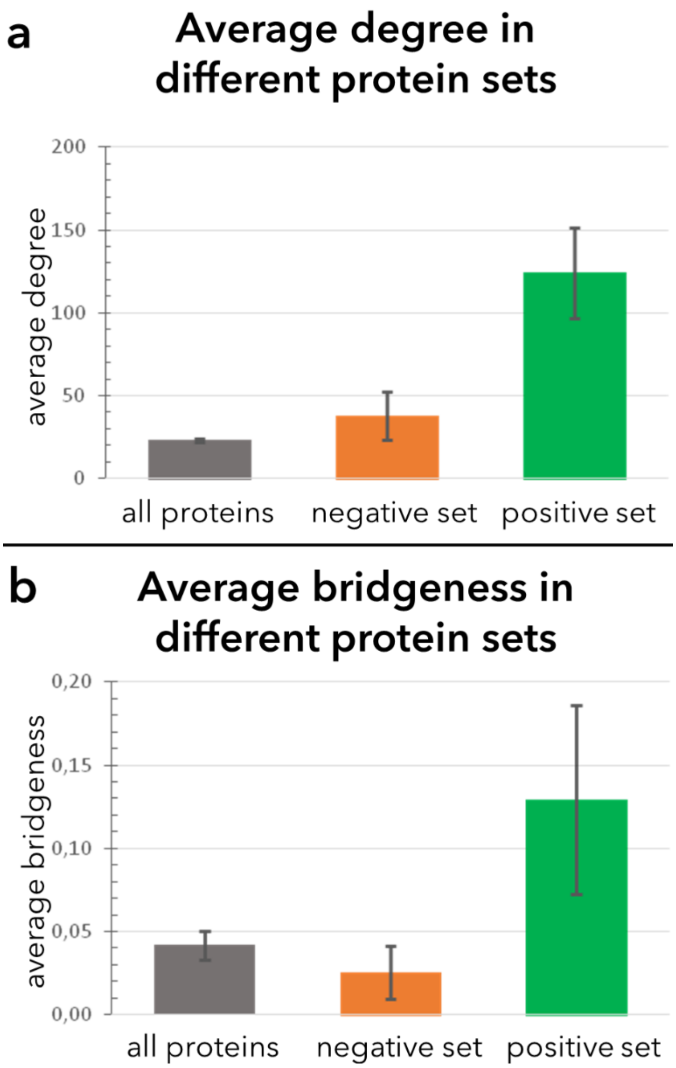


Figure 5. Degree and bridgeness of all proteins, positive and negative training sets. The mean ± standard deviation (SD) of the degree (panel a) and the bridgeness (panel b) values of all the 13 066 proteins as well as the 160 and 139 proteins of the positive and negative training sets. Degree and bridgeness values were calculated based on the ComPPI-derived human interactome having 151 889 interactions. (A) The average degree is 23.2, 37.9 and 124.1 for all proteins, the negative and the positive training sets, respectively. The average degree of the positive set is significantly higher than that of the other two sets ($p<0.05$, Student's two tailed t-test). (B) The average bridgeness is 0.04, 0.03 and 0.13 for all proteins, the negative and positive training sets, respectively. The average bridgeness of the positive set is significantly higher than that of the other two sets ($p<0.05$, Student's two-tailed t-test).

Similarly to previous studies we evaluated the XGBoost-selected feature sets by 5-fold cross-validation, and the predictive power by the area under the curve of the receiver operating characteristic curve (ROC AUC or shortly AUC). 5-fold cross-validation is a method where the training data is split into five random parts. While one part is used for evaluation of the predictions, the other four parts are used to train the XGBoost machine learning tool. For every feature set, we repeated this cross-validation process 100 times.

The XGBoost program characterized each feature used during the training part with an importance value. We relied only on the most important features, having an importance value greater than 0.02. The 15 GO features (from the initial 21 020) most suitable for the algorithm produced an average AUC of 0.916 (± 0.0046 standard deviation). To these 15 features we added the two interactome-derived features degree and bridgeness and this resulted in a final high performing model with the average AUC of 0.9207 (± 0.0056 standard deviation), which is even higher than an average AUC of 0.916 of the GO term based model. The ROC curves of 100 five-fold cross-validation runs of this final model showed a minimal, average and maximal AUC of 0.9047, 0.9207 and 0.9333, respectively (Figure 6).

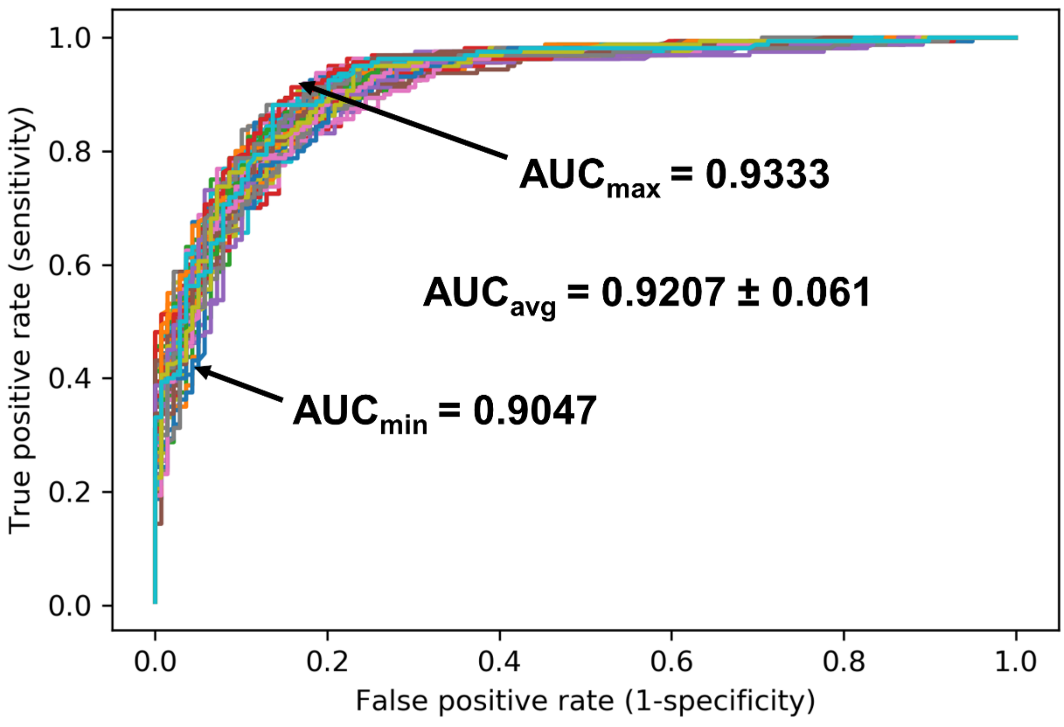


Figure 6. Performance of the XGBoost machine learning method on the final feature set. Each of the 100 different receiver operating characteristic (ROC) curves belong to a different 5 fold cross-validation run on the training set (containing 160 physiologically translocating and 139 non-translocating proteins). These ROC curves were plotted using the final feature set (see Table 1) selected earlier as described. The minimal, maximal and average area under the curve (AUC) were 0.9047, 0.9333 and 0.9207 (±0.0061 standard deviation), respectively.

DOI:10.14753/SE.2023.2788

Gene Ontology process (GO term name) **or interactome feature**	**Importance**	**Short biological explanation**
Parameters having a positive predictive value		
animal organ morphogenesis (GO:0009887)	2.68	Morphogenesis and other developmental processes are mostly regulated through complex networks of transcription factors, where translocation is often involved as a regulation step [53].
regulation of carbohydrate metabolic process (GO:0006109)	1.53	A lot of metabolic enzymes also function as protein kinases and translocate between cellular compartments playing a role e.g. in carcinogenesis [54].
cytoplasm (GO:0005737)	1.35	Large cellular compartments are often associated with proteins that translocate. Nucleo-cytoplasmic translocations play a key role in the regulation of transcription factors [53].
nuclear part (GO:0044428)	1.12	
negative regulation of cellular process (GO:0048523)	1.12	Negative regulatory mechanisms are frequently exerted by translocating proteins such as PTEN [55] or transcription factors.
plasma membrane part (GO:0044459)	0.70	Large cellular compartments are often associated with proteins that translocate. Cytosol-membrane translocations play a key role in the regulation of signalling pathways [56].
extracellular region (GO:0005576)	0.65	
cytosol (GO:0005829)	0.57	
spliceosomal complex (GO:0005681)	0.23	The spliceosome is composed of snRNPs translocating from the cytoplasm. Some spliceosome components are also involved in mRNA export [57, 58].
Parameters having a negative predictive value		
bridgeness value is lower than 0.000292 (bridgeness lower than 0.000292)	-0.36	Translocating proteins often bridge the two interactome modules (large protein complexes) of their two localisations. Therefore, their bridgeness values tend to be high ([59] and Mendik et al. 2019: Supplementary Figure S1 [24]).
degree is smaller than 62.5 (degree lower than 62.5)	-0.50	A reasonably high number of interaction partners often indicates a role in regulation and signal transduction. Many of these proteins are "date-hubs", which may undergo a translocation process. Nevertheless, too many partners could be a characteristics of a multi-compartmental housekeeping protein ([60] and Mendik et al. 2019: Supplementary Figure S1 [24]).
degree is smaller than 14.5 (degree lower than 14.5)	-0.54	
negative regulation of intracellular signal transduction (GO:1902532)	-0.61	If the translocation process becomes inhibited, it may often prevent signal transduction. Inhibition often occurs via sequestration by large protein complexes which usually have only one localisation [61].
myeloid cell differentiation (GO:0030099)	-0.74	Cell adhesion and membrane bound proteins play an important role in myeloid cell differentiation [62, 63]. Both protein categories are typically non-translocating proteins, which may over-compensate the role of translocating transcription factors in this process.
intrinsic component of membrane (GO:0031224)	-0.82	Intrinsic membrane components predominantly do not translocate to other major localisations.
system process (GO:0003008)	-0.91	A wide variety of proteins exert their system level biological functions (e.g. secretion of molecules) in a non-translocating manner: cell membrane channels, actin, myosin, etc.
single organismal cell-cell adhesion (GO:0016337)	-1.06	Cell adhesion proteins usually have a strictly limited location in the plasma membrane
bridgeness value is lower than 2.5e-06 (bridgness lower than 2.5e-06)	-1.10	Translocating proteins often bridge the two interactome modules (large protein complexes) of their two localisations. Therefore, their bridgeness values tend to be high ([59] and Mendik et al. 2019: Supplementary Figure S1 [24]).
protein complex (GO:0043234)	-1.24	Proteins often fulfil their roles in large protein mega-complexes. These complexes may assist other proteins to translocate, but their own components do not translocate.

Table 1. The feature set identified as best predictor by the XGBoost machine learning algorithm. Features selected by the XGBoost machine learning algorithm can be either interactome related network metrics or GO term-related functional parameters (first column). XGBoost assigns an importance score to each feature (as shown in the second column. Calculation detailed in Mendik et al, 2019: Supplementary Text S6 [24]). In the third column there is a general explanation as to why these features are biologically valid choices by the XGBoost machine learning algorithm.

Furthermore, both the precision-recall and Matthews correlation coefficient curves were evaluated and those further strengthened the high performance of our predictive model (details in Mendik et al, 2019: Supplementary Figure S3 [24]). All the data and computational codes needed to reproduce this methodology are available at https://github.com/kerepesi/translocatome_ml. The feature set of the final model is shown in Table 1. Features with positive importance values increase the probability of translocation, and the underlying logic is also explained in the table. On the contrary, features with a negative importance value decrease the probability of protein translocation. Using these features we calculated the translocation probability of each protein in the Translocatome.

Translocation probabilities were calculated with the help of the XGBoost program. The program takes into consideration the features characteristic of a certain protein and then based on the importance value of those features we calculated a Translocation Evidence Score (TES) which corresponds to the translocation probability of a given protein. The TES values follow a continuous distribution so we determined some cut-off values to translate those numerical values into biologically relevant categories. To define these cut-off values we used the F1-score (measures the performance of a binary classification being a harmonic average of precision and recall) which reached its peak at the TES value of 0.4487, so proteins with a smaller TES value were considered as non-translocating proteins (8665 proteins). Then we defined another threshold value at 0.6167, because no proteins in our negative training set had higher TES values than that, so we can propose that the proportion of false positive predictions in this subgroup is very low. So we named these proteins – with the highest TES values – high-confidence translocating proteins (1133 proteins). The proteins with TES values between 0.4487 and 0.6167 were considered as low-confidence translocating proteins (3268 proteins).

The Translocatome is the only established database of human translocating proteins and the extensive coverage of it enables further systems level analyses. To demonstrate the value and reliability of the predictions we assessed the top 40 proteins with the highest TES values. These proteins fall into four categories:

A. were already included in the manually curated 213 translocating protein set (12 proteins: PTEN, PTK2, FOXO3, GMNN, ATF2, MAPK1, GLI3, HRAS, AR, SMAD3, SMAD2 and HSP90AB1);
B. were previously shown to be translocating proteins but have not appeared in our Core Data of 213 proteins collected from keyword based searches (11 proteins: NF2, TULP3, SNCA, FGFR2, MTOR, GSK3B, EIF6, HDAC1, CARM1, CUL1 and RARB);
C. have not been described as translocating proteins yet, but from the literature we can conclude that their translocation is probable (one protein: TP63);
D. there is no information in the literature about their translocation (one protein: PRKRA).

Proteins in categories C and D are the most interesting ones, as novel discoveries may reside here. We highlighted this through the example of the p63 protein (Tumor protein 63), which physiologically resides in the nucleus of human cells [64], and it was not marked as a translocating protein in available databases. p63 is a transcription factor [64] and it plays an important role in the regulation of embryogenesis [65]. Beside this physiological functions p63 also appears in the cytoplasm of adenocarcinoma or prostate carcinoma cells, and this cytoplasmic localisation results in increased malignancy of these tumours [25, 66]. The XGBoost algorithm's prediction to signal p63 as a translocating protein thus could be understood and similar results should be verified and their regulational interactors should be better understood.

Altogether we've shown that the XGBoost machine learning algorithm is able to learn the features of translocating proteins from a wide feature set, and based on the identified discriminatory features the algorithm predicts further proteins' translocation probabilities with high efficiency. The result is an extensive database of human translocating proteins with a manually curated core dataset. The validation of the predicted translocations could shed light on important new regulatory roles of translocating proteins.

DOI:10.14753/SE.2023.2783

3.3 Enrichment of translocating proteins among signalling proteins

As detailed above in the first phase of my PhD studies we created an extensive database of human translocating proteins. During the second phase we focused on utilising this dataset and to prove the impact of protein translocations and in general compartment-specific subcellular dynamics on cellular behaviour.

The EMT is a biological process in which the role of protein translocations was already proven [32], but until our research no studies focused specifically on the regulatory role of network compartmentalisation. To see if EMT is a process where the observation of protein translocations is indeed possible first we proved that translocating proteins are enriched between human signalling and specifically EMT proteins. In the Translocatome 66% of the proteins are non-translocating proteins, with 25% low confidence and 9% high-confidence translocating proteins. Using GO terms we defined the set of general signalling proteins and EMT proteins, and observed if the same distribution of translocating proteins is true.

Among human signalling proteins we found that altogether a higher percentage of proteins are predicted as translocating (31% are low-confidence and 15% are high-confidence translocating proteins). Simultaneously, the percentage of non-translocating proteins (54%) is smaller. Similarly (but to an even greater extent), in the case of EMT proteins we found that 39% of the proteins are low-confidence translocating proteins and 33% are high-confidence translocating proteins whereas only 28% of the proteins are non-translocating proteins, based on the TES values of the Translocatome database (Figure 7). This observed distribution of translocating proteins is significantly ($p < 0.0001$, Chi-square test) different to what we observed between all proteins of the Translocatome. This underlines that translocating proteins in fact could have an important role in the regulation of certain signalling processes and EMT is a suitable model to observe those compartment-specific features.

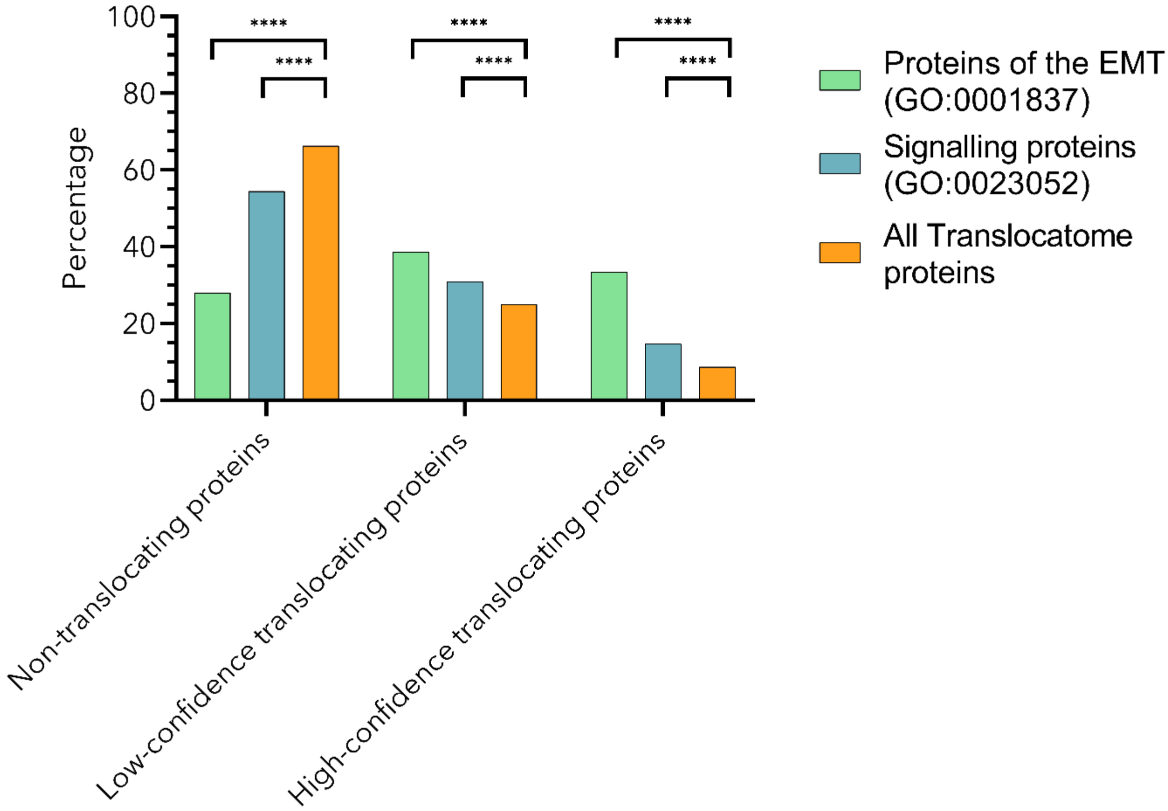


Figure 7. Translocating proteins are enriched among signalling and EMT proteins. The figure shows the distribution of proteins according to their translocation probability in different datasets (Translocatome proteins, signalling proteins and EMT proteins, see Mendik et al, 2022: Methods for the definition of these datasets [67]). The rate of high-confidence translocating proteins between the Signalling (15%) and EMT (33%) proteins is significantly higher than between the Translocatome proteins (9%). The same significant difference is true for low-confidence translocating proteins as well. The percentage of low-confidence translocating proteins is higher between Signalling (31%) and EMT proteins (39%) than between Translocatome proteins (25%). On the contrary, in the case of non-translocating proteins there is a significantly lower percentage of non-translocating proteins among signalling (54%) and EMT (28%) proteins than among Translocatome proteins (66%). (Further details of this analysis are discussed in Mendik et al, 2022: Methods [67]) ****: $p < 0.0001$, Chi-square test.

3.4 Compartmentalised Boolean model of the epithelial-mesenchymal transition

After demonstrating the overrepresentation of translocating proteins in EMT we focused on creating a compartmentalised network model of EMT, where the role of protein translocations can be assessed from a network biology perspective. In the compartmentalised network each protein is represented according to its subcellular localisation. If a protein can translocate between 2 subcellular localisations, then it is divided into 2 nodes and each node has only the interacting partners in that specific location (Figure 8). This process is the compartmentalisation of a network and it results in an end state where each protein's regulation will be represented according to its subcellular localisation as well. We always aimed to create Boolean rules which capture the experimentally validated biological functions, and for this study our primary aim was not to create a generalizable compartmentalisation process but to prove the additional value of compartment-specific interactions and thus regulatory relationships. Compartmentalisation is crucial because proteins can have utterly different functions in different subcellular organelles, e.g. in the case of translocating proteins.

To prove the additional value of compartmentalisation (our model), we used a 19-node EMT model published by Steinway et al. [39] as a benchmark and we compared our model to it. As mentioned in the introduction of this thesis I refer to their 19-node model as the "original EMT model".

The original EMT model had 19 nodes and 70 edges. Based on the data of the Translocatome 14 nodes out of the 19 were predicted as high-confidence translocating proteins in accordance with the previous GO based enrichment analysis and again underlying the importance of protein translocations in EMT. Based on a thorough manual curation process (detailed in Mendik et al, 2022: Methods [67]) we validated these translocations and included them in the compartmentalised model only if we could both validate the translocation itself and the fact that it is involved in EMT as well.

DOI:10.14753/SE.2023.2783

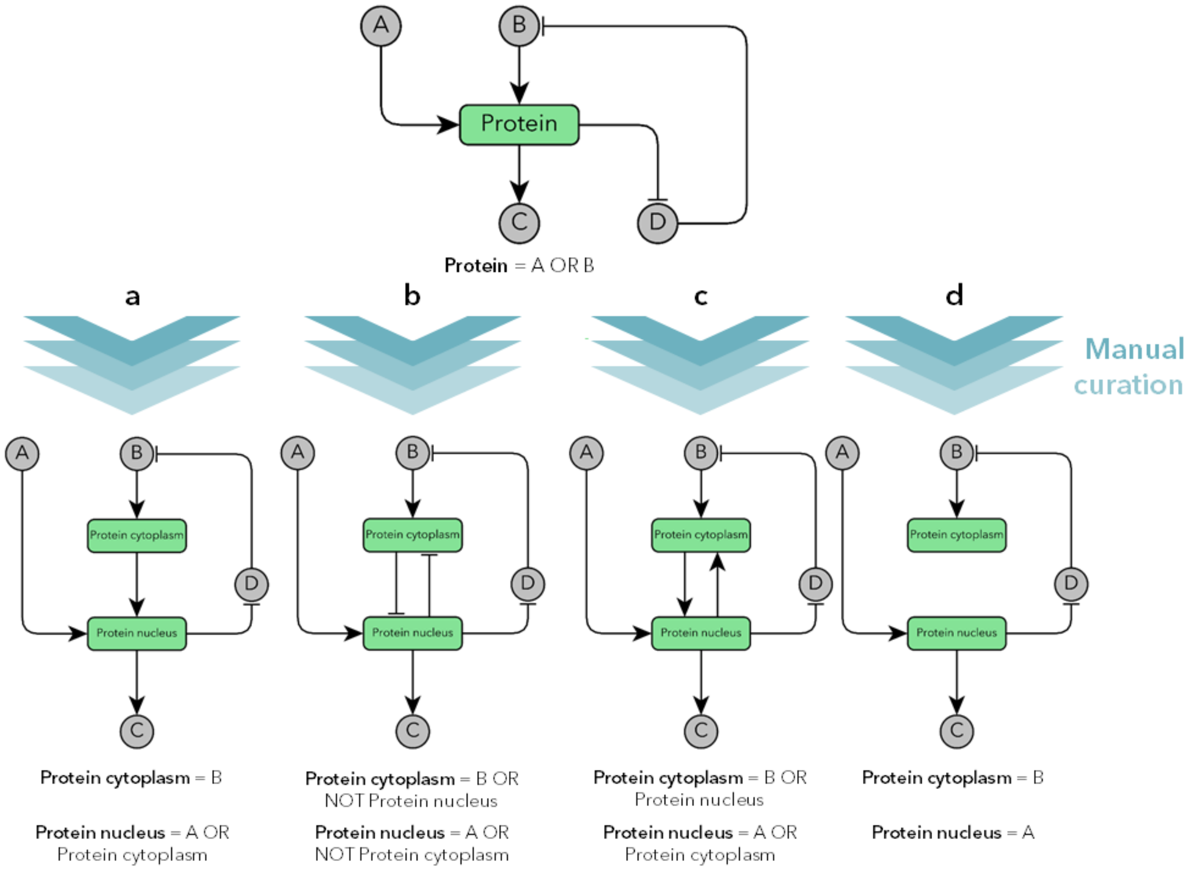


Figure 8. Creation of compartmentalised Boolean rules. In conventional network representation of signalling processes nodes represent proteins and edges are interactions between them (inhibitory or activatory), but their localisation specificity is not considered. In our compartmentalised Boolean model, we can systematically add this information, so the compartment-specific functions of proteins can be investigated. This figure shows a hypothetical explanatory example, where we highlighted a "translocating protein" with green. **(a)** One possibility is that after e.g. phosphorylation by B the protein translocates to the nucleus and this does not directly affect the cytoplasmic pool of that protein. **(b)** Another possibility was already contained in the original EMT model (in the case of β-catenin), where the nodes mutually inhibit each other. **(c)** Some transcription factors can upregulate their own expression which can act as a positive feedback, **(d)** but it is also possible that both the cytoplasmic and the nuclear pools of a protein have their own regulatory interactions but they don't affect each other directly. There are also other potential combinatorial possibilities discussed in details in Mendik et al, 2022: Supplementary Figure 7 [67].

Compartmentalised Boolean rules were created based on available literature data; 64 publications were reviewed for the 10 compartmentalised nodes and thus by node-duplications we extended the 19-node network into a 30-node network (the NOTCH node was divided into 3 subcellular nodes, since it has validated localisation and activity in the plasma membrane, in the cytosol and in the nucleus as well). This resulted in a compartmentalised Boolean model which was ready for further dynamic network analysis and to inspect the additional value provided by compartmentalisation and to validate it by comparing it to the available experimental results. The final model is defined in Mendik et al. 2022, Supplementary Data 3 [67] (I will refer to the node names as defined there).

Our work can be summarized as a workflow, where first the predicted translocations are identified (via the Translocatome database), then these are manually (or experimentally) validated and finally, compartmentalised Boolean rules are created. Computational analyses can be performed on this compartmentalised model and different outputs can be evaluated. We also created an online application (https://translocaboole.linkgroup.hu) where our model can be tailored to specific interests (Figure 9). The web application provides the opportunity to edit compartmentalised Boolean rules and to rerun computational analysis (the simulation runs with the default settings of 25 iterations and 20 000 steps, every node is perturbed). The server behind the webpage automatically reruns the simulations and users can download the final output files. Users who need more features and modifications can also opt to download the input files (from the webpage) and with those files they can use the relevant codes available in our GitHub repository (https://github.com/deriteidavid/compartmentalized_EMT_Boolean_model_Mendik_et_al_2021) to rerun the analyses. The website and GitHub tools make it possible to also use our workflow as a framework for similar future studies.

Our compartmentalised model is a discrete dynamic Boolean model that characterizes every node by a state variable (ON or OFF), corresponding to the node's activity and a Boolean function which describes the nodes' regulation. During a dynamic simulation an asynchronous updating algorithm randomly selects a node and updates its state according to its Boolean rules (more details in our respective publication). Boolean models of biological systems converge into stable states, called attractors, which qualitatively correspond to real biological phenotypes [38, 68]. To understand the biological validity of a Boolean model we can evaluate its attractors from a biological viewpoint.

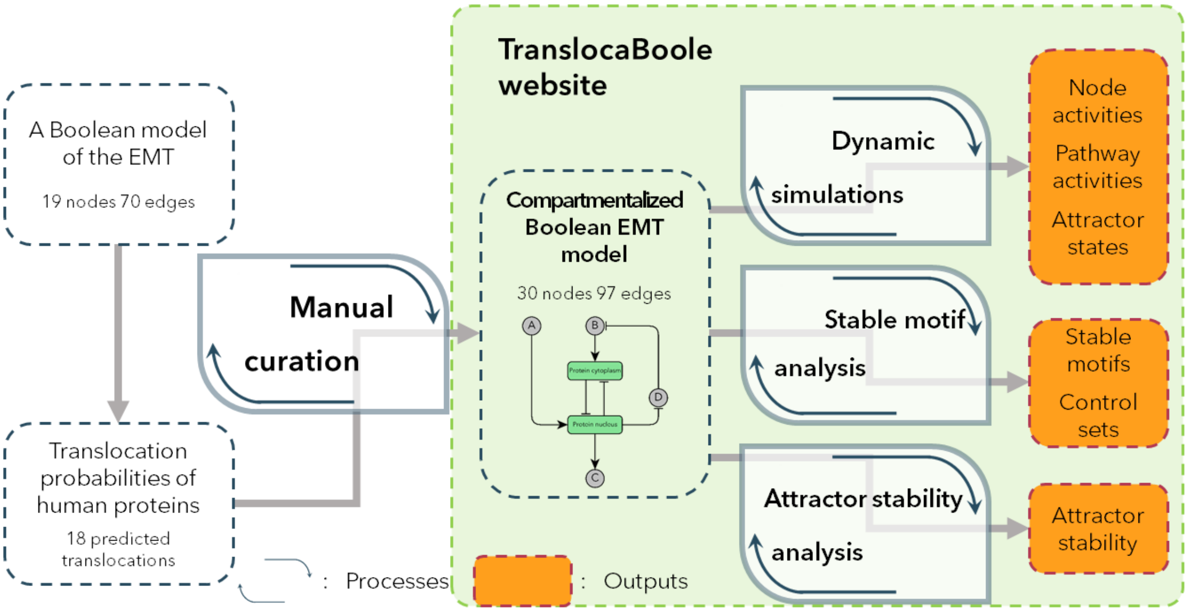


Figure 9. Workflow for the creation and evaluation of a compartmentalised Boolean model. Based on previously published Boolean models of the EMT and utilising the data available in the Translocatome it is possible to uncover translocating nodes that need to be compartmentalised. The compartmentalised model (with updated Boolean functions) was created through careful revision of available literature by manual curation. Dynamic simulation could be run on the model and the results could be analysed as different outputs on different levels, such as node or pathway activities or the identification of stable attractors of a system. Stable motifs and control sets of a model could be identified with additional analyses or attractor stability measurements could be executed. Users could modify our model, automatically rerun simulations with our standard settings (25 iterations and 20 000 steps, every node is perturbed) and download the results on the website: https://translocaboole.linkgroup.hu/

Our compartmentalised EMT model has seven attractors, the most important ones are the epithelial (E) and mesenchymal (M) attractors (Figure 10a). The attractors and stable motifs of our model were analysed using a state-of-the-art method [69] (see details in Mendik et al. 2022: Results [67]). The E and M attractors are polar opposite states, differing in every node, with the exception of SOS_GRB2. The remaining 5 attractors are intermediary states, which can be translated to hybrid states of the EMT which emerge during incomplete or partial EMT processes. Our model also successfully recapitulated the experimental findings used to validate the original EMT model (Table 2).

Experimental result	Reference	Recapitulated by the original EMT model	Recapitulated by the compartmentalised model
TGFβ signalling leads to SMAD complex formation, MAP kinase signalling, and AKT signalling.	[70]	yes	yes
Wnt signalling leads to nuclear localisation of β-catenin, AXIN2 induction, and suppression of the destruction complex	[71]	yes	yes
SHH signalling leads to induction of GLI transcription factors	[72]	yes	yes
miR200 inhibits TGFβ–driven EMT	[73]	yes	yes
E-cadherin suppressing transcription factors SNAI1, SNAI2, ZEB1, ZEB2 and TWIST1 induce EMT when acting together	[74, 75]	yes	yes
Constitutive SNAI1 or TWIST1 activation drives EMT	[74-77]	yes	yes

Table 2. Experimental outcomes reproduced by the compartmentalised and the original EMT model. The publication of Steinway et al. used the experimental results detailed in this table to prove the validity of their model. To ensure our compartmentalised model performs similarly we observed the same experimental results and concluded that it reproduces all experimental results as expected.

The analysis of the control sets of our model revealed that forcing the network into the epithelial state is more feasible (requires less complex interventions) if we also rely on compartment-specific perturbations. A control set of an attractor is a given group of nodes, which, when forced to the corresponding state, drive the whole system into a specific attractor. In the compartmentalised model the simplest control set of the epithelial state was a smaller set of nodes (5 nodes) than in the original EMT model (6 nodes). The simplest control set of our model contains compartmentalised nodes, so we could conclude that the perturbation of compartment-specific functions led to a simplification in the control of the system (corresponding data shown in Mendik et al. 2022: Supplementary Data 4 [67]).

We've also shown that in the presence of noise our model shows a somewhat improved stability profile where the stability of the epithelial attractor is improved (compared to the original EMT model), but still the system overwhelmingly favours the mesenchymal state (details in Mendik et al. 2022: Supplementary Note 2 and Supplementary Table 3 [67]).

To compare the two models head-to-head we used two simulation setups:

A. single node perturbations, where only one node's state was perturbed in the epithelial initial state. In this scenario TGFBR node was OFF, so we refer to these simulations as TGFBR OFF simulations.
B. in the second setup we also introduced single perturbations, but additionally already in the initial state we set the value of the TGFBR node permanently ON (TGFBR ON simulations), mimicking the biological scenario of active TGF-β signalling. In this setup we could inspect which nodes' perturbation could inhibit the TGF-β driven EMT.

These simulation setups were similar to wet-lab knock-out (KO) and knock-in (KI) experiments, but due to the compartmentalised nodes we were able to introduce compartment-specific perturbations.

We found differences between our and the original model in two cases: either NOTCH or MEK KI led to EMT in the original but not in the compartmentalised model. Experimental results suggest that Neurogenic locus notch homolog protein 1 (NOTCH) activation alone without the induction of TGFB was not sufficient to induce EMT [78, 79] so our model coincided better with these results. This improved behaviour stemmed from the fact that in our model we were able to better capture the regulatory events needed to activate the Zinc finger protein SNAI1 (SNAI1).

The results regarding MEK activation are more complex as experimental results verify both outcomes. Depending on the context MEK activation in itself is either satisfactory to induce EMT or not [80, 81], but it was important to note that in the case of TGFBR ON simulations only the compartmentalised model showed correctly that MEK inhibition could prevent TGFB induced EMT [82]. Further experimental results could shed light on more complex regulatory interactions and adding those to the model could clarify this ambiguous picture.

3.5 Compartment-specific functions of translocating proteins

Our new modelling approach also enabled the assessment of compartment-specific functions. During TGFBR OFF simulations GSK3B KO in the original model and GSK3B_cyto KO in the compartmentalised model both led to the mesenchymal phenotype. Experimental evidence confirms these outcomes as the inhibition of cytoplasmic GSK3B by LiCl led to EMT in ovarian adenocarcinoma [83]. But the compartmentalised model also uncovered that GSK3B_nuc KO did not lead to EMT, because GSK3B mainly localises to the cytoplasm and only translocates to the nucleus during EMT [83-85], thus it is expected that in the epithelial state the KO of GSK3B_nuc does not have an effect (Figure 10b).

Another compartment-specific result with GSK3B is that in the case of TGFBR ON simulations GSK3B were able to repress EMT, and this inhibitory effect is stronger in the nucleus (Figure 10c). Experimental results proved that the inhibition of the PI3K/AKT pathway suppressed EMT through the induction of GSK3B in hepatocellular carcinoma (HCC) [85]. The more robust inhibition in the nucleus is a consequence of the fact, that nuclear translocation of GSK3B prevents EMT through the downregulation of SNAIL transcription factor and that nuclear GSK3B is highly active relative to its cytoplasmic counterpart [86].

Beside GSK3B we also detected some compartment-specific functions of the GLI node. GLI KI in the original model and GLI_nuc KI in the compartmentalised model both led to EMT, which is expected as independent GLI activation can induce EMT [87]. Although in the compartmentalised model the KI of GLI_cyto did not result in EMT (Figure 10d), because in the absence of an upstream signal cytoplasmic GLI2 gets truncated into a transcriptional repressor form which inhibits GLI-induced gene transcription [88] and there is a simultaneous cytoplasmic sequestration of GLI1 [89].

DOI:10.14753/SE.2023.2783

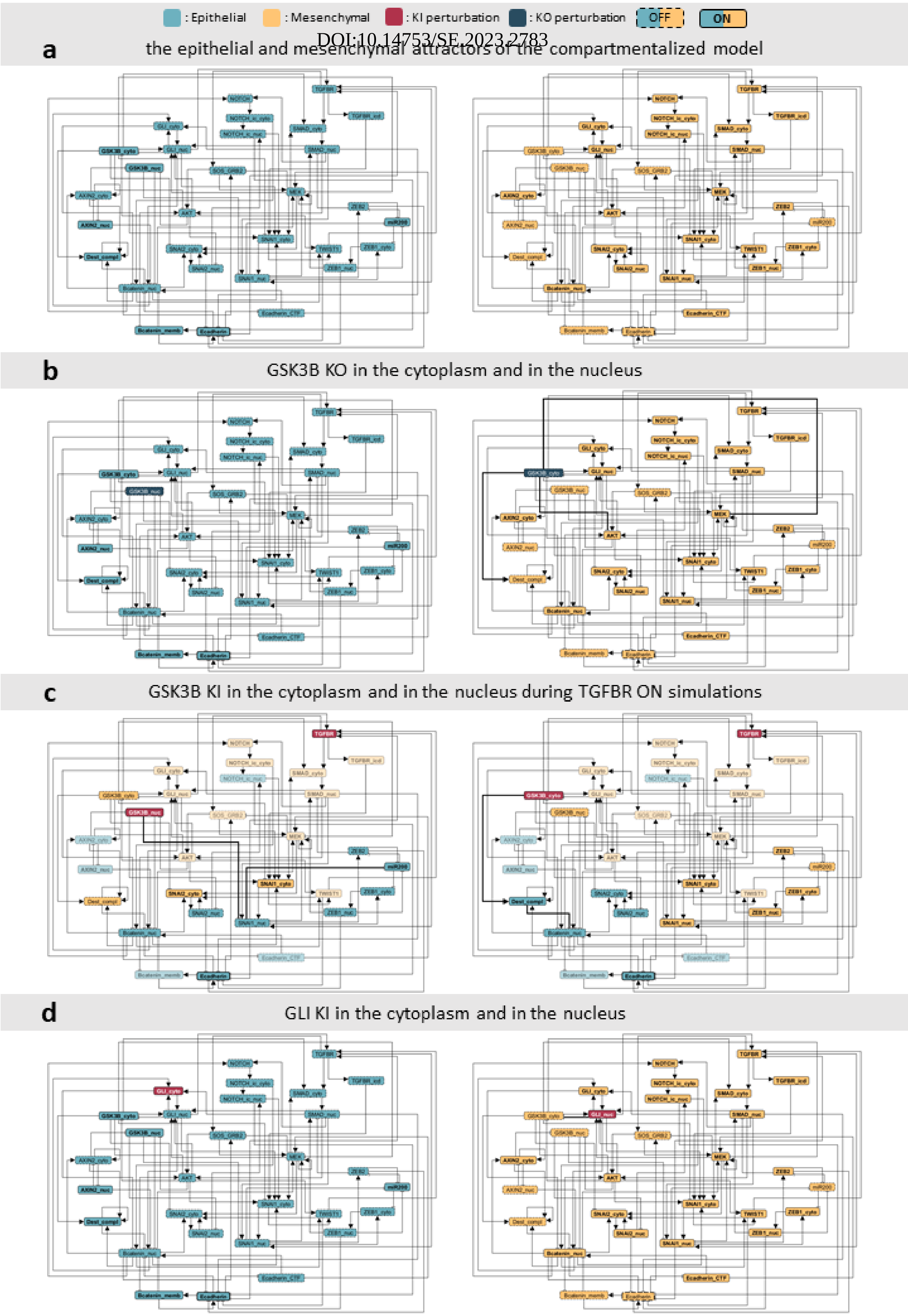


Figure 10. Attractors of the compartmentalised model showing localisation-specific functions of GSK3B and GLI. (a) The attractors corresponding to the epithelial (on the left) and mesenchymal (on the right) state. Perturbations were introduced to the initial epithelial state. **(b)** During TGFBR OFF simulations the cytoplasmic KO of GSK3B resulted in EMT, while the nuclear perturbation had no effect. The signal from GSK3B_cyto KO propagated through the loss of the Dest_compl and the activation of AKT and MEK. **(c)** TGFBR ON simulations showed that both the KI of the cytoplasmic and nuclear node of GSK3B inhibited the TGFB mediated EMT. The nuclear perturbation had a greater inhibitory effect by preventing the loss of miR200 and consequently the activation of ZEB1 and ZEB2. The nuclear perturbation also led to the inhibition of SNAI2_nuc and Bcatenin_nuc, despite the loss of the Dest_compl and the activation of SNAI2_cyto compared to the cytoplasmic perturbation. GSK3B_cyto functions both to stabilize the epithelial state and to inhibit TGFB mediated EMT, whereas GSK3B_nuc functions specifically to inhibit TGFB mediated EMT. **(d)** TGFBR OFF simulations show that the nuclear perturbation of GLI led to EMT, while the cytoplasmic perturbation alone was insufficient to destabilize the epithelial state. GLI proteins exert their main function in the nucleus and if sequestered in the cytoplasm, GLI2 gets truncated into a repressor form that further decreases transcriptional activity. The compartmentalised model reproduced these localisation-specific functions.

3.6 Analysing signalling pathway activities via network models

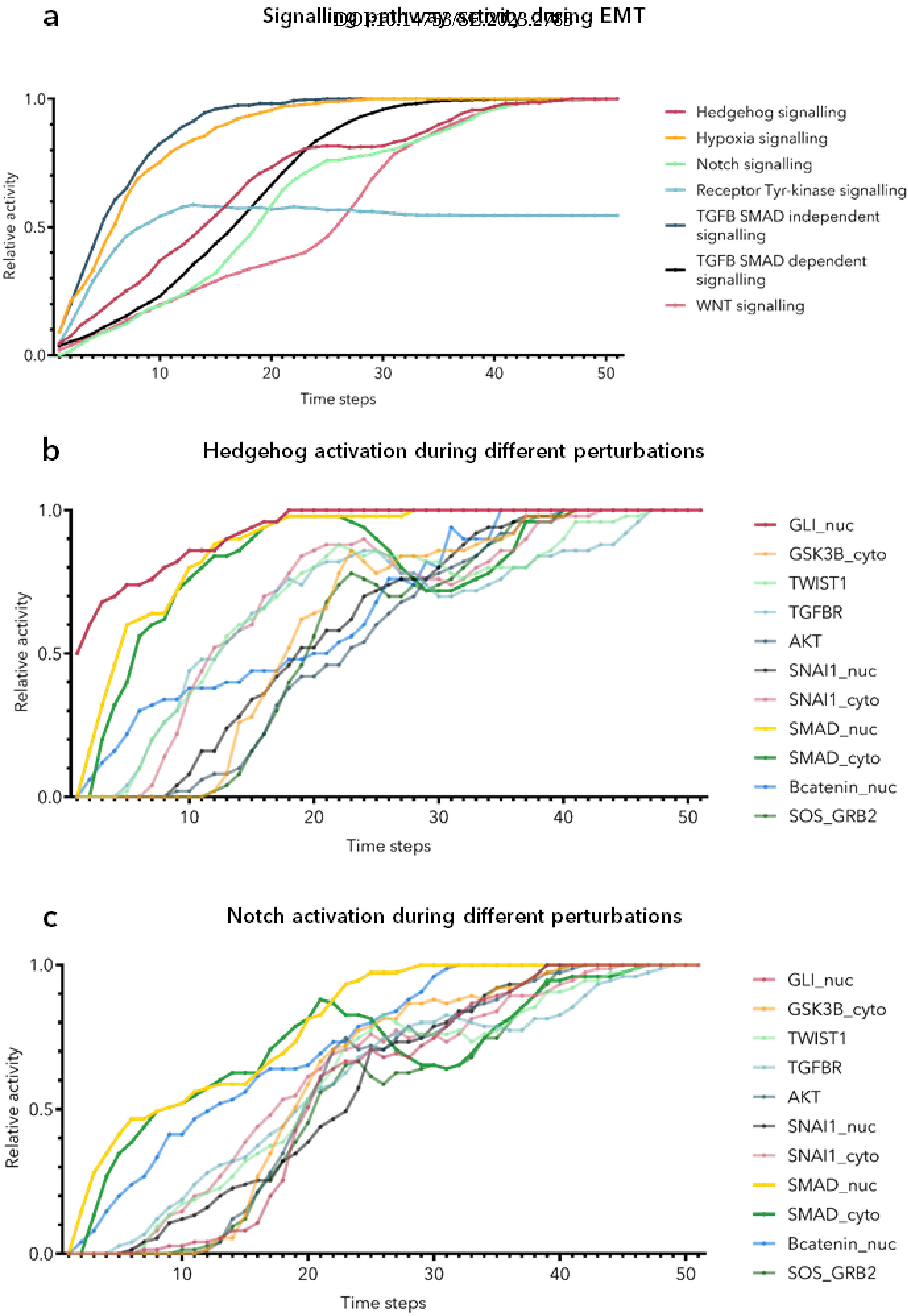
We also assessed the activities of signalling pathways that play a pivotal role in EMT. Based on a previous review [31] the most important pathways in EMT are:

A. TGF-β SMAD-dependent signalling
B. TGF-β SMAD-independent signalling
C. receptor tyrosine kinase (RTK) signalling
D. Wnt signalling
E. NOTCH signalling
F. Hedgehog signalling
G. Hypoxia signalling

This analysis recapitulated previous results of Steinway et al. [39] that Wnt and Hedgehog pathways are jointly activated during EMT but we also captured other synergistic functions of signalling pathways [39, 90]. As a validation of known crosstalk between TGF-β and Hedgehog signalling pathways [91, 92], the activation of SMAD proteins resulted in a quicker activation of the Hedgehog signalling pathway in our compartmentalised model. Similarly, TGF-β and Notch signalling also synergistically work together [93, 94] and that was also demonstrated with our signalling readout, as again the activation of SMAD proteins led to subsequent activation of the NOTCH pathway (Figure 11).

Figure 11. Signalling pathway activities during EMT. (a) This panel shows the activity of the main signalling pathways during EMT. There are 11 single node perturbations that led to EMT, here we show the average of the pathway activities that can be observed during these 11 perturbations. There was a difference in the kinetics of the activation of different pathways. Importantly, TGFB activation is still accompanied by the activation of the Hedgehog and Wnt pathways as captured by the original EMT model. **(b)** The activation of the Hedgehog signalling pathway for different perturbations. Each line symbolizes one specific perturbation which led to EMT and thus the activation of the Hedgehog pathway. We note here that the perturbation of the TGFB pathway member SMAD proteins (yellow and green) resulted in a quicker and more robust activation of the Hedgehog pathway. The most robust activation could be observed when we activated the nuclear pool of GLI proteins, with the perturbation of the GLI_nuc node (brown). **(c)** The activation of the Notch signalling pathway during different perturbations. The activation of SMAD proteins (green and yellow) resulted in a quick and robust activation of the Notch pathway which underlines the crosstalk between these pathways.

DOI:10.14753/SE.2023.2783



4. Discussion

During the first phase of my PhD studies we created an extensive database of human translocating proteins [24]. The database contains 213 manually curated translocating proteins which serve as a highly reviewed core of the database. Based on manually curated positive and negative learning sets we trained a machine learning algorithm to predict the translocation probability of a further ~13 000 proteins. These proteins are categorized as either high-confidence (1133 proteins) or low-confidence (3268 proteins) translocating proteins or non-translocating proteins (8665 proteins). We have also presented that the predictions of the Translocatome can be validated (*via* literature review) and are biologically valid. The performance of the prediction tool could be further enhanced in the future with the addition of other data layers, such as structural data (amino acid sequences), more complex network metrics data or combining these data with e.g. natural language processing (NLP) data.

The Translocatome is available at http://translocatome.linkgroup.hu and its user-friendly web-interface provides search options (for localisations, for UniProt identifiers, for protein names, for TES etc.) and download options. Thus the Translocatome data can be utilised in other studies that aim to either focus on protein specific discoveries or studies focusing on a general understanding of subcellular dynamics and the role protein translocations play in that.

Some limitations of the Translocatome could be resolved in the future, e.g. the extension of the database to incorporate other species (that are important in biological exploratory studies e.g. *Drosophila melanogaster* or *Caenorhabditis elegans* data), the further enrichment of the manually curated core of the database could enhance its predictive power and the implementation of a localisation based visualization tool may give an additional approach to our understanding of the data. The translocation of RNAs is also an emerging interesting topic [95] so incorporating those to our database would also be beneficial.

At the time of the development of the Translocatome most high-throughput methods that were used to generate proteome level localisation data could only produce data at the level of subcellular organelles. As a consequence of that Veres et al. [10] defined 6 major subcellular compartments during the development of the ComPPI database and a localisation tree which orders minor (smaller scale) localisations into those 6 major compartments. We followed the same logic to keep the interoperability of our database with the ComPPI database. The resolution of high-throughput subcellular localisation detection methods will improve, thus redefinition of that 6 major localisations will be necessary. Moreover, we know more and more about how membraneless organelles play a pivotal role in the regulation of cellular homeostasis [16, 17, 96, 97]. Adding these organelles – formed after liquid-liquid phase transition – to the Translocatome would uncover a new layer of complexity to understand cellular compartmentalisation.

In comparison with other existing databases (e.g. MoonProt or UniProt), those may also contain information on protein translocations but that is not always marked unequivocally. The 2.0 version of the MoonProt database [98] (accessed on 04/01/2018) contains 75 human proteins out of which 55 were translocating proteins based on literature data (these were incorporated in the manually curated core of the Translocatome). The other 20 moonlighting proteins achieved their multiple functions in the same subcellular compartment. Out of 20 239 human UniProt [22] proteins (accessed on 17 November 2017), based on their UniProt description or subcellular location data a translocation can be assumed in 1013 cases. But only 75 (35%) of the 213 manually curated core proteins of the Translocatome were included among the 1013 presumably translocating UniProt proteins, showing that UniProt's list can be greatly supplemented by Translocatome data. From the remaining 938 UniProt translocation candidates 25% and 34% were predicted in the Translocatome as high- and low-confidence translocating proteins, respectively. 31% of the 938 UniProt proteins were predicted as non-translocating while 10% of them were not part of the Translocatome database (because those proteins did not have an experimentally validated subcellular localisation in the ComPPI database).

During the second phase of my PhD studies we proved the applicability of the Translocatome to study subcellular dynamics and compartment-specific functions. We created a compartmentalised Boolean model based on Translocatome data and proved its validity [24]. The compartmentalised model reproduced previous key experimental outcomes and through the ability to better capture compartment-specific functions and regulations of proteins our model outperformed the model of Steinway et al. [39]

We have shown that the predictions of the Translocatome database can be used to tailor the generation of compartmentalised models and a generalizable workflow can be built (Figure 9), where first predicted translocations are detected with the help of the Translocatome database. In the next stage of the workflow the compartmentalisation of the model was implemented. In our case this was a mainly manual process in which we reviewed each prediction based on available literature data, and carefully modified the original EMT model. For compartmentalised models to become standard there would be a need to automate this manual part of the process. Currently there are no available tools that can replace the manual curation process, as biological expertise, the ability to judge the importance of information from different sources and oversee the whole model are not yet probable to be replaced by an automated algorithm. With the implementation of the https://translocaboole.linkgroup.hu website we aimed to create a workflow where all possible steps are automated and we hope the future solutions will enable this in the case of the manual curation process as well. The modular design of our workflow makes it possible to adapt to these future advances by swapping some elements of it for an automated version.

The complexity of the compartmentalisation process partly lies in the vast amount of combinations that are available to compartmentalise a node. In our paper we discussed all the possible node compartmentalisation scenarios (Mendik et al. 2022, Supplementary Figure 7 [67]). Even without accounting to the possible inputs and outputs there are already 9 scenarios, so simply testing all possible combinations algorithmically and then decide the valid setup based on the results would be impossible due to the large amount of available options.

The compartmentalisation process also raises an interesting question regarding mass preservation. As we duplicate the number of nodes (if a protein resides in two different compartments) for a translocating protein, and those nodes not always mutually inhibit each other so hypothetically there could be scenarios where two nodes corresponding to one protein are active. This seemingly doubles the amount of that protein in our model, whereas other proteins are only represented with one node. Importantly, the functionality of a protein is not always proportional with its concentration. Moreover, local concentration is important which may drastically change even without translocation. Translocation of a protein to an organelle where it has interaction partners with high affinity, would result in functional changes even if only a small amount of the protein is translocated and the original function in the other organelle may be preserved. In Boolean models the activity of a node signals the functionality of that given protein and that is not always proportional with the concentration of that protein (this topic is discussed in detail in Mendik et al. 2022, Supplementary Note 5 [67]).

Comparing our model to other previous models we can conclude that there are not many available system level models of the EMT [40]. Smaller scale kinetic and ordinary differential equation (ODE) models focused on a very limited number of nodes, and those were shown to be useful in deciphering the phase diagram of an EMT model [99], to specifically model the shuttling of β-catenin [100] or to analyse the bistable switches of EMT [101]. These models are useful in the quantification of dynamics, but due to their high computational demand cannot be scaled up to model processes on a system level. The model of Font-Clos et al. [102] also stemmed from the original EMT model of Steinway et al. [39] but used a slightly modified modelling approach. They were able to reach a more symmetric and more diverse representation of the attractor states with this modified approach and the addition of the LIF/KLF4 pathway, but their model also mostly highlights kinetic aspects of the model and not certain biological scenarios. During our work we extensively compared our model to the Boolean model of Steinway et al. [39] and we showed how our model succeeded it in certain aspects.

Our most important results uncovered some compartment-specific functions of proteins, namely those of the GSK3B and GLI proteins. GSK3B compartment-specifically inhibited EMT. Activation of GSK3B in the nucleus inhibited the activation of some key factors of EMT (Figure 10c). Almost simultaneously but independently from our research, Lee et al. [85] experimentally investigated EMT and they found in hepatocellular carcinoma (HCC) cells, that enhancing the nuclear translocation of GSK3B (i.e. the nuclear activation) suppressed EMT. Similarly, our compartment-specific results regarding GLI transcription factors coincide with other studies showing aberrant GLI1 activation in DNA damage and carcinogenesis and that GLI1 activates EMT due to its transcriptional activity mainly through SNAI1 [103]. Interestingly, there are promising therapeutic targets among Sonic Hedgehog (SHH) pathway members [104]. GLI antagonists (GANTs) and another inhibitor of GLI functions, Arsenic trioxide (ATO) have both been shown effective [104]. Our results also highlighted GLI as an important factor during EMT and this is in conjunction with the aforementioned experimental studies. Further analysis that incorporate more actors of the SHH pathway could shed light on novel, high impact interventions.

As EMT was lately described as a multifaceted process [30], to define it in its whole complexity and plasticity *in silico* methods need to adapt their readouts. We have shown that our signalling readout is able to recapitulate key signalling events, but readouts regarding functional traits of systems would also be important. Although some high throughput functional data is available in the GO database [43, 44] (biological process terms are annotated to proteins), those terms are not compartmentalised. In order to create readouts where compartmentalised models can be functionally characterized those high throughput datasets need to adapt to these changing needs (discussed in depth in Mendik et al. 2022, Supplementary Note 4 [67]).

5. Conclusions

These investigations conducted during my PhD studies significantly moved us closer to understand the role of translocating proteins in protein-protein interaction networks and to enable further system level studies to be carried out. Our new findings are the following:

1. We collected and characterized a gold standard set of 213 translocating proteins and 139 non-translocating proteins. Using the XGBoost machine learning algorithm a Translocation Evidence Score (TES) was assigned to 13.066 human proteins. Based on the TES values we have predicted 1133 and 3268 high- and low-confidence translocating proteins, respectively.

2. We created the Translocatome database (http://translocatome.linkgroup.hu) as a user-friendly tool showing these data. We demonstrated the use of the Translocatome database by the p63 protein, which wasn't marked as a translocating protein in available databases, but after targeted literature review we could validate that it has roles in different organelles and cytoplasmic p63 expression serves as a biomarker in prostate cancer.

3. Translocating proteins are enriched between signalling proteins and the same enrichment is observed in the case of the proteins of epithelial-mesenchymal transition (EMT).

4. We built a compartmentalised Boolean model of EMT and developed a user-friendly tool, https://translocaboole.linkgroup.hu for an automated workflow of the process.

5. Our *in silico* dynamic simulations on the compartmentalised model showed, that GSK3B compartment-specifically inhibits EMT and GLI transcription factors compartment-specifically drive EMT.

6. Summary

To summarize our efforts, we have created the Translocatome database, the first dedicated collection of human translocating proteins, with 213 manually curated core translocating proteins also including their interaction partners in the different subcellular localisations. Importantly, the Translocatome allowed the assessment of proteins' translocation probability by annotating a Translocation Evidence Score (TES) to 13 066 human proteins. These features are accessible via a webpage, in a user-friendly manner. The database allowed a better comprehension of protein translocation as a systems biology phenomenon, and it can be used as a discovery-tool of the field [24].

We have also presented that translocating proteins were enriched between proteins of the EMT. To model the compartment-specific functions of translocating proteins we created a compartmentalised Boolean dynamic and showed that GSK3B and GLI proteins, both alter the fate of EMT compartment-specifically. Based on our workflow future studies can also create compartmentalised models, aided by the algorithmic procedures available on the https://translocaboole.linkgroup.hu website and in the GitHub repository of our project. Our results underline that in order to model the physiological and pathological cellular behaviours and to rightly capture compartment-specific traits of proteins compartmentalised models should be used. This will also enhance the general understanding of subcellular dynamics [24]. As some translocating proteins are also important therapeutic targets [5, 6] the analysis of Translocatome data e.g. through studies that address the compartment-specific traits of biological processes and proteins may contribute to uncover new biomarkers and therapeutic targeting strategies ultimately to offer better future therapeutic options.

7. References

1. Gabaldón T, Pittis AA. (2015) Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes. Biochimie, 119: 262-268.
2. Barlan K, Rossow MJ, Gelfand VI. (2013) The journey of the organelle: teamwork and regulation in intracellular transport. Current Opinion in Cell Biology, 25: 483-488.
3. Wang X, Li S. (2014) Protein mislocalization: Mechanisms, functions and clinical applications in cancer. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 1846: 13-25.
4. Prager GW, Braga S, Bystricky B, Qvortrup C, Criscitiello C, Esin E, Sonke GS, Martínez G, Frenel J-S, Karamouzis M, Strijbos M, Yazici O, Bossi P, Banerjee S, Troiani T, Eniu A, Ciardiello F, Tabernero J, Zielinski CC, Casali PG, Cardoso F, Douillard J-Y, Jezdic S, McGregor K, Bricalli G, Vyas M, Ilbawi A. (2018) Global cancer control: responding to the growing burden, rising costs and inequalities in access. ESMO Open, 3: e000285.
5. Serrels A, Lund T, Serrels B, Byron A, McPherson Rhoanne C, von Kriegsheim A, Gómez-Cuadrado L, Canel M, Muir M, Ring Jennifer E, Maniati E, Sims Andrew H, Pachter Jonathan A, Brunton Valerie G, Gilbert N, Anderton Stephen M, Nibbs Robert JB, Frame Margaret C. (2015) Nuclear FAK controls chemokine transcription, Tregs, and evasion of anti-tumor immunity. Cell, 163: 160-173.
6. Frankowski KJ, Wang C, Patnaik S, Schoenen FJ, Southall N, Li D, Teper Y, Sun W, Kandela I, Hu D, Dextras C, Knotts Z, Bian Y, Norton J, Titus S, Lewandowska MA, Wen Y, Farley KI, Griner LM, Sultan J, Meng Z, Zhou M, Vilimas T, Powers AS, Kozlov S, Nagashima K, Quadri HS, Fang M, Long C, Khanolkar O, Chen W, Kang J, Huang H, Chow E, Goldberg E, Feldman C, Xi R, Kim HR, Sahagian G, Baserga SJ, Mazar A, Ferrer M, Zheng W, Shilatifard A, Aubé J, Rudloff U, Marugan JJ, Huang S. (2018) Metarrestin, a perinucleolar compartment inhibitor, effectively suppresses metastasis. Science Translational Medicine, 10: eaap8307.

7. Park S, Yang JS, Shin YE, Park J, Jang SK, Kim S. (2011) Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. Mol Syst Biol, 7: 494.
8. Laurila K, Vihinen M. (2009) Prediction of disease-related mutations affecting protein localization. BMC Genomics, 10: 122.
9. Wang J, Sun T, Meng Z, Wang L, Li M, Chen J, Qin T, Yu J, Zhang M, Bie Z, Dong Z, Jiang X, Lin L, Zhang C, Liu Z, Jiang R, Yang G, Li L, Zhang Y, Huang D. (2021) XPO1 inhibition synergizes with PARP1 inhibition in small cell lung cancer by targeting nuclear transport of FOXO3a. Cancer Letters, 503: 197-212.
10. Veres DV, Gyurko DM, Thaler B, Szalay KZ, Fazekas D, Korcsmaros T, Csermely P. (2015) ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis. Nucleic Acids Res, 43: D485-493.
11. Andrade MA, O’Donoghue SI, Rost B. (1998) Adaptation of protein surfaces to subcellular location Edited by F. E. Cohen. Journal of Molecular Biology, 276: 517-525.
12. Scott JD, Pawson T. (2009) Cell Signaling in space and time: Where proteins come together and when they’re apart. Science, 326: 1220-1224.
13. Dean KM, Palmer AE. (2014) Advances in fluorescence labeling strategies for dynamic cellular imaging. Nature Chemical Biology, 10: 512-523.
14. Day RN, Davidson MW. (2009) The fluorescent protein palette: tools for cellular imaging. Chemical Society Reviews, 38: 2887-2921.
15. Liu T-L, Upadhyayula S, Milkie DE, Singh V, Wang K, Swinburne IA, Mosaliganti KR, Collins ZM, Hiscock TW, Shea J, Kohrman AQ, Medwig TN, Dambournet D, Forster R, Cunniff B, Ruan Y, Yashiro H, Scholpp S, Meyerowitz EM, Hockemeyer D, Drubin DG, Martin BL, Matus DQ, Koyama M, Megason SG, Kirchhausen T, Betzig E. (2018) Observing the cell in its native state: Imaging subcellular dynamics in multicellular organisms. Science, 360: eaaq1392.
16. Hyman AA, Weber CA, Jülicher F. (2014) Liquid-liquid phase separation in biology. Annual Review of Cell and Developmental Biology, 30: 39-58.
17. Alberti S, Dormann D. (2019) Liquid–liquid phase separation in disease. Annual Review of Genetics, 53: 171-194.

18. Lang P, Yeow K, Nichols A, Scheer A. (2006) Cellular imaging in drug discovery. Nature Reviews Drug Discovery, 5: 343-356.

19. Gardy JL, Brinkman FSL. (2006) Methods for predicting bacterial protein subcellular localization. Nature Reviews Microbiology, 4: 741-751.

20. Nair R, Rost B. Protein subcellular localization prediction using artificial intelligence technology. In:Thompson JD, Ueffing M ,Schaeffer-Reiss C (szerk.), Functional Proteomics: Methods and Protocols. Humana Press, Totowa, NJ, 2008: 435-463.

21. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L, Breckels LM, Bäckström A, Danielsson F, Fagerberg L, Fall J, Gatto L, Gnann C, Hober S, Hjelmare M, Johansson F, Lee S, Lindskog C, Mulder J, Mulvey CM, Nilsson P, Oksvold P, Rockberg J, Schutten R, Schwenk JM, Sivertsson Å, Sjöstedt E, Skogs M, Stadler C, Sullivan DP, Tegel H, Winsnes C, Zhang C, Zwahlen M, Mardinoglu A, Pontén F, von Feilitzen K, Lilley KS, Uhlén M, Lundberg E. (2017) A subcellular map of the human proteome. Science, 356: eaal3321.

22. Consortium TU. (2020) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49: D480-D489.

23. Gut G, Herrmann MD, Pelkmans L. (2018) Multiplexed protein maps link subcellular organization to cellular states. Science, 361: eaar7042.

24. Mendik P, Dobronyi L, Hari F, Kerepesi C, Maia-Moco L, Buszlai D, Csermely P, Veres DV. (2019) Translocatome: a novel resource for the analysis of protein translocation between cellular organelles. Nucleic Acids Research, 47: D495-D505.

25. Dhillon PK, Barry M, Stampfer MJ, Perner S, Fiorentino M, Fornari A, Ma J, Fleet J, Kurth T, Rubin MA, Mucci LA. (2009) Aberrant cytoplasmic expression of p63 and prostate cancer mortality. Cancer Epidemiology Biomarkers Prevention, 18: 595-600.

26. Lundberg E, Borner GHH. (2019) Spatial proteomics: a powerful discovery tool for cell biology. Nature Reviews Molecular Cell Biology, 20: 285-302.

27. Wainstein E, Seger R. (2016) The dynamic subcellular localization of ERK: mechanisms of translocation and role in various organelles. Current Opinion in Cell Biology, 39: 15-20.
28. Nieto MA, Huang Ruby Y-J, Jackson Rebecca A, Thiery Jean P. (2016) EMT: 2016. Cell, 166: 21-45.
29. Yang J, Weinberg RA. (2008) Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. Dev Cell, 14: 818-829.
30. Yang J, Antin P, Berx G, Blanpain C, Brabletz T, Bronner M, Campbell K, Cano A, Casanova J, Christofori G, Dedhar S, Derynck R, Ford HL, Fuxe J, García de Herreros A, Goodall GJ, Hadjantonakis A-K, Huang RJY, Kalcheim C, Kalluri R, Kang Y, Khew-Goodall Y, Levine H, Liu J, Longmore GD, Mani SA, Massagué J, Mayor R, McClay D, Mostov KE, Newgreen DF, Nieto MA, Puisieux A, Runyan R, Savagner P, Stanger B, Stemmler MP, Takahashi Y, Takeichi M, Theveneau E, Thiery JP, Thompson EW, Weinberg RA, Williams ED, Xing J, Zhou BP, Sheng G, On behalf of the EMTIA. (2020) Guidelines and definitions for research on epithelial–mesenchymal transition. Nature Reviews Molecular Cell Biology, 21: 341-352.
31. Gonzalez DM, Medici D. (2014) Signaling mechanisms of the epithelial-mesenchymal transition. Science Signaling, 7: re8.
32. Chaw SY, Abdul Majeed A, Dalley AJ, Chan A, Stein S, Farah CS. (2012) Epithelial to mesenchymal transition (EMT) biomarkers – E-cadherin, beta-catenin, APC and Vimentin – in oral squamous cell carcinogenesis and transformation. Oral Oncology, 48: 997-1006.
33. Csermely P, Korcsmáros T, Kiss HJM, London G, Nussinov R. (2013) Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. Pharmacology & Therapeutics, 138: 333-408.
34. Hastings JF, O'Donnell YEI, Fey D, Croucher DR. (2020) Applications of personalised signalling network models in precision oncology. Pharmacology & Therapeutics, 212: 107555.
35. Hyduke DR, Palsson BØ. (2010) Towards genome-scale signalling-network reconstructions. Nature Reviews Genetics, 11: 297-307.

36. Albert I, Thakar J, Li S, Zhang R, Albert R. (2008) Boolean network simulations for life scientists. Source Code Biol Med, 3: 16.

37. Sherekar S, Viswanathan GA. (2021) Boolean dynamic modeling of cancer signaling networks: Prognosis, progression, and therapeutics. Computational and Systems Oncology, 1: e1017.

38. Wang R-S, Saadatpour A, Albert R. (2012) Boolean modeling in systems biology: an overview of methodology and applications. Physical Biology, 9: 055001.

39. Steinway SN, Zanudo JG, Ding W, Rountree CB, Feith DJ, Loughran TP, Jr., Albert R. (2014) Network modeling of TGFbeta signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and Wnt pathway activation. Cancer Res, 74: 5963-5977.

40. Burger GA, Danen EHJ, Beltman JB. (2017) Deciphering epithelial-mesenchymal transition regulatory networks in cancer through computational approaches. Front Oncol, 7: 162.

41. Mani M, Chen C, Amblee V, Liu H, Mathur T, Zwicke G, Zabad S, Patel B, Thakkar J, Jeffery CJ. (2015) MoonProt: a database for proteins that are known to moonlight. Nucleic Acids Res, 43: D277-282.

42. Chen T, Guestrin C. (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, doi:10.1145/2939672.2939785 pp. 785–794, Association for Computing Machinery, San Francisco, California, USA

43. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25: 25-29.

44. The Gene Ontology Consortium. (2016) Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research, 45: D331-D338.

45. Friedman JH. (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29: 1189-1232.

46. Kerepesi C, Daróczy B, Sturm Á, Vellai T, Benczúr A. (2018) Prediction and characterization of human ageing-related proteins by using machine learning. Scientific Reports, 8: 4094.
47. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. (2016) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research, 45: D353-D361.
48. Basit AH, Abbasi WA, Asif A, Gull S, Minhas FUAA. (2018) Training host-pathogen protein–protein interaction predictors. Journal of Bioinformatics and Computational Biology, 16: 1850014.
49. Chen X, Huang L, Xie D, Zhao Q. (2018) EGBMMDA: Extreme gradient boosting machine for miRNA-disease association prediction. Cell Death & Disease, 9: 3.
50. Zou LS, Erdos MR, Taylor DL, Chines PS, Varshney A, Parker SCJ, Collins FS, Didion JP, The McDonnell Genome I. (2018) BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. BMC Genomics, 19: 390.
51. Kovács IA, Palotai R, Szalay MS, Csermely P. (2010) Community landscapes: An integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. PLOS ONE, 5: e12528.
52. Barabási A-L, Albert R. (1999) Emergence of scaling in random networks. Science, 286: 509-512.
53. Smith JM, Koopman PA. (2004) The ins and outs of transcriptional control: nucleocytoplasmic shuttling in development and disease. Trends in Genetics, 20: 4-8.
54. Lu Z, Hunter T. (2018) Metabolic kinases moonlighting as protein kinases. Trends in Biochemical Sciences, 43: 301-310.
55. Stambolic V, Suzuki A, de la Pompa JL, Brothers GM, Mirtsos C, Sasaki T, Ruland J, Penninger JM, Siderovski DP, Mak TW. (1998) Negative regulation of PKB/Akt-dependent cell survival by the tumor suppressor PTEN. Cell, 95: 29-39.
56. Teruel MN, Meyer T. (2000) Translocation and reversible localization of signaling proteins: A dynamic future for signal transduction. Cell, 103: 181-184.

57. Zhou Z, Luo M-j, Straesser K, Katahira J, Hurt E, Reed R. (2000) The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. Nature, 407: 401-405.
58. Shaw DJ, Eggleton P, Young PJ. (2008) Joining the dots: Production, processing and targeting of U snRNP to nuclear bodies. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1783: 2137-2144.
59. Szalay-Bekő M, Palotai R, Szappanos B, Kovács IA, Papp B, Csermely P. (2012) ModuLand plug-in for Cytoscape: Determination of hierarchical layers of overlapping network modules and community centrality. Bioinformatics, 28: 2202-2204.
60. Ota M, Gonja H, Koike R, Fukuchi S. (2016) Multiple-localization and hub proteins. PLOS ONE, 11: e0156455.
61. Davies RG, Wagstaff KM, McLaughlin EA, Loveland KL, Jans DA. (2013) The BRCA1-binding protein BRAP2 can act as a cytoplasmic retention factor for nuclear and nuclear envelope-localizing testicular proteins. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1833: 3436-3444.
62. Paietta E. (1996) Adhesion molecules in acute myeloid leukemia. Leukemia Research, 20: 795-798.
63. Takagi S, Saito Y, Hijikata A, Tanaka S, Watanabe T, Hasegawa T, Mochizuki S, Kunisawa J, Kiyono H, Koseki H, Ohara O, Saito T, Taniguchi S, Shultz LD, Ishikawa F. (2012) Membrane-bound human SCF/KL promotes in vivo human hematopoietic engraftment and myeloid differentiation. Blood, 119: 2768-2777.
64. Levrero M, De Laurenzi V, Costanzo A, Gong J, Wang JY, Melino G. (2000) The p53/p63/p73 family of transcription factors: Overlapping and distinct functions. Journal of Cell Science, 113: 1661-1670.
65. Mills AA, Zheng B, Wang X-J, Vogel H, Roop DR, Bradley A. (1999) p63 is a p53 homologue required for limb and epidermal morphogenesis. Nature, 398: 708-713.
66. Narahashi T, Niki T, Wang T, Goto A, Matsubara D, Funata N, Fukayama M. (2006) Cytoplasmic localization of p63 is associated with poor patient survival in lung adenocarcinoma. Histopathology, 49: 349-357.

67. Mendik P, Kerestely M, Kamp S, Deritei D, Kunsic N, Vassy Z, Csermely P, Veres DV. (2022) Translocating proteins compartment-specifically alter the fate of epithelial-mesenchymal transition in a compartmentalized Boolean network model. npj Syst Biol Appl, 8: 19
68. Kauffman SA. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology, 22: 437-467.
69. Rozum JC, Gomez Tejeda Zanudo J, Gan X, Deritei D, Albert R. (2021) Parity and time reversal elucidate both decision-making in empirical models and attractor scaling in critical Boolean networks. Sci Adv, 7: eabf8124.
70. Ikushima H, Miyazono K. (2010) TGFβ signalling: a complex web in cancer progression. Nature Reviews Cancer, 10: 415-424.
71. Jho EH, Zhang T, Domon C, Joo CK, Freund JN, Costantini F. (2002) Wnt/beta-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway. Mol Cell Biol, 22: 1172-1183.
72. Gupta S, Takebe N, Lorusso P. (2010) Targeting the Hedgehog pathway in cancer. Ther Adv Med Oncol, 2: 237-250.
73. Korpal M, Lee ES, Hu G, Kang Y. (2008) The miR-200 family inhibits epithelial-mesenchymal transition and cancer cell migration by direct targeting of E-cadherin transcriptional repressors ZEB1 and ZEB2. J Biol Chem, 283: 14910-14914.
74. Sánchez-Tilló E, Liu Y, de Barrios O, Siles L, Fanlo L, Cuatrecasas M, Darling DS, Dean DC, Castells A, Postigo A. (2012) EMT-activating transcription factors in cancer: beyond EMT and tumor invasiveness. Cell Mol Life Sci, 69: 3429-3456.
75. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, Hartwell K, Onder TT, Gupta PB, Evans KW, Hollier BG, Ram PT, Lander ES, Rosen JM, Weinberg RA, Mani SA. (2010) Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. Proc Natl Acad Sci U S A, 107: 15449-15454.
76. Casas E, Kim J, Bendesky A, Ohno-Machado L, Wolfe CJ, Yang J. (2011) Snail2 is an essential mediator of Twist1-induced epithelial mesenchymal transition and metastasis. Cancer Res, 71: 245-254.

77. Dave N, Guaita-Esteruelas S, Gutarra S, Frias À, Beltran M, Peiró S, de Herreros AG. (2011) Functional cooperation between Snail1 and twist in the regulation of ZEB1 expression during epithelial to mesenchymal transition. J Biol Chem, 286: 12024-12032.
78. Natsuizaka M, Whelan KA, Kagawa S, Tanaka K, Giroux V, Chandramouleeswaran PM, Long A, Sahu V, Darling DS, Que J, Yang Y, Katz JP, Wileyto EP, Basu D, Kita Y, Natsugoe S, Naganuma S, Klein-Szanto AJ, Diehl JA, Bass AJ, Wong KK, Rustgi AK, Nakagawa H. (2017) Interplay between Notch1 and Notch3 promotes EMT and tumor initiation in squamous cell carcinoma. Nat Commun, 8: 1758.
79. Chanrion M, Kuperstein I, Barrière C, El Marjou F, Cohen D, Vignjevic D, Stimmer L, Paul-Gilloteaux P, Bièche I, Tavares Sdos R, Boccia GF, Cacheux W, Meseure D, Fre S, Martignetti L, Legoix-Né P, Girard E, Fetler L, Barillot E, Louvard D, Zinovyev A, Robine S. (2014) Concomitant Notch activation and p53 deletion trigger epithelial-to-mesenchymal transition and metastasis in mouse gut. Nat Commun, 5: 5005.
80. Kurimoto R, Iwasawa S, Ebata T, Ishiwata T, Sekine I, Tada Y, Tatsumi K, Koide S, Iwama A, Takiguchi Y. (2016) Drug resistance originating from a TGF-β/FGF-2-driven epithelial-to-mesenchymal transition and its reversion in human lung adenocarcinoma cell lines harboring an EGFR mutation. Int J Oncol, 48: 1825-1836.
81. Lemieux E, Bergeron S, Durand V, Asselin C, Saucier C, Rivard N. (2009) Constitutively active MEK1 is sufficient to induce epithelial-to-mesenchymal transition in intestinal epithelial cells and to promote tumor invasion and metastasis. Int J Cancer, 125: 1575-1586.
82. Buonato JM, Lazzara MJ. (2014) ERK1/2 blockade prevents epithelial-mesenchymal transition in lung cancer cells and promotes their sensitivity to EGFR inhibition. Cancer Res, 74: 309-319.
83. Mitra T, Roy SS. (2017) Co-activation of TGFβ and Wnt signalling pathways abrogates EMT in ovarian cancer cells. Cell Physiol Biochem, 41: 1336-1345.

84. Li S, Wang D, Zhao J, Weathington NM, Shang D, Zhao Y. (2018) The deubiquitinating enzyme USP48 stabilizes TRAF2 and reduces E-cadherin-mediated adherens junctions. FASEB J, 32: 230-242.
85. Lee S, Choi EJ, Cho EJ, Lee YB, Lee JH, Yu SJ, Yoon JH, Kim YJ. (2020) Inhibition of PI3K/Akt signaling suppresses epithelial-to-mesenchymal transition in hepatocellular carcinoma through the Snail/GSK-3/beta-catenin pathway. Clin Mol Hepatol, 26: 529-539.
86. Li J, Xing M, Zhu M, Wang X, Wang M, Zhou S, Li N, Wu R, Zhou M. (2008) Glycogen synthase kinase 3beta induces apoptosis in cancer cells through increase of survivin nuclear localization. Cancer Lett, 272: 91-101.
87. Zhang J, Tian XJ, Xing J. (2016) Signal transduction pathways of EMT induced by TGF-β, SHH, and WNT and their crosstalks. J Clin Med, 5: 41.
88. Niewiadomski P, Niedziółka SM, Markiewicz Ł, Uśpieński T, Baran B, Chojnowska K. (2019) Gli proteins: regulation in development and cancer. Cells, 8: 147.
89. Szczepny A, Wagstaff KM, Dias M, Gajewska K, Wang C, Davies RG, Kaur G, Ly-Huynh J, Loveland KL, Jans DA. (2014) Overlapping binding sites for importin β1 and suppressor of fused (SuFu) on glioma-associated oncogene homologue 1 (Gli1) regulate its nuclear localization. Biochem J, 461: 469-476.
90. Behan FM, Iorio F, Picco G, Goncalves E, Beaver CM, Migliardi G, Santos R, Rao Y, Sassi F, Pinnelli M, Ansari R, Harper S, Jackson DA, McRae R, Pooley R, Wilkinson P, van der Meer D, Dow D, Buser-Doepner C, Bertotti A, Trusolino L, Stronach EA, Saez-Rodriguez J, Yusa K, Garnett MJ. (2019) Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature, 568: 511-516.
91. Javelaud D, Alexaki VI, Dennler S, Mohammad KS, Guise TA, Mauviel A. (2011) TGF-β/SMAD/GLI2 signaling axis in cancer progression and metastasis. Cancer Res, 71: 5606-5610.
92. Javelaud D, Pierrat MJ, Mauviel A. (2012) Crosstalk between TGF-β and hedgehog signaling in cancer. FEBS Lett, 586: 2016-2025.
93. Lindsey S, Langhans SA. (2014) Crosstalk of oncogenic signaling pathways during epithelial-mesenchymal transition. Front Oncol, 4: 358.

94. Klüppel M, Wrana JL. (2005) Turning it up a Notch: cross-talk between TGF beta and Notch signaling. Bioessays, 27: 115-118.
95. Das S, Ferlito M, Kent OA, Fox-Talbot K, Wang R, Liu D, Raghavachari N, Yang Y, Wheelan SJ, Murphy E, Steenbergen C. (2012) Nuclear miRNA regulates the mitochondrial genome in the heart. Circulation Research, 110: 1596-1603.
96. Banani SF, Lee HO, Hyman AA, Rosen MK. (2017) Biomolecular condensates: organizers of cellular biochemistry. Nature Reviews Molecular Cell Biology, 18: 285-298.
97. Wang B, Zhang L, Dai T, Qin Z, Lu H, Zhang L, Zhou F. (2021) Liquid–liquid phase separation in human health and diseases. Signal Transduction and Targeted Therapy, 6: 290.
98. Chen C, Liu H, Zabad S, Rivera N, Rowin E, Hassan M, Gomez De Jesus SM, Llinás Santos PS, Kravchenko K, Mikhova M, Ketterer S, Shen A, Shen S, Navas E, Horan B, Raudsepp J, Jeffery C. (2020) MoonProt 3.0: an update of the moonlighting proteins database. Nucleic Acids Research, 49: D368-D372.
99. He P, Qiu K, Jia Y. (2018) Modeling of mesenchymal hybrid epithelial state and phenotypic transitions in EMT and MET processes of cancer cells. Scientific Reports, 8: 14323.
100. Schmitz Y, Rateitschak K, Wolkenhauer O. (2013) Analysing the impact of nucleo-cytoplasmic shuttling of β-catenin and its antagonists APC, Axin and GSK3 on Wnt/β-catenin signalling. Cellular Signalling, 25: 2210-2221.
101. Tian X-J, Zhang H, Xing J. (2013) Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition. Biophysical Journal, 105: 1079-1089.
102. Font-Clos F, Zapperi S, La Porta CAM. (2018) Topography of epithelial-mesenchymal plasticity. Proc Natl Acad Sci U S A, 115: 5902-5907.
103. Palle K, Mani C, Tripathi K, Athar M. (2015) Aberrant GLI1 activation in DNA damage response, carcinogenesis and chemoresistance. Cancers (Basel), 7: 2330-2351.
104. Rimkus TK, Carpenter RL, Qasem S, Chan M, Lo HW. (2016) Targeting the sonic hedgehog signaling pathway: Review of smoothened and GLI Inhibitors. Cancers (Basel), 8: 22.

8. Bibliography of the candidate's publications

8.1 Publications directly related to this thesis

Mendik P, Dobronyi L, Hari F, Kerepesi C, Maia-Moco L, Buszlai D, Csermely P, Veres DV. (2019) Translocatome: a novel resource for the analysis of protein translocation between cellular organelles. Nucleic Acids Research, 47: D495-D505.

Mendik P, Kerestely M, Kamp S, Deritei D, Kunsic N, Vassy Z, Csermely P, Veres DV. (2022) Translocating proteins compartment-specifically alter the fate of epithelial-mesenchymal transition in a compartmentalized Boolean network model. npj Syst Biol Appl, 8: 19

8.2 Publications indirectly related to this thesis

Csermely P, Kunsic N, Mendik P, Kerestely M, Farago T, Veres DV, Tompa P. (2020) Learning of signaling networks: Molecular mechanisms. Trends Biochem Sci, 45: 284-294

9. Individual contributions

- Manual curation of translocating proteins
- Designing the structure of the Translocatome database
- Biologically validating the results of the XGBoost algorithm (feature set, predicted translocations)
- Defining biologically valid protein categories and the logic behind each class
- Designing the structure and functionality of the Translocatome website
- Conceiving the idea of the compartmentalised Boolean modelling project and designing the appropriate study
- Carried out manual curation of predicted translocating proteins in EMT and being responsible for the creation of the Boolean rules in cooperation with Márk Kerestély
- Enrichment analysis of translocating proteins between signalling and EMT proteins
- Running dynamic simulations on the compartmentalised model and analysing results
- Creating data analysis and visualization and generation of some figures and tables
- Designing the structure and functionality of the Translocaboole website

10. Acknowledgements

This part of my thesis tries to list the people who I owe an enormous amount of gratitude because they significantly helped me to achieve my goals. Although I tried to collect everyone, but surely this list will not be complete. I believe that all human connections are special and there were probably a lot of interactions that I didn't even consider as relevant – or maybe at the time even considered to be a nuisance – but deep down those have also influenced my thinking and pushed me to find meaningful answers.

First of all, I am grateful to my family: my wife, my parents, my sisters, for supporting me and creating an environment where I could pursue my ambitions and supporting me when I encountered obstacles.

I'm lucky that I had the opportunity to work in such an excellent research group as the LINK-Group. I appreciate Prof Péter Csermely for creating this unparalleled environment – which is I believe very unique and enabled many young researchers to spread their wings – and for welcoming me with open arms to this community. He and Dániel Veres MD PhD introduced me to the enthralling world of networks and how to navigate this field with always keeping an eye on the clinical usability as well.

My experience would have been very different if I hadn't had the chance to work with such wonderful colleagues, I greatly appreciate the opportunity to work together with Levente Dobronyi, with whom we proved that a professional collaboration musn't negate fun and work can be enjoyed. I admire Márk Kerestély who has an incredible inner drive to find details and scientific truth, I'm happy that I could work with him. I also thank all the co-authors of our papers for their contributions.

I thank all current and previous Heads of the Department of Molecular Biology and other colleagues for providing the excellent background for these studies.

During my studies I have received funding via the following grants and sources: Hungarian National Research Development and Innovation Office [OTKA K115378]; New National Excellence Program of the Hungarian Ministry of Human Capacities (ÚNKP-16-2-41); Higher Education Institutional Excellence Programme of the Ministry of Human Capacities in Hungary '2018-1.2.1-NKP-00008: Exploring the Mathematical Foundations of Artificial Intelligence' grant; Thematic Excellence Programme

(Tématerületi Kiválósági Program, 2020-4.1.1.-TKP2020, TKP2021-EGA-24) of the Ministry for Innovation and Technology in Hungary, within the framework of the Molecular Biology thematic programme of the Semmelweis University; ÚNKP-20-III-2-SE-23 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation fund.